A NEW PERSPECTIVE ON INCOME EARNINGS USING AI

By

ANG SENG CHUN

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION SYSTEMS (HONOURS) INFORMATION SYSTEMS ENGINEERING

Faculty of Information and Communication Technology

(Kampar Campus)

JUNE 2024

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: **A NEW PERSPECTIVE ON INCOME EARNINGS USING AI**

Academic Session: 202406

I ANG SENG CHUN

(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,

(Author's signature) (Supervisor's signature)

Address:

No. 45, Laluan Wira Jaya
Timur 21, Taman Rapat Perdana
31350, Ipoh, Perak

Dr. Abdulkarim M. Jamal Kanaan Jebna
Supervisor's name

Date: 12 September 2024 **Date**: 12 September 2024

Universiti Tunku Abdul Rahman			
Form Title : **Sample of Submission Sheet for FYP/Dissertation/Thesis**			
Form Number: **FM-IAD-004**	Rev No.: **0**	Effective Date: **21 JUNE 2011**	Page No.: **1 of 1**

FACULTY/INSTITUTE* OF _INFORMATION AND COMMUNICATION TECHNOLOGY_

UNIVERSITI TUNKU ABDUL RAHMAN

Date: _12 September 2024_

SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS

It is hereby certified that ***Ang Seng Chun*** (ID No: ***20ACB02022***) has completed this final year project entitled "**A New Perspective on Income Earnings Using AI** " under the supervision of Dr Abdulkarim Kanaan Jebna (Supervisor) from the Department of Information Systems, Faculty of Information and Communication Technology.

I understand that University will upload softcopy of my final year project in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

(*ANG SENG CHUN*)

DECLARATION OF ORIGINALITY

I declare that this report entitled "**A NEW PERSPECTIVE ON INCOME EARNINGS USING AI**" is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature : 

Name : ANG SENG CHUN

Date : 12 September 2024

ACKNOWLEDGEMENTS

I would like to express my heartfelt appreciation and gratitude to my supervisor, Dr Abdulkarim Kanaan Jebna for giving me a great opportunity to accomplish the project that I wanted but was not capable of, guiding me through the hard times, and for being mentally and physically support me through this Adult Income Prediction project. He would always be there providing his advice to the problems that I had faced, no matter day or night. Also, I would like to appreciate my moderator, Mr. Cheang Kah Wai who have given me ideas to improve my project during the present stage. Without their guidance, I couldn't complete this project and finish this as my last semester in UTAR. Hence, once again, a million thanks to my supervisor and moderator.

To my dear housemates, parents and siblings, for their support and accompany, with me along the journey of this last final year project. Lastly, I must say thanks to my Ipoh friends and NGOs for their support and endless encouragement throughout the survey taken.

ABSTRACT

Income stands to be an important contribution to a country's economy, happiness index, and development growth. Understanding income dynamics enables policymakers to address the needs of different socioeconomic groups more effectively, improving financial freedom and overall quality of life. This project, titled "A New Perspective on Income Earnings Using AI," addresses the problem of inadequacies in predictive tools for Malaysian income dynamics. Existing predictive tools often fail to incorporate advanced AI techniques, which limits their effectiveness in providing accurate income predictions and insights. The primary objective of this project is to enhance income level prediction by leveraging machine learning and data mining techniques. Using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, we analyzed a dataset comprising 423 valid responses from residents of Ipoh, Malaysia, an area previously lacking such comprehensive income data. The results reveal significant economic insights and trends that were previously obscured, offering a nuanced understanding of income distribution. This project contributes to the advancement of AI-driven income analysis by creating an interactive dashboard that visualizes complex income data, thereby bridging the gap between high and low-income groups. The implications of this project are substantial for both policymakers and society. By providing a robust analytical tool, the dashboard supports more informed decision-making and enhances the ability to address income inequalities and economic disparities effectively.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

RM	Malaysian Ringgit
GDP	Gross Domestic Product
AI	Artificial Intelligence
GBC	Gradient Boosting Classifier
XGBOOST	Extreme Gradient Boosting
SVM	Support Vector Machine
PCA	Principal Component Analysis
QP	Quadratic Programming
SMO	Sequential Minimal Optimization
US	United States
USD	United States Dollar
UK	United Kingdom
API	Application Programming Interface
URL	Uniform Resource Locator
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets
JS	JavaScript
IDD	Income Distribution Database
WDD	Wealth Distribution Database
OECD	Organisation for Economic Co-operation and Development
PovCal	Poverty Calculator
Gini	Gini coefficient
SA	Sales Insight
CLI	Customer Lifetime Value
USP	Unique Selling Proposition
IP	Intellectual Property
CEO	Chief Executive Officer
SA	Sales Analyst
UI	User Interface
VP	Vice President

ML	Machine Learning
CRISP-DM	Cross-Industry Standard Process for Data Mining
KDD	Knowledge Discovery in Databases
SNN	Simulated Neural Networks
ANN	Artificial Neural Networks
PCA	Principal Component Analysis
KNN	K-Nearest Neighbours
Gantt	Gantt Chart
GB	Gradient Boosting
ROC	Receiver Operating Characteristic
CPU	Central Processing Unit
GPU	Graphics Processing Unit
RAM	Random Access Memory
OS	Operating System
GB	Gigabytes
BIOS	Basic Input/Output System
PhD	Doctor of Philosophy
NaN	Not a Number
RMSE	Root Mean Squared Error
DT	Decision Tree
EDA	Exploratory Data Analysis
UX	User Experience
B40	Bottom 40% (Income Group)
M40	Middle 40% (Income Group)
T20	Top 20% (Income Group)

Chapter 1

Introduction

In this chapter, we present the background and motivation of our research, our contributions to the field, and the outline of the thesis.

In today's era of utilizing artificial intelligence and machine learning, are impacting various industries and changing the way human solve problems. They allow us to analyze large amounts of data, identify patterns and make predictions like never before. Advances in artificial intelligence and machine learning allow us to use these technologies to automate tasks, improve workflows and make decisions based on data, thereby increasing efficiency and encouraging innovation. In this project, the goal is to understand the dynamics of income earnings as crucial as it directly impacts the lives of individuals across various strata of society. As one's income grows, the income effect predicts that people will begin to demand more. The dynamics of income earnings are integral to the lives of individuals, impacting their well-being, access to resources, and overall quality of life. Understanding these dynamics is essential for formulating policies that address the diverse needs of a population [1]. In the United States, they are concerned about the glaring wealth and income inequality. A study shows that stark contrast not only influences individual lives but also contributes to societal imbalances. In the rapidly evolving landscape of financial analysis, exploring innovative methodologies has become imperative. Income prediction can significantly aid governments in better understanding the welfare of their citizens by providing insights into economic disparities. For instance, if the prediction model identifies a concentration of low-income households in a specific region, the government can tailor social assistance programs, such as food subsidies, housing support, or educational grants, to address the unique needs of that community [2]. By exploring the intersection of artificial intelligence, machine learning, data mining, and income analysis provides a unique vantage point to address and potentially mitigate this formidable challenge [3]. The disparities in income levels can significantly influence access to education, healthcare, and opportunities for personal and local economic growth in a specific area. One compelling reason to address growing economic inequality is because it potentially reduces poverty globally. By gaining a deeper understanding of income dynamics, we can identify strategies to improve the socioeconomic status of marginalized communities, thereby contributing to overall global poverty reduction [4]. According to a paper titled "A machine learning approach

to improving occupational income scores", published by Martin Saavedra and Tate Twinam, they addressed potential bias induced by Occupational Score Index (OCCSCORE), employed a machine learning approach to develop a newly adjusted income score based on industry, occupation, race, sex, age, and state of residence, and have provided estimates closer to proper earning regressions [5]. Thus, by utilizing the adult income dataset, the relationship between the attributes mentioned above is proven to have a significant strong bond towards an area's economic level. From there, the policymakers could play a pivotal role in shaping the most suitable economic policies by understanding the insights of the finding in income earnings of the certain area. Beyond traditional methods, the research delves into other development tools for predicting income levels. By exploring innovative approaches, this project aims to enhance the accuracy and scope of income predictions, contributing to more effective economic planning.

1.1 Problem Statement and Motivation

Problem Statement 1: Underutilization of Machine Learning for Income Prediction

Current online income platforms are less use of various Machine Learning algorithms. For example, there are several online platforms on reviewing adult income named “PayScale”, “Glassdoor”, the dashboards didn’t implement Machine Learning to provide insight on income description, in fact the techniques used for the dashboards are descriptive statistic approach. According to the Pathrise resources, a review was written for Payscale dashboard. “The accuracy of Payscale's compensation data is the primary topic of discussion in online reviews. Many people believe that the Payscale salaries are not as high as the industry average. This gives the impression that they are paid significantly more than they actually are. This can be particularly annoying if the business uses Payscale services” [6]. Secondly for Glassdoor tool, "I've been looking at Glassdoor for years. It's gone downwards continuously and recently it's just become unuseable - I can't even get past the data collection screen because it can't agree that London is a valid city. Mandating data entry was already annoying even without the aggravation of not accepting it, but that's it I'm just heading to some other competitor sites which provide the same service with less fuss.” mentioned by Paul Oprea in Trustpilot’s reviews to Glassdoor [7]. The dashboards do not apply the use of data mining algorithms, causing deeper insights and more accurate predictions to be missing. The dashboards are more

towards collecting data and displaying the attributes linked from database for the visitors to the dashboard viewing.

Problem Statement 2: Inadequacies in Predictive Tools for Malaysian Income Dynamics

"Currently in Malaysia, the federal government revenue is only using forecasting. This can cause large forecasting errors. Though it can be overcome using predictive analytics. Since there are many machine learning methods available, the appropriate methods can be identified to do the prediction [8]" stated by Nadzira, Aliza Sarlan and Norshakirah Aziz, they mentioned in year 2022 that currently the Malaysia federal government did a lot forecasting works but also did can cause large error. Also, refer to the Malaysia Digital Economy Blueprint, officially published by the Malaysia Kementerian Ekonomi in year 2021, the plan does state in five years would want to enhance digital tools and technology towards digital economy or household income levels under the section named "My device" programme [9]. But predictive tool for Malaysian Income Level did not find throughout the blueprint and the internet by searching "Malaysia" "Predictive Tool" "Income analysis". The potential of machine learning to optimize revenue expectations remains underexploited, highlighting the gap in leveraging advanced algorithms to extract valuable insights from data sets relevant to Malaysian income analysis. There's an existing tool dashboard named OECD Income (IDD) and Wealth (WDD) Distribution Databases shows a world map that breaking down the poverty levels indicated (Gini Coefficient, Top 20% vs Bottom 20%, Relative Income Poverty) for various countries, example like US, Canada, Russia, China, etc. However, the dashboard does not provide the insights of poverty of Southeast Asia countries, especially for Malaysia.

Problem Statement 3: Absence of Visualization Interfaces/Features for User Data Exploration

"The visual components of a dashboard are referred to as the dashboard user interface design. It consists of the elements that facilitate user interaction with your dashboard, such as the bars, charts, and filters. Charts (bar, line, and pie charts), graphs, maps, and tables are examples of data visualizations" [10]. A website named SelectHub describes Salary.com as a compensation suite. It stated, "Nearly 53% of reviewers think that additional location-specific features and

compensation models would enhance the system dashboard. Most reviewers said that implementation is confusing, complex and requires sufficient training before users can leverage all features." [11]. Proved Salary.com, as one of the predictive income websites, does not have enough features or user friendly for user data exploration. A user-friendly dashboard will contain dynamic charts, graphs, and maps which enable users to explore and manipulate data, gaining a deeper understanding of the relationships between variables that affect income levels.

Problem Statement 4: Lack of Integration of AI in Providing Economic Insights

The potential of AI and machine learning to analyze large datasets and identify complex patterns in income data remains largely untapped. Research did by Pol Borrellas and Irene Unceta in 2021 declared that "While machine learning identifies patterns based on correlations rather than causal relations, poorly trained models may capture both signal and noise through confounding variables. This raises concerns about reliance on protected features like ethnicity and gender, even when not explicitly included. These variables can be redundantly encoded in rich feature sets, leading to potential biases in economic insights. For instance, the correlation between zip code, income, and ethnicity in some countries may result in erroneous decisions hindering access to resources, even with apparently unbiased training datasets." [12]. The statement above proofs that currently economic insights from machine learning field have yet fully been utilized. It relates to AI integration on providing economic insight based on income level are still lack in the market.

Motivation

The idea of "A new perspective on Income Earnings by AI" was got from a few news articles published by Sinchew News website on August 3, 2023.

The first news reported "Despite the fact that nearly 80% of low-income households or those with a monthly income of RM2,000 and below were assisted, a quarter were unable to access the cash disbursed. Outrageously, more than one-third of households earning more than RM10,000 a month received cash." and in response to this issue, the Bank Negara recommended that Malaysia should find a balanced approach to its future policy agenda,

including strengthening infrastructure, digitization, improving the efficiency of government assistance, and making access to information more equitable [13]. It brought out the issue of data analyze stage, accuracy lacked in targeting different income levels group is critical, and that shown a machine learning model needed to handle this problem.

The second news reported “Prime Minister Datuk Seri Anwar said he welcomed the announcement by German semiconductor company Infineon Technologies (Infineon) of its plan to build the world's largest 200mm silicon carbide plant in Kedah Kulim.” [14]. The keywords of “German Semiconductor company", "plan to build”, “in Kedah Kulim”, linked to some questions deep in mind, “How does the company decided to build plant in the specific location?” and “Why did the company chose this specific location to build?”. Followed by the news stated “economic”, it inspires the idea of the condition for targeting economic development in a certain area should clearly identify the SWOT analysis for that specific area. From this, an idea can be received by understanding the relationship between high income chance provided by the foreign company and identify insufficient demand and needs from that area are important and closely connected.

Summarize the ideas and concepts above, the main reason of conducting this project is to wish the final machine learning model could enhance economic development. Through deeply understanding income analysis, the model built would help policymakers to determine effective policies to lower local income inequality and boost the economic growth. By applying data mining techniques to income analysis, it would facilitate the development of fresh perspectives for study and exploration. By analyzing large datasets and identifying patterns in Malaysia’s area that were previously unknown, a deeper understanding of income analysis and how they impact individuals and societies can be discovered in future.

1.2 Objectives

Main Objective:

The project is going to utilize machine learning. The main purpose of machine learning is to develop and deploy a predictive dashboard that anticipates the income value from Ipoh citizens. This will help government policymakers make more informed decisions about economic issues by giving them accurate information. The milestones are planned to be done by following the structure below.

1. To form an income dataset values that collected from the Ipoh citizens so that it addresses the specific relationship between income and economic issues.
2. To develop a Machine Learning model that can predict the Ipoh income level using Data Mining Techniques.
3. To develop an interactive dashboard that provides graphs, histogram, box plots on showing clear picture of the income data to policymakers.
4. To develop a dashboard that can assist policymakers to make the most suitable effective decision on economic development direction based on the income dataset.

Sub-Objectives:

Data Collection and Preprocessing:

The goal of this phase is to gather relevant income data from residents of Ipoh. To achieve this, we will provide a Google Form for collecting data in an organized manner. This approach ensures data integrity and facilitates preliminary preprocessing. By using these forms, we can streamline the data collection process, making it easier to manage and analyze the information obtained from residents. Additionally, the structured format of the forms helps maintain consistency and accuracy in the collected data, laying a solid foundation for subsequent preprocessing steps.

Data Anonymization:

The goal of this phase is to ensure informed consent, adhere to data governance, and comply with ethical standards. To achieve this, we will utilize Google Form to facilitate secure data storage with encryption and access controls. By employing these platforms, we can protect individuals' privacy and guarantee ethical data handling throughout the course of an income analysis project. Furthermore, anonymizing the data through submitting an anonymous form allows us to minimize the attributes associated with each respondent. This approach builds confidence in our project's data handling procedures and demonstrates our commitment to ethical practices in research.

Algorithms Selection and Implementation:

The goal of this phase is to implement a diverse range of Data Mining Algorithms for income data analysis. To achieve this objective, we have selected a variety of algorithms to apply to the modeling process. These algorithms include Regression analysis, Decision Trees, Random Forest, Gradient Boosting, Support Vector Machines (SVM), Neural Networks (Deep Learning), K-Means Clustering, Association Rule Mining, Principal Component Analysis (PCA), and Ensemble Methods. Each algorithm is chosen based on its ability to contribute to accuracy, robustness, understanding model behavior, and handling complexity. By employing this comprehensive set of algorithms, we aim to explore various aspects of the income data and uncover valuable insights to inform decision-making processes.

Predictive Model Development:

The goal of this phase is to develop forecasting models that classify groups of demographics and estimate income levels. To achieve this objective, a structured process will followed in Data Mining Techniques, encompassing data exploration, preprocessing, validation, tuning, evaluation, and deployment. During the model training stage, we will employ Regression and Classification techniques based on the specific scenarios and desired output. This approach ensures a systematic and thorough development process, allowing us to build accurate and reliable predictive models capable of providing valuable insights into income demographics.

Dashboard Interface Creation:

The goal of this phase is to develop an interactive dashboard to serve as the tool's interface. To achieve this objective, we will utilize Python language to create a user-friendly interface tailored for government officials, policymakers, and data mining users. The dashboard will prioritize guided functionalities, clarity, and ease of use, ensuring that policymakers and data mining users can easily navigate and utilize the tool effectively. By employing these technologies and design principles, we aim to provide a visually appealing and intuitive interface that enhances user experience and facilitates informed decision-making processes.

Integration of Visualization Components:

The goal of this phase is to integrate various graphs, pies, and bar charts to enhance the user experience and facilitate data review. To achieve this objective, Data Visualization results will be extracted, and visualization components will be implemented to represent predictive model results. Additionally, other types of visualizations will be explored to provide richer insights. By integrating these components, users will have a more interactive and informative experience, enabling them to gain valuable insights from the data more effectively.

Performance Monitoring Functionality:

The goal of this phase is to develop functionality guidelines for users to monitor the model’s performance directly on the dashboard. To achieve this objective, users will be provided with tools to monitor the model's performance in real-time, enabling them to track how well it aligns with actual situations. Additionally, the functionality will allow for updates with fresh information or enhanced algorithms, ensuring that the model remains accurate and relevant over time. By implementing these features, users can continuously assess the model's effectiveness and make informed decisions based on its performance.

Documentation Feature:

The goal of this phase is to incorporate documentation features into the modeling process to ensure future reference. This involves adding a feature that allows users to record every step of the modeling process and download the results page. By providing transparency and facilitating future enhancements or revisions, this feature enables users to review and understand the modeling process comprehensively. Additionally, it allows for easy reference and documentation, ensuring that the project's progress and outcomes are well-documented for future use.

Testing and Validation:

The goal of this phase is to ensure everything works accurately by conducting extensive testing. This involves using test cases to identify and fix any problems and testing the entire system. Outcomes will be verified in comparison to expectations, and input will be gathered, especially from prospective end users such as policymakers and data mining users. Through comprehensive testing and validation, the aim is to guarantee the functionality and accuracy of the system.

1.3 Project Scope and Direction

Targeted User

The project scope for this tool is to assist the government and policymakers who in forecasting and preparing strategies or policies based on the income levels of individuals and citizens. After a period (setting four months) of collecting data from local citizens, targeting Ipoh area, the tool attempts to make it easier to identify and profile different groups within the population. Government policymakers who are looking for information to create rules and policies that meet the various needs of various demographics groups are expected to be the tool’s main users.

Type of Product

Finalizing the idea from existing research, journals documentation, the final product for this project would be defined as a predictive tool that utilizing exist Data Mining Techniques, helps to analyze the income data from Ipoh citizen and provide the algorithms result to the future

create interactive dashboard and could display various types of graphs, charts for visualization and enhance the decision support, empowering the policymakers with accurate insights for decision-making on the economic related issues.

Focus of the Project

The preliminary focus will be on gathering and preprocessing income data for analysis and implementing data mining algorithms for predictive modeling. The report focus will be on collected more data, searching more possible preprocessing method to lower down RMSE and discovering more dashboard development methods, creating visualization and serve it as the tool's interface, and integrating visualization components for effective data representation. Testing and Validation stage will be conducting thorough testing to ensure accuracy and functionality, targeted potential collecting feedback users are policymakers.

This tool would have contained the features as mentioned below:

1. Data Analysis

Utilize the existing Data Mining Algorithms to analyze income data from Ipoh citizens. The algorithms involved Regression analysis, Decision Trees, Random Forest. Gradient Boosting, Support Vector Machines (SVM), Principal Component Analysis (PCA).

2. Predictive Modeling

Develop predictive models based on the analyzed data to estimate income levels and categorize the various demographics groups. Developing a predictive model could briefly explain in these few steps. First Step is to identify the target variable, in the study, the variable would be "incomeLevel". Secondly, gather the relevant data that includes inputs and the target variable/attribute. Thirdly, explore and preprocess data to understand its characteristics and distributions, handle missing values and outliers, or irrelevant features, apply one hot encoding. Continuously proceed to split data, choose regression or classification, train the model, validate and tune, evaluate the model, interpret the results, and lastly deploy the model.

3. Dashboard Integration

Create a simple interactive dashboard to serve as the interface for the predictive tool and provide reports for the people who need to use and review. Mainly serve policymakers and data mining users to review. So, the design will be more towards to cleaner and clearer, easy to use, and provide guidelines after entering the dashboard and while manipulate the functionality there have. Also, another functionality of this dashboard in future is for the user to monitor the model's performance, so that the model could suit any real word issue, and that the dashboard needed to be updated with new data or improved algorithms. Lastly, provide the feature of document the modeling process of reproducibility and future reference.

4. Graphs, Histogram and Box Plots Visualization

Display various types of visualization like graphs and charts to represent the results of the predictive models. If possible, would consider developing more than three visualization types in future, to explore more valuable perspectives on the income data. The mission of providing pies, graphs and charts is to provide the information from the income data and from that, provide the insights for policymakers to make informed decisions on the local economic development.

1.4 Contributions

Understand our people, Malaysians' income level is important, from this aspect, it helps to get insights on solving the shortcomings of our country economic direction needs in every area. This project initiative represents to address critical inadequacies in existing predictive tools for Malaysian income analysis and the lack of integration of AI in providing economic insights. The introduction of a new artificial intelligence (AI) method that uses sophisticated machine learning and data mining techniques to assess and predict the income levels of Ipoh residents. Hence, dataset specifically needed to build by utilizing the collected data from Ipoh citizens because there is no existed dataset in Ipoh, even the entire nation, Malaysia. The project aims to create a predictive tool and strategically addresses the underutilization of machine learning in income anticipation for Ipoh citizens and provide a predictive tool for the government policymakers to make effective policy decisions. From that on the government would be able to make the most suitable policy decisions on shorting the distance between high-income profile citizens and low-income profile citizens.

1.5 Report Organization

This report is organized into 6 chapters: Chapter 1 Introduction, Chapter 2 Literature Review, Chapter 3 Methodology, Chapter 4 Data Analysis and Findings, Chapter 5 Discussion and Implications, and Chapter 6 Conclusion.

Chapter 1, Introduction, provides an overview of the project, including the problem statement, project background and motivation, project scope, project objectives, project contribution, highlights of project achievements, and the organization of the report.

Chapter 2, Literature Review, examines existing studies and data analysis methods relevant to the project. It evaluates the strengths and weaknesses of current methodologies and identifies gaps that the project aims to address.

Chapter 3, Methodology, details the approach and techniques used in collecting and analyzing the data. It includes a description of the data sources, the data collection process, and the analytical methods applied.

Chapter 4, Data Analysis and Findings, presents the results of the data analysis. This chapter provides a comprehensive review of the key findings and discusses how these findings relate to the research objectives.

Chapter 5, Discussion and Implications, interprets the results in the context of the project's objectives. It discusses the implications of the findings for policymakers and other stakeholders and suggests potential areas for further research or action.

Chapter 6, Conclusion, summarizes the key insights from the project, reflects on the achievements and contributions, and outlines the overall impact of the work. It also suggests directions for future research and potential improvements.

Chapter 2

Literature Review

In this section, we will review the various technologies that are relevant to our project on "A New Perspective on Income Earnings by AI." These technologies are essential for developing and implementing the AI models that will predict income levels and support policymakers.

2.1 Review of the Technologies

2.1.1 Programming Language

The one and only focus programming language involved in this project is Python. As Python has emerged as one of the leading programming languages for artificial intelligence (AI), data science, and machine learning (ML) applications [15]. Its simplicity, readability, and extensive library ecosystem make it a welcoming tool for developers and researchers to implement complex algorithms and create efficient models [15]. In this project, Python has been chosen for its compatibility with various machine learning frameworks and its flexibility in data manipulation and visualization [16].

Python's popularity in AI and ML stems from its community support, vast collection of open-source libraries, and its cross-platform capabilities [17]. Python’s easy-to-read syntax makes it accessible to developers with varying levels of expertise, allowing for faster prototyping and collaboration [18].

Python is particularly well-suited for this project due to the following reasons:

1. Data Handling: Python offers libraries such as Pandas and NumPy, which allow for efficient data manipulation, cleaning, and transformation [16]. These capabilities are essential when dealing with large income datasets for analysis and predictions, including retrieve dataset in pandas library [15].

Figure 2.1.1 pandas has been utilized to retrieve dataset from real python

```
Python
>>> import pandas as pd
>>> nba = pd.read_csv("nba_all_elo.csv")
>>> type(nba)
<class 'pandas.core.frame.DataFrame'>
```

2. Machine Learning: Scikit-learn, TensorFlow, and PyTorch are some of the most widely used machine learning libraries that integrate seamlessly with Python [19]. These libraries offer pre-built algorithms and tools for developing predictive models, making them crucial for implementing AI-based income prediction.

Figure 2.1.2 Scikit-learn has been implemented for feature selection and preprocessing from geeksforgeeks.org

```
Python
from sklearn.feature_selection import SelectKBest, chi2

# Apply SelectKBest with chi2
select_k_best = SelectKBest(score_func=chi2, k=2)
X_train_k_best = select_k_best.fit_transform(X_train, y_train)

print("Selected features:", X_train.columns[select_k_best.get_support()])
```

Output:

```
Selected features: Index(['petal length (cm)', 'petal width (cm)'], dtype='object')
```

3. Visualization: Python's Matplotlib and Seaborn libraries facilitate data visualization, enabling clear insights from complex datasets [19]. These tools are essential for understanding the distribution and trends within income data and presenting findings to policymakers effectively.

Figure 2.1.3 Matplotlib implemented for showing and specified diagram size from O'Reilly Media

```
# ------- file: myplot.py ------
import matplotlib.pyplot as plt
import numpy as np

x = np.linspace(0, 10, 100)

plt.plot(x, np.sin(x))
plt.plot(x, np.cos(x))

plt.show()
```

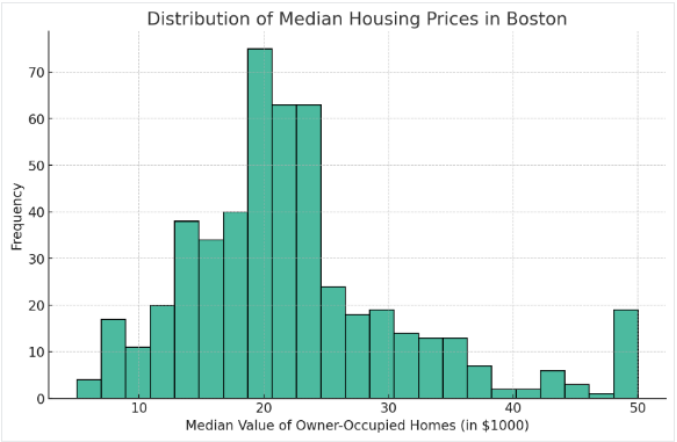
Figure 2.1.4 Seaborn for Histogram showcasing from datacamp

```
plt.figure(figsize=(10, 6))
sns.histplot(data=df, x='MEDV')
plt.title('Distribution of Median Housing Prices in Boston')
plt.xlabel('Median Value of Owner-Occupied Homes (in $1000)')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```

Explain code POWERED BY datalab

This command will display the histogram for the MEDV column.

Here's the generated plot from this code:



At the same time, this project development tools such as Jupyter Lab and Visual Studio Code were used in the Python Environment. These integrated development

environments (IDEs) offer flexibility and convenience for coding, debugging, and running Python scripts.

2.1.2 Algorithm

The three main algorithms applied in the analysis and prediction of income earnings: Linear Regression Models, Decision Trees, Random Forests, and Support Vector Machines (SVM). These algorithms are selected based on their ability to handle structured data, their interpretability, and their performance in predictive tasks. First and foremost, Linear Regression and Multiple Linear Regression are classical and widely used techniques for predicting continuous outcomes [57]. In this project such as income would be the continuous outcomes which suit the case of using Multiple Linear Regression. These models are valuable for understanding how individual and multiple factors influence income levels [56]. Secondly, Decision Trees and Random Forests are powerful algorithms for modeling complex relationships between features in structured data [55]. They are particularly well-suited to datasets with non-linear interactions between predictors and the target variable [58]. Third, Support Vector Machines (SVM) are known for their ability to find the optimal boundary between data points in classification tasks, but they can also be adapted for regression purposes in the form of Support Vector Regression (SVR) [60].

Linear Regression

This algorithm assumes a linear relationship between an independent variable such as years of experience and the target variable (income) [61]. It provides a clear and interpretable baseline for income prediction, allowing us to quantify how changes in one independent variable affect income.

Multiple Linear Regression

Extending beyond simple linear regression, multiple linear regression accounts for several predictors simultaneously such as education level, job type, and experience [20]. It is suitable for analyzing more complex interactions and is applied in this project to understand how different socioeconomic factors, in combination, affect income

levels. By incorporating multiple variables, this model gives a more nuanced view of income determinants.

Decision Trees

A decision tree divides the data into smaller, more manageable subsets (nodes) by selecting the most informative features. Each branch of the tree represents a decision rule based on a feature, leading to predictions about income [21]. The interpretability of decision trees makes them ideal for understanding the relationship between income and factors like job sector, experience, or education.

Random Forests

As an ensemble method, random forests combine the predictions of multiple decision trees, thus reducing the risk of overfitting and improving the robustness of the model [59]. In this project, random forests help identify the most important features influencing income while maintaining accuracy. The model's ability to handle high-dimensional data with multiple features makes it a powerful tool for income prediction.

Support Vector Regression (SVR)

SVR aims to fit a model within a margin of error that best approximates the relationship between features and the target variable. For income prediction, SVR can be applied to identify patterns in income distribution by adjusting the hyperplane to minimize prediction errors [23]. SVR is particularly useful in scenarios where the relationship between variables is complex and non-linear. Its ability to handle outliers and provide precise predictions makes it an ideal choice for modeling income variability across different demographic groups.

2.1.3 Summary of the Technologies Review

Table 2.1.1 Literature Python Libraries

Python Libraries Name	Relevant to the project
Pandas	A data analysis library used for manipulating the income dataset and preparing it for machine learning models.
NumPy	A numerical computation library used for efficient data processing and matrix operations.
Scikit-learn	A machine learning library used to build and evaluate income prediction models. It provides various algorithms such as decision trees, random forests, and regression models.
Matplotlib & Seaborn	Visualization libraries used for plotting data distributions, trends, and model outputs, making it easier to interpret the results of the predictive models.

Python's ecosystem, including its extensive libraries and development environments like Jupyter Lab and Visual Studio Code, offers the flexibility and power needed to develop robust AI models and interactive dashboards for income prediction in this project. The integration of Shiny for Python further enhances the project's capability to present actionable insights to policymakers through an intuitive, visual interface.

Table 2.1.2 Literature Python Algorithms

Algorithms	Relevant to the project
Linear Regression	Assumes a linear relationship between an independent variable (e.g., years of experience) and income. Provides a clear, interpretable baseline for predicting income.
Multiple Linear Regression	Extends linear regression by incorporating multiple predictors such as education level, job type, and experience to analyze the combined impact on income.

Decision Trees	Splits data into smaller subsets based on the most informative features, allowing for clear interpretation of income prediction based on factors like job sector, experience, and education.
Random Forests	An ensemble method combining predictions from multiple decision trees. Helps prevent overfitting and improves model robustness for income prediction, especially in high-dimensional data.
Support Vector Regression (SVR)	Aims to fit a model within a margin of error to best approximate income distribution. Particularly useful for complex, non-linear relationships and handling outliers.

2.2 Review of the Existing Systems/Applications

There are certain efforts have been made by the researchers in the past to use machine learning models to predict income levels.

2.2.1 Paper named "Artificial Intelligence and Economic Development: An Evolutionary Investigation and Systematic Review".

The paper named "Artificial Intelligence and Economic Development: An Evolutionary Investigation and Systematic Review" suggests that the adoption and integration of AI into operating systems aim to assist humans or even allow for utterly AI-driven decision-making, which indirectly informs policymakers by providing more efficient and potentially more informed decision-making processes.

The conclusion emphasizes the importance of policymakers paying attention to the safety risks and security concerns associated with AI technology, including ethical security, technical security, and data security. A suggested action for policymakers is to increase financial investment and policy protection to foster the development of the AI industry while also establishing legal provisions to address data privacy. Additionally, creating a professional code of ethics for AI to incorporate human ethical

guidelines and humanism is recommended, which could guide policymakers in developing policies that encourage ethical AI practices [24].

The project was likely an empirical investigation, targeting adult income data from the census to develop predictive models for income classification. The authors mentioned that in the initial stages of their project, they worked on the Census dataset and engaged in data preprocessing. This step likely involved cleaning the dataset, handling missing values, and transforming the data in preparation for analysis. Additionally, they focused on developing an understanding of the data and its useful features to explain the variances. This process is crucial for ensuring the data's quality and relevance for income classification. Chockalingam and colleagues delved into the Adult Dataset examining it thoroughly using a range of machine learning methods, like regression, naive Bayes, decision trees and more. They compared the capabilities of these models [25]. The project holds significance in the realm of data science and predictive modeling. By utilizing adult census data, the researchers aimed to predict income levels, which can have implications in various fields such as economics, sociology, and public policy. Moreover, the development of accurate income classification models can contribute to targeted interventions, resource allocation, and socioeconomic research.

2.2.2 Paper named “A Statistical Approach to Adult Census Income Level Prediction”.

In recent years, addressing income inequality has become a significant concern worldwide, with particular attention to the United States. Various studies have explored the application of machine learning and data mining techniques to predict income levels, aiming to contribute to the reduction of economic disparity. One notable contribution in this field is the paper titled "A Statistical Approach to Adult Census Income Level Prediction" by Navoneel Chakrabarty and Sanket Biswas.

Chakrabarty and Biswas present a comprehensive analysis of income prediction utilizing the UCI Adult Dataset. Their study builds upon previous research efforts, which have explored diverse machine learning models for income prediction tasks. For

instance, Chockalingam et al. conducted an extensive comparative analysis of machine learning models, including logistic regression, decision trees, support vector machines, and gradient boosting, among others. Similarly, Bekena employed the random forest classifier algorithm to predict income levels, while Topiwalla utilized complex algorithms such as XGBOOST and stacking of models for enhanced accuracy [26].

In their study, Chakrabarty and Biswas address key preprocessing steps, including handling missing values, encoding categorical features, shuffling the dataset, and splitting it into training and testing sets. They adopt the gradient boosting classifier (GBC) as the learning algorithm, a powerful ensemble method that sequentially constructs predictors to correct errors from previous iterations. The authors meticulously tune the GBC model using grid search algorithm to optimize hyperparameters, achieving impressive results [26].

The implementation details highlight the utilization of Python's Scikit-Learn machine learning toolbox and plotting libraries for data preprocessing, model development, and visualization. Notably, the results obtained by Chakrabarty and Biswas showcase significant performance metrics, including high training and validation accuracies, recall, precision, and F1-score. Moreover, the area under the receiver operator characteristic curve (AUROC) indicates strong predictive capability of the model [26].

In conclusion, the paper by Chakrabarty and Biswas presents a robust statistical approach to adult census income level prediction, demonstrating the effectiveness of ensemble learning techniques, specifically gradient boosting, in addressing income inequality. Their study contributes to the existing literature by achieving a validation accuracy that surpasses previous benchmarks. Furthermore, the authors propose avenues for future research, such as hybrid models incorporating machine learning and deep learning, to further improve predictive accuracy without compromising efficiency.

2.2.3 Paper named “Income Prediction Using Decision Tree”.

The paper on "Income Prediction Using Decision Tree" on Kaggle provides a comprehensive overview of applying decision tree methodology to predict individual income levels. The focus is on classifying an individual's income and determines the factors that contribute to variations in income levels. This paper aims to understand the role of variables in income prediction modeling and evaluates the predictive performance of decision trees. Bekena used the random forest classifier algorithm to forecast income levels for individuals [27]. The decision tree methodology is applied to the Adult Income dataset obtained from the UCI Machine Learning Repository. The dataset from Kaggle, titled "Adult Census Income," consists of 15 features and more than 32,000 rows for income prediction. The prediction task involves determining whether a person's income exceeds 50k or not, portraying a binary classification problem.

2.2.4 Paper named “Machine learning on UCI adult data set using various classifier algorithms and scaling up the accuracy using extreme gradient boosting”.

Topiwalla employed algorithms such as XGBOOST and random forest along with model stacking techniques to improve prediction accuracy. This involved combining models, like stacking on XGBOOST and SVM stacking on logistic regression [28]. The paper explores the application of various machine learning algorithms on the UCI Adult data set, aiming to enhance accuracy using extreme gradient boosting. The study discusses the use of classifiers such as the Gradient Boosting Classifier Model, Logistic Regression, and Support Vector Machines, among others, to predict census income levels and address imbalanced classification. Additionally, the paper highlights the significance of the eXtreme Gradient Boosting Algorithm, XGBoost, in achieving improved accuracy and scalability in machine learning models. Furthermore, the authors emphasize the potential of gradient-boosted decision trees and deep neural networks in handling tabular data and addressing

classification challenges. Overall, the paper underscores the importance of leveraging advanced machine learning algorithms, like XGBoost, for efficient data analysis, high accuracy, and scalability in various applications.

2.2.5 Paper named “Income Prediction via Support Vector Machine”.

The paper titled "Income Prediction via Support Vector Machine" was presented by Alina Lazar at the International Conference on Machine Learning and Applications (ICMLA) in 2004. The main focus of the paper was on the application of Support Vector Machine (SVM) for the prediction of income [29]. The use of SVM, a supervised learning model, for income prediction highlights the significance of machine learning techniques in the domain of economic prediction. The paper likely delves into the methodology, dataset, features, and results related to the implementation of SVM for income prediction. This contribution would have provided valuable insights into the potential use of machine learning algorithms in predicting income levels. The application of Support Vector Machine (SVM) for income prediction reflects the growing interest in leveraging machine learning techniques for economic forecasting and analysis. By employing SVM, a powerful and flexible algorithm, researchers aim to build accurate models that can predict income levels based on relevant features or attributes. The utilization of SVM in this context underscores the potential for advanced computational methods to contribute to the development of predictive models in socio-economic domains, thus paving the way for enhanced decision-making and resource allocation. Lazar utilized principal component analysis (PCA). Support vector machine (SVM) approaches to create and assess income prediction data derived from the U.S. Census Bureaus Current Population Survey.

2.2.6 Paper named “Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques”.

The paper "Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques" explores the application of diverse classification techniques on the Adult Income Census dataset from the Kaggle website. It aims to predict whether an individual's income exceeds $50,000 per year based on demographic and socioeconomic characteristics [30]. The dataset provides valuable information on economic trends and income distribution, enabling various real-world applications such as targeted marketing, credit risk assessment, and public policy analysis. The project involves multiple steps such as data preprocessing, exploratory data analysis, applying machine learning models, and assessing model performance, showcasing the significance and complexity of income prediction using census data. The census income dataset, credited to Ronny Kohavi and Barry Becker, contains 14 input variables, with the classes '>$50K' and '<=$50K' for binary classification, rendering the dataset as an imbalanced classification task. It involves personal details such as education level to predict an individual's income level, showcasing the varied characteristics and factors contributing to income levels. The project employs machine learning algorithms such as Perceptron, H2O.ai, and supervised learning methods to build accurate income prediction models, ensuring the robustness and reliability of the prediction. Furthermore, the paper delves into the significance and use cases of the census income dataset in diverse industries such as healthcare, real estate, education, non-profit, retail, insurance, and marketing, highlighting the widespread applicability of the dataset for a multitude of business and societal needs. The census income data set thus proves to be a valuable resource to understand income inequality and to design policies for fostering economic growth and reducing poverty.

Amidst the technical complexities, the paper also emphasizes the social and economic relevance of census data, portraying its pivotal role in shaping public policies, business strategies, and societal welfare. It underlines the significance of demographic and socioeconomic characteristics in predicting income levels, thereby contributing to the broader understanding of income distribution and economic trends.

Overall, the paper provides a comprehensive overview of utilizing census data for income prediction and highlights its implications in various domains, reinforcing the importance of leveraging such datasets for societal and economic developments.

2.2.7 Paper named “Predicting Earning Potential using the Adult Dataset”.

Haojun Zhu's study "Predicting Earning Potential using the Adult Dataset" presents a robust exploration into predicting earning potential based on the Adult Income Dataset. The research report delves into the fundamental aspects of building a predictive model and the statistical approach used in predicting income ranges.

Zhu elaborates on the statistical approach and the use of tree-based models, which notably perform well in predicting income ranges for classification problems. The study establishes the role of supervised learning in the prediction task, particularly classification [31]. Logistic Regression is highlighted as one of the easiest and most commonly used supervised machine learning algorithms for categorical classification. The report provides a comprehensive explanation of the logistics behind Logistic Regression that is well-suited for beginners entering the Machine Learning domain.

Moreover, the research emphasizes the model's evaluation process, providing an in-depth overview of the accuracy values. It demonstrates the significance of achieving a 76% accuracy, which Zhu deems positive for a machine learning prediction model, particularly in the context of a classification regression problem.

In summary, "Predicting Earning Potential using the Adult Dataset" offers valuable insights into machine learning techniques for predicting earning potential. It stands as a commendable resource for learners, providing a detailed step-by-step guide, as well as valuable advice for aspiring data scientists, emphasizing on the importance of understanding the mathematical background while leveraging Python libraries for resolving complex computational problems. The research serves as a foundational resource for individuals interested in the machine learning field, offering guidance on where to begin and how to progress in their journey.

2.2.8 Summary of the Existing Papers

Table 2.2.1 Overview Table for Seven Papers

Paper's Name	Proposed Algorithms	Findings
Artificial Intelligence and Economic Development: An Evolutionary Investigation and Systematic Review [24]	Not mentioned	Emphasizes the importance of policymakers paying attention to the safety risks and security concerns associated with AI technology
A Statistical Approach to Adult Census Income Level Prediction [26]	Logistic regression, gradient boosting classifier	Presents a comprehensive analysis of income prediction utilizing the UCI Adult Dataset
Income Prediction Using Decision Tree [27]	Decision tree, random forest	Focuses on the classification of individual income levels using the Adult Income dataset
Machine learning on UCI adult data set using various classifier algorithms and scaling up the accuracy using extreme gradient boosting [28]	XGBOOST, random forest, GBC, logistic regression, support vector machines, model stacking	Explores the application of various machine learning algorithms on the UCI Adult data set and highlights the significance of leveraging advanced machine learning algorithms
Income Prediction via Support Vector Machine [29]	SMO, SVM	Focuses on the application of SVM for the prediction of income and highlights the significance of machine learning techniques
Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques [30]	Not specified	Explores the application of diverse classification techniques on the Adult Income Census dataset

Predicting Earning Potential using the Adult Dataset [31]	Logistic regression	Elaborates on the statistical approach and the use of logistic regression for predicting earning potential based on the Adult Income Dataset

Analysis for the proposed algorithms in the papers [(Frequency of been used in different papers)]:

Logistic regression (3):

The algorithm is widely used across the papers for income prediction. It is chosen due to its simplicity and efficiency in binary classification tasks, such as predicting income levels exceeding a threshold. It exhibits consistent performance and is often employed as a baseline model for comparison with other advanced techniques.

Decision Tree (1):

The Decision Tree algorithm is extensively utilized in the context of income prediction tasks. It serves as a fundamental machine learning technique for creating predictive models based on demographic and socioeconomic features. Decision Trees offer a straightforward and interpretable model for predicting individual income levels. They are effective in capturing complex decision boundaries and relationships within the data, making them suitable for socioeconomic analysis. By segmenting the data based on feature tests at each node, Decision Trees provide a hierarchical structure that aids in understanding the criteria used for income prediction. This algorithm is valued for its versatility and applicability to various income prediction scenarios. The papers highlight how Decision Trees play a crucial role in predicting income levels by uncovering statistical significance between features and income categories. They are instrumental in identifying key predictors of income levels based on the given datasets. Decision Trees are particularly effective in explaining the reasoning behind income predictions, providing insights into the factors influencing individual income levels. Despite their interpretability and simplicity, Decision Trees may require pruning

techniques to prevent overfitting and ensure optimal predictive performance, emphasizing the need for careful model tuning in income prediction tasks.

Random Forest (2):

Random Forest, an ensemble technique, is applied in income prediction tasks to improve predictive accuracy by generating multiple decision trees based on different subsets of the data. This algorithm is used to address issues of overfitting inherent in single Decision Trees by aggregating diverse models to produce a more robust prediction. Random Forest leverages majority voting or averaging of individual tree outputs to enhance overall predictive performance and mitigate the shortcomings of individual trees. The papers demonstrate that Random Forest significantly enhances predictive accuracy in income prediction scenarios by combining the strengths of multiple decision trees to produce more reliable predictions. The ensemble nature of Random Forest reduces variance and generalizes well to unseen data, leading to improved performance in predicting income levels above specified thresholds. Research highlights the effectiveness of Random Forest in socioeconomic analysis, showcasing lower misclassification rates and improved accuracy in predicting income categories.

Gradient Boosting Classifier, GBC (2):

GBC is extensively utilized in income prediction tasks to refine decision boundaries and optimize predictive models through iterative error correction. This algorithm is known for boosting predictive accuracy by aggregating weak learners into a strong learner, thereby improving classification performance in machine learning tasks. GBC is applied to overcome income inequality challenges and achieve high accuracy in predicting income levels based on demographic and socioeconomic attributes. The research emphasizes the significant role of GBC in enhancing predictive accuracy and scalability in income prediction models. It showcases how GBC excels in refining decision boundaries and reducing errors with each iteration. GBC is particularly effective in addressing income prediction challenges, demonstrating improved

efficiency in data analysis and socioeconomic applications. The algorithm's iterative nature and emphasis on error reduction contribute to its success in achieving high predictive accuracy and optimizing machine learning models for income prediction scenarios.

Extreme Gradient Boosting, XGBOOST (1):

XGBOOST, an extreme gradient boosting algorithm, is a powerful tool employed to enhance prediction accuracy and scalability in various income prediction applications. This algorithm optimizes predictive performance by minimizing prediction errors through boosting techniques and iterative model improvement. XGBOOST is favoured for its ability to handle large datasets, making it ideal for addressing complex data analysis tasks and optimizing predictive models for income prediction. In the paper, XGBOOST emerges as a significant contributor to improving the efficiency and accuracy of income prediction models. It showcases the effectiveness of handling extensive datasets and minimizing prediction errors through boosting mechanisms. XGBOOST plays a vital role in improving scalability, accuracy, and overall predictive performance in socioeconomic analysis and economic prediction tasks. The algorithm's ability to iteratively refine models and enhance predictive accuracy highlights its importance in addressing income prediction challenges and optimizing machine learning approaches for socioeconomic applications.

Support Vector Machine, SVM (2):

Support Vector Machine is applied for the prediction of income levels and is mainly highlighted for its significant role in machine learning techniques within the domain of economic prediction. Employing SVM and its variations adds to the diversity of machine learning techniques in predicting income levels with high precision.

Sequential Minimal Optimization, SMO (2):

Sequential Minimal Optimization is highlighted for its ability to train Support Vector Machines (SVMs) faster by breaking down large Quadratic Programming (QP) problems into smaller constituent problems, allowing for analytical solutions to be obtained. SMO is particularly advantageous for sparse data sets, as it requires linear memory with respect to the training set size, making it suitable to handle extensive data. This algorithm is used to optimize training of SVMs, enhancing their efficiency and practicality for various applications. Therefore, SMO constitutes a significant part of the proposed algorithms, demonstrating its relevance in the context of machine learning on the UCI adult data set.

Model Stacking (1):

A unique approach, model stacking, is applied to improve prediction accuracy by leveraging varied algorithmic outputs. This method is employed to enhance predictive performance by combining outputs from multiple machine learning algorithms and thereby scaling up the accuracy of predictions which is relevance for the context of complexity on the UCI adult data set.

Summary:

The analysis examines the usage of various machine learning algorithms for income prediction across multiple papers. Logistic Regression emerges as a widely employed method due to its simplicity and effectiveness in binary classification tasks, often serving as a benchmark model. Decision Trees are extensively utilized for their interpretability and ability to capture complex data relationships, providing valuable insights into income prediction. Random Forests address overfitting issues by aggregating multiple decision trees, demonstrating enhanced predictive performance. Gradient Boosting Classifier and Extreme Gradient Boosting algorithms refine decision boundaries and minimize prediction errors iteratively, showcasing high accuracy and scalability in income prediction tasks. Support Vector Machines are valued for their

precision in economic prediction, while Sequential Minimal Optimization enhances efficiency in training SVMs, particularly for sparse datasets. Model Stacking emerges as a novel approach to improving prediction accuracy by combining outputs from multiple algorithms. Both GBC and XGBOOST stand out for their effectiveness in addressing income inequality and achieving high prediction accuracy, contributing significantly to optimizing machine learning for socioeconomic applications.

2.3 Review of the relevant existing Dashboards/Tools

There are six tools or dashboards can be reviewed for this project, the tools included PayScale, Salary.com, Glassdoor, Predictive Insights, World Bank's PovCal and OECD's Income Distribution Database.

2.3.1 PayScale

PayScale provides salary information based on job title, location, and experience level. This tool provides the insights for employers and employees, better understand the appropriate salary for each position with the aid of Payscale. Their primary goal when working with businesses is to make sure that wages are commensurate with market value in order to preserve equity and employee retention [32].

Figure 2.3.1 Average salary by job title and location made by PayScale

The PayScale offers insightful information that might help with the project's issue statements, example shown in figure 2.3.1 PayScale is a popular tool for gaining insights into individual salaries and is a valuable benchmarking tool due to its ability to provide estimates of potential salaries. Adding features akin to PayScale's customized pay snapshot could improve user experience and yield more accurate predictions, helping to address the underuse of machine learning for income anticipation and the shortcomings in predictive tools for Malaysian income dynamics. Also, the lack of user-friendly features or interfaces for user data exploration is addressed by PayScale's user-friendly interface, which lets users compare salaries by a number of parameters. The project's prediction tool can enable users to effectively explore and manipulate income data by incorporating comparable interactive elements.

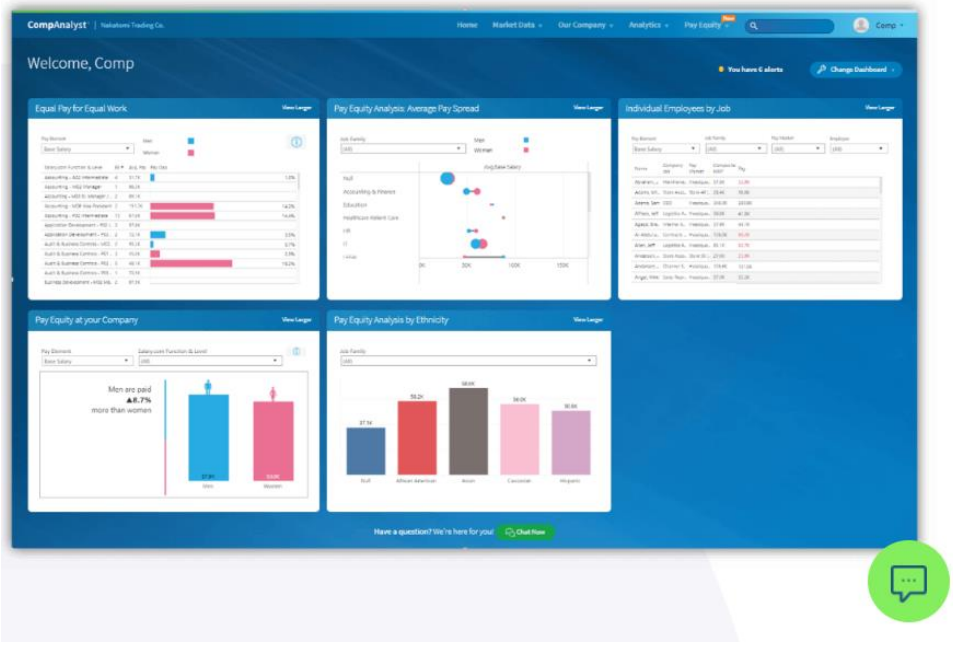
Furthermore, PayScale's approach aligns with addressing the lack of integration of AI in providing economic insights, emphasizing the importance of benchmarking salaries against market values. Integrating machine learning algorithms that PayScale might leverage for its personalized pay snapshots can contribute to more informed economic insights and decision-making. Additionally, PayScale's strategy emphasizes the significance of benchmarking salaries against market values, which is in line with addressing the lack of AI integration in providing economic insights. Including machine learning algorithms, which PayScale may use for its customized pay snapshots, can help make better economic judgments and insights.

2.3.2 Salary.com

Similar to PayScale, Salary.com also provides salary information with details on bonuses, benefits, and other compensation. Features for employers included conpensation decisions, survey pricing intelligence, and consulting works whereas for employees included salary research, personla salary report, job available listings [33]. About 90% of users have left positive reviews for Salary.com, highlighting its prompt and amiable customer service. One of the platform's strongest points is how easily 88% of users can customize the visuals and widgets on their dashboards. Users frequently highlight the validity of analysis and data collection tools while praising the consistency of data insights. Sixty-three percent of reviewers commend the platform's functionality,

which includes customizable features like a minimum wage tool and relo wizard. Highlights include collaboration features that facilitate efficient data sharing and benchmarking, though 63% of respondents mention a learning curve that is lessened by the vendor's support and training materials.

Figure 2.3.2 Customizable dashboard from Salary.com



Salary.com offers a model for improving user experience and engagement in online income platforms with its emphasis on providing excellent customer support and customizable dashboards. The dependable data insights and functionalities as shown in figure 2.3.2 of the platform, along with its helpful features and collaborative tools, point to potential solutions for addressing the underutilization of machine learning and predictive tools for the analysis of Malaysian income dynamics. The difficulties with creating a user-friendly interface, as demonstrated in the instance of Salary.com, emphasize how crucial it is to include dynamic graphs, charts, and maps for efficient user data exploration.

Furthermore, Salary.com's advantages support the necessity of utilizing AI to provide economic insights by highlighting the possible advantages of utilizing machine learning to examine sizable datasets and spot intricate patterns in income data. Online income platforms can improve their predictive tools, user interfaces, and overall efficacy in providing useful economic insights based on income levels by integrating

features that draw inspiration from Salary.com, such as customizable dashboards and dependable data insights.

2.3.3 Glassdoor

Comparison to PayScale and Salary.com, Glassdoor has the worst reviews from users due to the fact that it doesn't emphasize the data integrity and does not handle well on the data collection caused crowdsourced. In another word, overfitting happened. Evidence proof in below, figure 2.3.3 and figure 2.3.4.

Figure 2.3.3 Complaints from Paul Oprea on Trustpilot platform regarding data collection

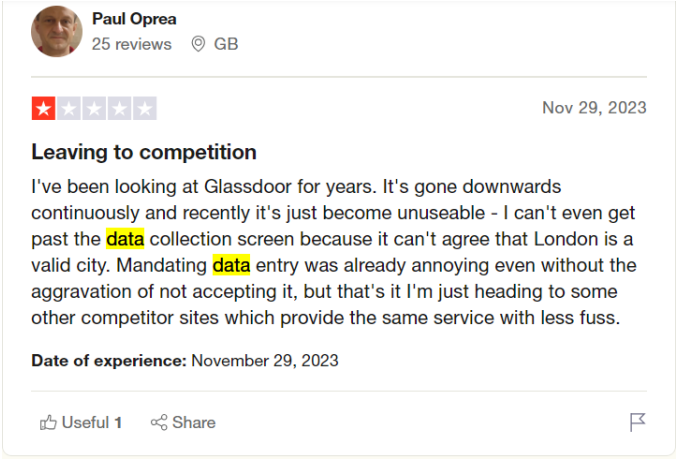
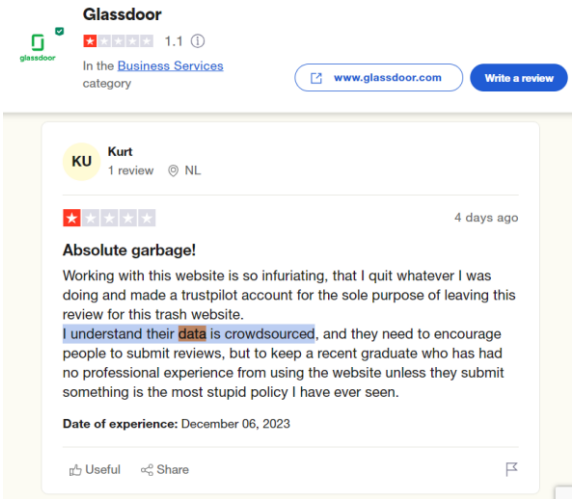


Figure 2.3.4 Complaints from Kurt on Trustpilot platform regarding data crowdsourced



In this Glassdoor review, it reminds of the importance of data collection step has to be verified from the original source, or trustworthy agencies, government, to ensure the accuracy result made from the model.

2.3.4 Predictive Insights

An inventive platform called Predictive Insights provides a plethora of potent features that are intended to transform business decision-making. Organizations can optimize staffing and stock levels, track performance across multiple levels, and identify various customer types and their values with the help of modules like Demand Forecasting, Sales Insight, and Customer Lifetime Value available on the platform [34]. This guarantees cost-effectiveness, targeted efforts at the most valuable customers, and strategic decision-making. For companies looking to increase productivity and profitability, the smooth integration of these modules offers a complete solution.

Comparatively, Predictive Insights uses machine learning algorithms obviously for analyzing customer data. In the realm of underutilization of machine learning for income anticipation, the Sales Insight module of the platform which highlights patterns and anomalies is a prime example of how machine learning algorithms should be implemented. This has the potential to revolutionize already-existing online income platforms, such as Glassdoor and PayScale, by providing deeper insights and more precise forecasts in addition to descriptive statistics.

Predictive Insights establishes a standard for utilizing machine learning techniques in order to address shortcomings in the predictive tools for Malaysian income dynamics. The Malaysian government can utilize the Demand Forecasting and Sales Insight modules of the platform as a model to determine suitable machine learning techniques for revenue forecasting, thereby mitigating forecasting errors. Example visualization in Figure 2.3.5 shown Predictive Insights implement Data Visualization on the staff management and Heatmap for Correlation Coefficient on figure 2.3.6.

Figure 2.3.5 A photo shown Predictive Insights implement Data Visualization on the staff management.

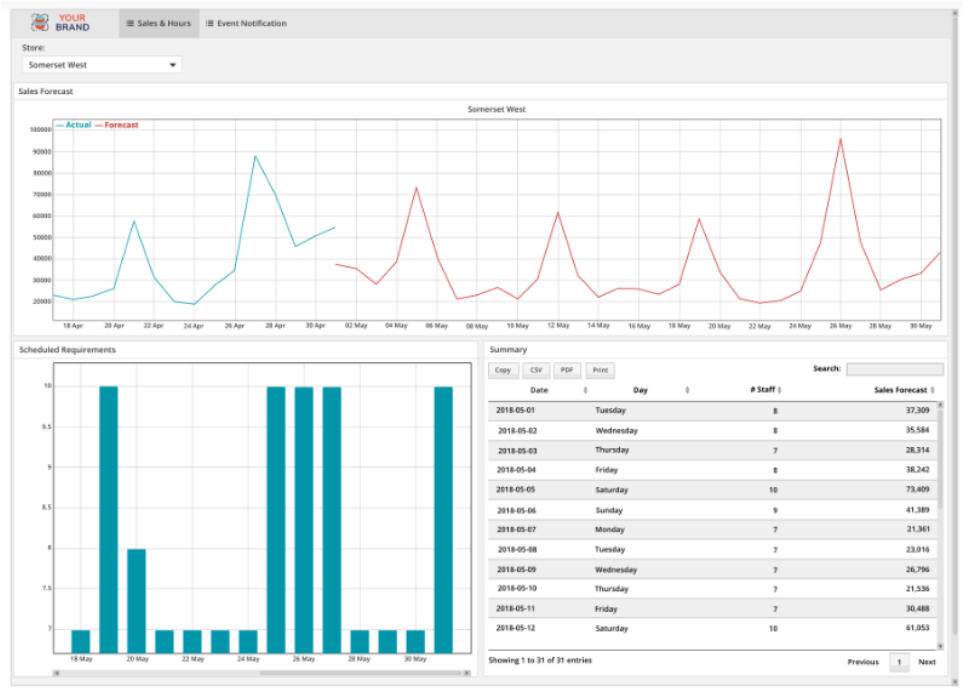
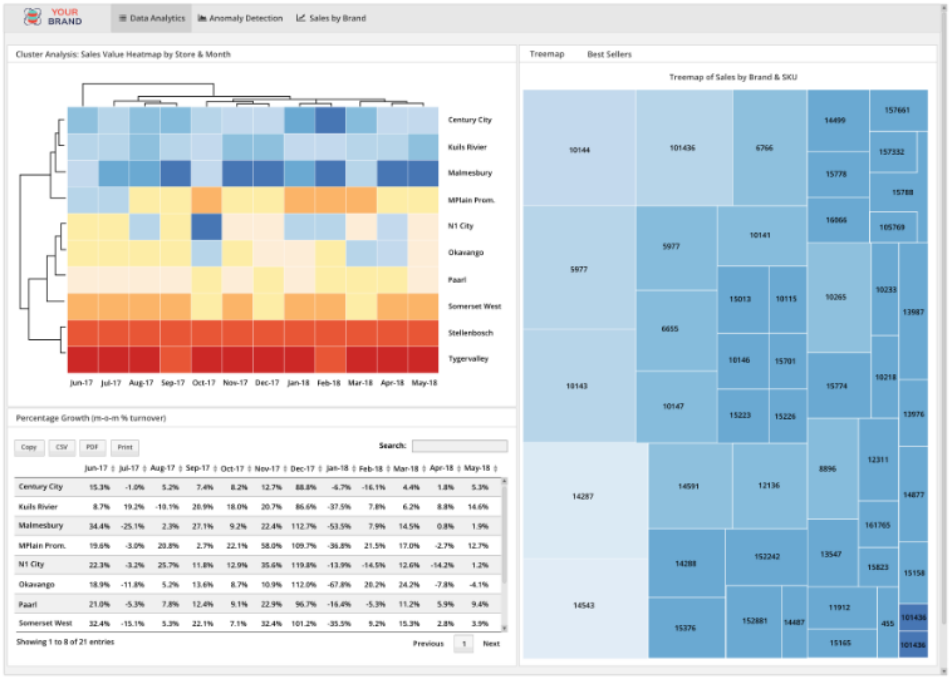


Figure 2.3.6 A photo shown Predictive Insights implement Heatmap for Correlation Coefficient on the staff management.



Regarding the lack of integration of AI in providing economic insights, Predictive Insights' method of finding patterns in big datasets can be used as a guide. Online income platforms can close the current gap in the market by combining machine learning and AI to overcome biases and provide more accurate economic insights based on income levels. Predictive insights components can be integrated into current predictive income platforms to greatly improve their functionality and offer users more useful features.

2.3.5 **World Bank's PovCal**

This tool has a poverty predictive model used by the World Bank to estimate global poverty rates. It considers various factors, including income and household composition and is able to provide a global perspective on poverty rates. Compared to the previous few tools, this valuable tool is designed for policymakers, but not directly applicable to individual income prediction. Various interesting visualization tools are shown below in Figure 2.3.7, Figure 2.3.8, and Figure 2.3.9.

Figure 2.3.7 The dashboard shown Poverty trend in a share of the population.

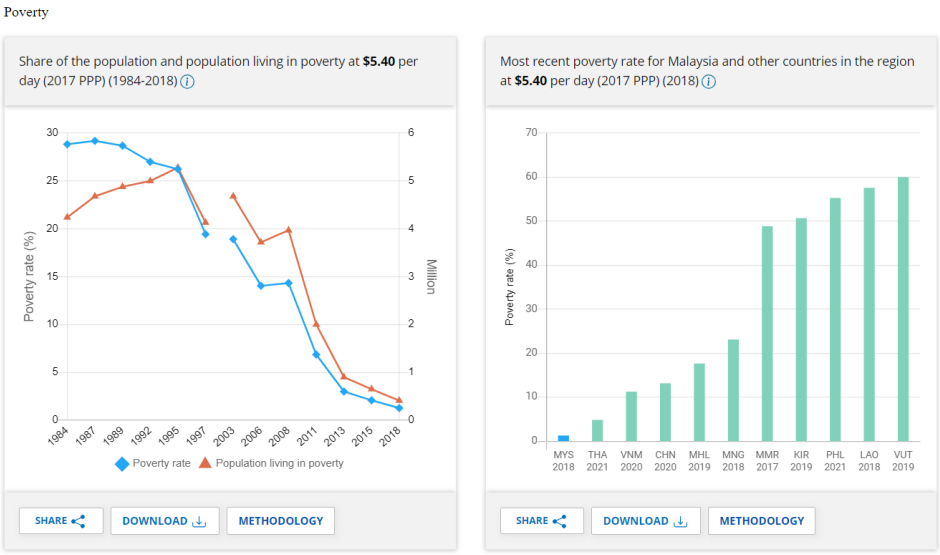


Figure 2.3.8 The dashboard shown Inequality trend in year 1984 to 2018.

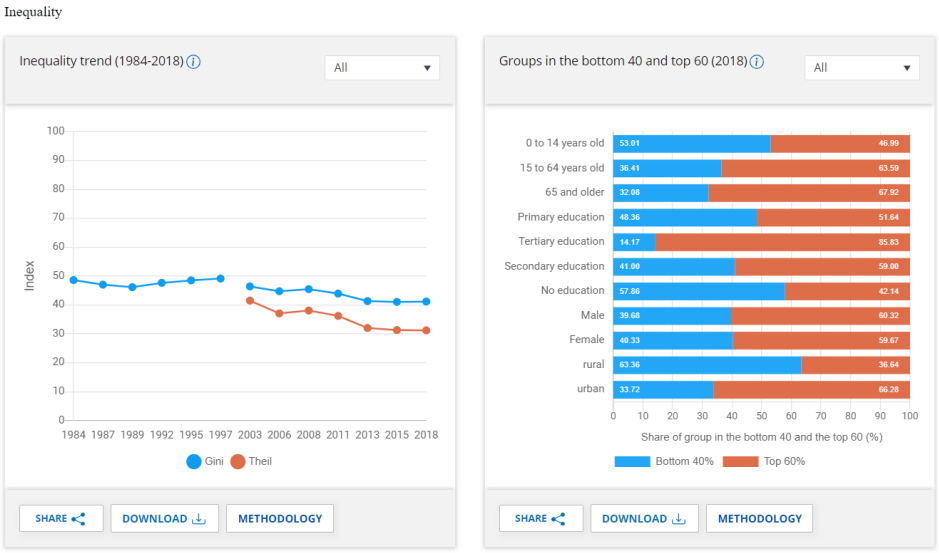
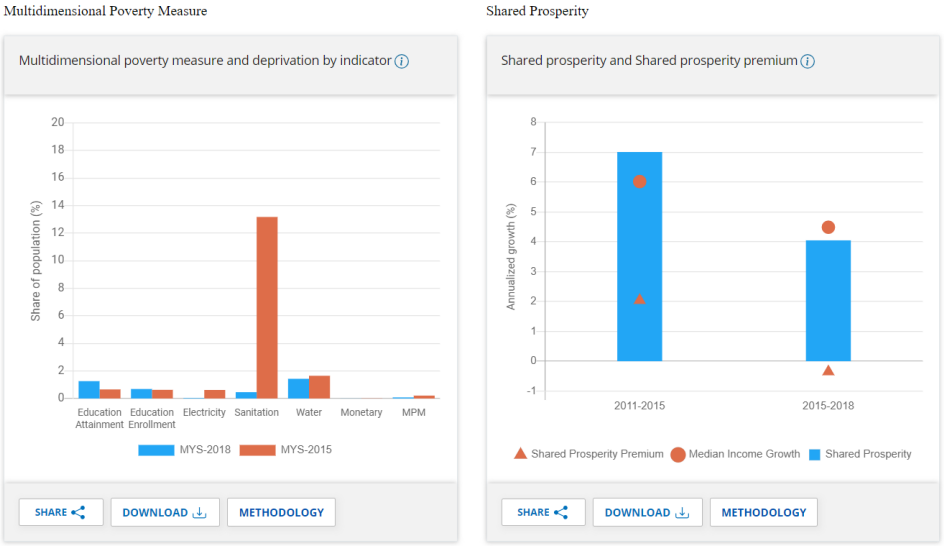


Figure 2.3.9 The dashboard shown Multidimensional Poverty Measure in a share of the population.

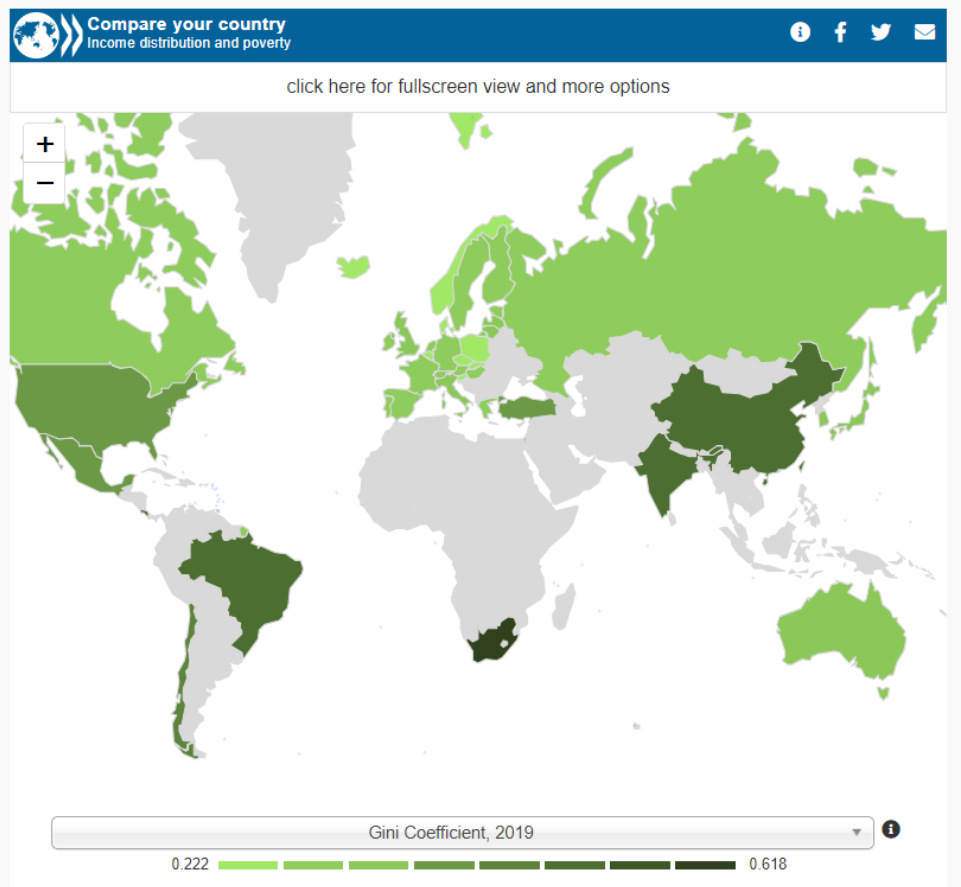


The platform owned by The World bank fully utilizes the data visualization implementation and interactive interface for user to review data. Also, the algorithms it used is machine learning, but it does not show attributes as to uphold data privacy and allows to download model results for documentation purposes [35].

2.3.6 OECD's Income Distribution Database

OECD Wealth Distribution Database (WDD) and Income Distribution Database (IDD) benchmarked and tracked economic inequality from all around the world as key resources. The IDD offers thorough data on trends in poverty and income inequality, along with regular updates and comprehensive details on numerous metrics and important indicators. Data from several countries are included in the June 6, 2023, updates, which shows changes in the risk of poverty and how it affects various age groups. The IDD also provides insights into income inequality and levels in specific contexts, covering metropolitan and regional areas. The database includes a range of research papers and publications covering topics like joint distribution of household income, wealth, and consumption, living wages, nowcasting, and preliminary estimates of income inequality based on microsimulation techniques.

Figure 2.3.10 Interactive maps dashboard shows gini coefficient in year 2019 provided by the OECD website



In Figure 2.3.10, the interactive maps dashboard not only calculated the Gini coefficient for poverty from different countries but display in a well interactive interfaces where user able to touch and move the map to explore. OECD iLibraies do not analyze Malaysia gini index on the poverty level. But in the survey conducted by them, named "OECD Economic Surveys: Malaysia 2021" have discussed about the priority of reducing income inequality forced Malaysia's social protection is still weak, and researched that "The consumption of lower-income households is more susceptible to an economic slowdown" [36]. It has proved that the relationship between income level and economic development exists and provides us insights on figuring output attributes for in Data Collection. Other than that, it offers insights into income distribution trends in OECD countries like US, China, Russia, but its extensive data resource for income analysis is unclear whether the capabilities on individual prediction provided. The only disadvantage of OECD-iLibrary is fewer guidelines and explanations on the use of the dashboard, giving less insight and copilot on getting ideas on utilizing the library, due to its complex menu and oversimplified design. But the iLibrary's dashboard does provide the ability to filter data, select pertinent attributes, and investigate various scenarios. Hence overall it is new and fresh for the user to explore ideas between income levels with countries.

2.3.7 Summary for Six Predictive Dashboards

Table 2.3.1 Overview Table for Six Predictive Dashboards

Tool's Name	**Overall Functions**	**Limitations**
PayScale [32]	- Provides salary information based on job title, location, and experience level. - Offers insights for employers and employees. - Customizable pay snapshots.	- Limited focus on machine learning utilization. - Lack of integration of AI for economic insights.

Salary.com [33]	- Offers salary information with details on bonuses, benefits, and compensation. - Customizable dashboards.	- Potential learning curve for users. - Lack of extensive AI integration for economic insights.
Glassdoor	- Provides salary and job information based on user contributions. - Poor reviews due to data integrity issues.	- Overfitting due to crowdsourced data. Lack of emphasis on data integrity.
Predictive Insights [34]	- Offers modules for demand forecasting, sales insight, and customer lifetime value. - Utilizes machine learning for data analysis.	- Limited mention of specific limitations. - Potential for overreliance on machine learning algorithms.
World Bank's PovCal [35]	- Predicts global poverty rates based on income and household composition. - Provides visualization tools for policymakers.	- Not directly applicable to individual income prediction. - Limited insight into specific data attributes for user privacy reasons.
OECD's Income Database [36]	- Provides comprehensive data on trends in poverty and income inequality globally. - Offers insights into income distribution. Interactive visualization.	- Limited guidelines on dashboard usage. - Complex menu and oversimplified design may hinder user experience.

Summary:

None of the tools above is using predictive model on income level to gain insights. Most of the tools are either providing Statistical Histogram that involved retrieve data from database, do not apply prediction on income level but filtering for statistical data and job and income information that store in database. Fortunately, only one tool applied Prediction model, which is World Bank's Povcal, but it focuses on household composition and global poverty rates, instead of Income level for certain area, especially Ipoh, Malaysia.

Chapter 3

Proposed Methods/Approaches

The processes of the project were divided into six phases which were project proposal, survey literature review, data collection, data pre-processing, model training selection building and data training, and lastly test dataset prediction.

3.1 Proposed Methodology

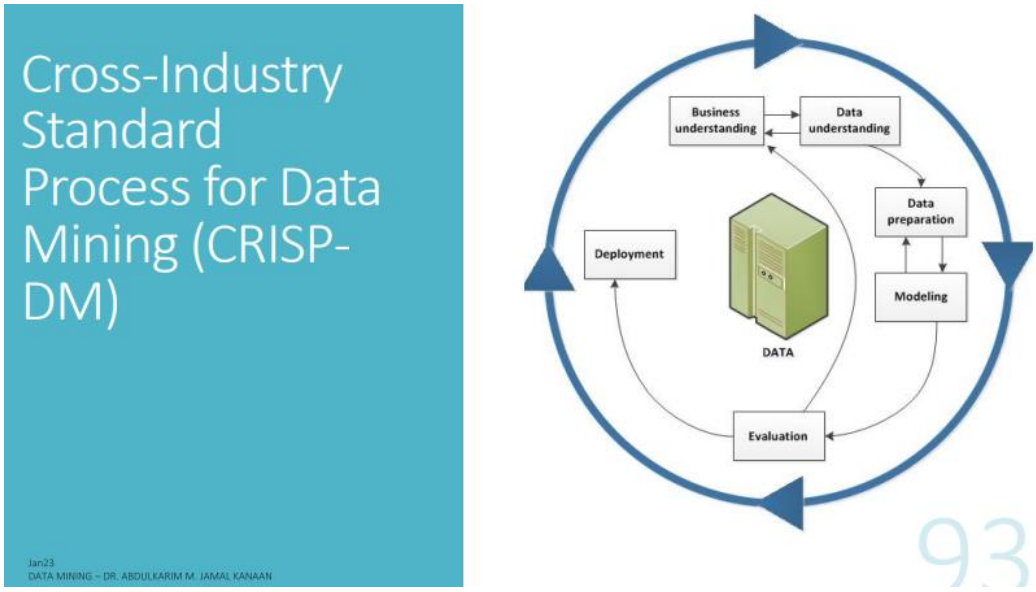
3.1.1 Data Mining

In the context of computer science, "Data mining" can also be referred to as knowledge extraction from data, data/pattern analysis, data archaeology, and data dredging in the context of computer science. The nontrivial extraction of implicit, previously unknown, and potentially useful information from database data is referred to as data mining, also known as knowledge discovery in databases (KDD). Data mining is necessary in order to retrieve valuable information from big datasets and apply it to forecasts or improved decision-making [37]. Summarize up the beneficials, hence, this project will be using Data Mining Techniques to discover the income level and figure out the insights for local economic development direction.

3.1.2 Cross-Industry Standard Process for Data Mining (CRISP-DM)

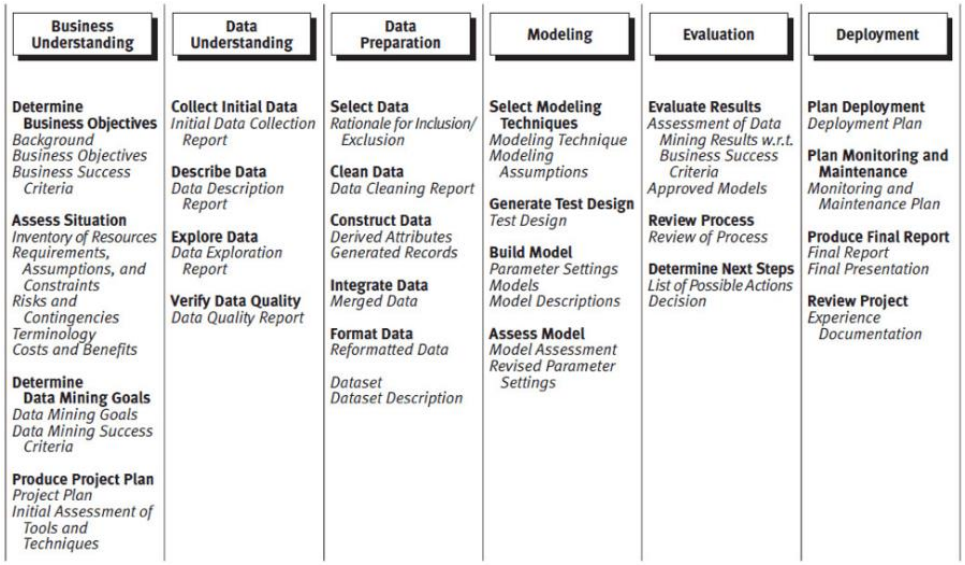
The purpose of using CRISP-DM methodology in the project because it is suitable to serve as the primary framework for conducting the various phases of the model development. It guarantees an organized and methodical approach to tasks related to data mining and machine learning. Six sequential phases stages make up this tried-and-true method, and each one adds to the project's final success. Since then, it has become the most common method for data mining, analysis, and data science projects [38]. Figure 3.1.1 shown a picture of the process of CRISP-DM from Data Mining Techniques.

Figure 3.1.1 A picture of the process of CRISP-DM from Data Mining Techniques



An outline of the phases is shown in Figure 3.1.2 along with generic tasks (bold) and outputs (italic). We go into further detail about each generic task and its results in the sections that follow. We concentrate on task summaries and output overviews [39]. As of the stages from the figure 3.1.2 will be applied on the project.

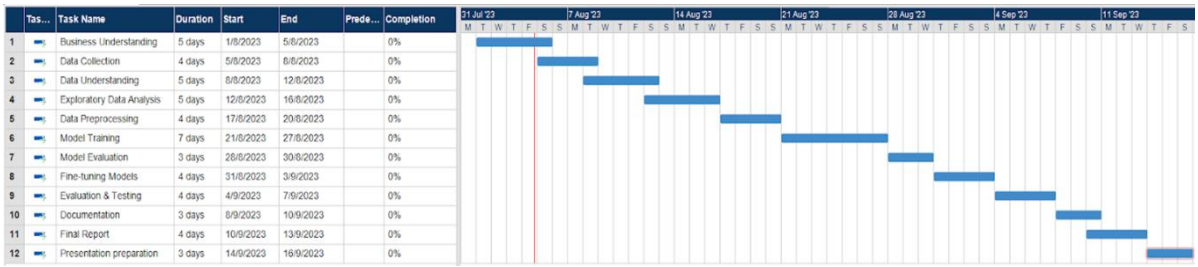
Figure 3.1.2 The list down list from every phrase of CRISP-DM from Data Mining Techniques



3.1.3 Business Understanding

Knowing the business inside and out will help you pinpoint exactly what the current problems are, how to assess them and find a solution or solutions, and what tactics to use to meet the objectives of the company [40]. It helps to gain a clear understanding of the project's objectives and requirements from a business perspective, similar to Chapter 2 in this project. The goals have been defined as to develop a predictive tool for anticipating income levels in Ipoh citizens. Background information identified the needs and expectations of government policymakers which is a tool for helping provide insights in economic development in local area. This section mentions the business goal, problem statement, current solutions to the problem, motivation, data mining goals, Gantt chart and development tools needed to use. A draft Gantt chart for the project duration approximately shown below in Figure 3.1.3.

Figure 3.1.3 Gantt Chart that review the project duration



	Tas...	Task Name	Duration	Start	End	Prede...	Completion
1		Business Understanding	5 days	1/8/2023	5/8/2023		0%
2		Data Collection	4 days	5/8/2023	8/8/2023		0%
3		Data Understanding	5 days	8/8/2023	12/8/2023		0%
4		Exploratory Data Analysis	5 days	12/8/2023	16/8/2023		0%
5		Data Preprocessing	4 days	17/8/2023	20/8/2023		0%
6		Model Training	7 days	21/8/2023	27/8/2023		0%
7		Model Evaluation	3 days	28/8/2023	30/8/2023		0%
8		Fine-tuning Models	4 days	31/8/2023	3/9/2023		0%
9		Evaluation & Testing	4 days	4/9/2023	7/9/2023		0%
10		Documentation	3 days	8/9/2023	10/9/2023		0%
11		Final Report	4 days	10/9/2023	13/9/2023		0%
12		Presentation preparation	3 days	14/9/2023	16/9/2023		0%

3.1.4 Data Understanding

The primary goal of data understanding is to gain general insights about the data that may be useful for subsequent steps in the data analysis process, but data understanding should not be driven solely by the goals and methods to be used in subsequent steps [41]. It is the first step before started to data cleaning and preprocessing, gets a basic understanding of the structure, limitations, and the available dataset. In this project, it is important to explore the income dataset of Ipoh citizens to examine its variables, patterns and data quality.

To ensure accurate predictions while building the model, it is crucial to decide on appropriate questions for the respondents as it directly affects the input data. Hence, several studies on different papers from literature review section. Multiple questions have been selected which

defined as required to be included in the survey for income predictions and inputs have been modify and reamend to Malaysia's conditions. The attributes included:

Table 3.1.1 Original Attributes Reference [42]

Attributes	**Types of data**
Age	Integer
Workclass	Categorical
Fnlwgt	Integer
Education Level	Categorical
Education-num	Integer
Marital Status	Categorical
Occupational	Categorical
Relationship	Categorical
Race	Categorical
Gender	Binary
Capital-gain	Integer
Capital-loss	Integer
Hours-per-week	Integer
Native-country	Categorical
Income	Binary

Capital Gain and Capital Loss

Capital gain refers to the profit earned from the sale of an investment or asset. It occurs when the selling price of the asset is higher than its original purchase price. For example:

If you buy a stock for $100 and sell it later for $150, you have a capital gain of $50 ($150 - $100 = $50). Similarly, if you purchase a property for $200,000 and sell it for $250,000, your capital gain is $50,000 ($250,000 - $200,000 = $50,000).

Whereas for Capital gain, refers to the profit earned from the sale of an investment or asset. It occurs when the selling price of the asset is higher than its original purchase price. For example:

If you buy a stock for $100 and sell it later for $150, you have a capital gain of $50 ($150 - $100 = $50). Similarly, if you purchase a property for $200,000 and sell it for $250,000, your capital gain is $50,000 ($250,000 - $200,000 = $50,000).

Both capitals are beneficial to understand whether the respondent participant in an investment or other source of income. It would help to emphasize the various reasons causing income level in the area.

Exclude Native Country and Final Weight

'Native-Country' has been excluded, due to it is clear that the research target area would be Ipoh instead of among countries. Whereas for 'Fnlwgt' (Final Weight), this attribute also has been excluded from the list due to Final Weight is a format from US Census Bureau to calculate their census each people weight in the survey, it's not applicable in Malaysia and each state could have different weights. Hence, final weight calculation will not be used in this project.

Modified attributes shown as below.

Table 3.1.2 Attributes after reamend

Attributes	**Types of data**
Age	Integer
Employment Status	Categorical
Education Level	Categorical
Marital Status	Categorical
Occupation	Categorical
Own-Child	Categorical
Ethnicity	Categorical
Gender	Categorical
Capital-gain	Integer
Capital-loss	Integer
Hours-per-week	Integer
Current Monthly Income	Integer

Additional attributes added as below.

Table 3.1.3 New Attributes after consideration

Attributes	**Types of data**
Number of children	Integer
Current Residency	Categorical
Current Working Place	Categorical
Current Working Environment	Categorical
Job position classification	Categorical
Years of Experience	Integer
Additional professional certifications	Categorical
Number of professional certifications	Integer
Additional income or investments	Categorical
Capital gain	Categorical
Capital loss	Categorical
Satisfaction towards current income	Categorical

Additional Questions for understanding the need of AI Adoption on Economics Planning

1. Are you familiar with artificial intelligence (AI) and its applications?
2. How does Artificial Intelligence apply in your area of work?
3. What can be or is already the purpose of AI adoption in your organization (select multiple options if applicable)?
4. What is your level of understanding of Artificial Intelligence and Allied Technologies and the application domain?
5. Is there a need to establish national centers on innovation, research, and entrepreneurship-focused on AI and allied technologies to allow knowledge-based socio-economic development in the country?
6. Is there a need to establish national centers on innovation, research, and entrepreneurship-focused on AI and allied technologies to allow knowledge-based socio-economic development in the country?
7. Do you think income analysis provided by AI can provide government policymakers with more accurate information to make more informed decisions on economic issues?

Question 1 to Question 5 was taken from LinkedIn, Sohail Khan, Ph.D.'s personal profile page, article titled 'Sector-Specific Questionnaires for Identifying Challenges regarding AI Adoption in Pakistan'[43]. He is an expert in power system, ex-scientist Austrian Institute of Technology and Associate Prof @ Pak-Austria Fachochschule, university in Pakistan. The study's goal is to collect data on how to develop competencies in artificial intelligence and related technologies at the national level to support socioeconomic development and the Digital Pakistan Vision. For this purpose, it is necessary to assess the current state of AI adoption in various sectors of society and predict how it can be improved [43]. In the study, Sohail Khan addressed several topics relied on the specific industries which have AI implemented. The purpose of conducting the surveys were to gather and compiled to consolidate a complete pathway to the AI adoption to each industry. He declared that the complied data would facilitate a gap analysis, enabling policymakers to discern areas where improvements are necessary and changes need to be made. Hence, armed with these insights, the government can create and implement policies to promote AI development and innovation.

Whereas for Question 6 and Question 7, there are no supportive articles from papers or expert, but questions that seek for Ipoh working adults' opinion on agreeing further development for AI adoption and analysis, especially focus on Income Prediction, Economy related issues.

Final structure of the Data Collection

After strict screening through a pool of questions, 32 questions in total have been selected and sequenced in four sections accordingly.

The sections namely 'Section 1: Demographic Information', 'Section 2: Occupational Information', 'Section 3: Income Earnings' and lastly 'Section 4: AI and Technology Adoption on Economics Planning Policy'.

Section 1 contains 9 questions such as

"Age", "Gender", "Ethnicity", "Highest level of education", "Marital status", "Own-child", "Number of children", "Current Residency" and "Current Working Place".

Section 2 contains 8 questions such as

“Have you worked in Ipoh before?”, " Current employment status”, “What is your current working environment?”, “Occupation ", “What is your job position classification?”, “Years of Experience”, “Have you pursued any additional professional certifications related to YOUR OCCUPATION?” and “Number of Professional Certifications”.

Section 3 contains 8 questions such as

“Working Hours Per Week”, “Current monthly income (RM)”, “Do you have any additional sources of income or investments?”, “Capital Gain”, “Capital Gain (RM)”, “Capital Loss”, “Capital Loss (RM)” and “Are you satisfied with your current income level?”.

Section 4 contains 7 questions such as

“Are you familiar with artificial intelligence (AI) and its applications?”, “ How does Artificial Intelligence applies in your area of work?”, “What can be or is already the purpose of AI adoption in your organization (select multiple options if applicable)?”, “What is your level of understanding of Artificial Intelligence and Allied Technologies and the application domain?”, “Is there a need to establish national centers on innovation, research, and entrepreneurship-focused on AI and allied technologies to allow knowledge-based socio-economic development in the country?”, “Would you be willing to use AI-powered tools or platforms to get insights and understand more about improving your financial planning or income optimization?” and “Do you think income analysis provided by AI can provide government policymakers with more accurate information to make more informed decisions on economic issues?”.

3.1.5 Data Collection & Data Description

Random Sampling has been applied to this project’s data collection due to the data collection didn’t record any direct expose identity of respondents, and the survey link was sent to

surrounding friends and family, social media such as Instagram, Facebook Group for Ipoh Citizens and anonymous street people.

As of now, the data collection has begun for four months, and a total of 428 responses were received, and 423 of it are valid reponses.

This dataset is about the income level information of Ipoh working adults and factors influencing Income Earnings in Ipoh. Moreover, it also consists of various attributes and information of the person's income, occupation, education information. Below are the summarized lists of the dataset attributes and its corresponding values.

Numerical Attributes:

Table 3.1.4 Table for numerical attributes

No.	Attributes	Values
1	**Age**	Int64 {20-60 years old}
2	**Number of Children**	Int64 {0-4}
3	**Working Hours Per Week**	Float64 {0-78}
4	**Current monthly income (RM)**	Int64 {0-100000}
5	**Capital Gain (RM)**	Float64 {0-200000}
6	**Capital Loss (RM)**	Float64 {0-52000}
7	**Are you satisfied with your current income level?**	Int64 {1-5}

Categorical Attributes:

Table 3.1.5 Table for object attributes

No.	Attributes	Values
1	**Gender**	**Object** {Male, Female}
2	**Ethnicity**	**Object** {Malay, Chinese, Indian, Bangladeshi}
3	**Highest level of education**	**Object** {Primary School, Secondary School, Diploma, Foundation, Bacheloers, Master, PhD}

4	**Marital Status**	**Object** {Single, never married., , Married civilian spouse. , Separated from spouse but not divorced., Divorced, Widowed, Married but spouse is not currently living with them., Married to a spouse who is in the Armed Forces.}
5	**Own-child**	**Object** {Unmarried, Yes, No}
6	**Current Residency**	**Object {**Ipoh, KL, George Town, Johor Bahru (JB), Shah Alam, Petaling Jaya (PJ), Kuching (Sarawak), Melaka City, Cyberjaya, Singapore, SG, Kampar}
7	**Current Working Place**	**Object** {Ipoh, KL, George Town, Johor Bahru (JB), Shah Alam, Petaling Jaya (PJ), Kota Kinabalu (Sabah), Melaka City, Cyberjaya, Singapore, Overseas, Perlis, NaN, Tapah, SG}
8	**Have you worked in Ipoh before?**	**Object** {Yes, No}

9	**Current employment status**	**Object** {Never-worked, Unemployed, Local government (e.g., city, district, municipality), State government entity, Federal government, Self-employed, incorporated (running a registered business), Self-employed, not incorporated (e.g.…, Employed by a private sector (compa…}
10	**What is your current working environment?**	**Object** {Physical, Work from Home, Hybrid}

11	**Occupation**	**Object** {Engineers, Medical Professionals, Bankers and Finance Professionals, Teachers and Educators, Information Technology (IT) Professional, Sales and Marketing Professionals, Manufacturing Workers, Government Officials and Civil Servants, Tourism and Hospitality Workers, Entrepreneurs and Business Owners, Wholesale Chinese Medicine, F&B, Company Secretary, Recruitment Consultant, Construction, Chef, Neet, Management, Auditor, Health care, Accounts executive, Executive, Security & Safety, Pedlar, Human resource, Constructors, Admin, Construction, Student, Finance Controller, Digital marketing, Art and Design field, assistance administrator, Accounts}

12	**What is your job position classification?**	**Object** {Senior Level Executive, Mid to Senior Level Manager, Mid-Level, Junior Employee, Entry Level, Freelancer/Consultant, Student/Researcher, Director Owner, Staff nurse, None of these above}
13	**Years of Experience**	**Object** {0-43, 5 years}
14	**Have you pursued any additional professional certifications related to YOUR OCCUPATION?**	**Object** {Yes, No}
15	**Number of Professional Certifications**	**Object** {0 – 8, -, Null}
16	**Do you have any additional sources of income or investments?**	**Object** {Yes, No}
17	**Capital Gain**	**Object** {Yes, No}
18	**Capital Loss**	**Object** {Yes, No}
19	**Are you familiar with artificial intelligence (AI) and its applications?**	**Object** {Yes, No, Maybe}

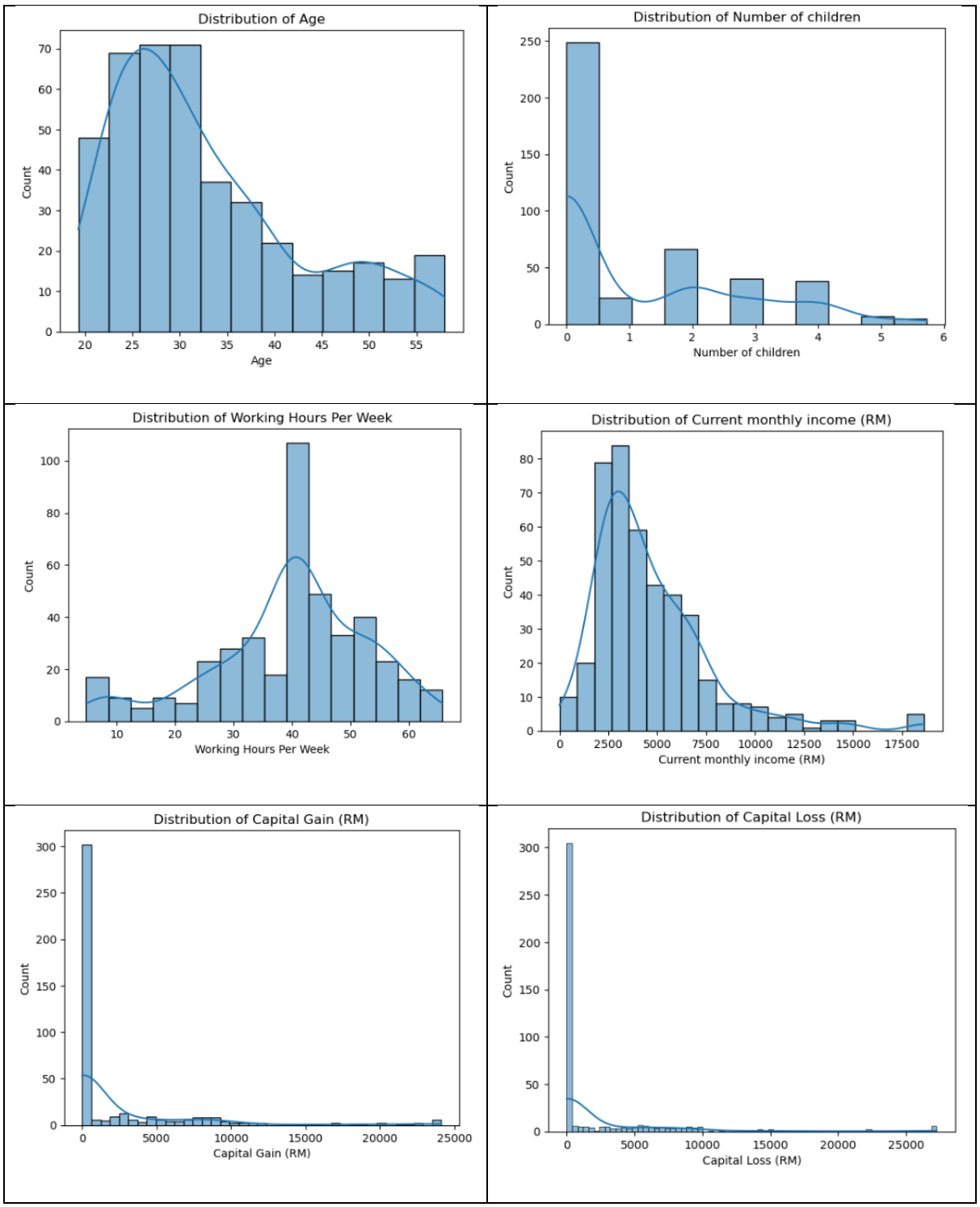
20	**How does Artificial Intelligence applies in your area of work?**	**Object** {Financial Planning/Forecasting, Market Predication, Risk Analysis, KYC/AML process implementation, Customer Awareness & Acquiesce, Did not use, How to cheat money more efficiency, ChatGPT, Research works, Planning for content, For personal knowledge}
21	**What can be or is already the purpose of AI adoption in your organization (select multiple options if applicable)?**	**Object** Operational efficiency, Cost reduction, Creation of new value streams, Work-style reforms, Cheat, No, 0}
22	**What is your level of understanding of Artificial Intelligence and Allied Technologies and the application domain?**	**Object** {ı have sufficient research experience to be considered an expert, Yes, I have a working knowledge of the basic concepts and terms, I have limited knowledge about the Subject, I have heard the terms but don't understand them, No knowledge at all}
23	**Is there a need to establish national centers on innovation, research, and entrepreneurship-focused on AI and allied technologies to allow knowledge-based socio-economic development in the country?**	**Object** {Existing initiatives are sufficient, and they shall be equitably funded Urgently needed for organizing a central hub for AI-related initiatives for developing a comprehensive ecosystem}

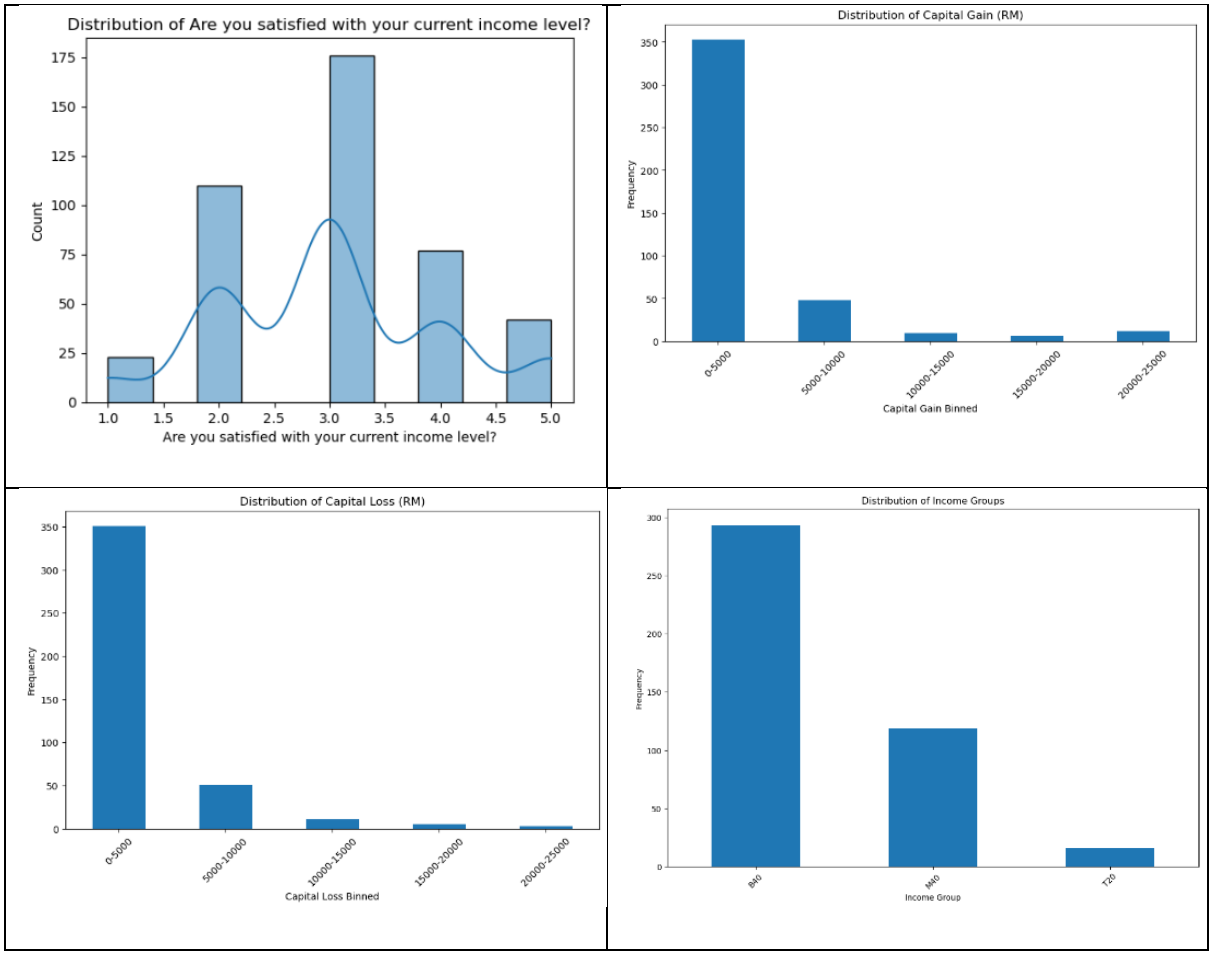
24	**Would you be willing to use AI-powered tools or platforms to get insights and understand more about improving your financial planning or income optimization?**	**Object** {Yes, No}
25	**Do you think income analysis provided by AI can provide government policymakers with more accurate information to make more informed decisions on economic issues?**	**Object** {Yes, No}

Data Exploration and Visualization

The numerical data existed in our dataset are Age, Number of children, Working Hours Per Week, Currently Monthly Income(RM), Capital Gain (RM) and Capital Loss (RM), Are you satisfy with your current income level, binned Catpital Gain (RM) & binned Capital Loss (RM), and Income Groups. As shown in Figure 3.1.4, we visualized them in the form of histogram. From the figure, we can clearly see that the Currently Monthly Income(RM) and Capital Gain (RM) are capped. To determine the values at which the variables are capped, the following code is executed.

Figure 3.1.4 Histogram for checking Distribution is Skew, Symmetric, and if there are any Outliers





We have displayed the line within the histogram to determine the skewedness, symmetry, and presence of any outliers in the data distribution. As shown in Figure 3.1.4, other than Working Hours Per Week and Are You satisfied with your current income level, the rest attributes are skewed to left. To further checking the outliers, we had explored the numerical value in the form of boxplot.

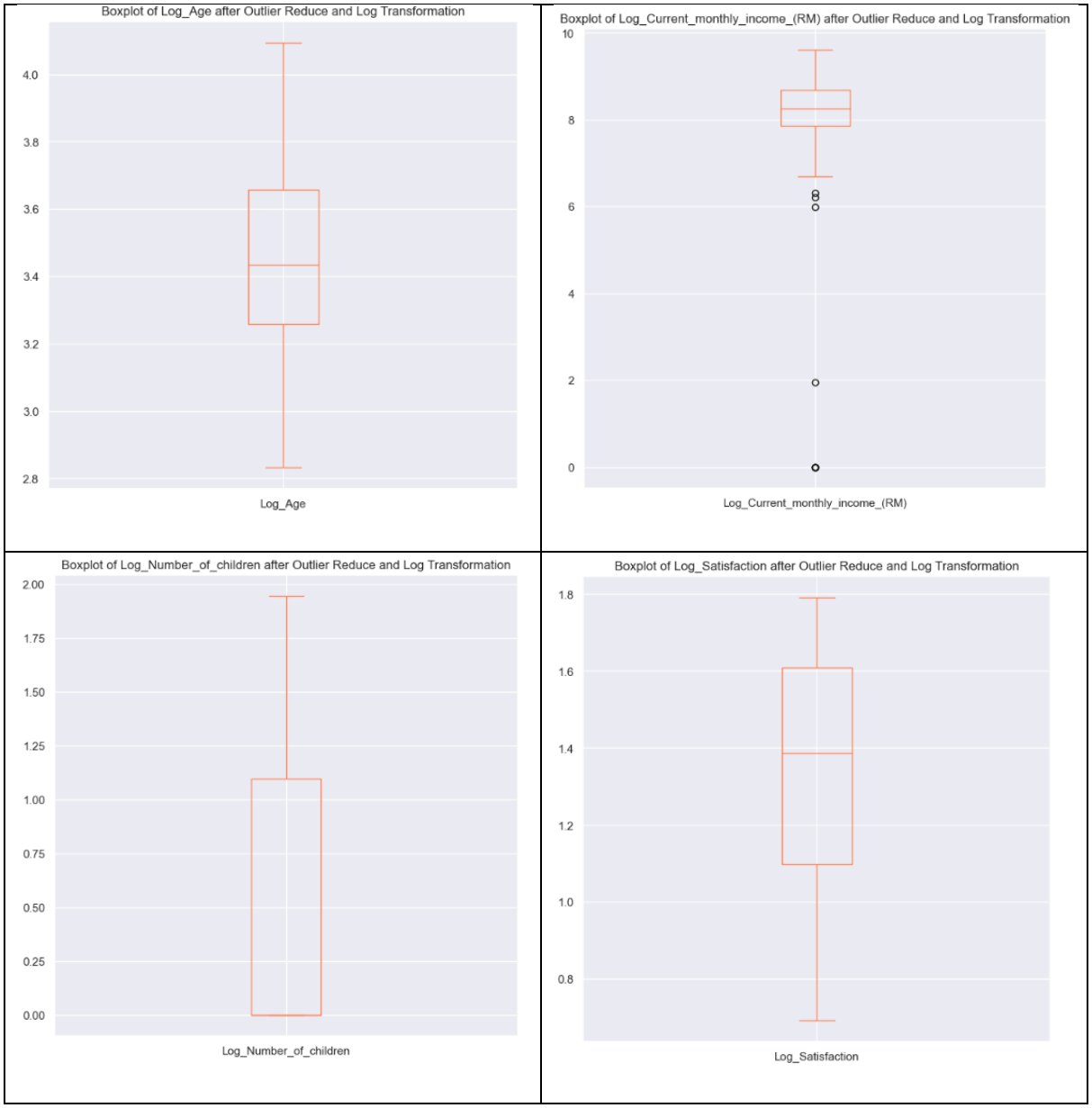
Besides, below figure shown the capped values for Currently Monthly Income (RM) and Capital Gain (RM) have been checked

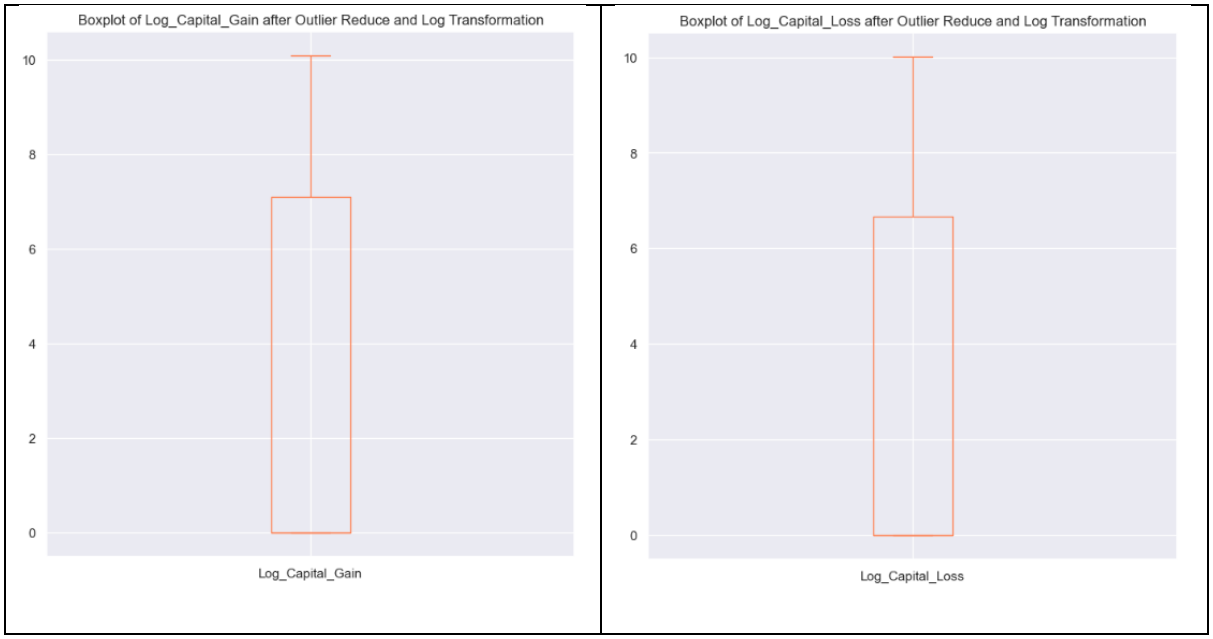
Figure 3.1.5 Code for checking the capped values for Currently Monthly Income (RM) and Capital Gain (RM)

```
# Capped value for 'Current monthly income (RM)'
print("Current monthly income (RM) is capped at min value:", df['Current monthly income (RM)'].min())

# Capped value for 'Capital Gain (RM)'
print("Capital Gain (RM) is capped at max value:", df['Capital Gain (RM)'].max())
```

```
Current monthly income (RM) is capped at min value: 1.6200000000000028
Capital Gain (RM) is capped at max value: 24113.78000000001
```

Figure 3.1.6 Reduced Outliers Boxplot for Age, Number of children, Working Hours Per Week, Currently Monthly Income (RM), Capital Gain (RM) and Capital Loss (RM)





Now, we can see from Figure 3.1.6 that Age, Number of children, Working Hours Per Week , Capital Gain (RM) and Capital Loss (RM) do not have any outliers whereas for Currently Monthly Income (RM) still appear outliers, this is because outliers were found too many in this attributes.

Figure 3.1.7 Count plot for Gender

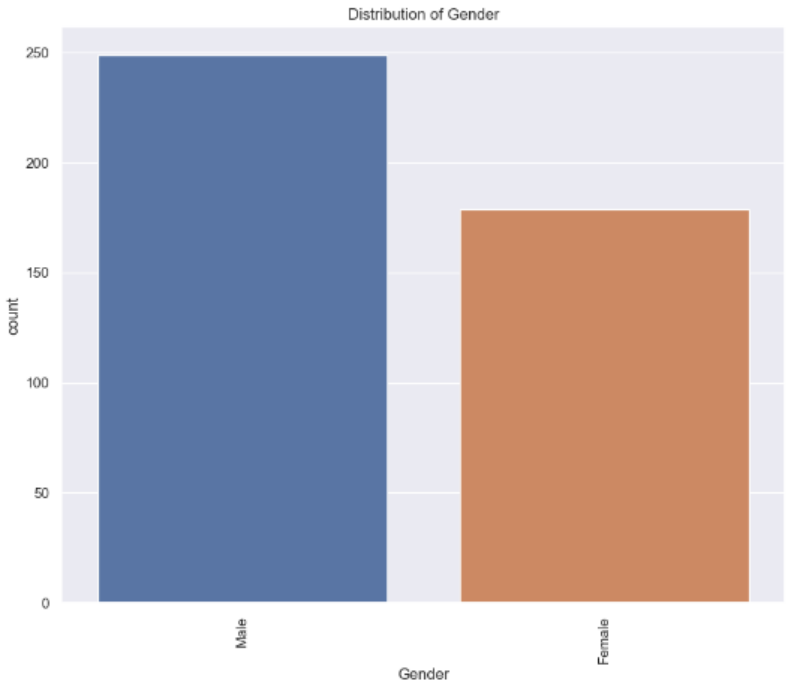


Figure 3.1.8 Count plot for Ethnicity

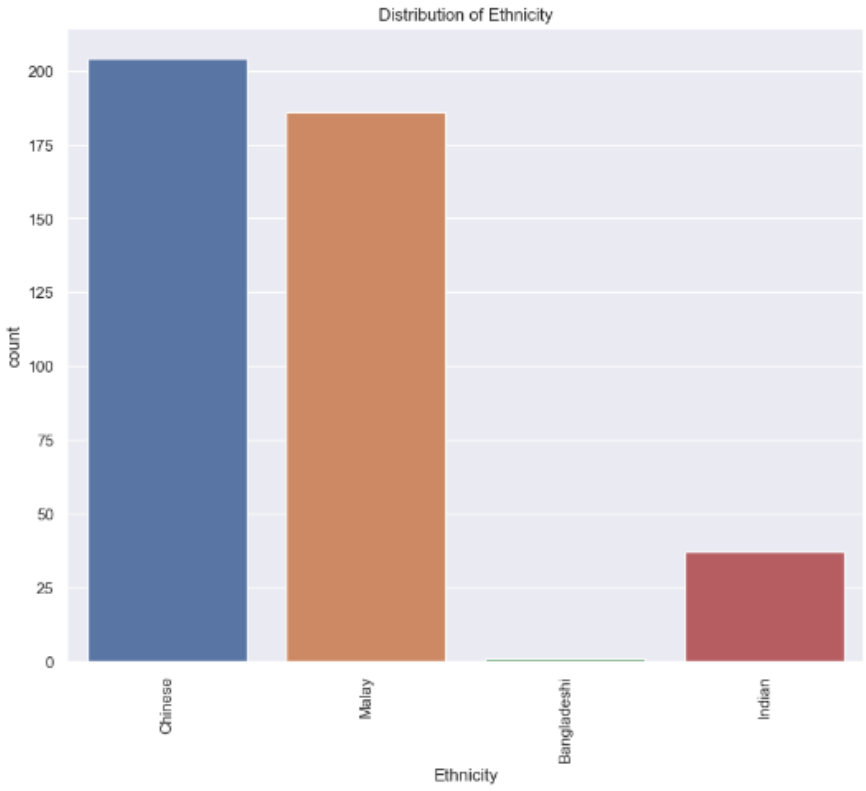


Figure 3.1.9 Count plot for Highest level of Education

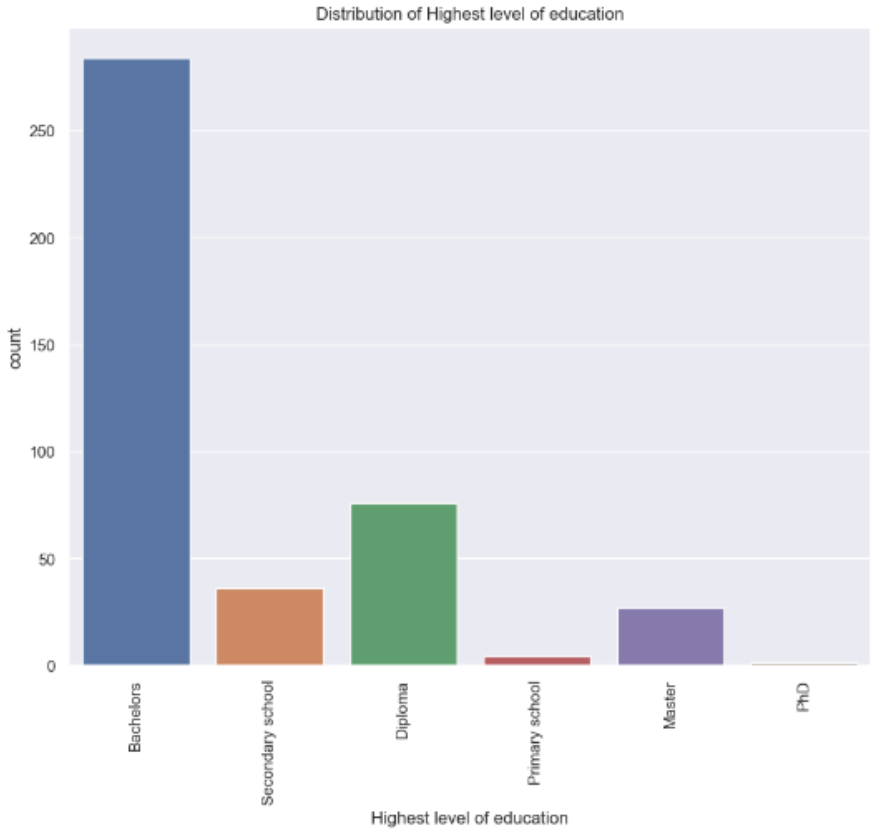


Figure 3.1.10 Count plot for Marital Status

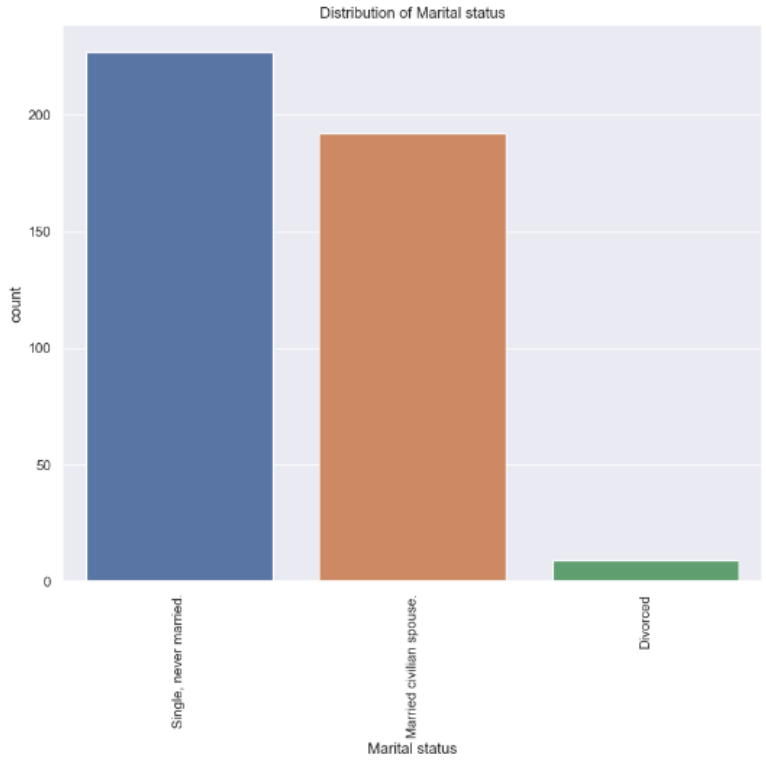


Figure 3.1.11 Count plot for Own-child

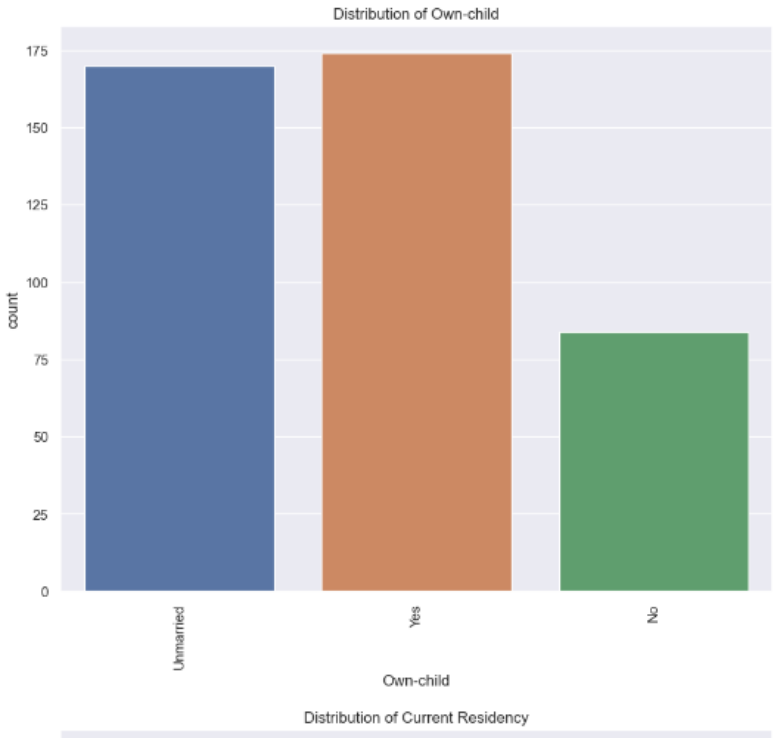


Figure 3.1.12 Count plot for Current Residency

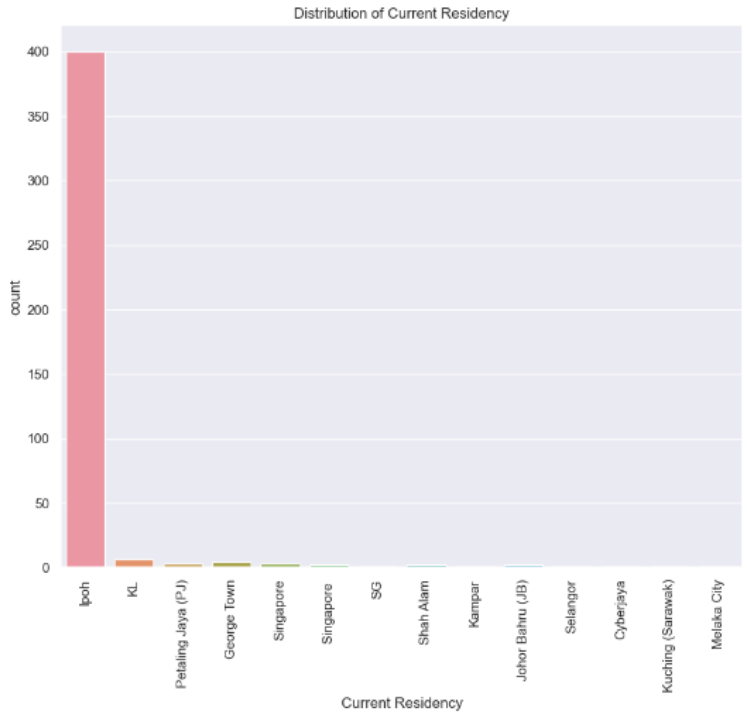


Figure 3.1.13 Count plot for Current Working Place

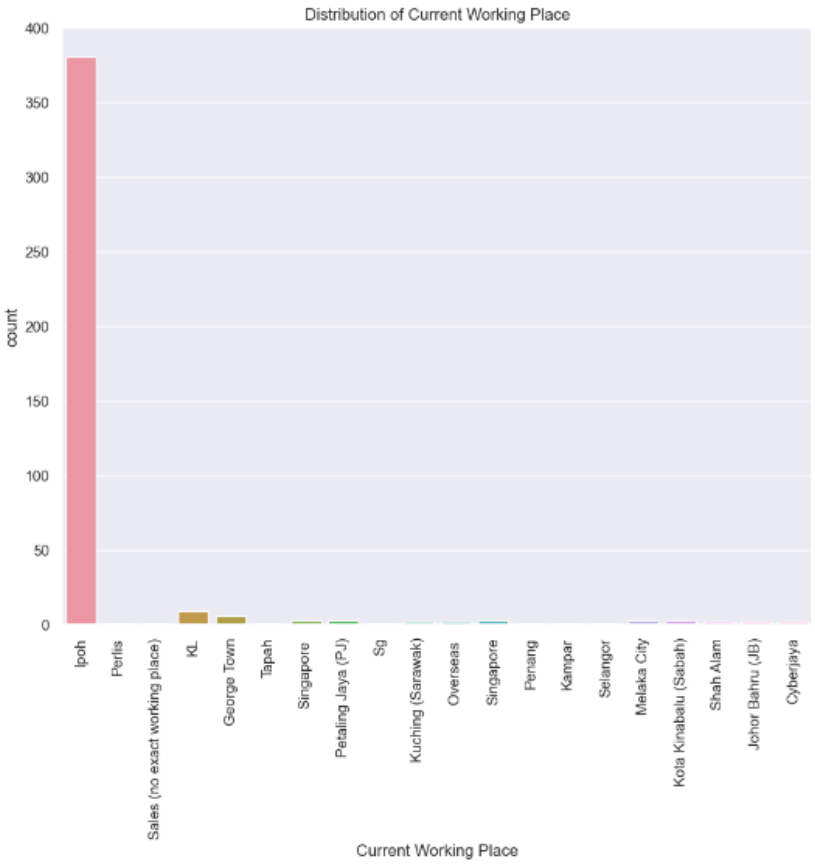


Figure 3.1.14 Count plot for lived in ipoh before

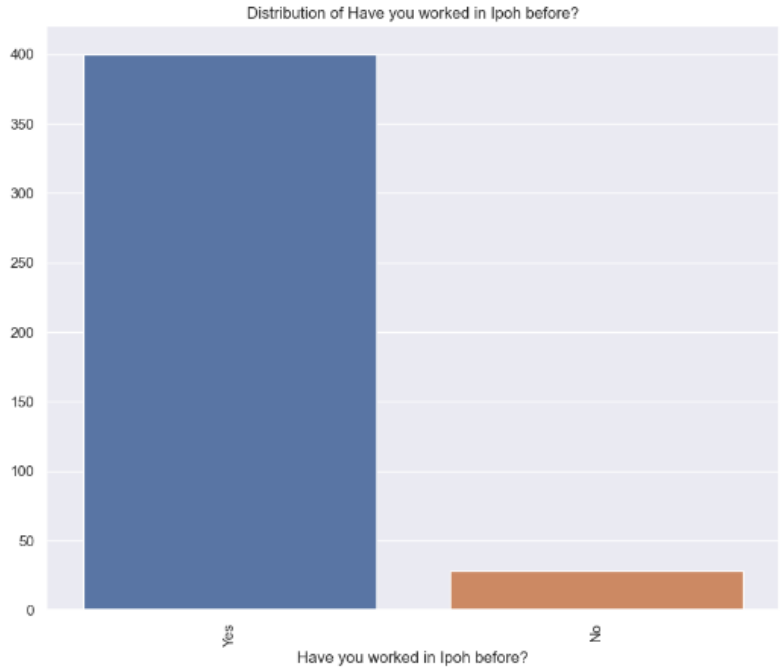


Figure 3.1.15 Count plot for Current Employment Status

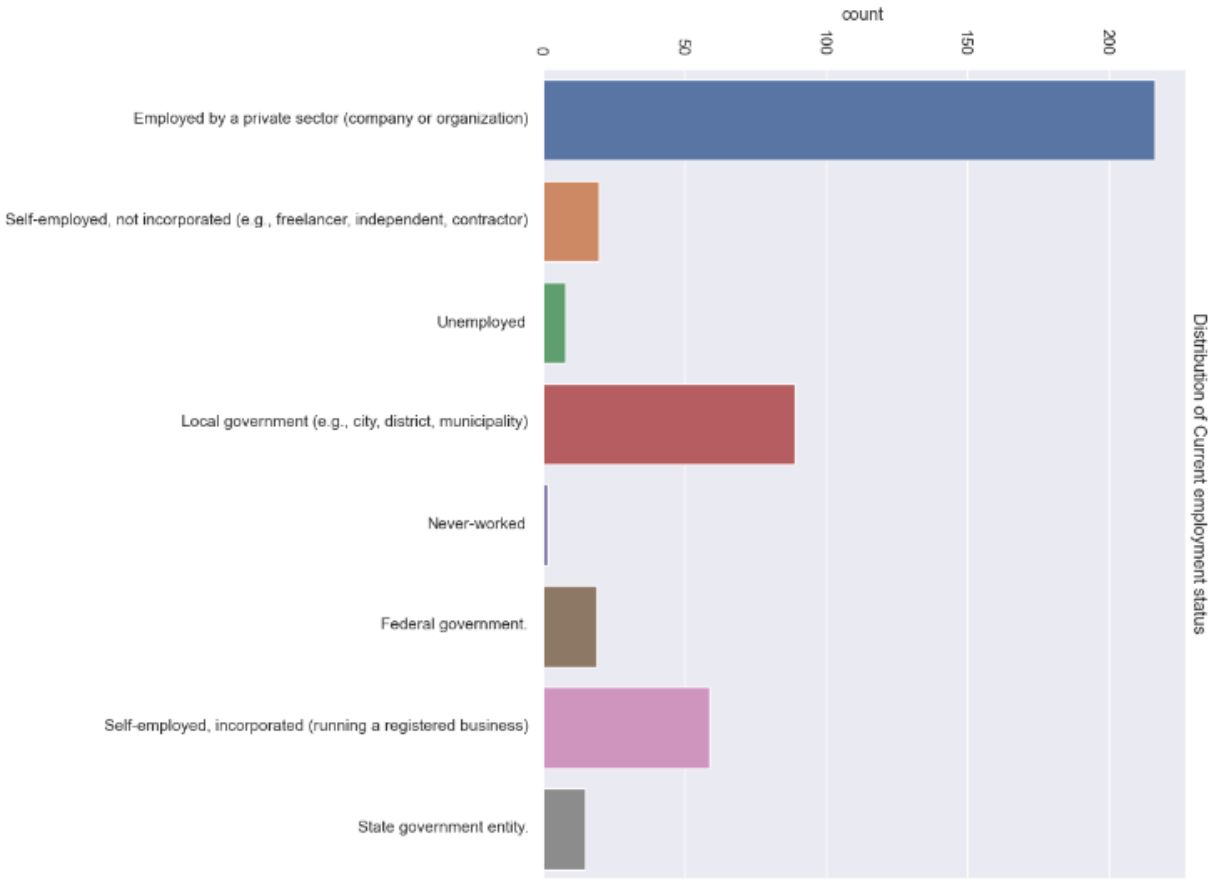


Figure 3.1.16 Count plot for Current Working Environmental

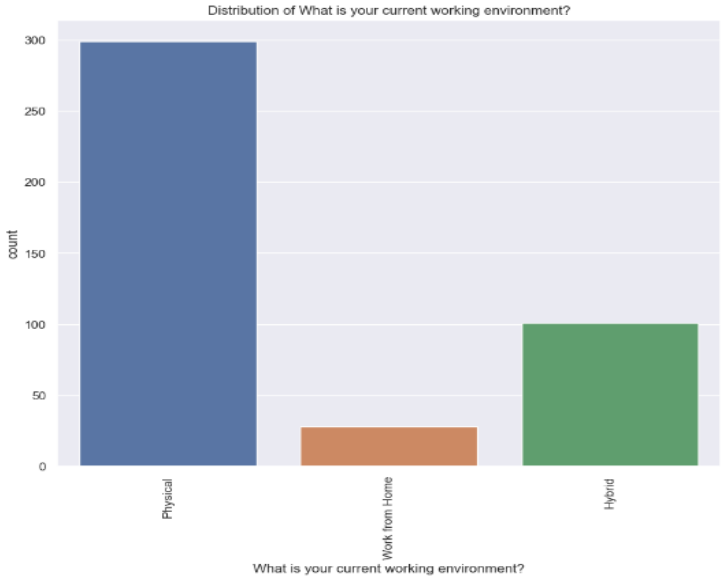


Figure 3.1.17 Count plot for Occupation

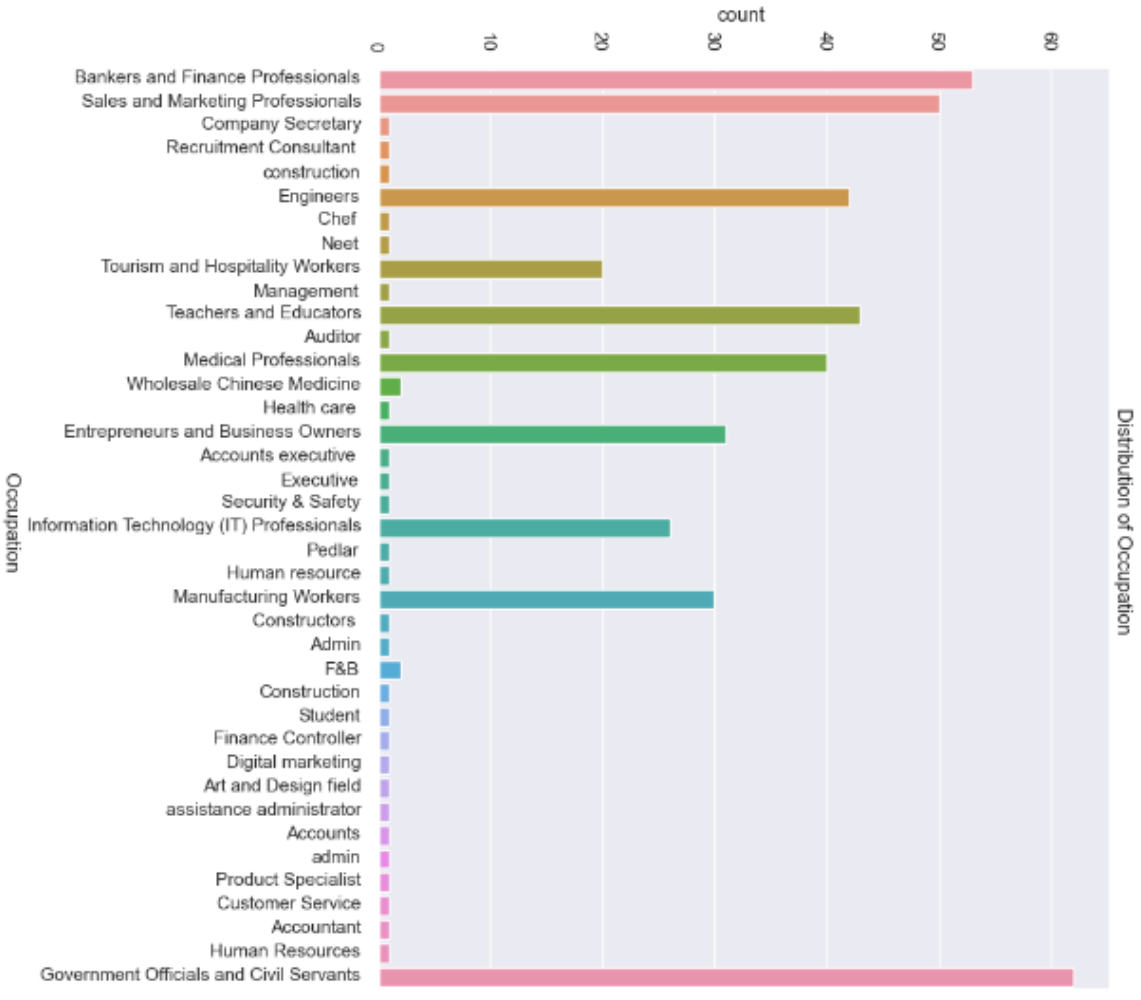


Figure 3.1.18 Count plot for Job Position Classification

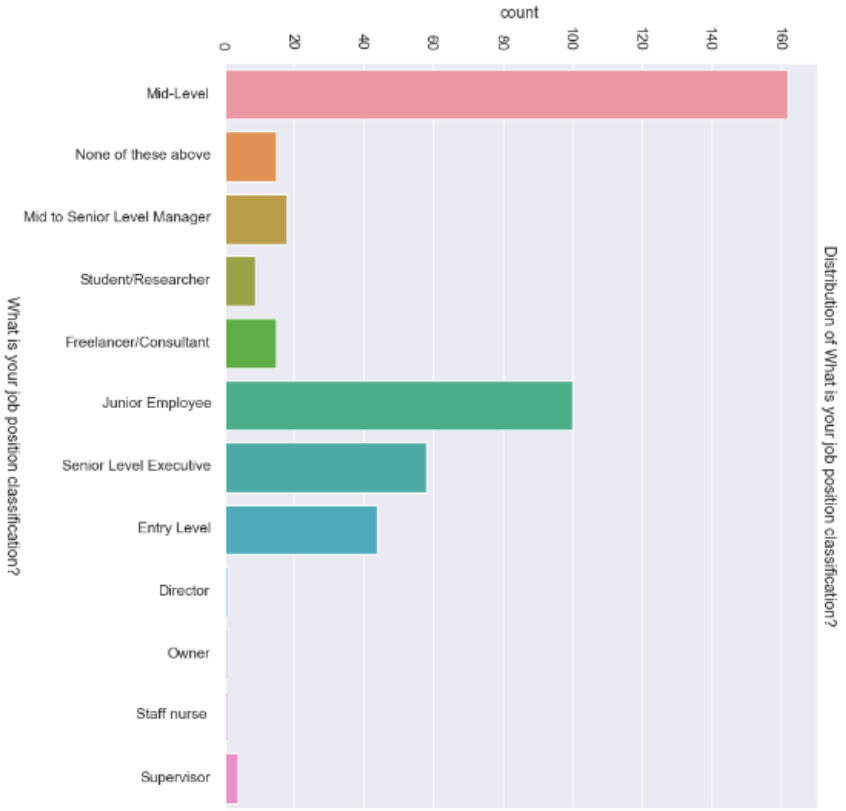


Figure 3.1.19 Count plot for Years of Experience

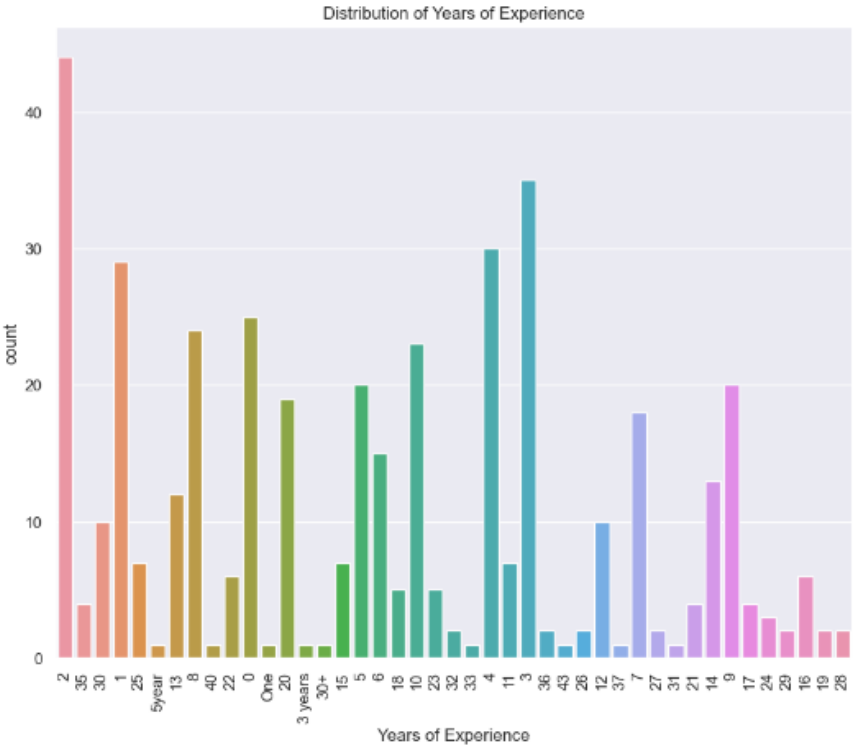


Figure 3.1.20 Count plot for Additional Certificate

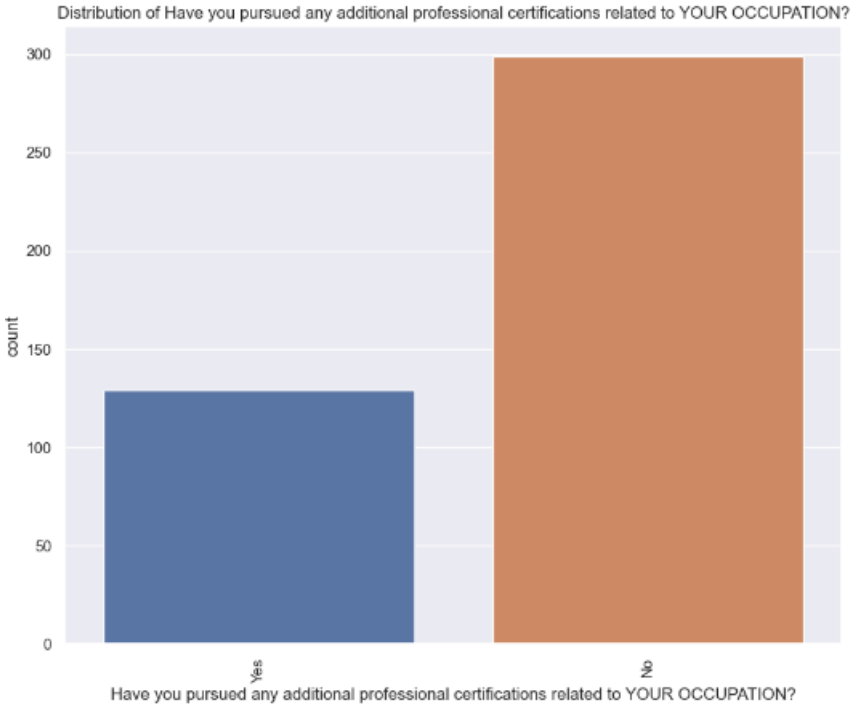


Figure 3.1.21 Count plot for Number of Professional Certificate

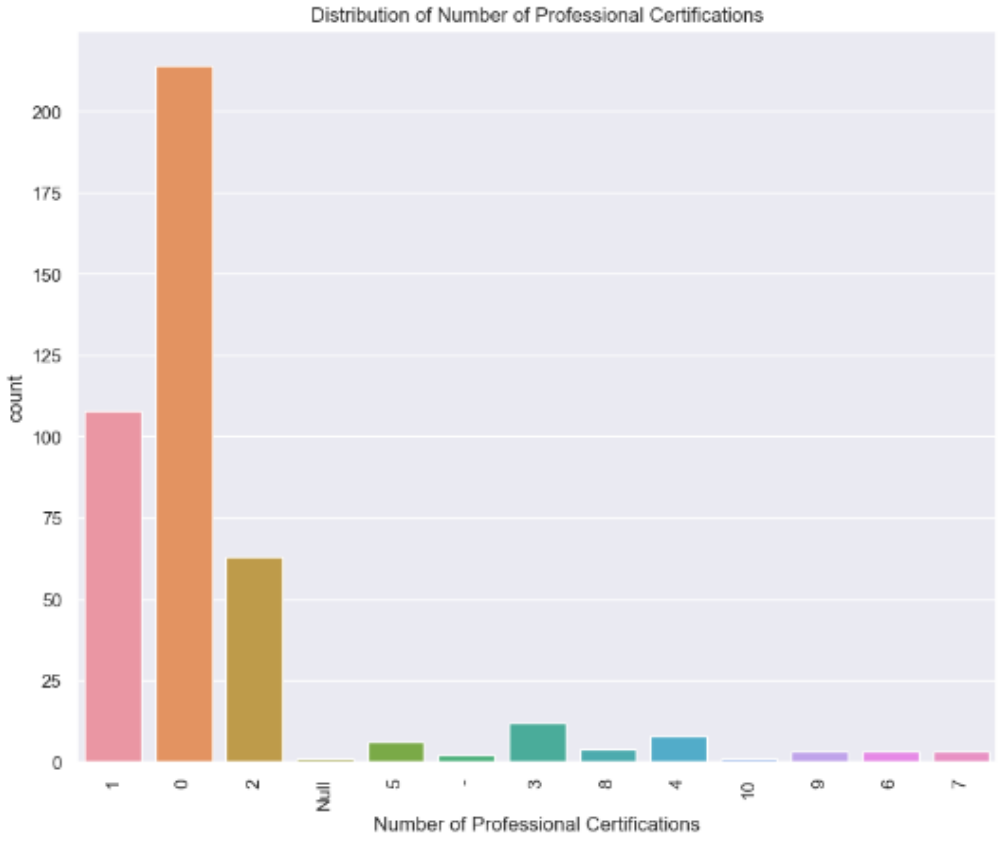


Figure 3.1.22 Count plot for Additional Income or Investment

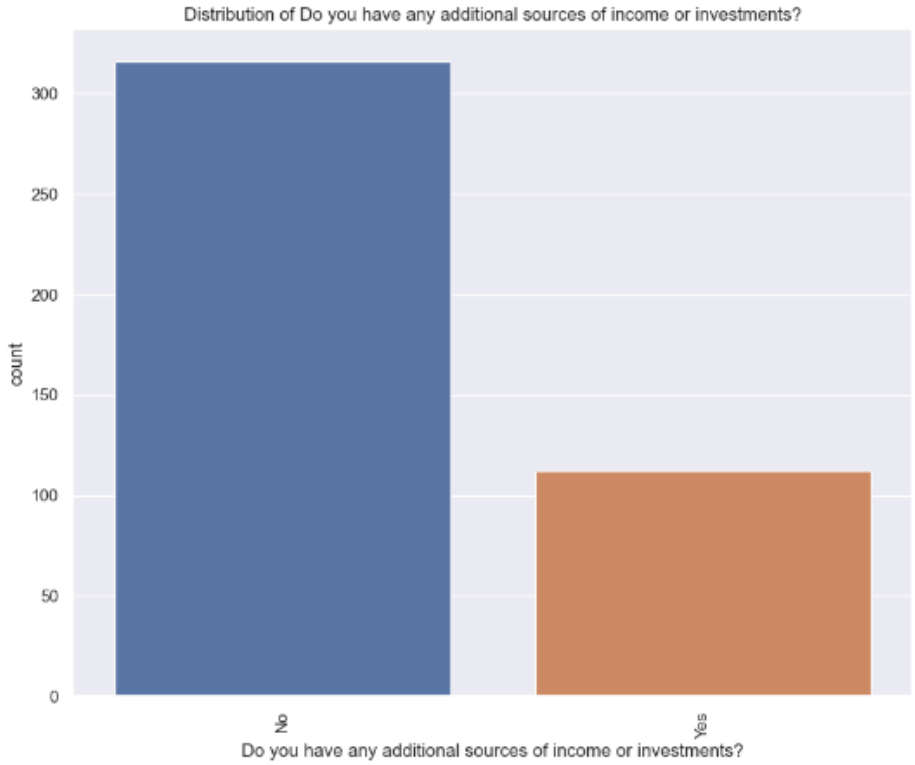


Figure 3.1.23 Count plot for Capital Gain

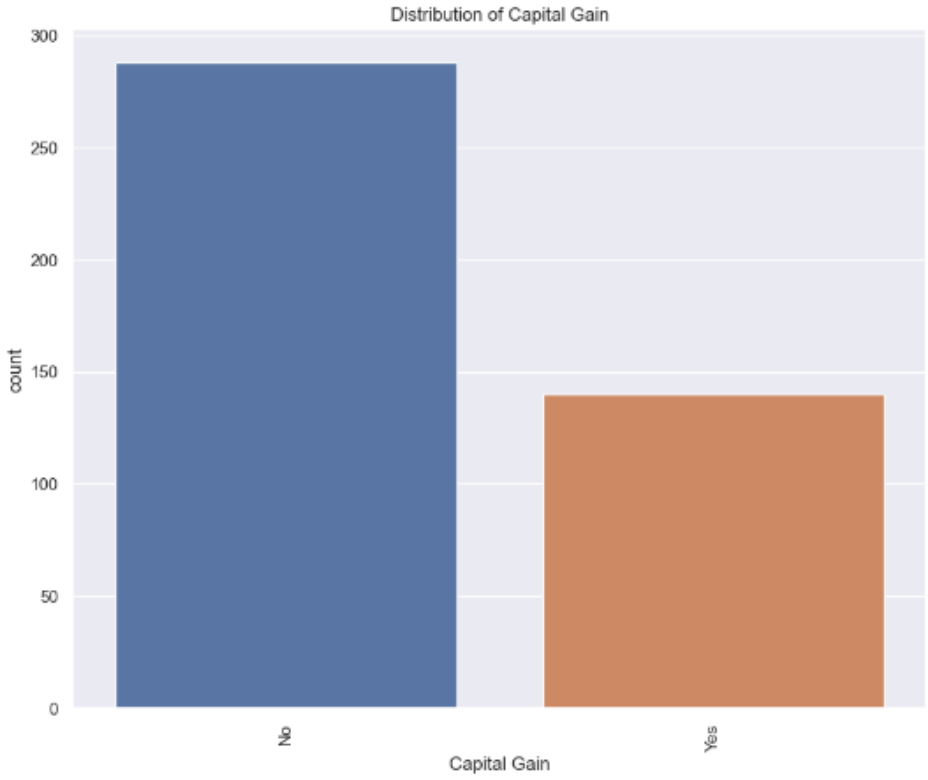
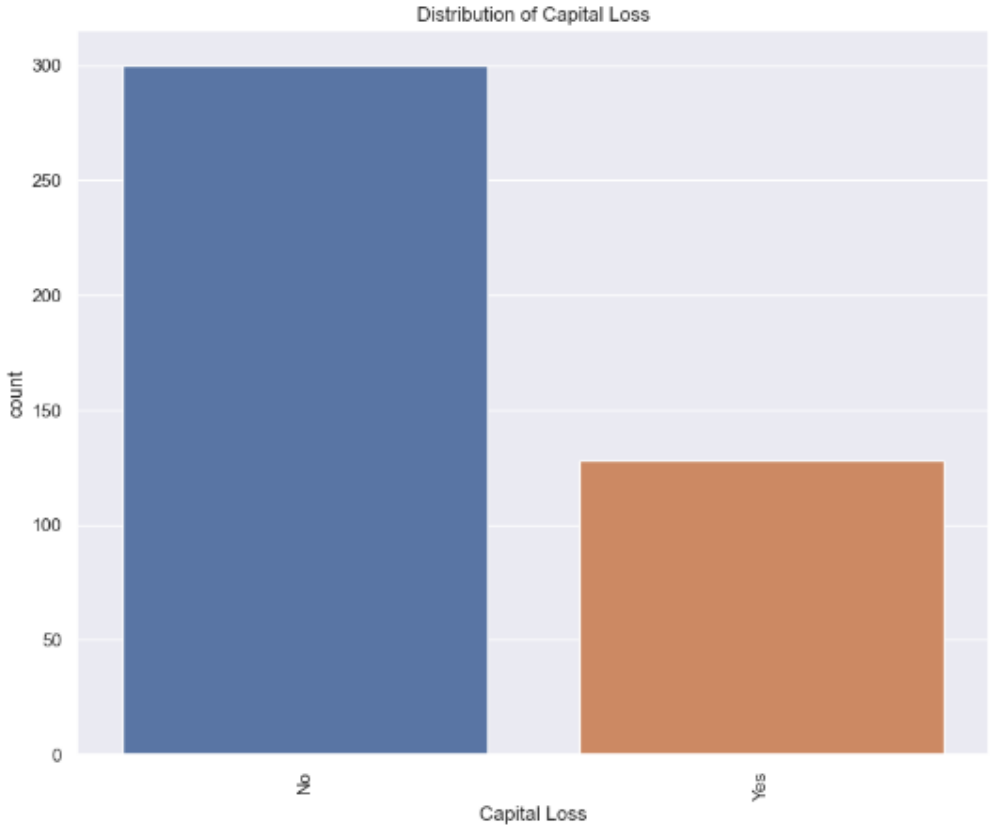


Figure 3.1.24 Count plot for Capital Loss



Following the visualization of the numerical variables, we used a count plot to visualize the categorical variables. The categorical variables in our dataset are Gender, Ethnicity, Highest level of education, Marital status, Own-child, Current Residency, Current Working Place, Have you worked in Ipoh before?, Current employment status, What is your current working environment?, Occupation, What is your job position classification?, Years of Experience, Have you pursued any additional professional certifications related to YOUR OCCUPATION?, Number of Professional Certifications, Do you have any additional sources of income or investments?, Capital Gain, Capital Loss.

Figure 3.1.25 Heatmap for Correlation Coefficient

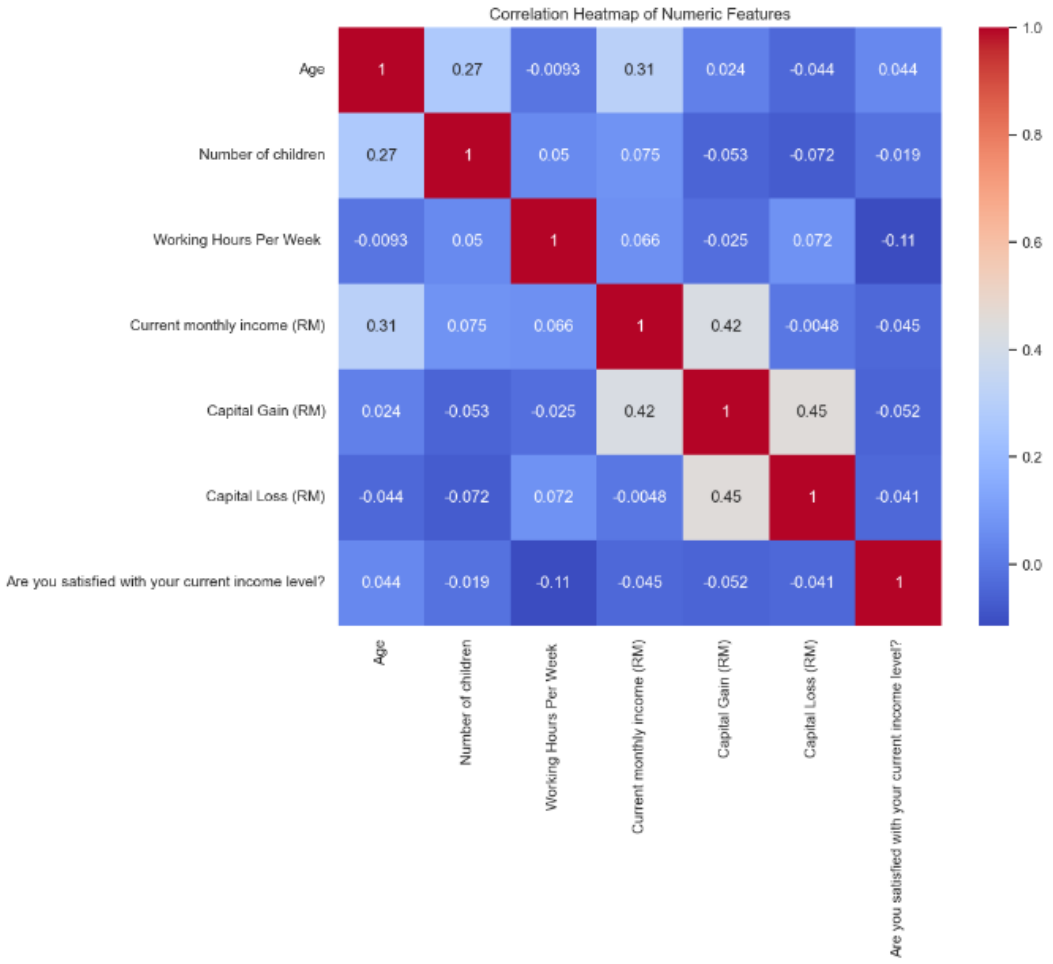
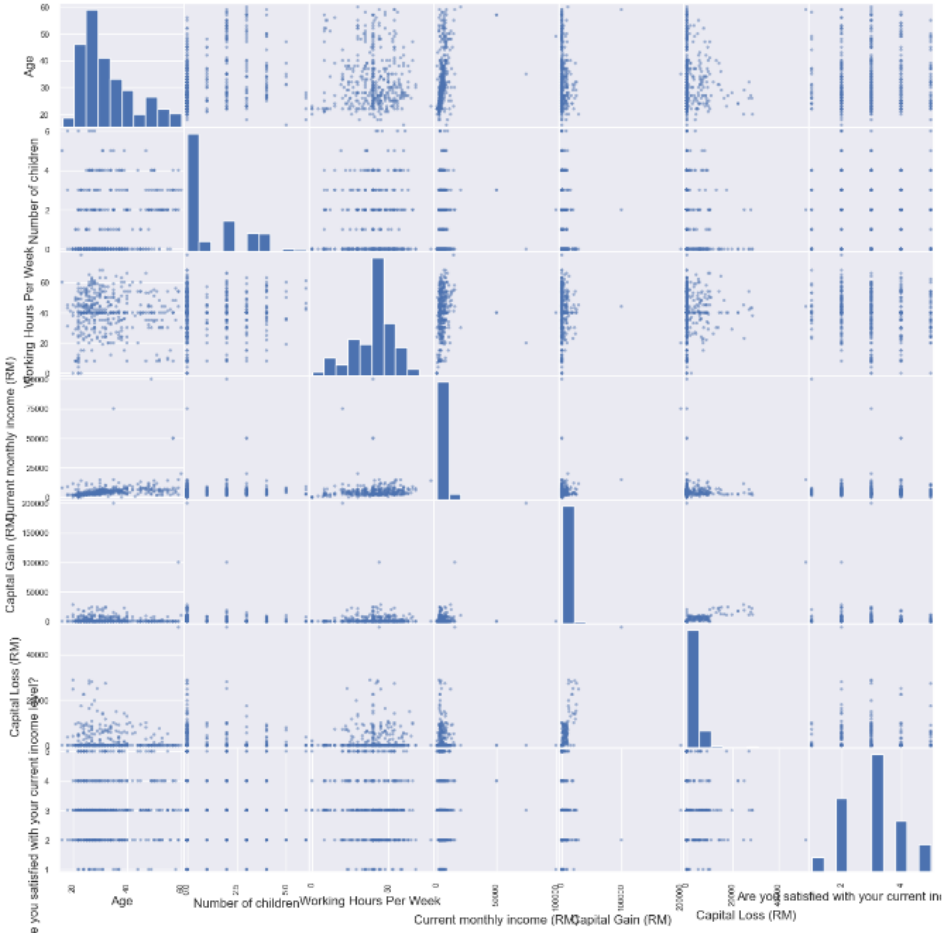


Figure 3.1.26 Correlation coefficient between numerical variables and Current Monthly Income (RM)

```
Current monthly income (RM)                          1.000000
Capital Gain (RM)                                    0.422928
Age                                                  0.308438
Number of children                                   0.074748
Working Hours Per Week                               0.066053
Capital Loss (RM)                                   -0.004758
Are you satisfied with your current income level?   -0.044548
Name: Current monthly income (RM), dtype: float64
```

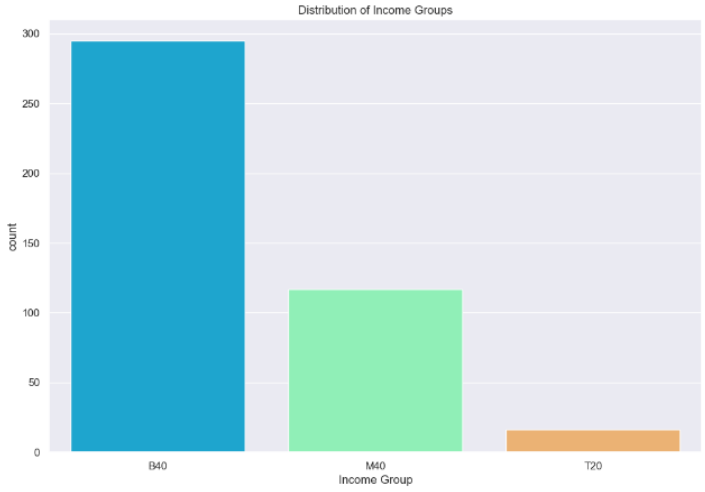
The dependence between each variable was then examined by looking at the correlation coefficient between numerical variables. In order to make the results easier to understand, we sorted the correlation once more after displaying the heatmap initially. We deduced from Figure 3.1.26 that temperature is a highly helpful characteristic for predicting Monthly Income.

Figure 3.1.27 Scatter Matrix of Numerical Variables



A scatter matrix is explored to further study on the relationship between the data. From the scatter matrix, we confirmed that temperature is a good predictor for Current Monthly Income (RM)

Figure 3.1.28 Count plot for Current Monthly Income (RM) categorised in Income Levels

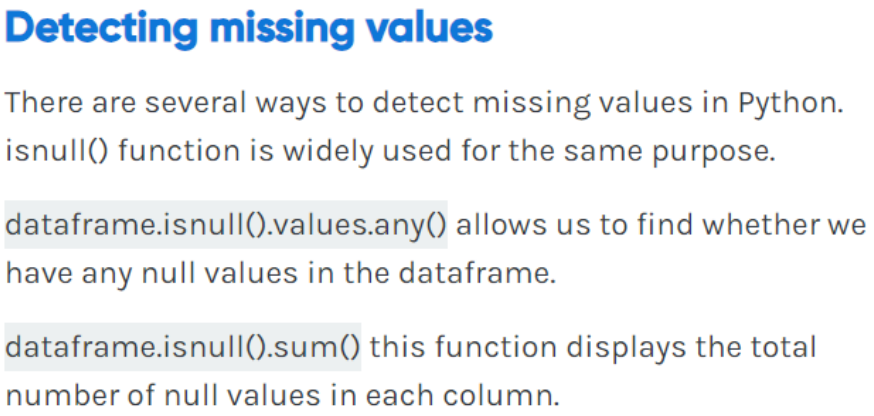


A count plot explored for Currently Monthly Income. It is clearly seen that B40 Ipoh citizens are nearly 290. Following by M40 around 120, and T20 not more than 30.

3.1.6 Data Preparation

The process of cleaning and transforming raw data prior to processing and analysis is known as data preparation. It is an important step before processing that frequently involves reformatting data, making data corrections, and combining datasets to enrich data. Good data preparation allows for efficient data analysis, reduces the possibility of errors and inaccuracies in the processed data, and improves user accessibility for all processed data [44]. This stage helps to handle missing values, outliers, and irrelevant features from the income data. Example function in figure 3.2.1 that displays null values in each column "dataframe.isnull().sum()" or "dataframe.isnull().values.any()" [45].

Figure 3.2.1 Detect Nan code above provided from Neenopal website



whereas one of the example functions handles null value is by using median, codes as shown in figure 3.2.2,

Figure 3.2.2 Median code above provided from Neenopal website

```
#Replacing using median
median = df['column_name'].median()
df['column_name'].fillna(median, inplace=True)
```

other option will have mean, replaced with the number. Common techniques in making data cleaning are one-hot encoding that allows categorical variables in models convert to numerical input, data scaling that set specific range between 0-1 by using libraries scikit-learn, pandas and numpy [46].

Figure 3.2.3 Example outputs One Hot Encoding above provided from Geeksforgeeks.org

Output:

	Emploee_ID	Remarks_Good	Remarks_Great	Remarks_Nice	Gender_Female	Gender_Male
0	45	0	0	1	0	1
1	78	1	0	0	1	0
2	56	0	1	0	1	0
3	12	0	1	0	0	1
4	7	0	0	1	1	0
5	68	0	1	0	1	0
6	23	1	0	0	0	1
7	45	0	0	1	1	0
8	89	0	1	0	0	1
9	75	0	0	1	1	0
10	47	1	0	0	1	0
11	62	0	0	1	0	1

One-Hot encoded columns of the dataset

Figure 3.2.4 Code provided from Medium Website

```
from sklearn.preprocessing import MinMaxScaler

# instance of the MinMaxScaler
sc = MinMaxScaler()

# fit the scaler to the data
sc.fit(data)

# transform the data using the scaler
data_scaled = sc.transform(data)
```

Codes in Figure 3.2.3 and Figure 3.2.4 are all helps for improved predictive model performance.

3.1.7 Predictive Modeling

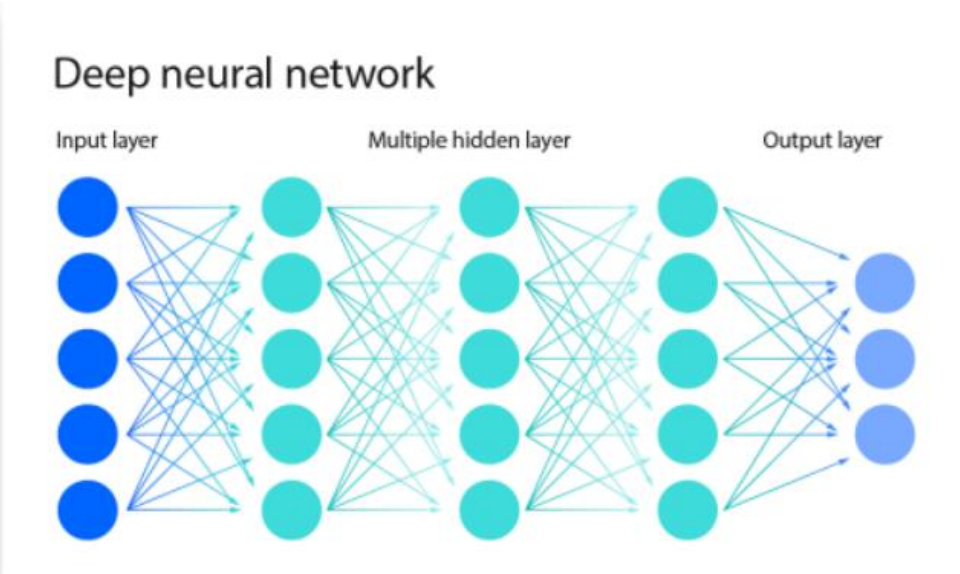
"Predictive modeling is a form of data mining that analyses historical data with the goal of identifying trends or patterns and then using those insights to predict future outcomes," explained Donncha Carroll a partner in the revenue growth practice of Axiom Consulting Partners [47]. Predictive models can be categorized in a variety of ways, and combining different model types can often yield the best outcomes. The key difference is between

supervised and unsupervised models. Modeling applied in the model are decision tree, neural networks, SVM, from Supervised learning algorithms whereas unsupervised learning algorithms will use K-Means Clustering, KNN. The reason for using these algorithms to build models upon the dataset can be seen below.

3.1.8 Neural Networks

Deep learning algorithms are based on neural networks, which are also referred to as simulated neural networks (SNNs) or artificial neural networks (ANNs). Neural networks are a subset of machine learning. Their nomenclature and organization are derived from the human brain, emulating the communication patterns of biological neurons [48]. While the project involved deep learning algorithms, it has affirmatory become machine learning and caused the work related to Artificial Intelligence.

Figure 3.2.5 A concept of Neural Networks provided by IBM website

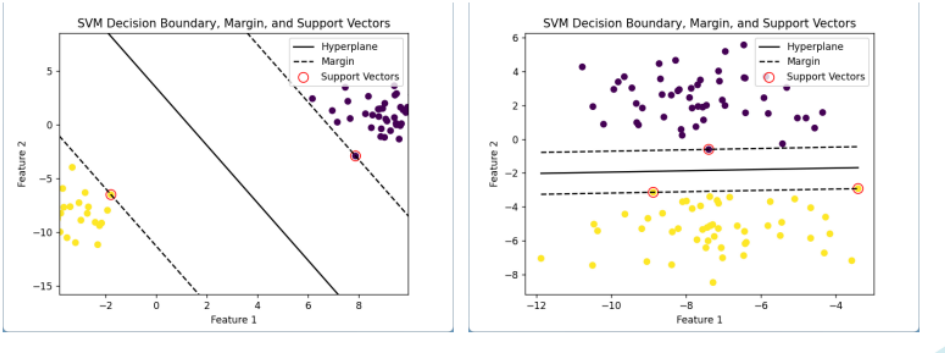


In figure 3.2.5, it drafts the structure of Neural Networks, reason using Neural Networks, due we need the model that can learn directly from raw income-related data, and deep neural networks can perform end-to-end learning, able to map raw income inputs to predictions without intermediate manual processing.

3.1.9 Support Vector Machines (SVM)

SVM is an effective supervised algorithm that performs best on complex but smaller datasets. Although Support Vector Machines, also known as SVMs, are useful for classification as well as regression applications, their performance is generally greatest in the former [49]. The reason for using SVM is because it aims to find a clear margin of separation between classes, which can lead to better generalization.

Figure 3.2.6 A picture of SVM apply margin on separation classes provided by Data Mining Techniques



In figure 3.2.6, Support vector machines can provide clear decision boundaries in situations where groups or classes at different income levels need to be identified.

3.1.10 K-Means Clustering

Data assigned to distinct groups (clusters) based on shared characteristics among observations within a given group is known as clustering [50]. As an illustration,

Figure 3.2.7 above is an example of K-Means Clustering by Intuitive Tutorials

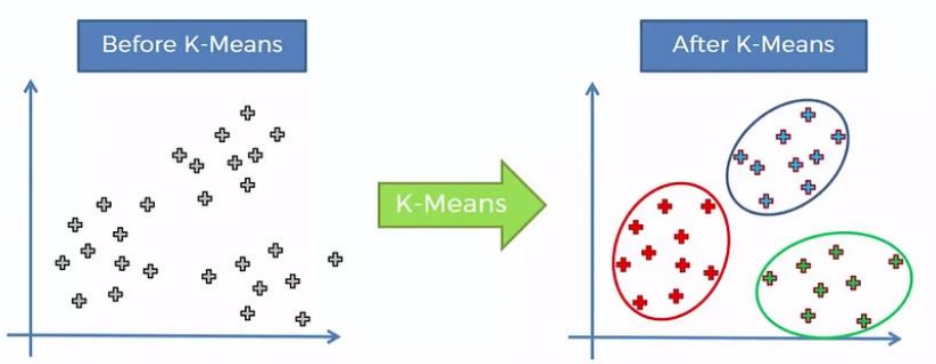


Photo by Google Images

K-Means is an algorithm for unsupervised learning that groups data points into clusters according to similarities. K-Means clustering can assist in locating natural clusters within the income dataset if it contains innate patterns or groups that aren't clearly labeled. For example, in figure 3.2.7, it may identify different sections or groups within the income dataset, possibly exposing trends or associations that could be useful to policymakers. Following the identification of clusters, you can profile each group to learn about the traits and dynamics connected to various income brackets. At the end, by identifying income clusters, policymakers can use clustering to get specific insights that help them customize strategies and policies.

3.1.11 KNN

Based on the similarities in the text vectors it discovers during training, the KNN algorithm divides the text vectors into the categories of positive and negative reviews. In order to assign new data points to appropriate categories based on their similarity to a specific set of text vectors, the KNN algorithm learns a mathematical function [51].

Figure 3.2.8 Example KNN algorithm work provided by Intuitive Tutorials

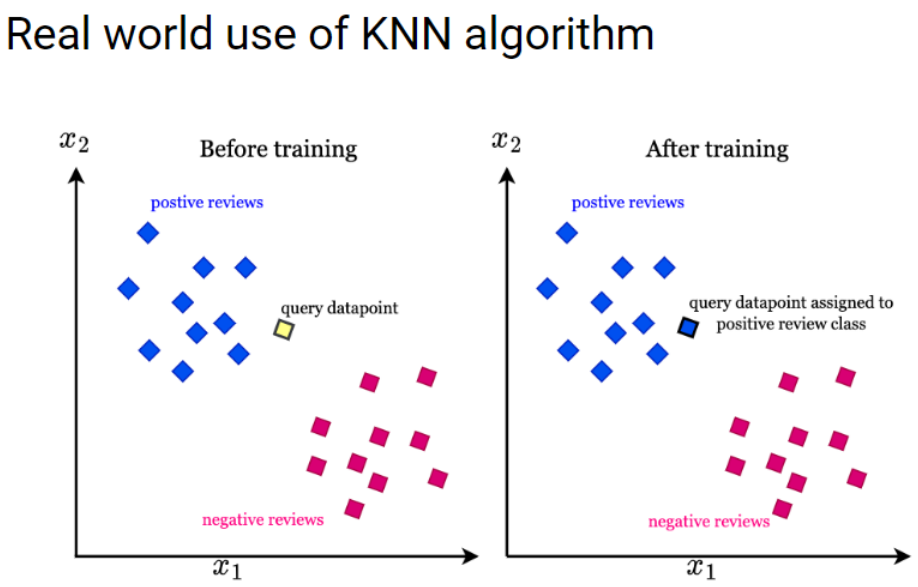


Figure 3.2.8 KNN takes into account the characteristics of individuals that are similar in the dataset to provide personalized predictions. It is appropriate for situations where income levels may be impacted by nearby or comparable people since it can identify local patterns in the data. KNN can respond to changes in income dynamics over time because it can adjust well to changes in the dataset.

3.1.12 Model Evaluation

Model evaluation is the process of analyzing a machine learning model's performance and strengths and weaknesses using various evaluation metrics. Model evaluation is crucial for determining a model's effectiveness in the early stages of research and is involved in model monitoring [52].

Accurately assessing the performance of the income prediction model is crucial to ensure its effectiveness and reliability. Several key metrics, including accuracy, precision, recall, F1 score, and AUC-ROC, will be used to evaluate the model's ability to accurately predict income levels. Additionally, cross-validation techniques like k-fold cross-validation will be employed to prevent overfitting and enhance the model's generalization to unseen data.

Furthermore, the performance of various machine learning algorithms, such as regression, decision trees, random forest, SVM, KNN, neural networks, and ensemble methods,

will be compared to identify the most effective algorithm for this specific task. Importantly, the evaluation process will be aligned with the broader business or project objectives.

Test cases mimicking real-world scenarios will be developed and chosen evaluation metrics will be ensured to resonate with the goals of government policymakers. Ultimately, the goal is to provide valuable insights that will contribute to informed economic decisions.

3.1.13 Deployment

At the end of project, deployment stage examines the predictive tool's scalability in light of prospective increases in data volume and user interactions. It ensures the implemented solution is capable of meeting demands in the real world.

While deploy the model by showing them the ipynb file might affects the policymaker's user experience. To improve user experience by creating a dashboard's user interface. For the benefit of policymakers and data mining users who will be interacting with the tool, make it clear, easy to use, and educational.

Other than that, provide the predictive tool with thorough guideline on how to use all its features, including how to interpret results and comprehend visualizations and provide avenues for support to users who might need help or have questions when using the dashboard.

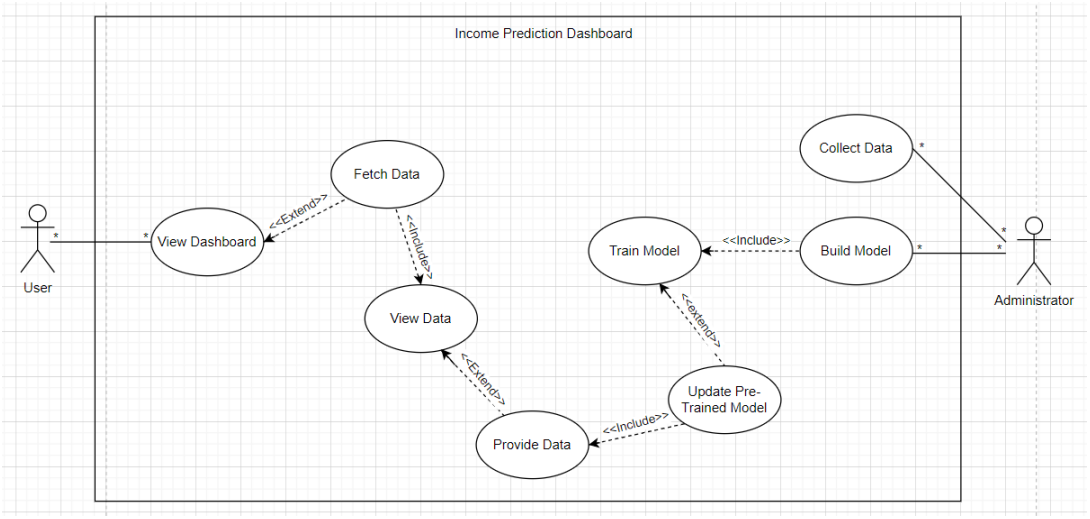
In conclusion, the processes of the project were divided into six phases which were project proposal, survey literature review, data collection, data pre-processing, model training selection building and data training, and lastly test dataset prediction.

3.2 System Design Diagram

3.2.1 Use Case Diagram

First of all, an administrator will need to perform data collection outside of the system. Once the new dataset is received, the next step is data preprocessing. After that, models for prediction need to be selected for the next building stage. Once the models are built, they need to be trained. If the administrator chooses to update the pre-trained models on the dashboard, the dashboard will be updated accordingly, providing data for users to view. Users can select features by clicking different function buttons on the dashboard.

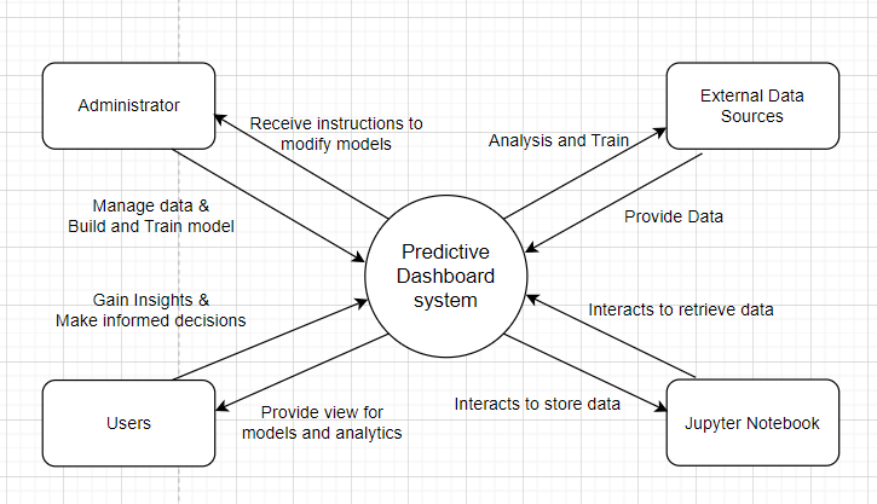
Figure 3.2.9 Use Case Diagram



3.2.2 Context Diagram

The Administrator interacts directly with the Predictive Dashboard System to perform tasks such as data preprocessing, model building, and model training. Users interact with the Predictive Dashboard System through its user interface to access predictions and visualizations. External Data Sources provide raw data to the Predictive Dashboard System for analysis. The Jupyter Lab Database may store datasets and other resources used by the system during data preprocessing and model building.

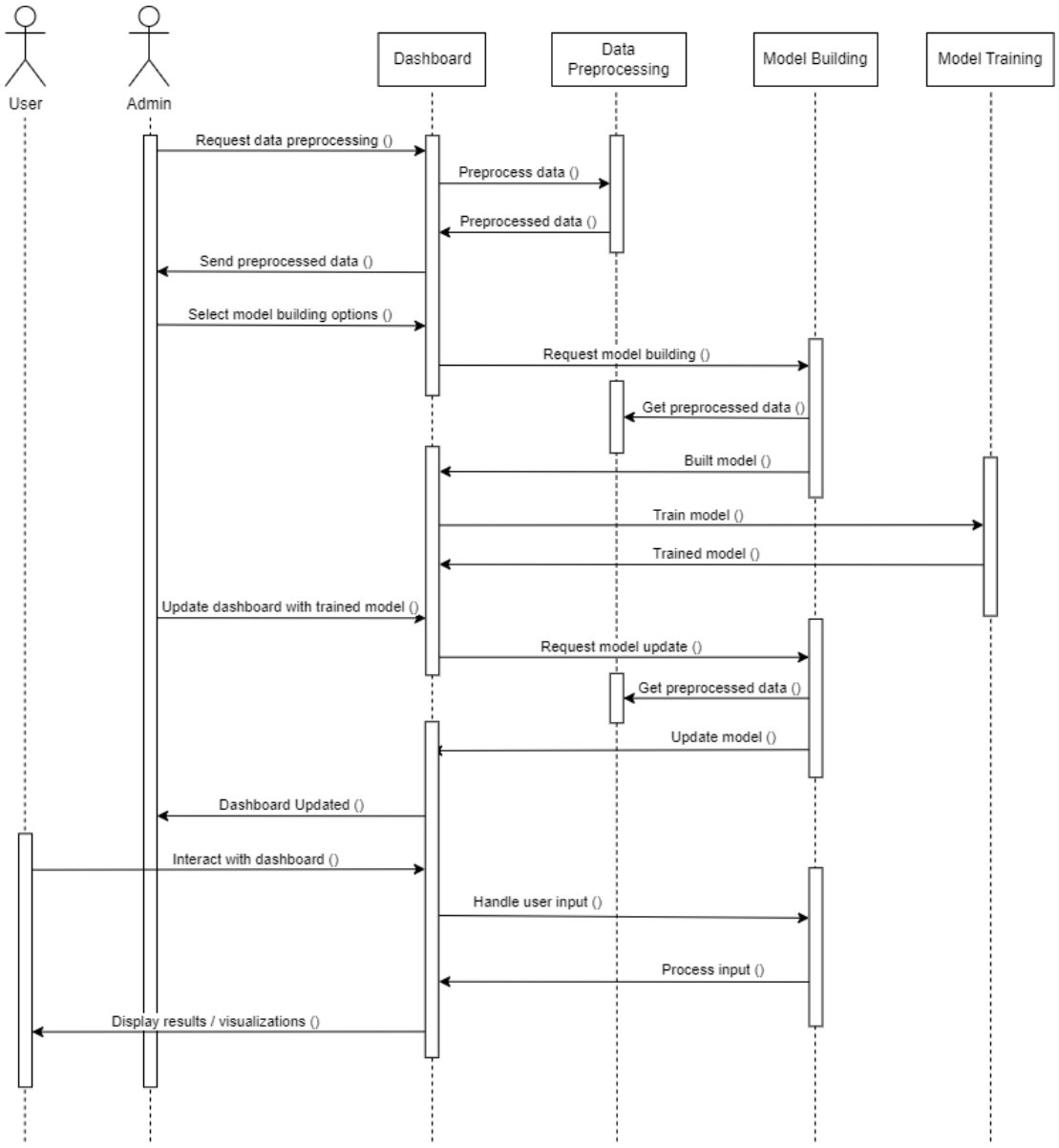
Figure 3.2.10 Context Diagram



3.2.3 Sequence Diagram

The sequence diagram illustrates the interactions within the Predictive Dashboard System, highlighting the flow of actions between the administrator, dashboard, data preprocessing module, model building module, model training module, and user.

Figure 3.2.11 Sequence Diagram



The sequence diagram illustrates the interactions within the Predictive Dashboard System, highlighting the flow of actions between the administrator, dashboard, data preprocessing module, model building module, model training module, and user.

Data Preprocessing:

The sequence begins with the Administrator requesting data preprocessing from the Dashboard. The Dashboard forwards the request to the Data Preprocessing module. Data Preprocessing preprocesses the data and returns it to the Dashboard. The Dashboard sends the preprocessed data back to the Administrator.

Model Building and Training:

The Administrator selects model building options through the Dashboard. The Dashboard initiates model building by requesting it from the Model Building module. Model Building retrieves the preprocessed data from the Data Preprocessing module. After building the model, it sends the built model to the Dashboard. The Dashboard then triggers model training by sending the model to the Model Training module. Model Training trains the model and returns the trained model to the Dashboard.

Model Update:

The Administrator decides to update the dashboard with the trained model. The Dashboard requests a model update from the Model Building module. Model Building retrieves the preprocessed data and updates the model accordingly. The updated model is sent back to the Dashboard, completing the update process.

User Interaction:

Users interact with the dashboard by selecting various options. The Dashboard handles user input and communicates with the Model Building/Training modules accordingly. The Model Building/Training modules process the input and provide results or visualizations back to the Dashboard. Finally, the Dashboard displays the results/visualizations to the user.

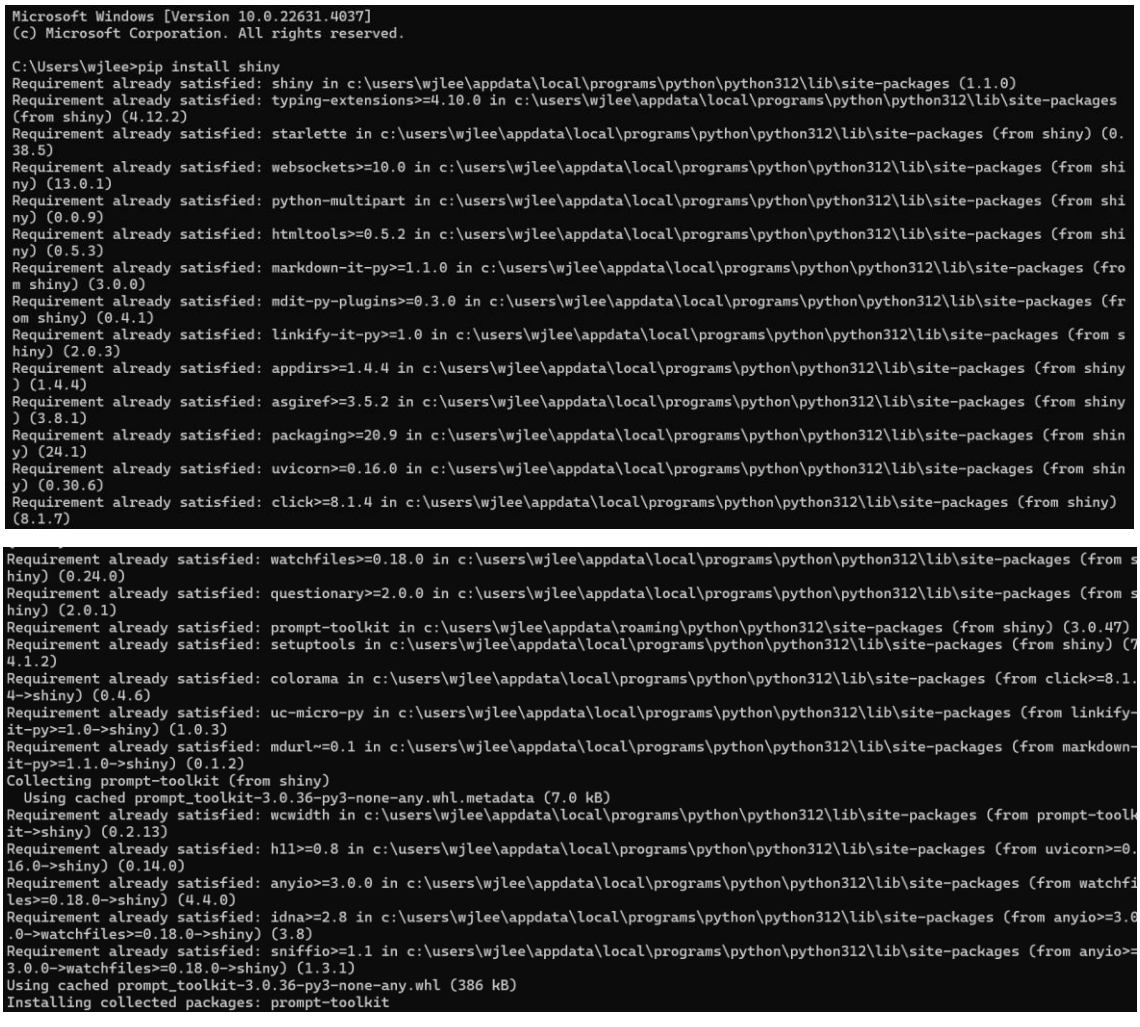
Chapter 4

System Design

4.1 Command Prompt

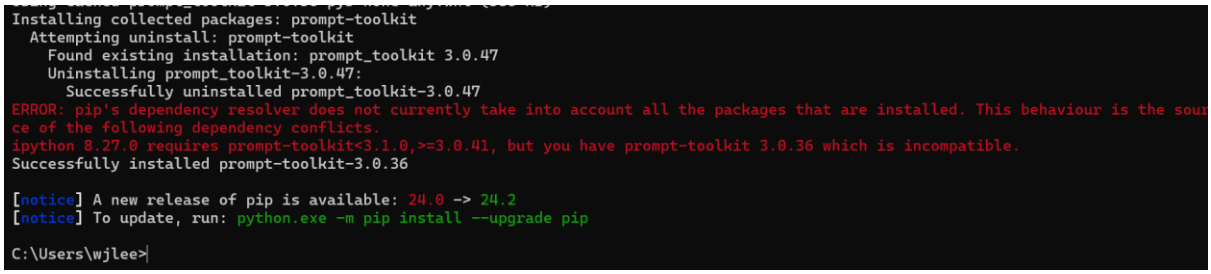
First step of creating the dashboard, Shiny for python is required to be stored on the local system as shown in figure 4.1.1.

Figure 4.1.1 Installation Shiny Step 1



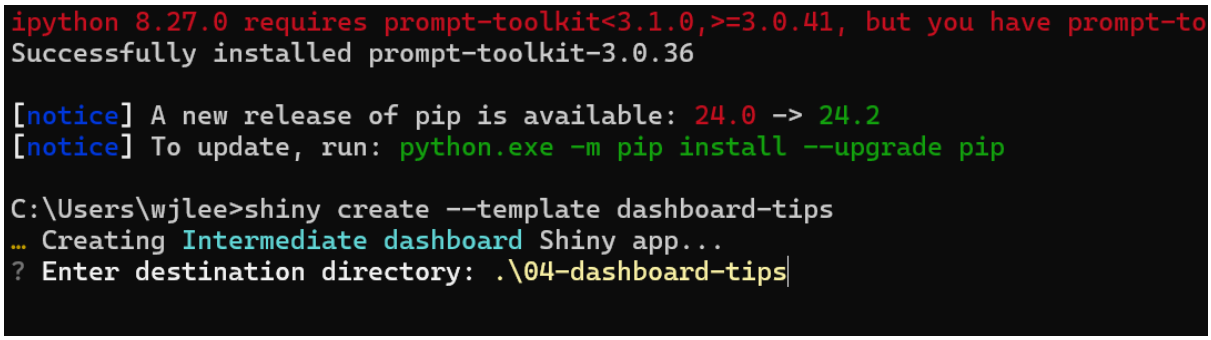
Once the installation has been successfully installed, the “python.exe -m pip install --upgrade pip” as shown in Figure 4.1.2 would be ask to the user to upgrade the pip package installer to its latest version, but it’s not necessarily to do so because it has not affected much to the installed version of python in your laptop

Figure 4.1.2 Installation Shiny Step 2



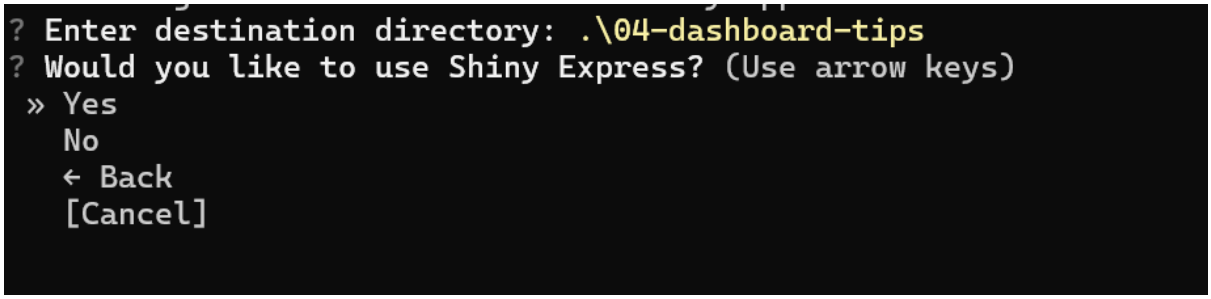
Next step, “shiny create --template dashboard-tips” as shown in figure 4.1.3 is required to insert as to create a folder and template in your local device.

Figure 4.1.3 Folder and template create in local device



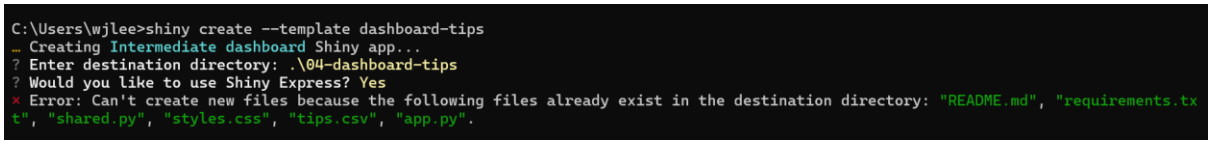
As by clicking “Enter”, following question will be asked as shown in figure 4.1.4

Figure 4.1.4 Shiny Express



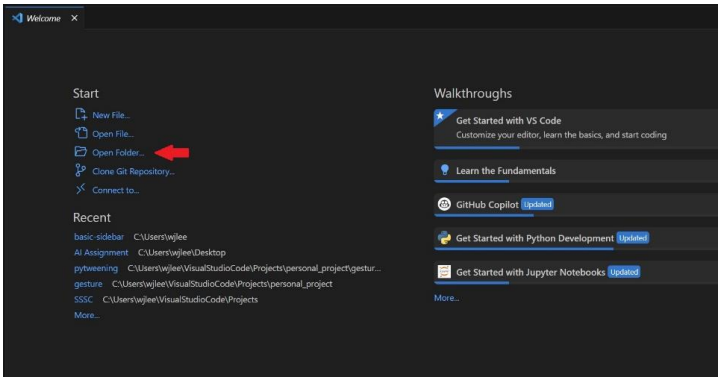
By clicking “Yes”, the files such as "README.md", "requirements.txt", "shared.py", "styles.css", "tips.csv", "app.py" will be created as the CSS file and Navigation file and these will be installed in the destination directory. If same command used on the same directory, error will be appeared as shown in Figure 4.1.5.

Figure 4.1.5 File created in same directory



Once the steps reach this stage, developers can leave the command prompt aside, and proceed to the Visual Studio Code or other web development tools. Whereas for this project, “Visual Studio is chosen as the tools develop environment. Next, open up the Visual Studio Code, and clicked the “Open folder" button as shown in Figure 4.1.6.

Figure 4.1.6 Visual Studio Code



The folder which needs to be select is under the C Drive > Users as shown in Figure 4.1.7. and Figure 4.1.8

Figure 4.1.7 C Drive

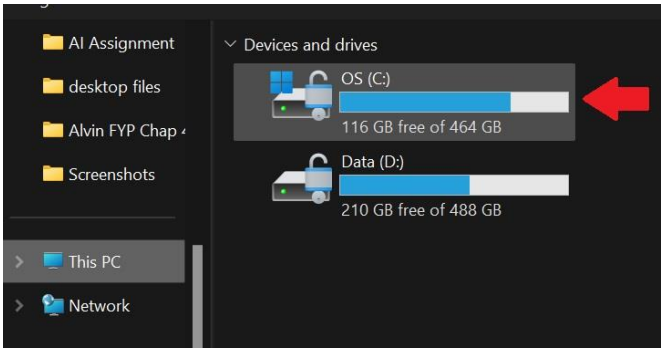
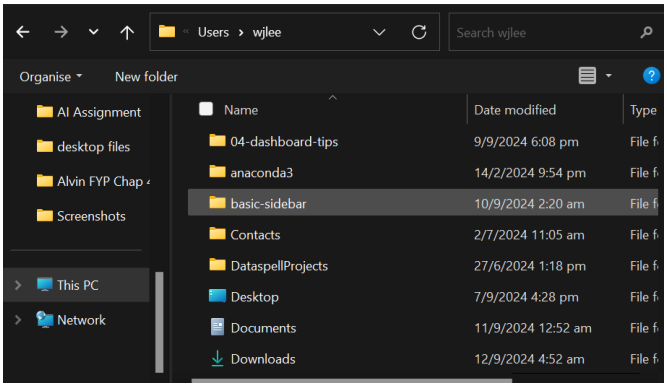


Figure 4.1.8 Folder needed to be select on project



4.2 Features of the dashboard

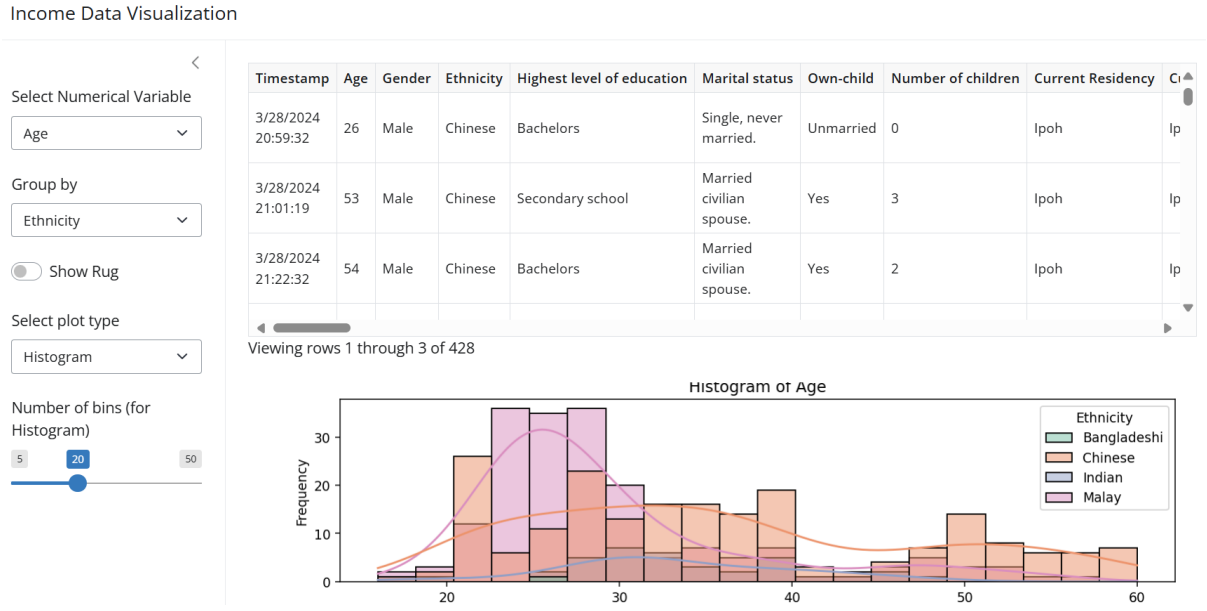
Finally, developer can start to develop dashboard by using shiny python. The dashboard was created by using Visual Studio Code but do remember that extension for Shiny is needed to be added into VS Code libraries.

In the Income Prediction dashboard, these functions from Shiny Libraries are making important features as its help user to determine the categories of information will be displayed in the output. Functions such as “ui.input_select”, “ui.input_switch”, and “ui.input_slider” have been applied to this Income Prediction Dashboard’s sidebar as shown in figure 4.1.9 and figure 4.1.10

Figure 4.1.9 Example code working on the dashboard

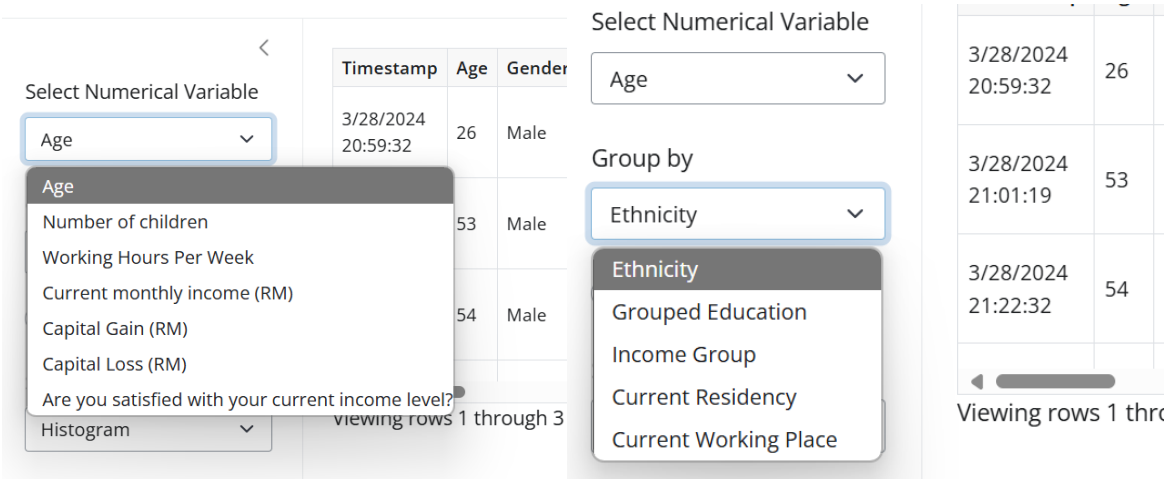
```
# UI setup for your dashboard
ui.page_opts(title="Income Data Visualization", fillable=True)

with ui.sidebar():
    # Numerical variables for numeric features like income, capital gain, etc.
    ui.input_select("var", "Select Numerical Variable", choices=df.select_dtypes(include=np.number).columns.
    # Categorical variables like Ethnicity, Education, and Residency
    ui.input_select("group_by", "Group by", choices=['Ethnicity', 'Grouped Education', 'Income Group', 'Curr
    ui.input_switch("show_rug", "Show Rug", value=False)  # Optional rug plot for distributions
    ui.input_select("plot_type", "Select plot type", choices=["Histogram", "Boxplot", "Correlation Matrix",
    ui.input_slider("bins", "Number of bins (for Histogram)", min=5, max=50, value=20)  # For histogram bin
```

Figure 4.1.10 Income Prediction Dashboard



The selection under “Select Numerical Variable” and “Group by” is to let user to choose the category they would like to refer with from the dataset. Sample shown as below in Figure 4.1.11.

Figure 4.1.11 Selector for “Select Numerical Variable” and “Group by”



When user pointed on the input, the list of categories attribute from dataset will be listed out on the dashboard, and user is able to see different output based on different input combination selected.

4.3 Functions of the dashboard

The code implementation includes two core functions, data() and visualizations(), respectively in figure 4.2.1 and 4.2.2 both of which are designed to render the dataset and generate visualizations based on user inputs.

Data Display with data()

The Data() function is responsible for displaying the dataset in a structured format. It utilizes the @render.data_frame decorator, which suggests that the function is integrated into a web-based application environment, likely to display the dataset interactively. The function simply returns the dataset (df), allowing users to inspect the raw data before performing any analysis or visualization through shared.py file.

Figure 4.2.1 Display data()

```
@render.data_frame
def data():
    return df  # Displaying the dataset
```

Generating Visualizations with visualizations()

The Visualizations() function plays a crucial role in generating plots based on user preferences. This function also is decorated with @render.plot, indicating that it renders visual output such as charts or graphs as shown in figure 4.2.2 It dynamically responds to the following user inputs:

Plot Type: This allows the user to specify the kind of plot they wish to generate (e.g., histogram, bar chart).

Group By: This is a categorical variable (e.g., 'Ethnicity') used to segment the data within the plot, enhancing the analysis by enabling a comparison across different groups.

Variable: This is the numerical variable (e.g., 'Age' or 'Income') that will be plotted.

Figure 4.2.2 Render.Plot call function on Visualization()

```
@render.plot
def visualizations():
    plot_type = input.plot_type()
    group_by = input.group_by()  # Categorical variable like 'Ethnicity'
    var = input.var()  # Numerical variable like 'Age', 'Income', etc.

    if plot_type == "Histogram":
        # Seaborn requires long-form data when using 'hue'.
        if group_by:
            # Check if the group_by column exists and is categorical
            if group_by in df.columns:
                # Make sure it's treated as a categorical column
                if df[group_by].dtype.name != 'category':
                    df[group_by] = df[group_by].astype('category')

                sns.histplot(data=df, x=var, hue=group_by, bins=input.bins(), kde=True, palette="Set2")
            else:
                plt.text(0.5, 0.5, f"Column '{group_by}' not found in data", fontsize=12, ha='center', va='center')
        else:
            sns.histplot(df[var], bins=input.bins(), kde=True, color="skyblue")

        if input.show_rug():
            sns.rugplot(df[var], color="black", alpha=0.25)

        plt.title(f'Histogram of {var}')
        plt.xlabel(var)
        plt.ylabel('Frequency')
```

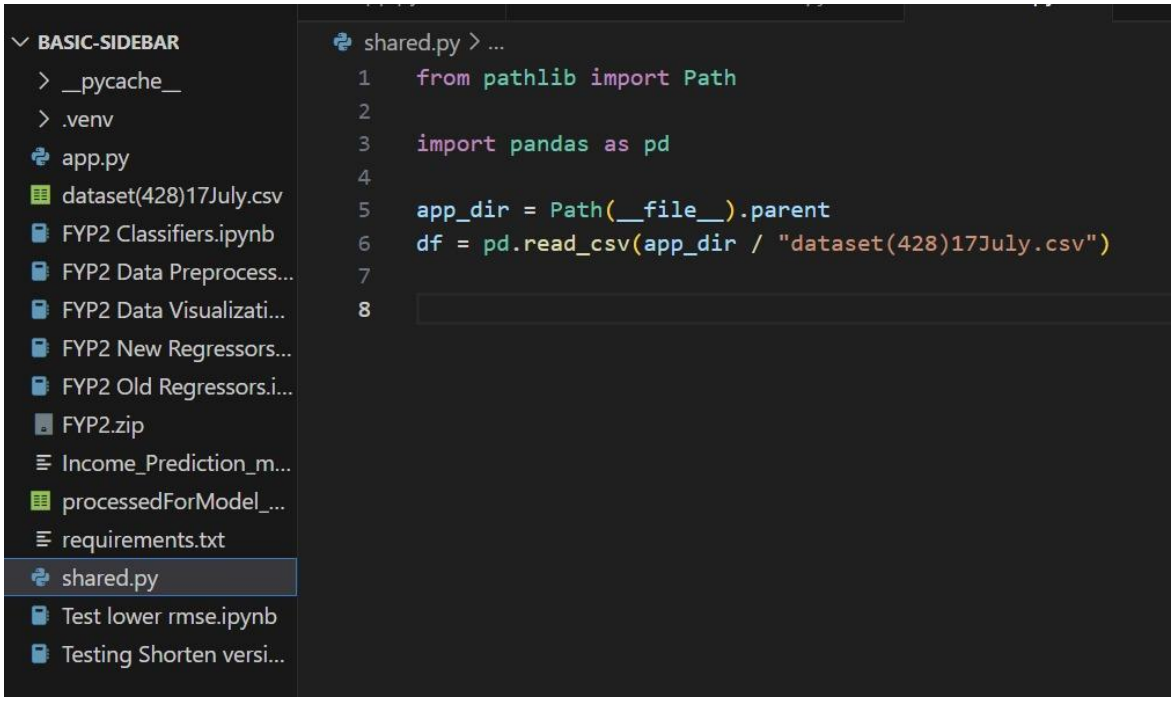
The function is designed to handle user inputs efficiently. If the plot type is a "Histogram" and a group_by variable is specified, the function first verifies whether the group_by column exists in the dataset. Additionally, it checks whether the variable is correctly classified as a categorical

type. If not, it converts the column to a categorical format using the astype('category') method to ensure proper grouping in the plot.

4.3 Dataset pathway

In Shiny for python case, dataset required to make a navigation in a file called “shared.py” as its provide a reliable way of loading the dataset, regardless of where the script is being executed from. By dynamically referencing the directory of the script (shared.py) as shown in figure 4.3.1.

Figure 4.3.1 Shared.py



The way doing this helps making the module portability. The dataset can be easily located without hardcoding an absolute path, which would break if the file is moved to a different directory or machine. This relative path approach guarantees the dataset is found as long as it resides in the same directory as shared.py. Another advantage is that it is ease of use in Shared Environments. If multiple scripts or parts of the application need to access the same dataset, centralizing the dataset loading in shared.py simplifies the process. Other scripts or modules can simply import the shared df DataFrame, rather than each script handling file loading separately. This reduces duplication and makes the codebase easier to maintain.

In the shared.py file, this code efficiently locates and loads the dataset (dataset(428)17July.csv) relative to the location of the script, using the pathlib and pandas libraries. By determining the

directory where the script is stored with Path(_file_).parent, the file ensures that the correct path is dynamically constructed, allowing the dataset to be loaded reliably and without needing to hardcode file paths. This approach enhances portability and maintainability, making the code robust when deployed or shared across different environments.

4.4 Model Pipeline

On the other hand, beside dashboard designed avoiding complexity, the dataset also has been processed by using machine learning pipeline and it built to streamline the data preprocessing and modeling workflow. The pipeline automates several stages of data preparation, transforming both numerical and categorical features, and integrates these transformations with a machine learning algorithm for predicting income. This approach ensures consistency, reduces code duplication, and allows for easy model tuning through hyperparameter search.

4.4.1 Pipeline Structure

The pipeline in this project consists of two main components

1. **Preprocessing Pipeline**

This component handles the transformations required for both numerical and categorical features. The dataset contains a mix of numerical, nominal (categorical without ordinal relationship), and ordinal categorical features. The preprocessing pipeline ensures that each of these types of data is treated appropriately before feeding it into the model.

2. **Model**

A RandomForestRegressor is the chosen machine learning algorithm in this pipeline. This algorithm is embedded within the pipeline and is tuned using cross-validation and randomized search.

4.4.2 Detailed Breakdown of the Pipeline

The pipeline used in this project is designed to handle both data preprocessing and model training in a structured, automated fashion as shown figure 4.4.1. It ensures that each type of feature, whether numerical, nominal categorical, or ordinal categorical, is appropriately transformed before being passed to the machine learning model. This multi-step pipeline reduces complexity and maintains consistency across different stages of the workflow, from data cleaning to model evaluation.

Figure 4.4.1 Pipeline structure in Income Prediction

```
[84]: from sklearn.compose import ColumnTransformer
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import OrdinalEncoder, StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer

num_attribs = [
    "Age",
    "Number of children",
    "Working Hours Per Week",
    "Satisfaction_Income",
]

nom_cat_attribs = [
    "Gender",
    "Ethnicity",
    "Marital_Status_Grouped",
    "Own_Child_Grouped",
    "Location",
    "Work_Ipoh_Before",
    "Working_Environment",
    "Occupation",
    "Certs_Occupation",
    "Additional_SourceIncomeInvest",
    "Experience_Binned"

]

ord_cat_attribs = [
    "Job_Position_Grouped",
    "Highest_Education",
    "Employment_Status_Grouped",
    "Certifications_Grouped"
]

num_pipeline = make_pipeline(
    SimpleImputer(strategy="median"),
    StandardScaler()
)

nom_cat_pipeline = make_pipeline(
    SimpleImputer(strategy="most_frequent"),
    OneHotEncoder(drop='first', handle_unknown='ignore', sparse=False)
)

ord_cat_pipeline= make_pipeline(
    SimpleImputer(strategy="most_frequent"),
    OrdinalEncoder(handle_unknown='use_encoded_value', unknown_value=-1)
)

preprocessing = ColumnTransformer([
    ("num", num_pipeline, num_attribs),
    ("nom_cat", nom_cat_pipeline, nom_cat_attribs),
    ("ord_cat", ord_cat_pipeline, ord_cat_attribs),
])
```

Numerical Feature Preprocessing

Numerical features such as “Age”, “Number of children”, “Working Hours Per Week”, “Satisfaction_Income” and “Years_of_Experience”.

For these features, the pipeline applies the following transformations:

a. Imputation: Missing values in numerical columns are imputed using the median strategy (SimpleImputer(strategy='median')), which is a robust approach, especially in the presence of outliers.
b. Scaling: After imputation, the numerical features are standardized using StandardScaler(), ensuring that all features have a mean of 0 and a standard deviation of 1, which is important for models sensitive to feature scaling, like tree-based models.

Nominal Categorical Feature Preprocessing

Nominal categorical features such as “Working_Environment”, “Occupation”, “Certs_Occupation”, “ Additional_SourceIncomeInvest”, “Experience_Binned”

For nominal (unordered) categorical features, the following steps are applied:

a. Imputation: Missing values are filled with the most frequent value in the column (SimpleImputer(strategy='most_frequent')).
b. One-Hot Encoding: Categorical values are then transformed using OneHotEncoder(), which converts each category into a binary vector. This method allows the model to handle categorical data without introducing any implicit order or ranking.

Ordinal Categorical Feature Preprocessing

Ordinal categorical features such as "Job_Position_Grouped", "Highest_Education", "Employment_Status_Grouped" and "Certifications_Grouped"

Since these features have an inherent order (e.g., levels of education, employment status), the following transformations are applied:

i. Imputation: Missing values are again imputed with the most frequent value (SimpleImputer(strategy='most_frequent')).
ii. Ordinal Encoding: The OrdinalEncoder() is used to assign numerical values to the categories, maintaining the order of the categories. The parameter handle_unknown='use_encoded_value', unknown_value=-1 is included to handle any unseen categories that may appear in test data by encoding them with a value of -1.

Random Forest Regressor

After preprocessing, the transformed data is passed to a RandomForestRegressor, a convenient tree-based ensemble model that is highly capable of handling both numerical and categorical data. The RandomForestRegressor was chosen for its ability to handle complex, non-linear relationships and its robustness to overfitting due to the ensemble nature of decision trees.

The random forest model was integrated directly into the pipeline as the final estimator. The model was trained using the preprocessed data, and hyperparameters such as max_features were optimized using RandomizedSearchCV.

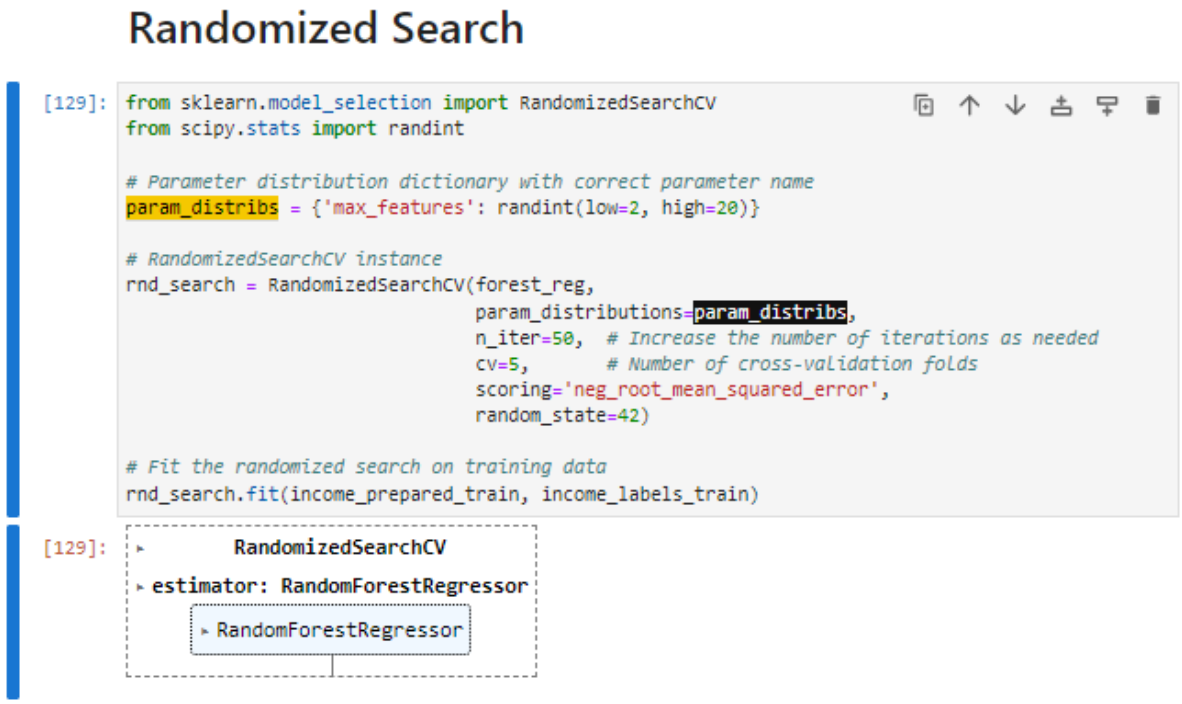
Hyperparameter Tuning

To further improve model performance, a hyperparameter search was conducted using RandomizedSearchCV. The following parameter was tuned:

i. max_features: The number of features considered by the model for splitting at each node, with values ranging between 2 and 20. The RandomizedSearchCV was used with 10 iterations (n_iter=10) and 3-fold cross-validation (cv=3). The scoring metric used

was the negative root mean squared error (neg_root_mean_squared_error), which aligns with the goal of minimizing prediction error as shown in figure 4.4.1

Figure 4.4.2 RandomizedSearchCV



The pipeline-based approach employed in this project offers a robust, efficient, and scalable solution for preprocessing complex datasets and training machine learning models. By automating the data preparation steps and model training in a single workflow, the pipeline ensures consistency across different stages, reduces the risk of data leakage, and allows for easy hyperparameter tuning through cross-validation.

This pipeline will be critical in ensuring that the model is properly validated and ready for deployment.

Chapter 5

System Implementation

5.1 Hardware Setup

The hardware involved in this project separated into two laptops. First computer issued for the process of data mining that able to work with Jupyter lab. From the Ipoh working adult dataset to obtain the relationship analysis and insight on the income level, then it also used for applying prediction towards the Monthly Income attribute.

Table 5.1.1 Specifications of first laptop

Description	Specifications
Model	HP Pavilion Laptop 15-cs3xxx
Processor	Intel(R) Core (TM) i5-1035G1
Operating System	Windows 10
Graphic	Intel(R) UHD Graphics
Memory	12.00 GB
Storage	475GB

However, this project required a second laptop prepared for installing Shiny for python due to the first laptop operating system does not support and effected by OSError: [WinError 2].

Table 5.1.2 Specifications of second laptop

Description	Specifications
Model	VivoBook S13 X330FA_S330FA
Processor	Intel(R) Core (TM) i7-8565U
Operating System	Windows 10
Graphic	INTEL(R) UHD GRAPHICS 620
Memory	8.00 GB
Storage	475GB

5.2 Software Setup

There is total three development tools been used on setting up this Income Prediction Models:

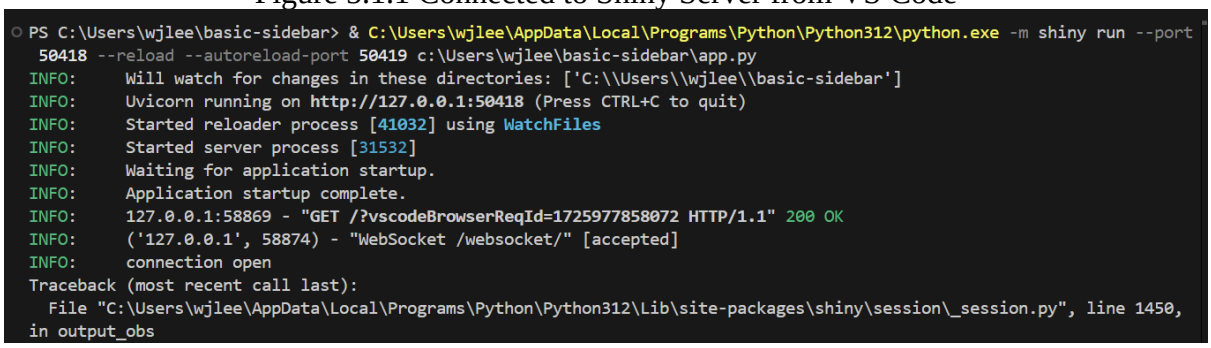
5.2.1 Jupyter Lab

This environment provides an interactive interface for executing Python code in a cell-based format. It is highly suitable for iterative development, testing AI models, and conducting data analysis in real-time [53]. Jupyter Lab also supports visualization directly in the notebook, making it a convenient tool for exploring income data.

5.2.2 Visual Studio Code (VS Code)

VS Code is a lightweight yet powerful IDE with rich extensions for Python development [54]. Its features, such as IntelliSense and debugging tools, enhance productivity. For this project, VS Code is particularly useful for managing larger codebases and deploying models, especially when integrating with Shiny for Python. Extension of Shiny Python are required to add into VS code libraries, only with that VS Code would be only able to execute the server process as shown in figure 5.1.1.

Figure 5.1.1 Connected to Shiny Server from VS Code

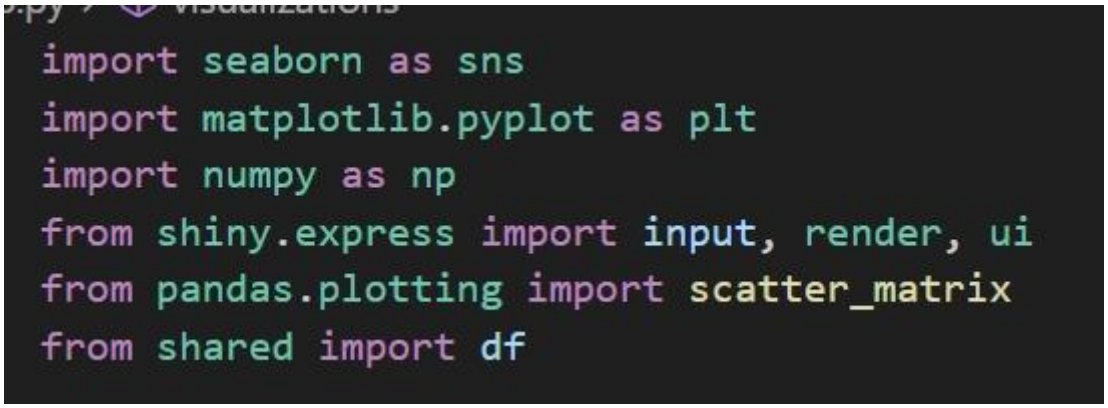


5.2.3 Shiny for Python

While Shiny is traditionally known for its R version, the Shiny for Python extension is utilized in this project to build interactive web applications that allow users, including policymakers, to explore AI-driven income predictions visually. Shiny for Python integrates seamlessly with existing Python scripts and provides an interface for displaying interactive data visualizations and dashboards [22]. Shiny's use in this project allows for Visualization Dashboard Creation

and this enables non-technical users to interact with complex income data and predictive models in a visual and intuitive way, and also Real-Time Data Interaction, by allowing users to adjust parameters and see the results in real-time, Shiny helps to democratize the insights gained from AI models [23]. The libraries needed to be installed when initializing the interpreter in VS Code, libraries included such as shiny.express, pandas.plotting, shared, and matplotlib.pyplot in figure 5.2.1.

Figure 5.2.1 Libraries installation through terminal to interpreter.



5.3 Libraries used

This table outlines the Python libraries employed in the development of the income prediction system. Each library serves a specific function, contributing to various stages of the project, from data handling to model implementation and visualization:

Table 5.3.1 Libraries used in this project

Library	Description
Pandas	Used for data manipulation and cleaning, enabling effective handling of missing values and preparation of the dataset.
NumPy	Provided tools for numerical operations and efficient management of arrays and matrices, foundational for mathematical computations.

Scikit-learn	Central to machine learning tasks, offering tools for data preprocessing, model training, and evaluation. Included algorithms like Linear Regression, Decision Trees, and Random Forests.
Matplotlib	Utilized for basic plotting capabilities to visualize data distributions and model performance.
Seaborn	Enhanced data visualizations with more informative and aesthetically pleasing plots.
Shiny for Python	Enabled the development of interactive web applications for dynamic interaction with model predictions and insights.
Plotly Express	Used for creating interactive visualizations to explore results interactively.

5.4 Algorithms used

This table summarizes the primary algorithms implemented for predicting income. Each algorithm was selected based on its suitability for handling structured data and its performance in regression tasks:

Table 5.3.2 Algorithms applied on modelling

Algorithm	Description
Linear Regression	Modeled the linear relationship between predictors and income, providing a baseline for understanding individual variable effects.
Multiple Linear Regression	Extended linear regression to incorporate multiple predictors, analyzing complex interactions between factors affecting income levels.
Ridge Regression	Regularized the linear regression model to address multicollinearity and enhance stability.

Decision Trees	Divided data into subsets based on informative features, offering clear interpretability of how factors like job sector and experience impact income.
Random Forests	An ensemble method combining multiple decision trees, improving model robustness and accuracy while handling high-dimensional data.
RandomizedSearchCV	Optimized model performance by exploring various hyperparameter settings for algorithms like Random Forests.
GridSearchCV	Fine-tuned model parameters to improve accuracy and performance through an exhaustive search of specified parameter values.

5.5 Libraries for Data Preprocessing

This table provides details on the specific libraries and tools used for preprocessing the data, preparing it for modeling:

Table 5.3.3 Libraries on Data Preprocessing

Library	Description
SimpleImputer	Handles missing values by imputation strategies.
OrdinalEncoder	Encodes categorical features as ordinal integers.
OneHotEncoder	Converts categorical variables into a one-hot numeric array.
MinMaxScaler	Scales features to a given range, often between 0 and 1.
StandardScaler	Standardizes features by removing the mean and scaling to unit variance.

ColumnTransformer	Applies different preprocessing pipelines to different subsets of features.
Pipeline	Chains together preprocessing and modeling steps into a single workflow.

Chapter 6

System Evaluation and Discussion

6.1 Model Evaluation Train & Test Set

6.1.1 Regression Model

Below are the Root Mean Squared Error number according to different method applied

Dataset Amount: 128

Table 6.1.1 RMSE before implemented Pipeline on Regression Models

Method Applied	RMSE
Linear Regression	**12586.0668**
Decision Tree Regression	**16986.1690**
Grid Search CV	**14023.8655**
Test Set	**4618.3175**
Random Forest Regression	**7667.6716**

Dataset Amount: 428 (Implemented Pipeline)

Table 6.1.2 RMSE after implemented Pipeline on Regression Models

Method Applied	RMSE
Ridge Regression	**Train:** 6284.5327
	Test: 5615.6765
Decision Tree Regression	**Train: 0**.0
Random Forest Regression	**Train:** 2730.952
	Test: 2227.1685
Grid Search CV	**Train:** 24.74
	Test: 2562.09
Final RMSE	**Test: 4614.2244**

Linear Regression has been changed to Ridge due to its complexity, if above model building train with Linear Regression, the final rmse would be fixed at 2313

Important Features selected from Random Forest

Figure 6.1.1 Important Features selected from Random Forest

```
Top 10 most important features:
                                      feature  importance
0                                    num__Age    0.497370
4                    num__Years_of_Experience    0.138672
37               ord_cat__Job_Position_Grouped    0.077522
2                  num__Working Hours Per Week    0.045145
17             nom_cat__Occupation_Engineering    0.040789
30  nom_cat__Additional_SourceIncomeInvest_Yes    0.030874
3                    num__Satisfaction_Income    0.030046
38                ord_cat__Highest_Education    0.018470
18  nom_cat__Occupation_Finance and Accounting    0.009772
39          ord_cat__Employment_Status_Grouped    0.009733
```

6.1.2 Classification Model

Below are the Performance according to different parameters applied

Dataset Amount: 428 (Implemented Pipeline)

Table 6.1.3 Accuracy of Random Forest Classification (Without Randomized Search CV)

Method Applied	RMSE
Random Forest Classifier	**Train** Accuracy: 1.0 CV Score: 0.84 RMSE : 0.0
	Test Accuracy: **0.75** CV Score: 0.71 RMSE : 0.56

With Randomized Search CV

Table 6.1.4 Accuracy of Random Forest Classification (With Randomized Search CV)

Test Case (Parameter)	RMSE
Test Case 1 {'bootstrap': True, 'criterion': 'gini', 'max_depth': 25, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 3, 'n_estimators': 179}	**Train** Accuracy: 1.0 CV Score: 0.8313 RMSE : 0.0
	Test Accuracy: **0.7529** CV Score: 0.6941 RMSE : 0.5839
Test Case 2 (bootstrap= False, criterion= 'entropy', max_depth= 15, max_features= 'sqrt', min_samples_leaf= 5, min_samples_split= 9, n_estimators= 155)	**Train** Accuracy: 1.0 CV Score: 0.8137 RMSE : 0.3263
	Test Accuracy: **0.7529** CV Score: 0.6941 RMSE : 0.6326
Test Case 3 (bootstrap= False, criterion= 'gini', max_depth= 17, max_features= 'sqrt', min_samples_leaf= 1, min_samples_split= 9, n_estimators= 111)	**Train** Accuracy: 1.0 CV Score: 0.8285 RMSE : 0.1804
	Test Accuracy: **0.7529** CV Score: 0.7176 RMSE : 0.5530

Test Case 4 (bootstrap= True, criterion= 'gini', max_depth= 25, max_features= 'log2', min_samples_leaf= 1, min_samples_split= 3, n_estimators= 179)	**Train** Accuracy: 1.0 CV Score: 0.8226 RMSE : 0.0
	Test Accuracy: **0.7529** CV Score: 0.7059 RMSE : 0.5636
Test Case 5 (bootstrap= False, criterion= 'gini', max_depth= 17, max_features= 'sqrt', min_samples_leaf= 1, min_samples_split= 9, n_estimators= 111)	**Train** Accuracy: 1.0 CV Score: 0.8314 RMSE : 0.1632
	Test Accuracy: **0.7529** CV Score: 0.7294 RMSE : 0.5841
Test Case 6 (bootstrap= True, criterion= 'gini', max_depth= 37, max_features= 'log2', min_samples_leaf= 1, min_samples_split= 2, n_estimators= 120)	**Train** Accuracy: 1.0 CV Score: 0.8284 RMSE : 0.0
	Test Accuracy: **0.7529** CV Score: 0.7176 RMSE : 0.5636

All the test cases above shown 0.75 is best accuracy in test sets income prediction.

Among all the test cases, test case 5 is the best parameter as its factors as below:

Train Accuracy: 1.0 (same across all test cases)

Train RMSE: 0.1632 (one of the lowest values)

Test Accuracy: 0.7529 (same across all test cases)

Test CV Score: 0.7294 (highest among the test cases)

Test RMSE: 0.5841 (acceptable given the higher CV score)

Best Parameter (Test Case 5):

{

'bootstrap': False,

'criterion': 'gini',

'max_depth': 17,

'max_features': 'sqrt',

'min_samples_leaf': 1,

'min_samples_split': 9,

'n_estimators': 111

}

6.2 System Testing Dashboard and Result

Test Case 1: Working Hours of Week vs. Current Residency (Boxplot)

Objective: To verify that the dashboard can generate a boxplot for the numerical variable "Working Hours of Week" grouped by the categorical variable "Current Residency."

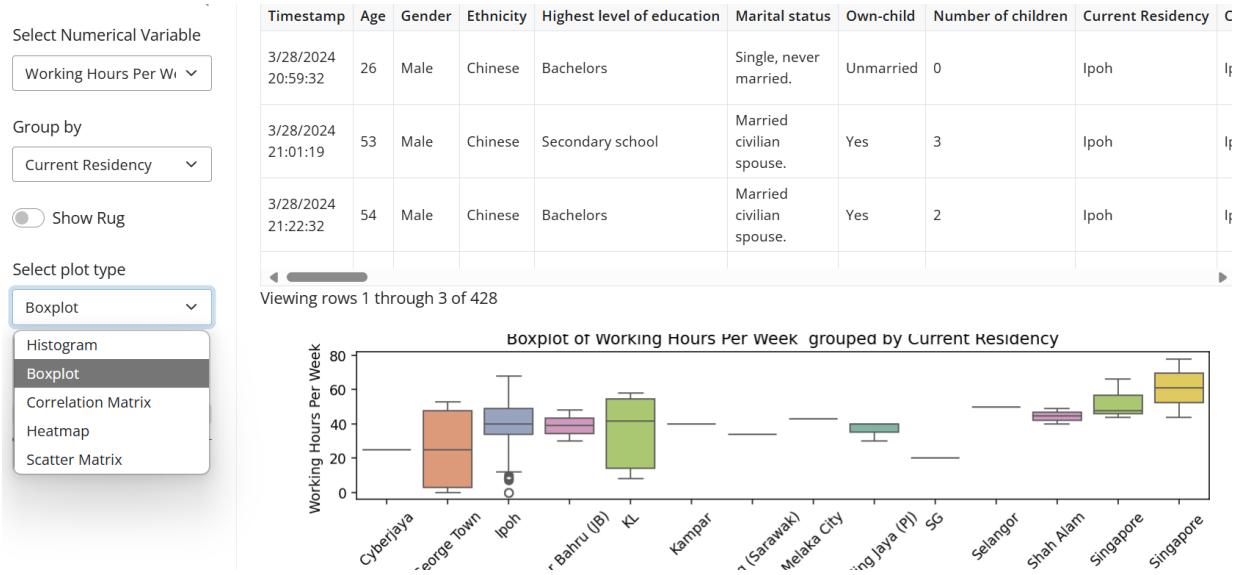
Select Numerical Variable: “Working Hours of Week”

Group by : “Current Residency”

Plot Type : “Boxplot”

Expected Outcome: A boxplot should display the distribution of "Working Hours of Week" for each category of "Current Residency," showing median, quartiles, and any outliers.

Figure 6.2.1 Test Case 1



Actual Result: The boxplot is generated successfully with clear groupings by "Current Residency" (Figure 6.2.1). Outliers and quartile ranges are visible as expected. No errors encountered.

Test Case 2: Age vs. Ethnicity (Histogram)

Objective: To ensure the dashboard generates a histogram of the numerical variable "Age" grouped by the categorical variable "Ethnicity."

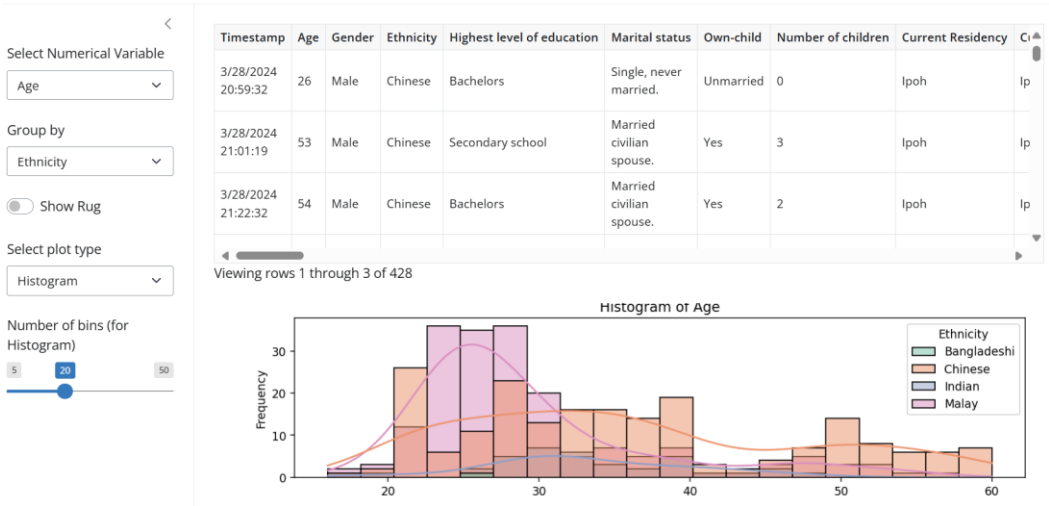
Select Numerical Variable: “Age”

Group by : “Ethnicity”

Plot Type : “Histogram”

Expected Outcome: A histogram should appear, displaying the frequency distribution of ages across the different ethnic groups, with the option to group by "Ethnicity."

Figure 6.2.2 Test Case 2



Actual Result: The histogram is generated successfully, with proper grouping and color differentiation for each ethnicity (Figure 6.2.2). No errors encountered during generation.

Test Case 3: Current Monthly Income (RM) vs. Ethnicity (Histogram)

Objective: To test the ability of the dashboard to generate a histogram for "Current Monthly Income (RM)" grouped by "Ethnicity."

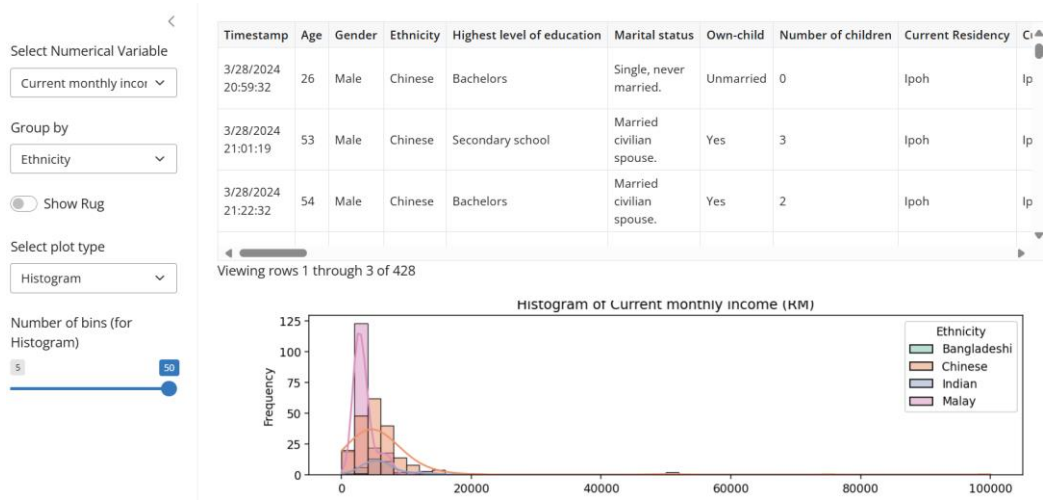
Select Numerical Variable: “Current Monthly income (RM)”

Group by : “Ethnicity”

Plot Type : “Histogram”

Expected Outcome: A histogram should display the distribution of "Current Monthly Income (RM)" across different ethnic groups, with clear segmentation.

Figure 6.2.3 Test Case 3



Actual Result: The histogram is successfully generated (Figure 6.2.3). The distribution of incomes across ethnic groups is clearly shown with distinct groupings by color. No errors encountered.

Test Case 4: Satisfaction with Current Income Level vs. Ethnicity (Boxplot)

Objective: To verify that the dashboard can produce a boxplot for the variable "Are you satisfied with your current income level?" grouped by "Ethnicity."

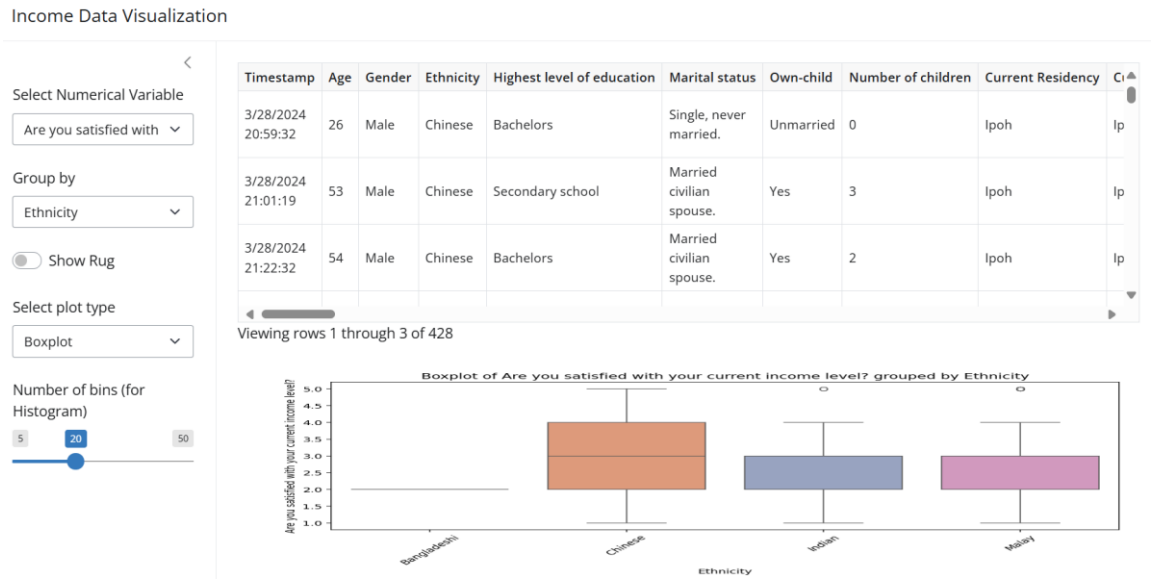
Expected Outcome: A boxplot should show the distribution of satisfaction levels across the different ethnic groups.

Select Numerical Variable: “Are you satisfied with your current income level?”

Group by : “Ethnicity”

Plot Type : “Boxplot”

Figure 6.2.4 Test Case 4



Actual Result: The boxplot is successfully generated (Figure 6.2.4). Each ethnic group is represented, with satisfaction levels distributed appropriately. No errors encountered.

There is also cases where combination would not be needed, for example Scatter Matrix and Heatmap.

Test Case 5: Scatter Matrix for Numerical Variables

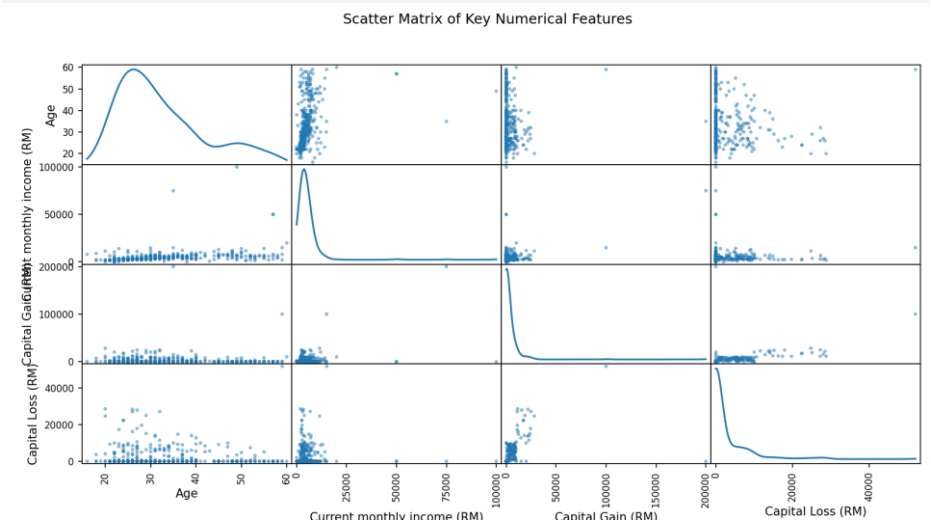
Objective: To check whether the dashboard can generate a scatter matrix for any selected numerical variables.

Expected Outcome: A scatter matrix showing pairwise relationships between selected numerical variables should be generated.

Select Numerical Variable: (Any Numerical Variable)

Plot Type : "Scatter Matrix"

Figure 6.2.5 Test Case 5



Actual Result: The scatter matrix is successfully generated (Figure 6.2.5). Relationships between all selected numerical variables are visible in a grid format. No errors encountered.

Test Case 6: Heatmap for Correlation Between Numerical Variables

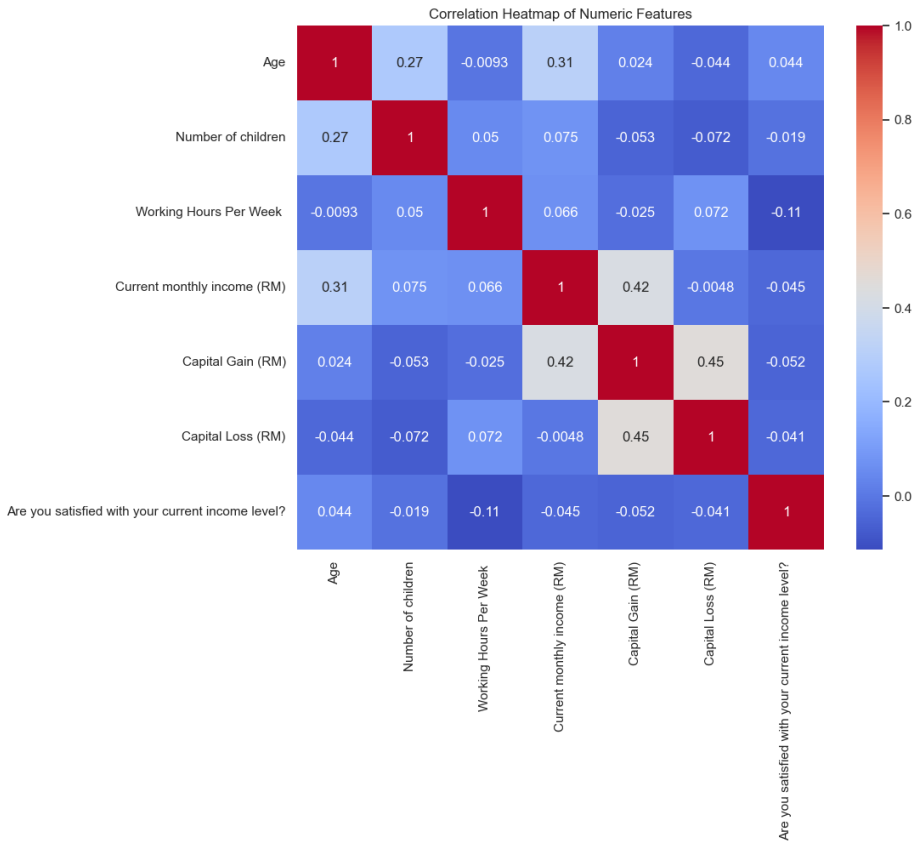
Objective: To ensure that the dashboard can generate a heatmap showing the correlation between any numerical variables selected.

Expected Outcome: A heatmap should display the correlation between the selected numerical variables, with a color gradient representing correlation strength.

Select Numerical Variable: (Any Numerical Variable)

Plot Type : "Heatmap"

Figure 6.2.6 Test Case 6



Actual Result: The heatmap is successfully generated (Figure 6.2.6). Correlations between the numerical variables are displayed correctly, with the color gradient clearly showing the correlation values. No errors encountered.

Test Case 7a: Show Rug and Bins Slider in 37

Objective: To verify that the dashboard can toggle the rug plot and adjust the number of bins for the histogram visualization.

Select Numerical Variable: "Age"

Group by : "Ethnicity"

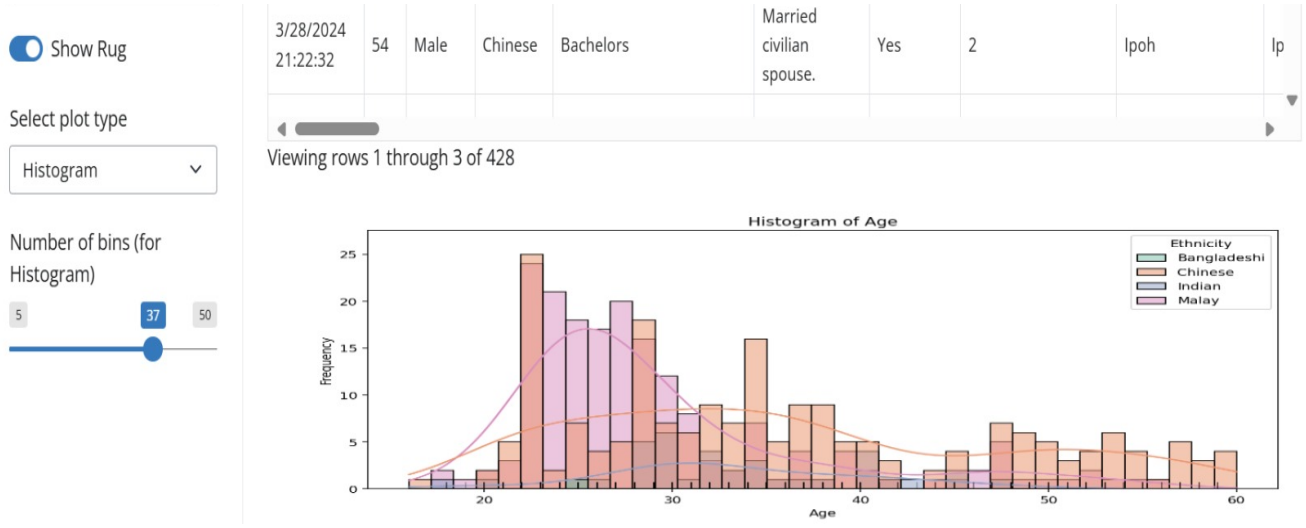
Plot Type : "Histogram"

Show Rug : ON

Number of bins (for Histogram): 37 of 50

Expected Outcome: The histogram should display the distribution of "Age" grouped by "Ethnicity" with 37 bins. The rug plot should appear at the bottom of the histogram, showing the exact locations of individual data points along the x-axis.

Figure 6.2.7 Test Case 7a



Actual Result: The histogram is successfully generated with 37 bins (as shown in the figure 6.2.7). The rug plot is visible along the x-axis, and the distribution is clearly grouped by different ethnicities. The histogram color shades represent each ethnicity, and the bin count adjustment works as expected. No errors encountered.

Test Case 7b: Show Rug and Bins Slider in 20

Objective: To verify that the dashboard can toggle the rug plot and adjust the number of bins for the histogram visualization.

Select Numerical Variable: "Age"

Group by : "Ethnicity"

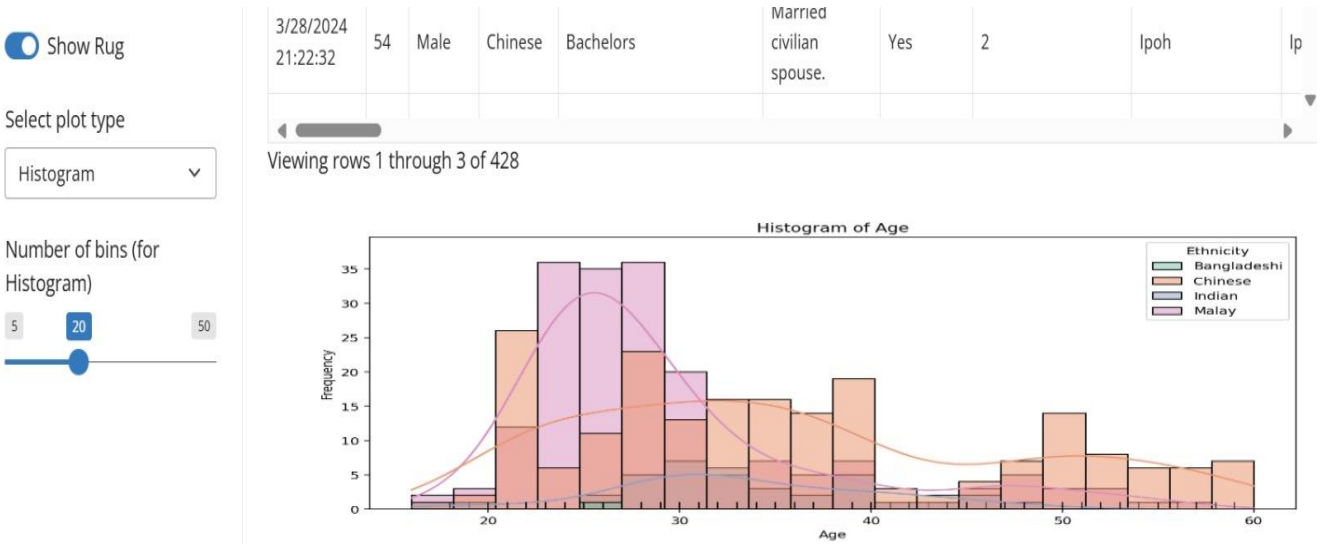
Plot Type : "Histogram"

Show Rug : ON

Number of bins (for Histogram): 20 of 50

Expected Outcome: The histogram should display the distribution of "Age" grouped by "Ethnicity" with 20 bins. The rug plot should appear at the bottom of the histogram, showing the exact locations of individual data points along the x-axis.

Figure 6.2.8 Test Case 7b



Actual Result: The histogram is successfully generated with 20 bins (as shown in the figure 6.2.8). The rug plot is visible along the x-axis, and the distribution is clearly grouped by different ethnicities. The histogram colour shades represent each ethnicity, and the bin count adjustment works as expected. No errors encountered.

Besides that, the dashboard also shown the entire dataset collected from Ipoh citizen as shown in figure 6.2.9 to know Data

Figure 6.2.9 Dataset Display

Timestamp	Age	Gender	Ethnicity	Highest level of education	Marital status	Own-child	Number of children	Current Residency	C
					spouse.				
3/29/2024 20:21:32	22	Female	Chinese	Bachelors	Single, never married.	No	0	KL	Kl
3/29/2024 22:22:56	47	Female	Malay	Diploma	Married civilian spouse.	Yes	3	Ipoh	Ip
3/29/2024 22:26:23	27	Female	Malay	Bachelors	Single, never married.	Unmarried	0	Ipoh	Ip

Viewing rows 19 through 21 of 428

6.3 Project Challenges

Model Overfitting

In the course of developing the predictive model, addressing model overfitting presented significant challenges. Overfitting occurs when the model performs exceptionally well on training data but fails to generalize effectively to unseen data. To tackle this issue, we had to implement rigorous cross-validation and adjust model parameters meticulously. Techniques such as regularization and hyperparameter tuning were essential to balance the model's complexity and improve its generalization capabilities. Despite these efforts, it remained challenging to find the optimal balance that prevents the model from fitting noise rather than underlying patterns.

Pipeline Integration and Consistency

Ensuring consistent pipeline integration was another considerable challenge. The pipeline needs to seamlessly integrate data preprocessing, feature engineering, and model training stages to ensure that transformations applied during training are consistent with those applied during evaluation. This integration required careful alignment of all components within the pipeline, including the proper handling of categorical and numerical features. The complexity of managing various preprocessing steps and parameter settings across different stages added to the difficulty. Any misalignment or inconsistency in this integration can lead to inaccurate model performance and unreliable predictions.

6.4 Objective Evaluation

1. Income Dataset Collection from Ipoh Citizens

The first objective of this project was to collect and form an income dataset from Ipoh citizens that specifically addresses the relationship between income and various economic factors. This objective has been successfully achieved at the amount of 423 valid responses. The dataset I compiled includes comprehensive information such as income, age, gender, ethnicity, marital status, education level, and number of children. These attributes enable a detailed analysis of how income is influenced by demographic and socioeconomic variables. The dataset provides a foundation for exploring key economic relationships.

2. Development of a Machine Learning Model for Predicting Ipoh Income Levels

The second objective focused on developing a machine learning model capable of predicting income levels using data mining techniques. This objective has been met successfully. I experimented with several machine learning models and conducted multiple test cases to optimize performance. The final model achieved a reasonable accuracy of approximately 0.75 on the test dataset and demonstrated a strong ability to generalize predictions for income levels. The models have been fine-tuned for parameters such as criterion, max_depth, bootstrap, and n_estimators, achieving satisfactory performance metrics such as RMSE and cross-validation scores. These results indicate that the model is reliable in predicting income levels based on the collected dataset.

3. Interactive Dashboard for Data Visualization

The third objective was to create an interactive dashboard that offers clear visual representations of the income dataset through graphs, histogram and box plots. This objective has been fully accomplished. The dashboard provides various interactive features, such as the ability to select numerical variables, adjust groupings, and choose different plot types (e.g., heatmaps, histograms, scatter plots). These visualizations are designed to offer policymakers an easy-to-understand view of the income distribution and related factors within the Ipoh population. The dashboard not only visualizes the data effectively but also allows for deeper analysis of trends and patterns.

4. Dashboard to Assist Policymakers in Decision-Making

The final objective was to develop a dashboard that can assist policymakers in making informed decisions regarding economic development, based on the income dataset. This objective has largely been achieved. The interactive nature of the dashboard, combined with the clarity of the visualizations, ensures that policymakers can easily interpret the data and use it to identify trends and inform policy decisions. The inclusion of various graphs, pie charts, and bar charts provides a comprehensive view of the income data, offering valuable insights into economic inequalities, demographic distributions, and potential areas for development.

Overall, the objectives of the project have been successfully achieved. The income dataset collected from Ipoh citizens has been effectively formed with the amount of 423 valid reponses, providing a solid foundation for analysis. A machine learning model has been developed and optimized, accurately predicting income levels based on various demographic and socioeconomic factors. Additionally, an interactive dashboard was successfully created, offering clear and intuitive visualizations such as graphs, histogram and box plots, enabling policymakers to better understand the income data. The dashboard also aids in informed decision-making regarding economic development, addressing the needs of policymakers in assessing income distribution and its implications.

Chapter 7

Conclusion and Recommendation

7.1 Conclusion

In summary, this project has successfully collected and analyzed income data from Ipoh citizens, offering valuable insights into the Ipoh's economic landscape. Through the application of data mining techniques and predictive analytics, we identified significant economic factors influencing income levels, such as education, employment type, and working hours through 423 valid responses. These insights led to the development of a machine learning model that forecasts income levels with improved accuracy, highlighting areas where economic interventions could have the most impact.

The creation of an interactive dashboard has further augmented the project's utility by providing policymakers with a user-friendly platform to visualize complex income data. This dashboard enables the identification of important trends, disparities, and the underlying factors driving income inequality. As a result, policymakers are better equipped to make informed, data-driven decisions that address socioeconomic challenges and promote sustainable economic growth in Ipoh.

Overall, this project delivers a valuable tool for economic planners, allowing them to craft more effective and informed development strategies based on a comprehensive analysis of local income data. It stands as a significant contribution to understanding and addressing income disparities, ultimately supporting the creation of a more equitable and prosperous economic environment in Ipoh.

7.2 Recommendation

While the dashboard effectively visualizes income data, it can be further enhanced by incorporating features that provide specific, actionable insights and recommendations for policymakers. By integrating additional indicators and highlighting key economic trends or anomalies, the dashboard could more directly guide decision-making. Future iterations could focus on improving these decision-support features to better serve the needs of economic planners.

Additionally, geospatial analysis could be integrated into the dashboard, enabling policymakers to visualize income disparities across different regions within Ipoh. By mapping income distribution on an interactive map, economic planners can more easily identify areas that require targeted interventions or development initiatives. This would enhance the dashboard's capability to inform location-specific policies, improving the precision of economic development efforts.

Moreover, incorporating a predictive analytics feature would provide an even more powerful tool for decision-making. This feature could simulate potential policy outcomes based on current income data, allowing policymakers to forecast the impact of initiatives such as tax reforms, subsidies, or investment in infrastructure. By enabling data-driven predictions, the dashboard would help ensure that economic strategies are not only reactive but also proactive, optimizing resource allocation and improving overall economic resilience.

Finally, expanding the dataset to include a broader geographical area beyond Ipoh could significantly enhance the relevance of the model. By collecting and analyzing data from surrounding regions, policymakers can gain insights into income trends on a wider scale. This expansion would also enable comparisons between Ipoh and other regions, facilitating a more comprehensive understanding of economic disparities and allowing for more holistic economic planning at the state level.

Implementing these enhancements would not only elevate the functionality of the dashboard but also ensure that it becomes an indispensable tool for economic development and policy formulation in the region.

REFERENCES

[1] M. Taylor and A. Ersoy, “Understanding dynamics of local and regional economic development in emerging economies,” Ekonomska Istrazivanja-economic Research, vol. 25, no. 4, pp. 1079–1088, Jan. 2012, doi: 10.1080/1331677x.2012.11517549. [Accessed: Dec. 7, 2023].

[2] A. M. Cabello, “Piketty, Thomas (2014): Capital in the Twenty-First Century. Cambridge: The Belknap Press of Harvard University Press,” Methaodos.Revista De Ciencias Sociales, vol. 3, no. 2, Oct. 2015, doi: 10.17502/m.rcs.v3i2.92. [Accessed: Dec. 7, 2023].

[3] T. H. Davenport and D. J. Patil, HBR's Guide to Data Science: The New Frontier of Innovation, Competition, and Productivity, Harvard Business Review Press, 2012. [Accessed: Dec. 7, 2023].

[4] B. Milanovic, Global Inequality: A New Approach for the Age of Globalization, Harvard University Press, 2016. [Accessed: Dec. 7, 2023].

[5] M. Saavedra and T. Twinam, “A machine learning approach to improving occupational income scores,” Explorations in Economic History, vol. 75, p. 101304, Jan. 2020, doi: 10.1016/j.eeh.2019.101304. [Accessed: Dec. 4, 2023].

[6] A. MacPherson and A. MacPherson, “A review of Payscale as a salary comparison tool – 2023 update,” Pathrise Resources, Nov. 08, 2023. https://www.pathrise.com/guides/a-review-of-payscale-as-a-salary-comparison-tool/#:~:text=The%20main%20reviews%20of%20Payscale,company%20is%20a%20Payscale%20customer.

[7] Trustpilot, “Glassdoor Reviews,” Trustpilot, Dec. 09, 2023. https://www.trustpilot.com/review/www.glassdoor.com

[8] N. Noor, A. Sarlan, and N. Aziz, “Revenue Prediction for Malaysian Federal Government Using Machine Learning Technique,” Revenue Prediction for Malaysian Federal Government Using Machine Learning Technique, Feb. 2022, doi: 10.1145/3524304.3524337.

[9] F. Lee and F. Lee, “Through MyDIGITAL, this is what M’sian education should look like in 10 years,” Vulcan Post, May 10, 2021. https://vulcanpost.com/745607/malaysia-digital-economy-blueprint-mydigital-education/

[10] Admin, “Best Dashboard User interface design for effective data analysis,” ChartExpo, Nov. 24, 2023. https://chartexpo.com/blog/dashboard-user-interface-design#

[11] “Salary.com pricing, demo & reviews (December 2023).” https://www.selecthub.com/p/compensation-management-software/salary-com/#:~:text=Salary.com%20has%20a%20'great,3%20recognized%20software%20review%20sites.

[12] P. Borrellas and I. Unceta, “The challenges of machine learning and their economic implications,” Entropy, vol. 23, no. 3, p. 275, Feb. 2021, doi: 10.3390/e23030275.

[13] 星洲网, “世界银行报告 | 25%穷人没受惠 三分一万元户有份 马疫情援金派发欠周 - 国内 - 封面头条,” 星洲网 Sin Chew Daily Malaysia Latest News and Headlines, Aug. 03, 2023. [Online]. Available: https://www.sinchew.com.my/news/20230803/nation/4880311

[14] 星洲网, “吉将建最大碳化硅工厂 安华：英飞凌将追加 250 亿投资 - 国内 - 即时国内,” 星洲网 Sin Chew Daily Malaysia Latest News and Headlines, Aug. 03, 2023. [Online]. Available: https://www.sinchew.com.my/news/20230803/nation/4880377

[15] A. Velimirovic, “Top 10 Python Machine Learning libraries,” *phoenixNAP Blog*, Aug. 15, 2024. https://phoenixnap.com/blog/python-machine-learning-library [Accessed: Sep. 10, 2024].

[16] GeeksforGeeks, “Best Python libraries for Machine Learning,” *GeeksforGeeks*, Aug. 08, 2024. https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning/ [Accessed: Sep. 10, 2024].

[17] D. Sheremetov, “The Python advantage: Why it’s the top choice for AI and ML,” *Onix*, Jul. 24, 2024. https://onix-systems.com/blog/python-is-best-for-ai-ml-and-deep-learning [Accessed: Sep. 10, 2024].

[18] T. Karl, "6 reasons Why is Python used for machine Learning," *New Horizons*, Jan. 03, 2024. https://www.newhorizons.com/resources/blog/why-is-python-used-for-machine-learning [Accessed: Sep. 10, 2024].

[19] G. Knez, "25 Python Machine Learning Libraries | Sunscrapers," *Sunscrapers*, Oct. 23, 2023. [Online]. Available: https://sunscrapers.com/blog/python-machine-learning-libraries-for-data-science/ [Accessed: Sep. 10, 2024].

[20] ITVidz, "Python data visualization, data analysis in Jupyter Lab (Pandas and PyWedge)," *YouTube*. Oct. 24, 2023. [Online]. Available: https://www.youtube.com/watch?v=P1P_bwi-GhU [Accessed: Sep. 10, 2024].

[21] Visual Studio Code, "Getting Started with Python in VS Code (Official Video)," *YouTube*. Aug. 12, 2024. [Online]. Available: https://www.youtube.com/watch?v=D2cwvpJSBX4 [Accessed: Sep. 10, 2024].

[22] PyData, "Joe Cheng - Shiny for Python: Interactive apps and dashboards made easy-ish | PyData NYC 2022," *YouTube*. Jan. 23, 2023. [Online]. Available: https://www.youtube.com/watch?v=ijRBbtT2tgc [Accessed: Sep. 10, 2024].

[23] Alex The Analyst, "Building a Fully Interactive Web App using Shiny for Python," *YouTube*. Jul. 09, 2024. [Online]. Available: https://www.youtube.com/watch?v=zv1nfZTYpio [Accessed: Sep. 10, 2024].

[24] Y. Qin, X. Zhang, X. Wang, and M. Škare, "Artificial Intelligence and Economic Development: An Evolutionary investigation and Systematic review," Journal of the Knowledge Economy, Mar. 2023, doi: 10.1007/s13132-023-01183-2.

[25] Vidya Chockalingam, Sejal Shah and Ronit Shaw: "Income Classification using Adult Census Data", https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a120.pdf

[26] "A Statistical approach to adult Census income level Prediction," *IEEE Conference Publication | IEEE Xplore*, Oct. 01, 2018. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8748528

[27] Nikhil Sai, Income prediction using Decision tree, Kaggle, Aug. 20, 2019. https://www.kaggle.com/code/jnikhilsai/income-prediction-using-decision-tree

[28] Mohammed Topiwalla: “Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting”, University of SP Jain School of Global Management.

[29] A. Lazar, "Income Prediction via Support Vector Machine," in Proceedings of the International Conference on Machine Learning and Applications (ICMLA), Louisville, KY, USA, 16-18 Dec. 2004.

[30] Chet Lemon, Chris Zelazo and Kesav Mulakaluri: “Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques”, https://cseweb.ucsd.edu/ jmcauley/cse190/reports/sp15/048.pdf.

[31] H. Zhu, “Predicting earning potential using the adult dataset,” *Predicting Earning Potential Using the Adult Dataset*. https://rstudio-pubs-static.s3.amazonaws.com/235617_51e06fa6c43b47d1b6daca2523b2f9e4.html

[32] Payscale, “Why Payscale: Learn why we are the industry leader,” Payscale - Salary Comparison, Salary Survey, Search Wages, Dec. 21, 2022. https://www.payscale.com/why-payscale/?tk=nav

[33] Salary.com, “Salary.com homepage,” Salary.com, Apr. 01, 2021. https://www.salary.com/

[34] Predictive Insights, “Our solutions - predictive insights,” Predictive Insights - Machine Learning PRODUCT DESIGN STUDIO, Nov. 08, 2023. https://predictiveinsights.net/solutions/

[35] “Poverty and Inequality platform,” Poverty and Inequality Platform. https://pip.worldbank.org/#

[36] “OECD Economic Surveys: Malaysia 2021 | READ online,” *oecd-ilibrary.org*. https://read.oecd-ilibrary.org/economics/oecd-economic-survey-malaysia-2021_cc9499dd-en#page26

[37] GeeksforGeeks, “KDD process in data mining,” *GeeksforGeeks*, May 23, 2023. https://www.geeksforgeeks.org/kdd-process-in-data-mining/

[38] N. Hotz, "What is CRISP DM? - Data Science Process Alliance," *Data Science Process Alliance*, Jan. 19, 2023. https://www.datascience-pm.com/crisp-dm-2/

[39] "CRISP-DM 1.0. Step-by-step data mining guide - PDF Free Download." https://docplayer.net/202628-Crisp-dm-1-0-step-by-step-data-mining-guide.html

[40] Z. Luna, "CRISP-DM Phase 1: Business Understanding - Analytics Vidhya - Medium," Medium, Jan. 06, 2022. [Online]. Available: https://medium.com/analytics-vidhya/crisp-dm-phase-1-business-understanding-255b47adf90a

[41] M. R. Berthold, C. Borgelt, F. Höppner, and F. Klawonn, "Data understanding," in Texts in computer science, 2010, pp. 33–79. doi: 10.1007/978-1-84882-260-3_4.

[42] "UCI Machine Learning Repository." https://archive.ics.uci.edu/dataset/2/adult

[43] S. Khan PhD, "Sector-Specific Questionnaires for Identifying Challenges regarding AI Adoption in Pakistan," Jun. 09, 2022. https://www.linkedin.com/pulse/sector-specific-questionnaires-identifying-challenges-khan-ph-d-/?trackingId=I9yLZKrqT9ugG2zBUAH1Rw%3D%3D

[44] Talend, "What is Data Preparation? Processes and Example," Talend - a Leader in Data Integration & Data Integrity. https://www.talend.com/resources/what-is-data-preparation/#:~:text=Good%20data%20preparation%20allows%20for,data%20more%20accessible%20to%20users.

[45] NeenOpal, "Dealing with outliers and missing values in a dataset." https://www.neenopal.com/dealing-with-outliers-and-missing-values-in-a-dataset.html

[46] W. H. Khoong, "Why scaling your data is important - CodeX - Medium," Medium, Jan. 21, 2023. [Online]. Available: https://medium.com/codex/why-scaling-your-data-is-important-1aff95ca97a2#:~:text=Data%20scaling%20is%20the%20process,the%20performance%20of%20the%20algorithm.

[47] G. Lawton, J. M. Carew, and E. Burns, "predictive modeling," Enterprise AI, Jan. 21, 2022. https://www.techtarget.com/searchenterpriseai/definition/predictive-modeling

[48] "What are Neural Networks? | IBM." https://www.ibm.com/id-en/topics/neural-networks

[49] A. Saini, "Guide on Support Vector Machine (SVM) Algorithm," Analytics Vidhya, Oct. 27, 2023. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

[50] A. Kalla, "K means clustering - DataDrivenInvestor," Medium, Dec. 08, 2021. [Online]. Available: https://medium.datadriveninvestor.com/k-means-clustering-4a700d4a4720

[51] R. S. H, "K-Nearest Neighbors Algorithm - Intuitive tutorials," Intuitive Tutorials, Apr. 05, 2023. https://intuitivetutorial.com/2023/04/07/k-nearest-neighbors-algorithm/

[52] "What is Model Evaluation? | Domino Data Science Dictionary." https://domino.ai/data-science-dictionary/model-evaluation

[53] H. Zhou, "The relationship between education level and wages: Based on Regression model analysis," *Advances in Economics Management and Political Sciences*, vol. 62, no. 1, pp. 204–210, Dec. 2023, doi: 10.54254/2754-1169/62/20231346.

[54] R. Fiagbe, "Classification of adult income using decision tree," *STARS*. https://stars.library.ucf.edu/data-science-mining/3/

[55] S. P. M. W. and C. Penfold, "7 Decision trees and random forests | An Introduction to Machine Learning," Oct. 07, 2020. https://cambiotraining.github.io/intro-machine-learning/decision-trees.html

[56] shiny "Benchmarking regression algorithms for income prediction modeling," *IEEE Conference Publication | IEEE Xplore*, Dec. 01, 2015. https://ieeexplore.ieee.org/document/7424087

[57] S. Taylor, "Multiple Linear regression," *Corporate Finance Institute*, Jul. 10, 2024. https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/

[58] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. 2009. doi: 10.1007/978-0-387-84858-7.

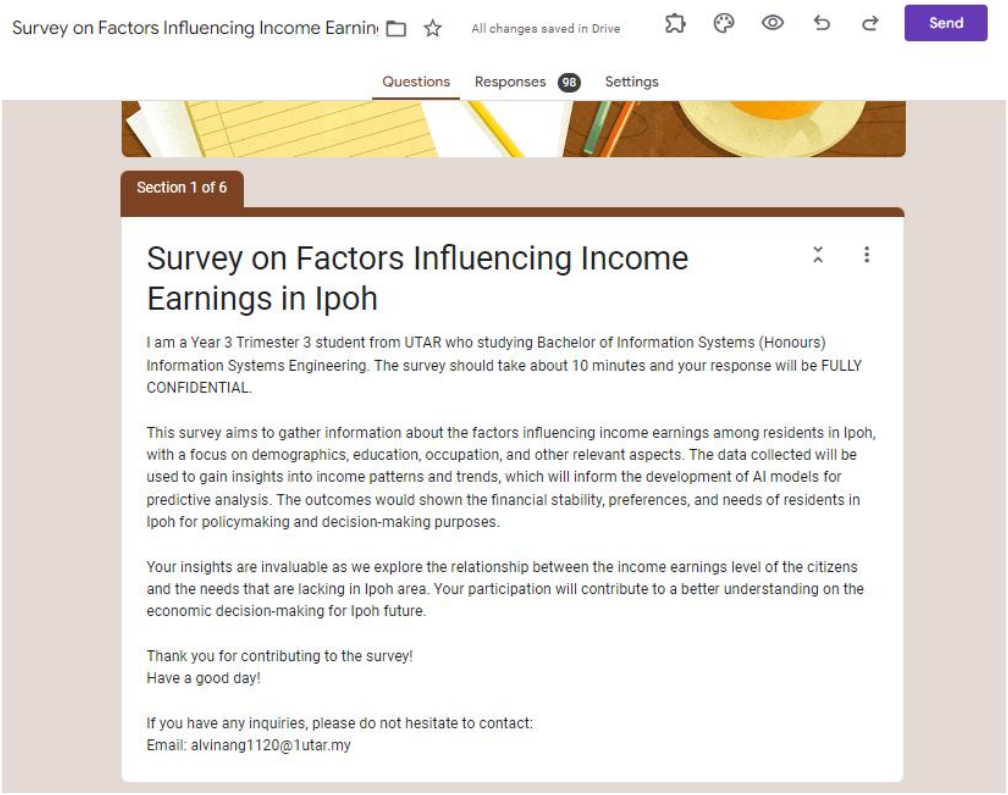
[59] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Jan. 2001, doi: 10.1023/a:1010933404324.

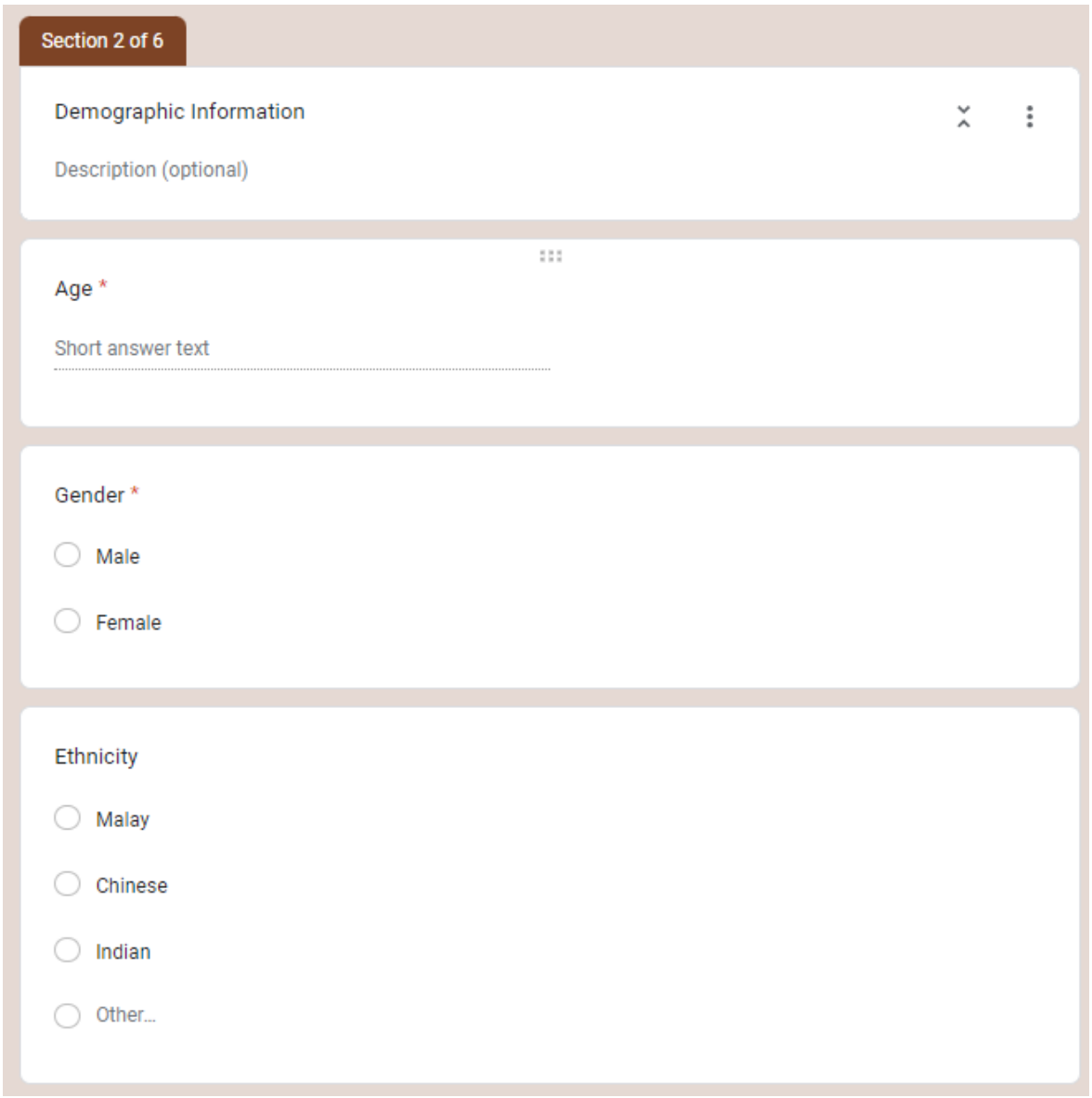
[60] F. Tabsharani, “support vector machine (SVM),” *WhatIs*, Aug. 03, 2023. https://www.techtarget.com/whatis/definition/support-vector-machine-SVM

[61] Y. Kulkarni, “Simple Linear Regression model to predict the Salary based on Years of Experience.,” *Medium*, Jul. 26, 2023. [Online]. Available: https://medium.com/@yash_kulkarni/simple-linear-regression-model-to-predict-the-salary-based-on-years-of-experience-9ed93a02c5ef

APPENDIX

Questionnaire Sample




Section 2 of 6
Demographic Information
Description (optional)
Age *
Short answer text
Gender *
Male
Female
Ethnicity
Malay
Chinese
Indian
Other...

Highest level of education *

- ○ Primary school
- ○ Secondary school
- ○ Diploma
- ○ Foundation
- ○ Bachelors
- ○ Master
- ○ PhD

Marital status *

- ○ Single, never married.
- ○ Married civilian spouse.
- ○ Separated from spouse but not divorced.
- ○ Divorced
- ○ Widowed
- ○ Married but spouse is not currently living with them.
- ○ Married to a spouse who is in the Armed Forces.

Own-child *

- ○ Unmarried
- ○ Yes
- ○ No

Number of children *

Short answer text

Current Residency *

- ○ Ipoh
- ○ KL
- ○ George Town
- ○ Johor Bahru (JB)
- ○ Shah Alam
- ○ Petaling Jaya (PJ)
- ○ Kuching (Sarawak)
- ○ Melaka City
- ○ Cyberjaya

- ○ Other...

Current Working Place

- ○ Ipoh
- ○ KL
- ○ George Town
- ○ Johor Bahru (JB)
- ○ Shah Alam
- ○ Petaling Jaya (PJ)
- ○ Kuching (Sarawak)
- ○ Kota Kinabalu (Sabah)
- ○ Melaka City
- ○ Cyberjaya
- ○ Other...

Section 3 of 6

Occupational Information

Description (optional)

Have you worked in Ipoh before?

- Yes
- No

Current employment status *

- Never-worked
- Unemployed
- Local government (e.g., city, district, municipality)
- State government entity.
- Federal government.
- Self-employed, incorporated (running a registered business)
- Self-employed, not incorporated (e.g., freelancer, independent, contractor)
- Employed by a private sector (company or organization)

What is your current working environment? *

- Physical
- Work from Home
- Hybrid

Occupation *

- Engineers
- Medical Professionals
- Bankers and Finance Professionals
- Teachers and Educators
- Information Technology (IT) Professionals
- Sales and Marketing Professionals
- Manufacturing Workers
- Government Officials and Civil Servants
- Tourism and Hospitality Workers
- Entrepreneurs and Business Owners
- Other...

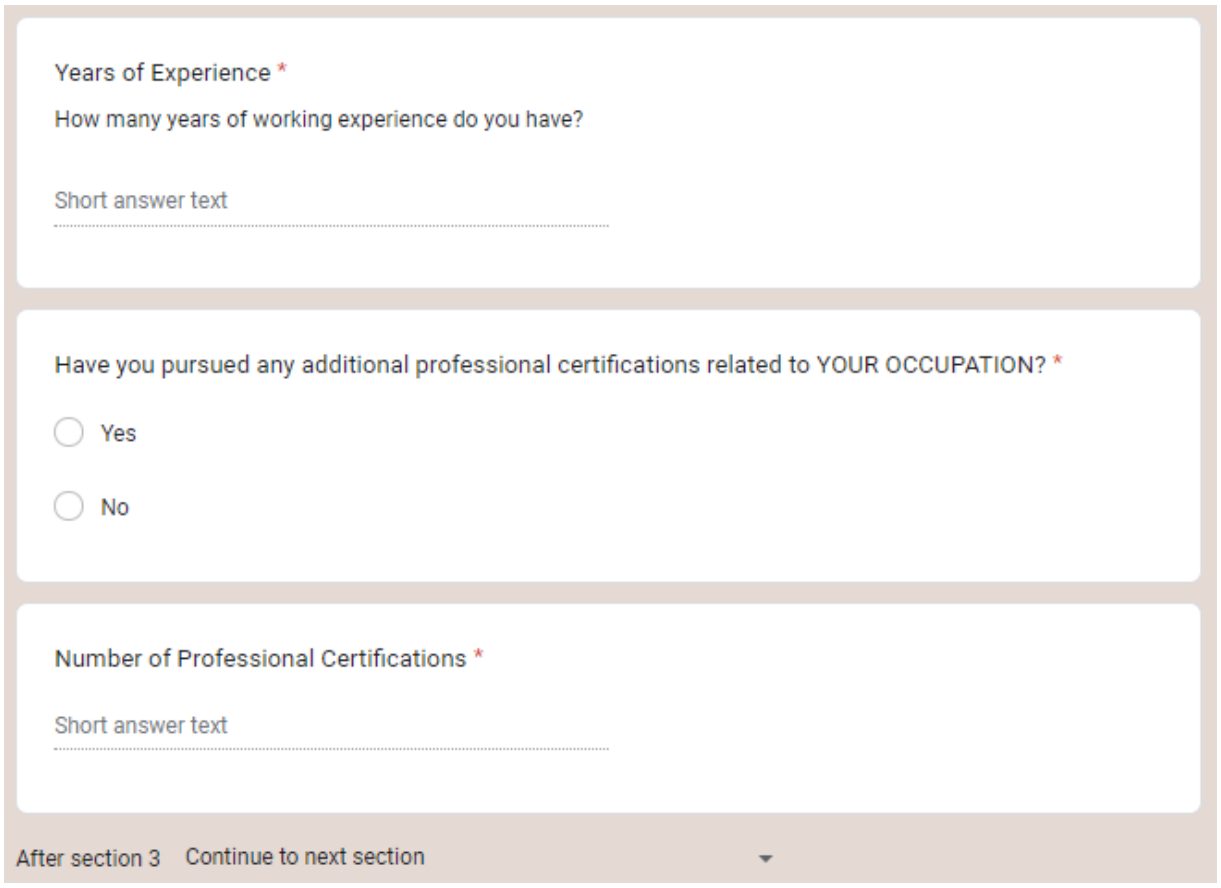
What is your job position classification? *

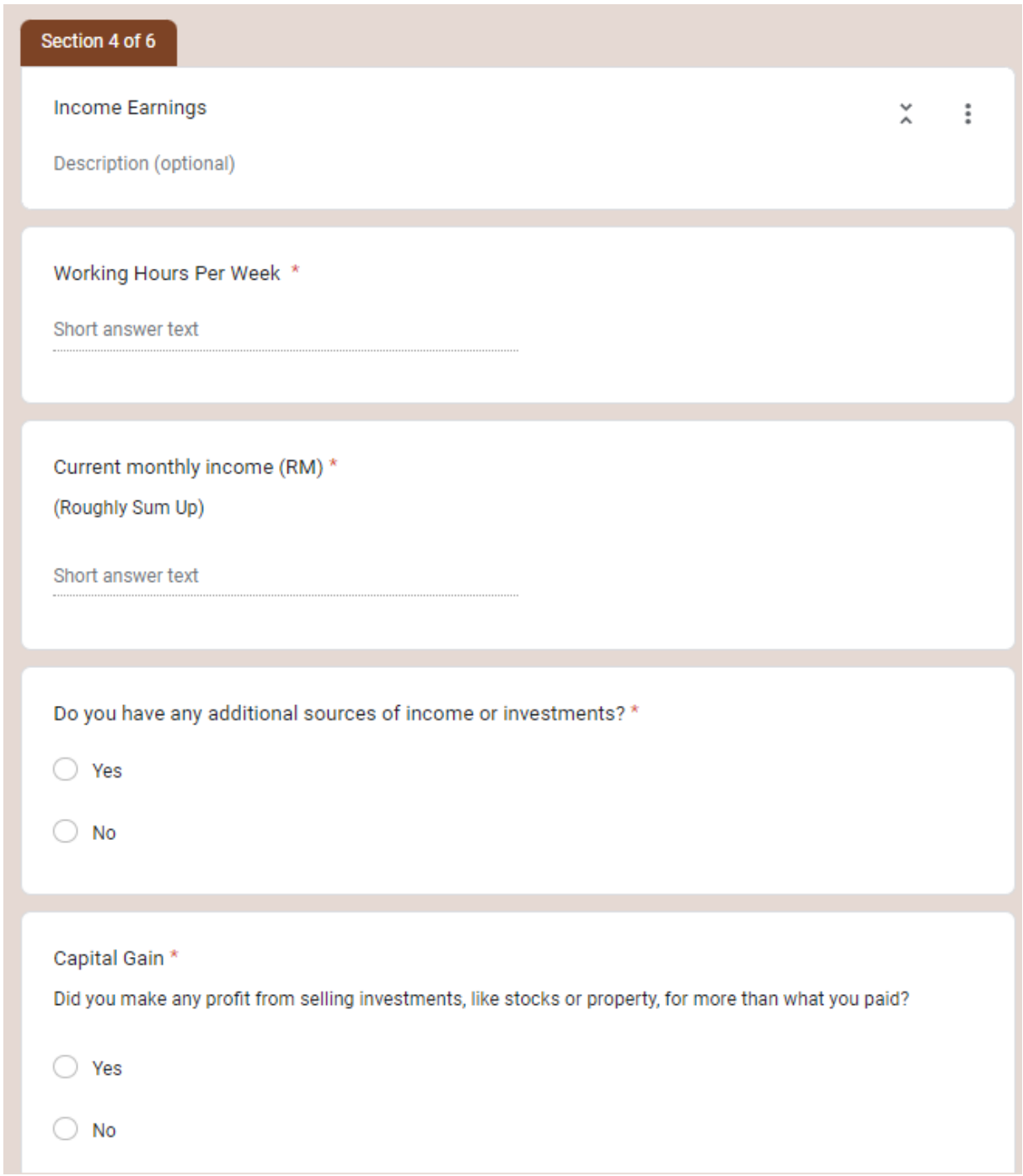
- Senior Level Executive
- Mid to Senior Level Manager
- Mid-Level
- Junior Employee
- Entry Level
- Freelancer/Consultant
- Student/Researcher
- None of these above
- Other...

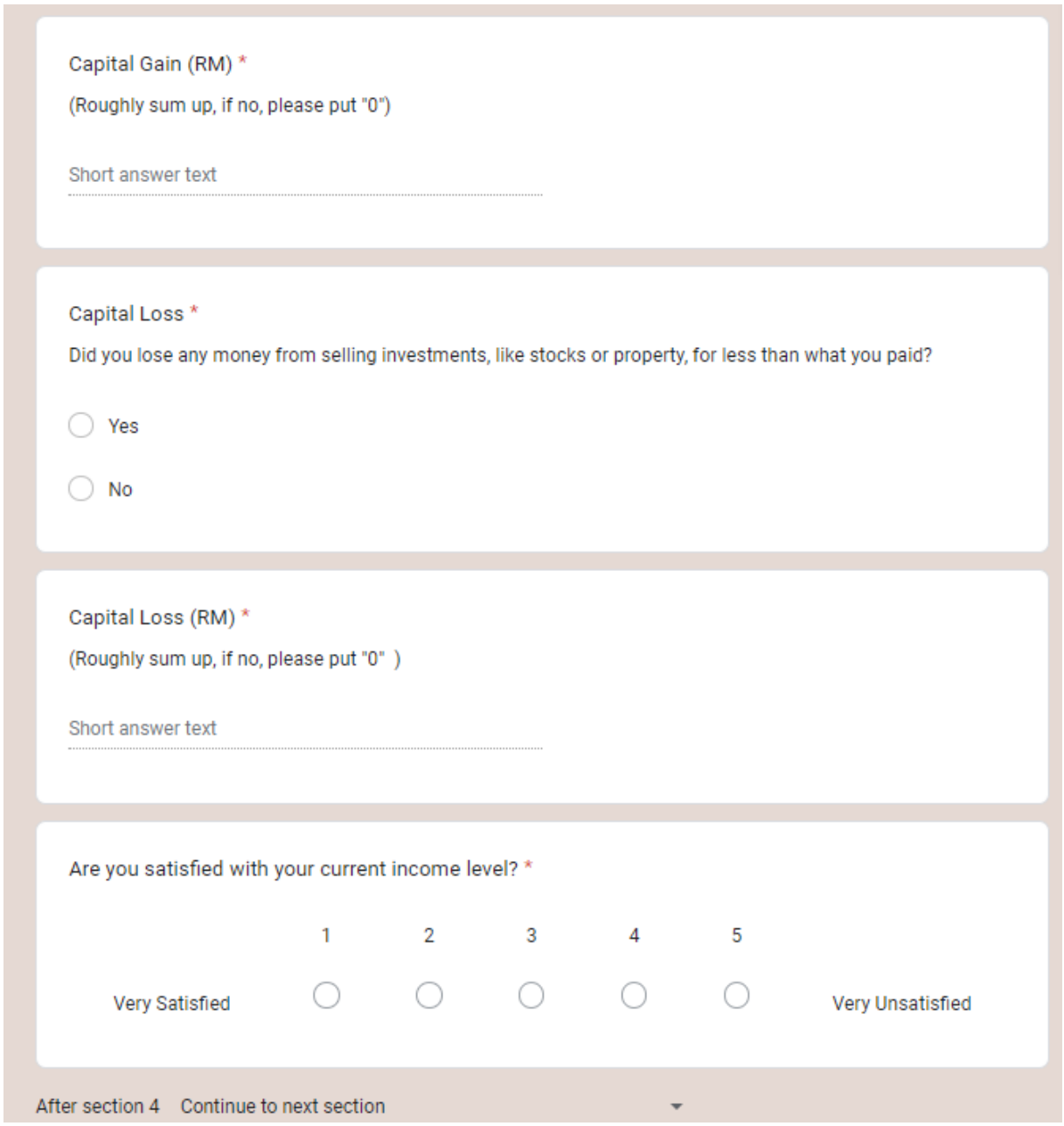
Years of Experience *

How many years of working experience do you have?

Short answer text


Years of Experience *
How many years of working experience do you have?
Short answer text
Have you pursued any additional professional certifications related to YOUR OCCUPATION? *
Yes
No
Number of Professional Certifications *
Short answer text
After section 3
Continue to next section


Section 4 of 6
Income Earnings
Description (optional)
Working Hours Per Week *
Short answer text
Current monthly income (RM) *
(Roughly Sum Up)
Short answer text
Do you have any additional sources of income or investments? *
Yes
No
Capital Gain *
Did you make any profit from selling investments, like stocks or property, for more than what you paid?
Yes
No



Capital Gain (RM) *

(Roughly sum up, if no, please put "0")

Short answer text

Capital Loss *

Did you lose any money from selling investments, like stocks or property, for less than what you paid?

Yes

No

Capital Loss (RM) *

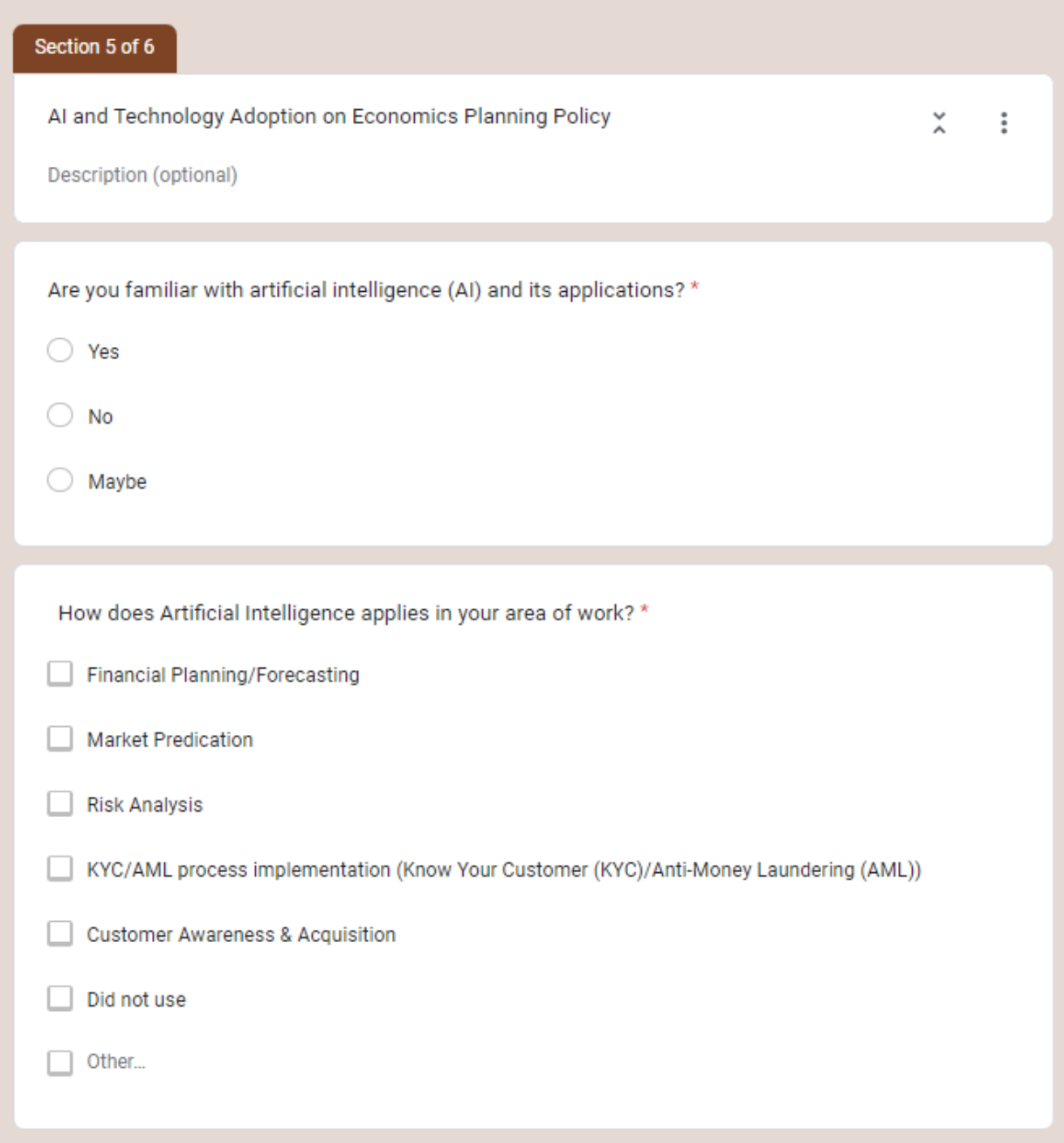
(Roughly sum up, if no, please put "0")

Short answer text

Are you satisfied with your current income level? *

	1	2	3	4	5	
Very Satisfied						Very Unsatisfied

After section 4 Continue to next section


Section 5 of 6
AI and Technology Adoption on Economics Planning Policy
Description (optional)
Are you familiar with artificial intelligence (AI) and its applications? *
Yes
No
Maybe
How does Artificial Intelligence applies in your area of work? *
Financial Planning/Forecasting
Market Predication
Risk Analysis
KYC/AML process implementation (Know Your Customer (KYC)/Anti-Money Laundering (AML))
Customer Awareness & Acquisition
Did not use
Other...

What can be or is already the purpose of AI adoption in your organization (select multiple options if applicable)? *

- [ ] Operational efficiency
- [ ] Cost reduction
- [ ] Creation of new value streams from an existing application
- [ ] Work-style reforms
- [ ] Other...

What is your level of understanding of Artificial Intelligence and Allied Technologies and the application domain? *

- ( ) I have sufficient research experience to be considered an expert
- ( ) Yes, I have a working knowledge of the basic concepts and terms
- ( ) I have limited knowledge about the subject
- ( ) I have heard the terms but don't understand them
- ( ) No knowledge at all

Is there a need to establish national centers on innovation, research, and entrepreneurship-focused on AI and allied technologies to allow knowledge-based socio-economic development in the country? *

- Existing initiatives are sufficient, and they shall be equitably funded
- Urgently needed for organizing a central hub for AI-related initiatives for developing a comprehensive eco...

Would you be willing to use AI-powered tools or platforms to get insights and understand more about improving your financial planning or income optimization? *

- Yes
- No

Do you think income analysis provided by AI can provide government policymakers with more accurate information to make more informed decisions on economic issues? *

- Yes
- No

Section 6 of 6

Thank You for Your Participation!

Description (optional)

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: 4, 1	**Study week no.: 3**
Student Name & ID: Ang Seng Chun, 20ACB02022	
Supervisor: Dr Abdulkarim Kanaan Jebna	
Project Title: A NEW PERSPECTIVE ON INCOME EARNINGS USING AI	

1. WORK DONE

- Respondent from the google form increased from 128 to 323.
- Studied 2 similar papers have done on predicting income, seeking for lower RMSE solution

2. WORK TO BE DONE

- Keep seeking more respondents to increase the dataset
- Searching more solutions on decreasing the RMSE of Income Prediction Regression Model.

3. PROBLEMS ENCOUNTERED

- Worrying that dataset respondents do not reach the target of 500 at the end.

4. SELF EVALUATION OF THE PROGRESS

- Have to work harder on seeking respondents

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: 4, 1	**Study week no.: 5**
Student Name & ID: Ang Seng Chun, 20ACB02022	
Supervisor: Dr Abdulkarim Kanaan Jebna	
Project Title: A NEW PERSPECTIVE ON INCOME EARNINGS USING AI	

1. WORK DONE

- Respondent from the google form increased from 323 to 428.
- Studied 4 similar papers have done on predicting income, seeking for lower RMSE solution

2. WORK TO BE DONE

- Searching more solutions on decreasing the RMSE of Income Prediction Regression Model.
- Started creating Classification models for the Income Prediction

3. PROBLEMS ENCOUNTERED

- Regression model seems do not get any lower of RMSE: 5687

4. SELF EVALUATION OF THE PROGRESS

- Data preparation and preprocessing are important as it affects a lot to the models

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: 4, 1	**Study week no.: 6**
Student Name & ID: Ang Seng Chun, 20ACB02022	
Supervisor: Dr Abdulkarim Kanaan Jebna	
Project Title: A NEW PERSPECTIVE ON INCOME EARNINGS USING AI	

1. WORK DONE

- Pipelines implemented, RMSE lower to 4357
- 1 Classification models created

2. WORK TO BE DONE

- Searching more solutions on decreasing the RMSE of Income Prediction Regression Model.
- Creating more Classification models

3. PROBLEMS ENCOUNTERED

- Categories data for pipeline seems facing obstacles as the data is contain too many varieties.
- Regression model seems do not get any lower of RMSE: 4357

4. SELF EVALUATION OF THE PROGRESS

- Meeting supervisor is important as he able to guide you and provide you some ideas to solve the current problems

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: 4, 1	**Study week no.: 9**
Student Name & ID: Ang Seng Chun, 20ACB02022	
Supervisor: Dr Abdulkarim Kanaan Jebna	
Project Title: A NEW PERSPECTIVE ON INCOME EARNINGS USING AI	

1. WORK DONE

- Categories data into Numerical, Nominal, Ordinal completed
- 3 Classification models created

2. WORK TO BE DONE

- Remove least relevant attributes to seek for lower RMSE

3. PROBLEMS ENCOUNTERED

- Transformation proceed to model faced errors

4. SELF EVALUATION OF THE PROGRESS

- Preprocessing really and truly important among all the model training.

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: 4, 1	**Study week no.: 10**
Student Name & ID: Ang Seng Chun, 20ACB02022	
Supervisor: Dr Abdulkarim Kanaan Jebna	
Project Title: A NEW PERSPECTIVE ON INCOME EARNINGS USING AI	

1. WORK DONE

- Refine the preprocessing stage
- Report draft for chapter 1 and chapter 2

2. WORK TO BE DONE

- Dashboard building

3. PROBLEMS ENCOUNTERED

- Seeking for lower RMSE for Regression Models.

4. SELF EVALUATION OF THE PROGRESS

- Would need to speed up the process lowering RSME

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: 4, 1	**Study week no.: 11**
Student Name & ID: Ang Seng Chun, 20ACB02022	
Supervisor: Dr Abdulkarim Kanaan Jebna	
Project Title: A NEW PERSPECTIVE ON INCOME EARNINGS USING AI	

1. WORK DONE

- Applied OneHotEncoder, drop=first, tried evaluation on train and test set on classification
- Report draft for chapter 3 and partially chapter 4

2. WORK TO BE DONE

- Seeking for more possibility on lower RMSE through shorten the preprocessing process

3. PROBLEMS ENCOUNTERED

- Error appeared where when I try to preprocessing.fit_transform my train set and transform my test set, it appears something is missing in my test set

4. SELF EVALUATION OF THE PROGRESS

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: 4, 1	**Study week no.: 12**
Student Name & ID: Ang Seng Chun, 20ACB02022	
Supervisor: Dr Abdulkarim Kanaan Jebna	
Project Title: A NEW PERSPECTIVE ON INCOME EARNINGS USING AI	

1. WORK DONE

- Error preprocessing.fit_transform imbalance amount attributes between train and test solved
- Dashboard for Income Data Visualization built

2. WORK TO BE DONE

- Restructure for chapter 1,2,3
- Draft on Chapter 4, 5, 6, 7

3. PROBLEMS ENCOUNTERED

- Dashboard partially worked, required more time to solve the bugs and errors on VS Code.

4. SELF EVALUATION OF THE PROGRESS

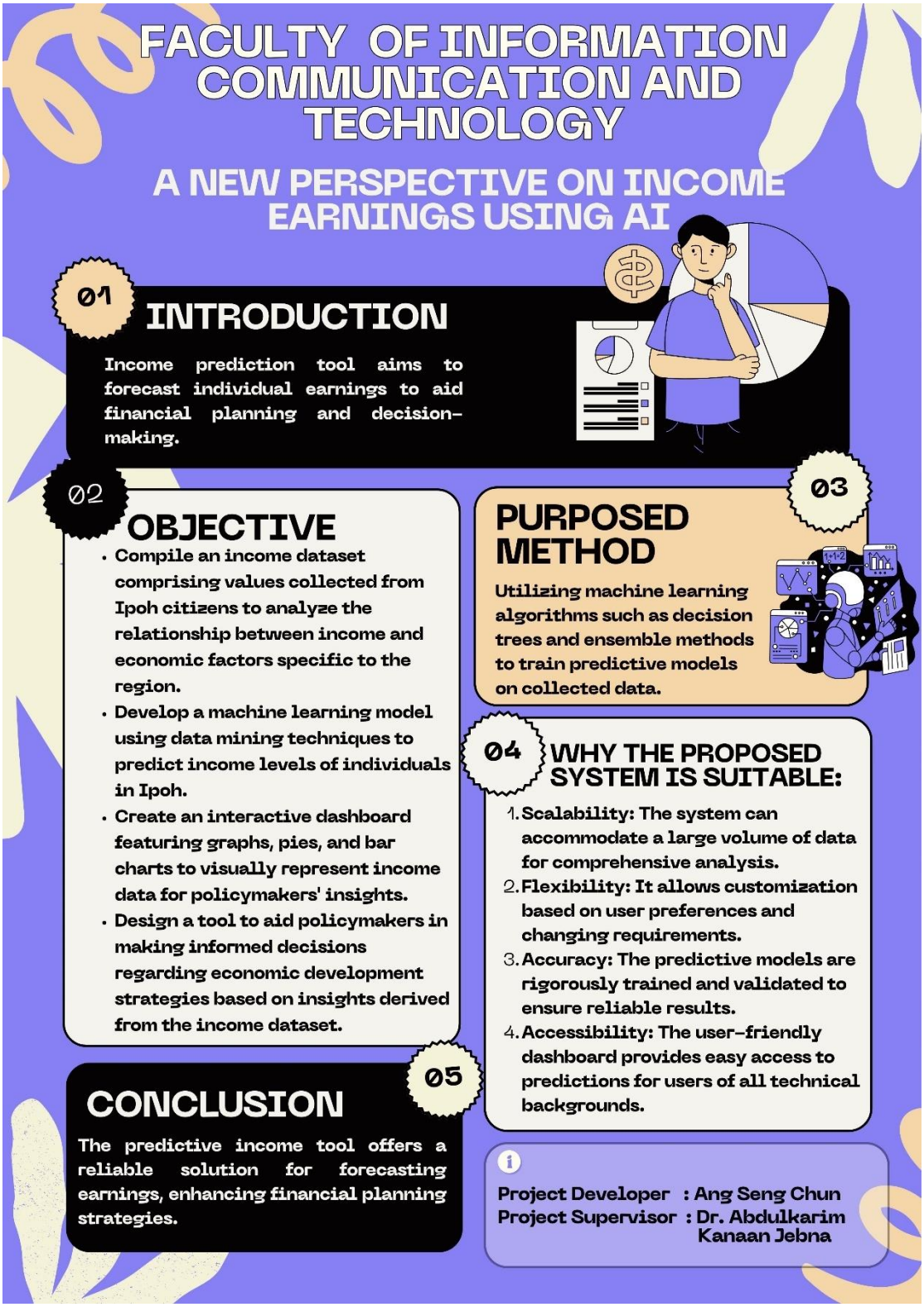
- Need to rush the report now!

Supervisor's signature

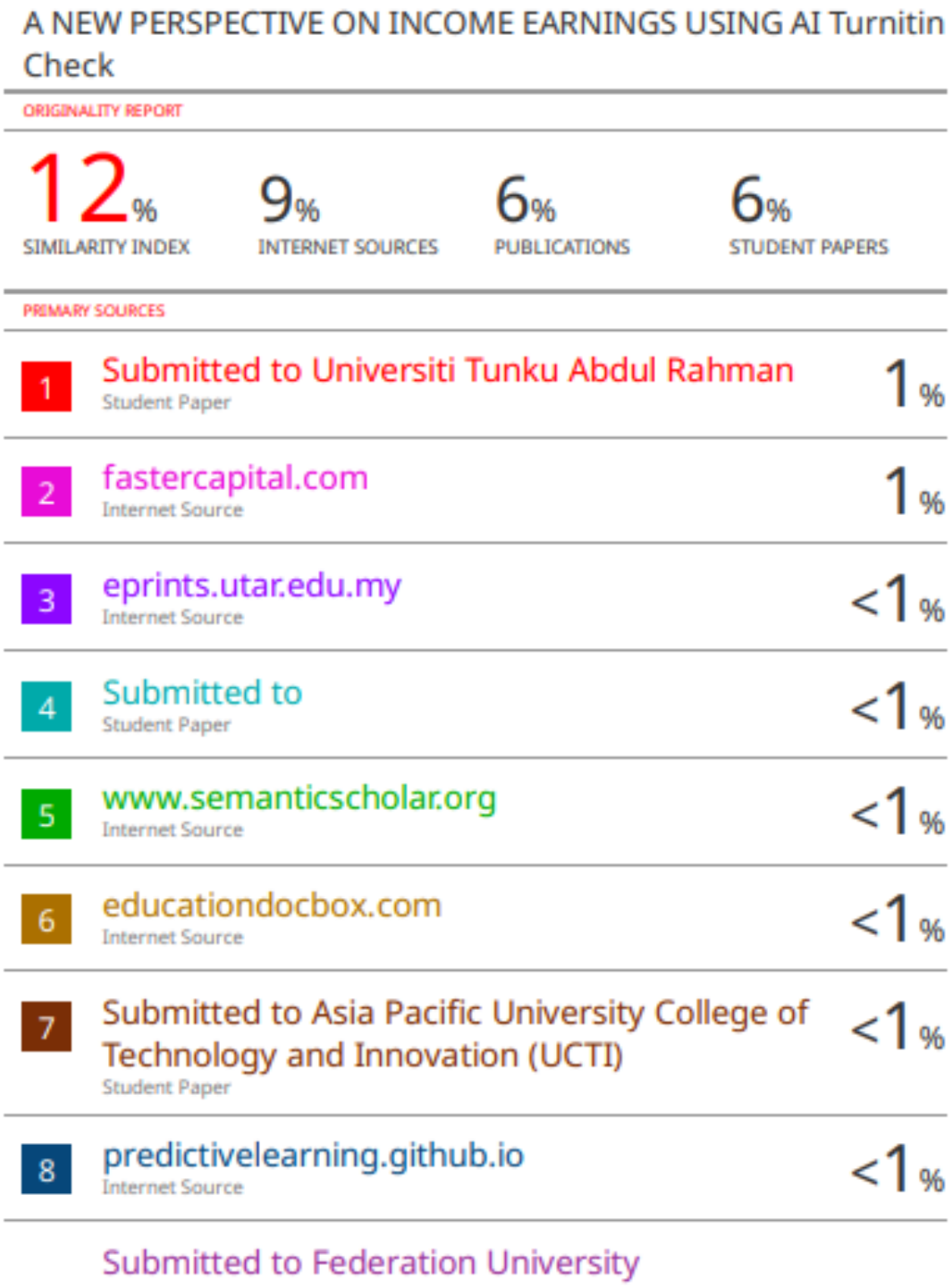
Student's signature

PLAGIARISM CHECK RESULT

POSTER



PLAGIARISM CHECK RESULT



A NEW PERSPECTIVE ON INCOME EARNINGS USING AI Turnitin Check

ORIGINALITY REPORT

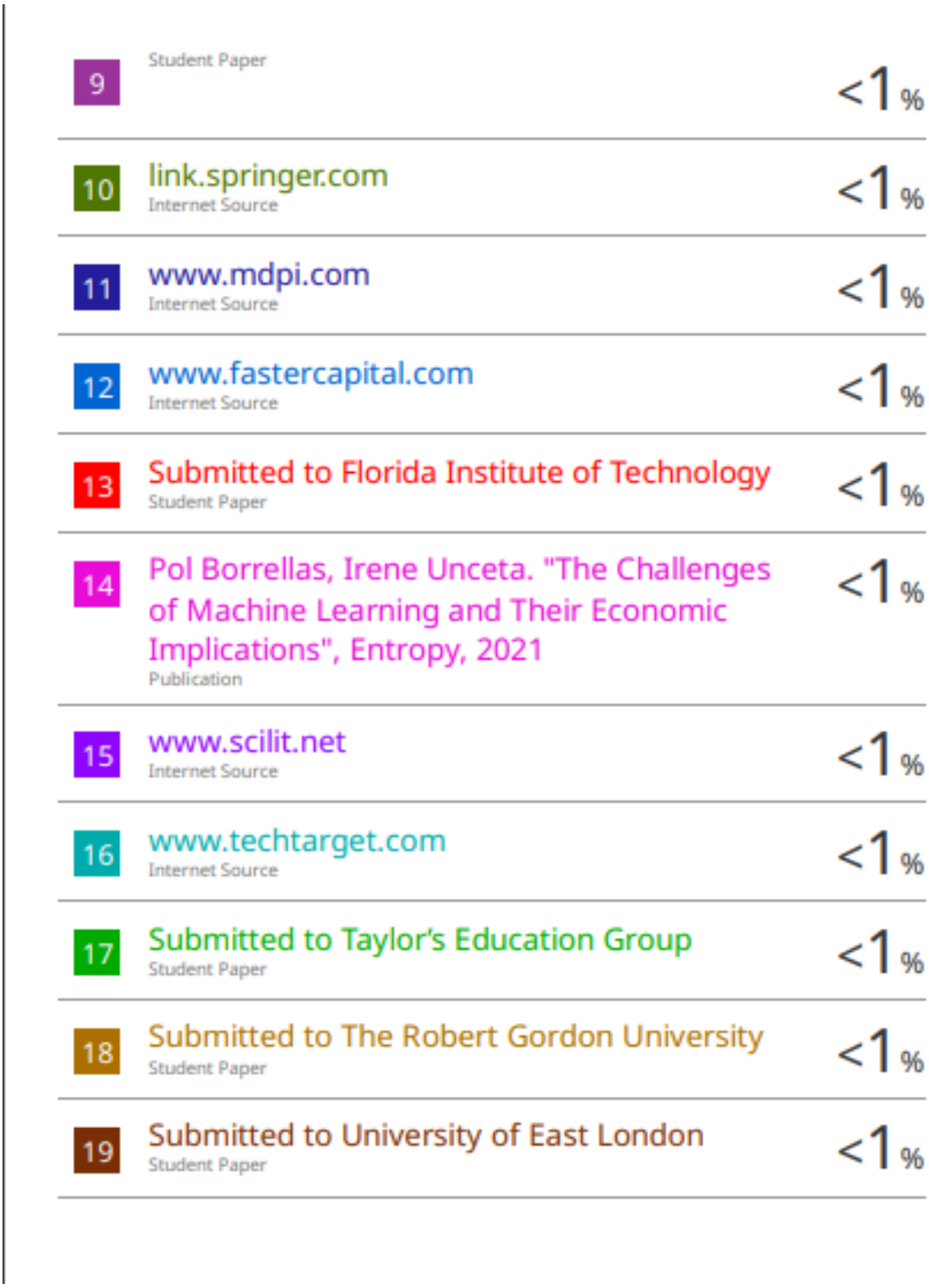
12%	9%	6%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universiti Tunku Abdul Rahman Student Paper	1%
2	fastercapital.com Internet Source	1%
3	eprints.utar.edu.my Internet Source	<1%
4	Submitted to Student Paper	<1%
5	www.semanticscholar.org Internet Source	<1%
6	educationdocbox.com Internet Source	<1%
7	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%
8	predictivelearning.github.io Internet Source	<1%

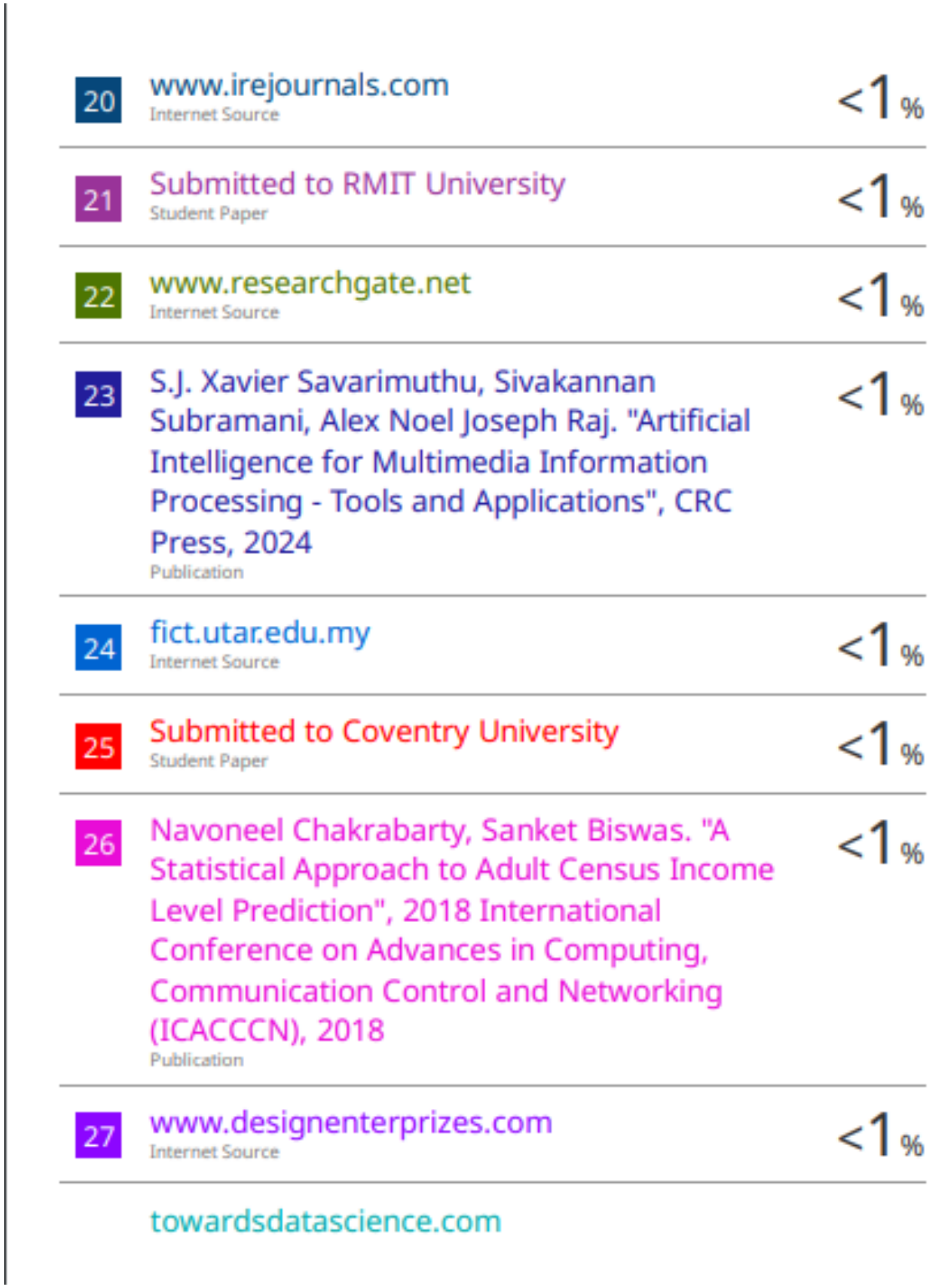
Submitted to Federation University

PLAGIARISM CHECK RESULT

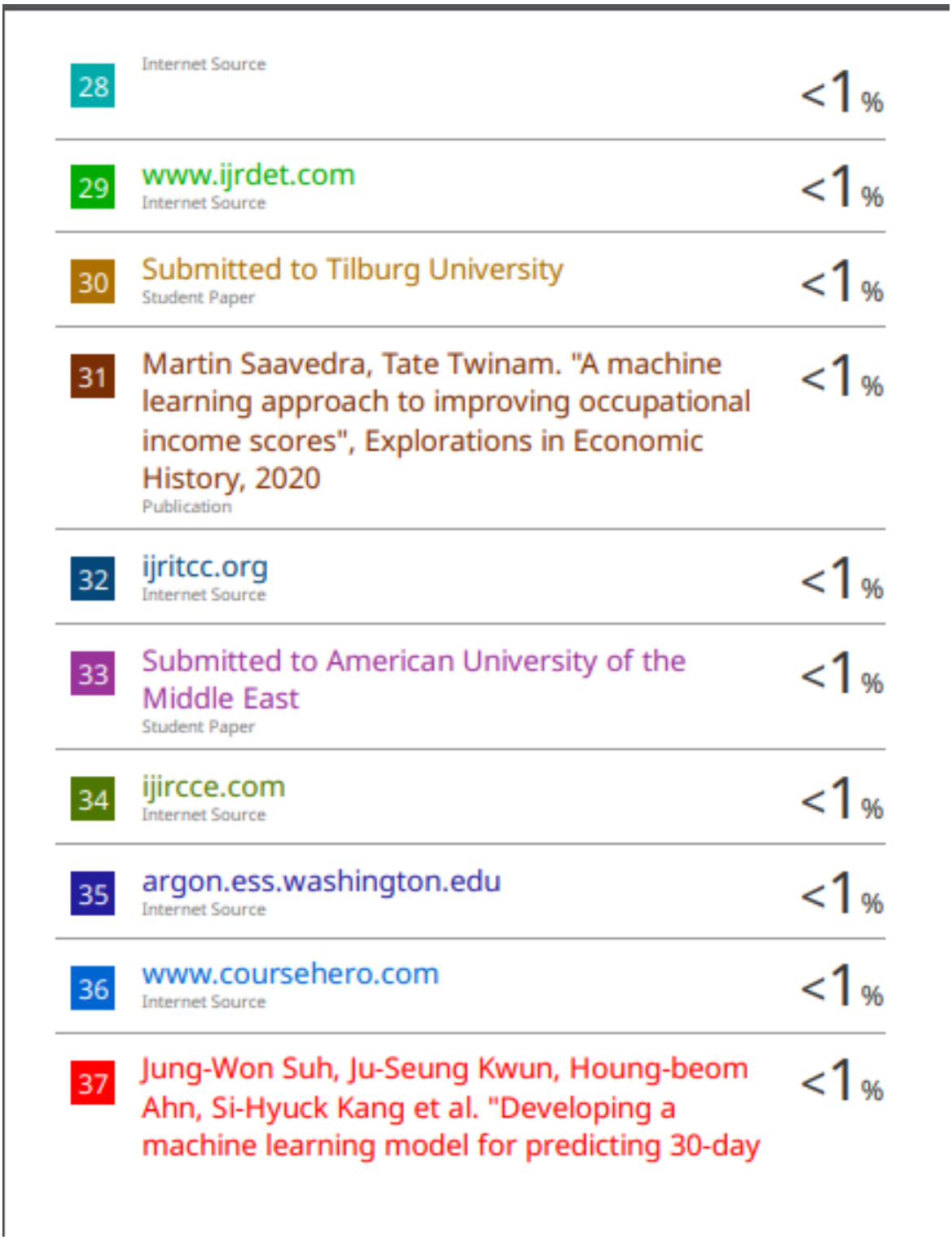


9	Student Paper	<1%
10	link.springer.com Internet Source	<1%
11	www.mdpi.com Internet Source	<1%
12	www.fastercapital.com Internet Source	<1%
13	Submitted to Florida Institute of Technology Student Paper	<1%
14	Pol Borrellas, Irene Unceta. "The Challenges of Machine Learning and Their Economic Implications", Entropy, 2021 Publication	<1%
15	www.scilit.net Internet Source	<1%
16	www.techtarget.com Internet Source	<1%
17	Submitted to Taylor's Education Group Student Paper	<1%
18	Submitted to The Robert Gordon University Student Paper	<1%
19	Submitted to University of East London Student Paper	<1%

PLAGIARISM CHECK RESULT

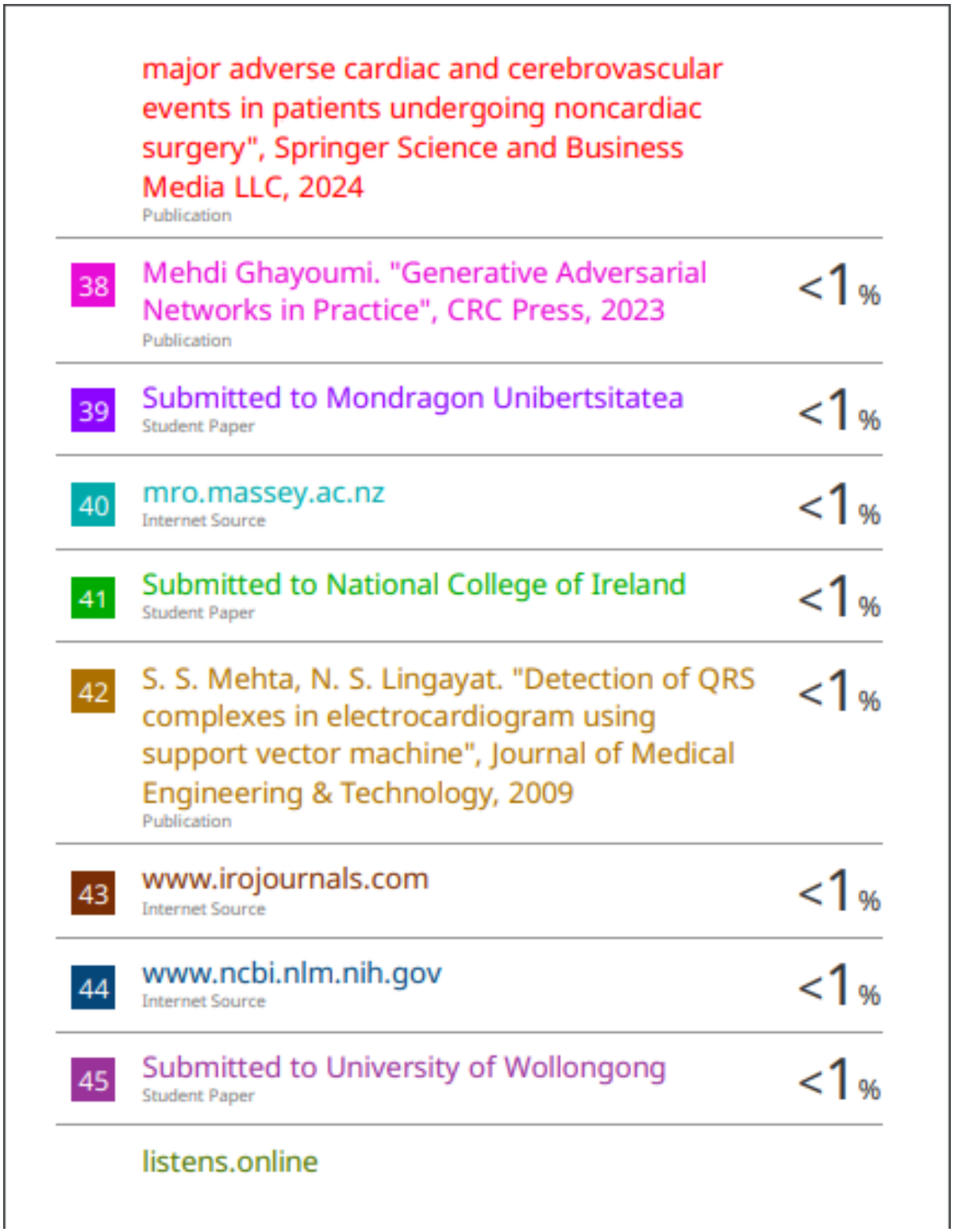


PLAGIARISM CHECK RESULT

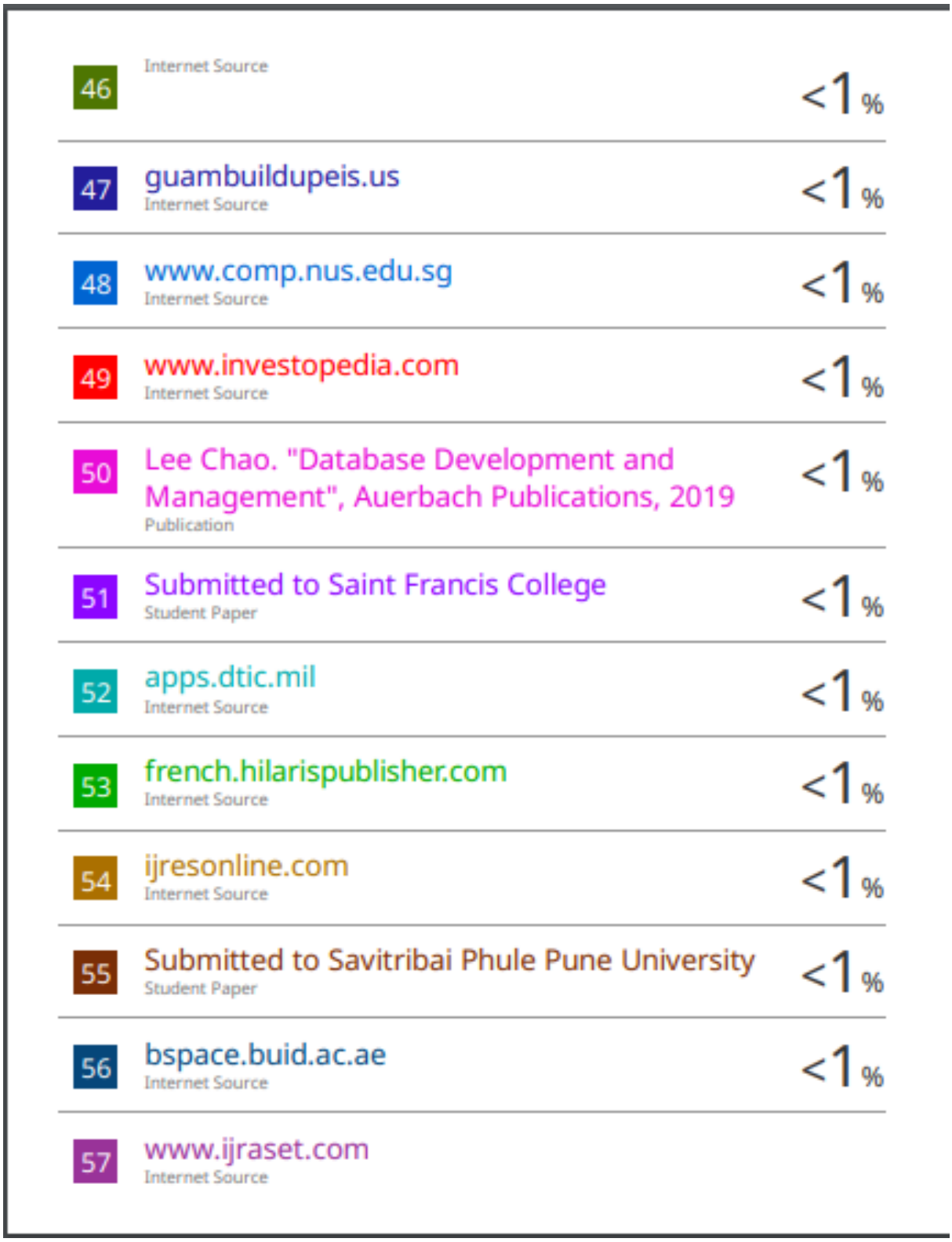


28	Internet Source	<1%
29	www.ijrdet.com Internet Source	<1%
30	Submitted to Tilburg University Student Paper	<1%
31	Martin Saavedra, Tate Twinam. "A machine learning approach to improving occupational income scores", Explorations in Economic History, 2020 Publication	<1%
32	ijritcc.org Internet Source	<1%
33	Submitted to American University of the Middle East Student Paper	<1%
34	ijircce.com Internet Source	<1%
35	argon.ess.washington.edu Internet Source	<1%
36	www.coursehero.com Internet Source	<1%
37	Jung-Won Suh, Ju-Seung Kwun, Houng-beom Ahn, Si-Hyuck Kang et al. "Developing a machine learning model for predicting 30-day	<1%

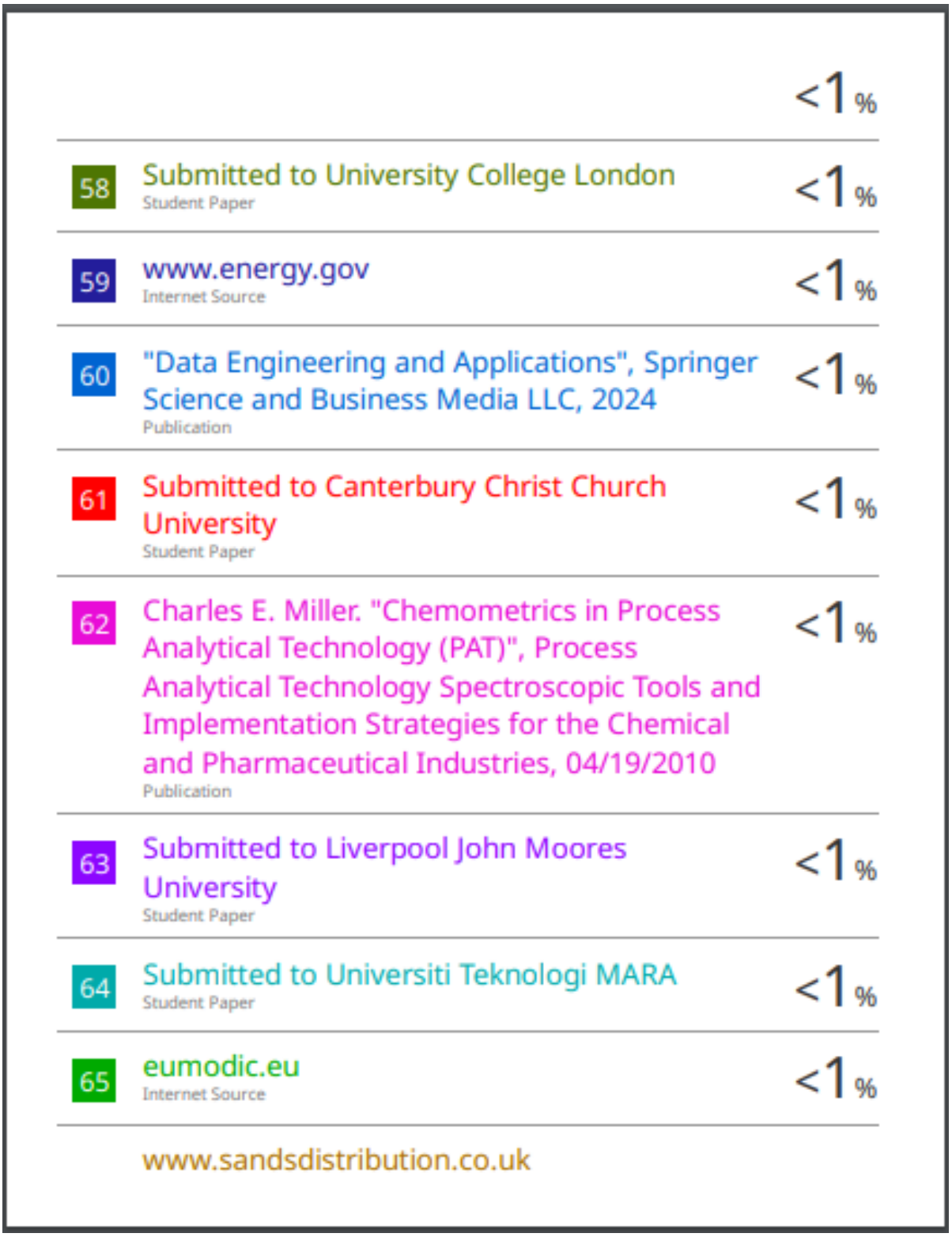
PLAGIARISM CHECK RESULT



PLAGIARISM CHECK RESULT

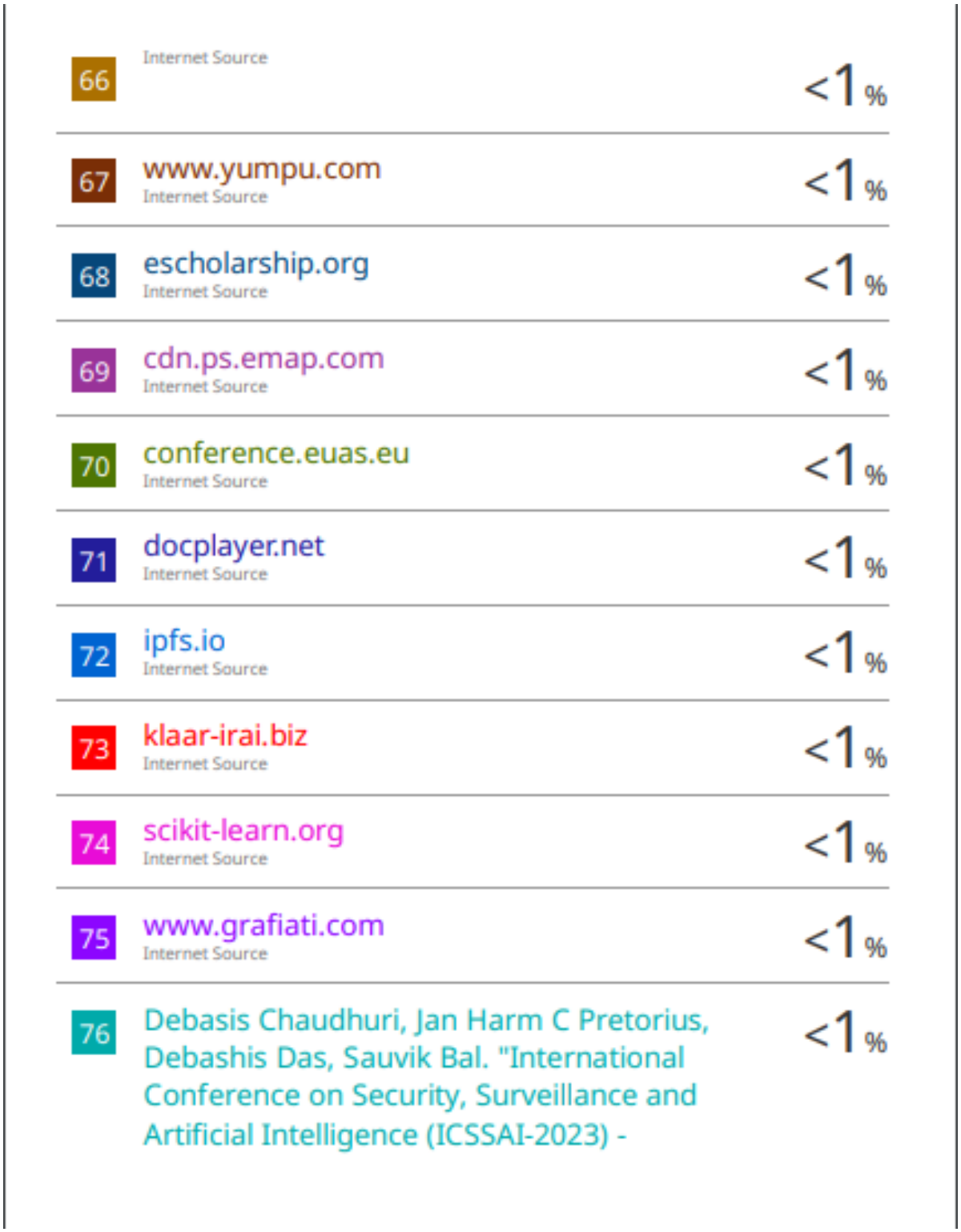


PLAGIARISM CHECK RESULT

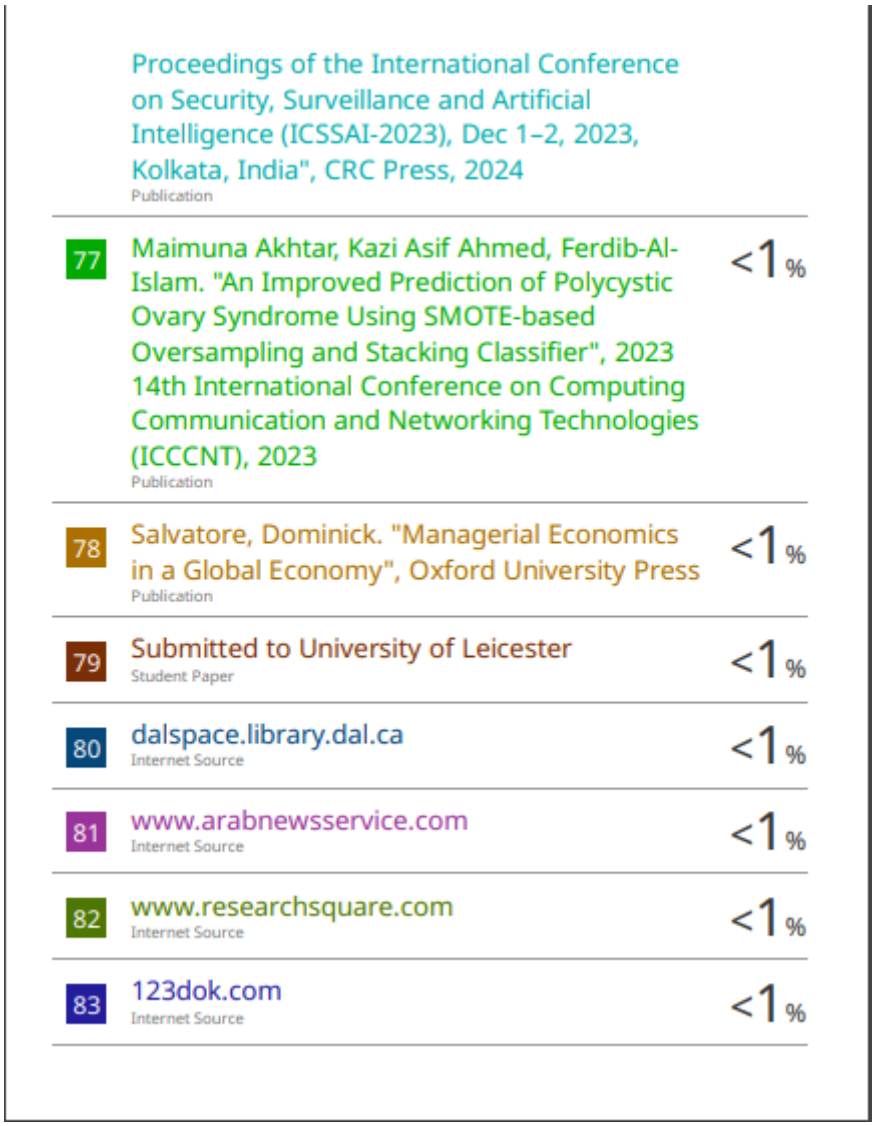


		<1 %
58	Submitted to University College London Student Paper	<1 %
59	www.energy.gov Internet Source	<1 %
60	"Data Engineering and Applications", Springer Science and Business Media LLC, 2024 Publication	<1 %
61	Submitted to Canterbury Christ Church University Student Paper	<1 %
62	Charles E. Miller. "Chemometrics in Process Analytical Technology (PAT)", Process Analytical Technology Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries, 04/19/2010 Publication	<1 %
63	Submitted to Liverpool John Moores University Student Paper	<1 %
64	Submitted to Universiti Teknologi MARA Student Paper	<1 %
65	eumodic.eu Internet Source	<1 %
	www.sandsdistribution.co.uk	

PLAGIARISM CHECK RESULT

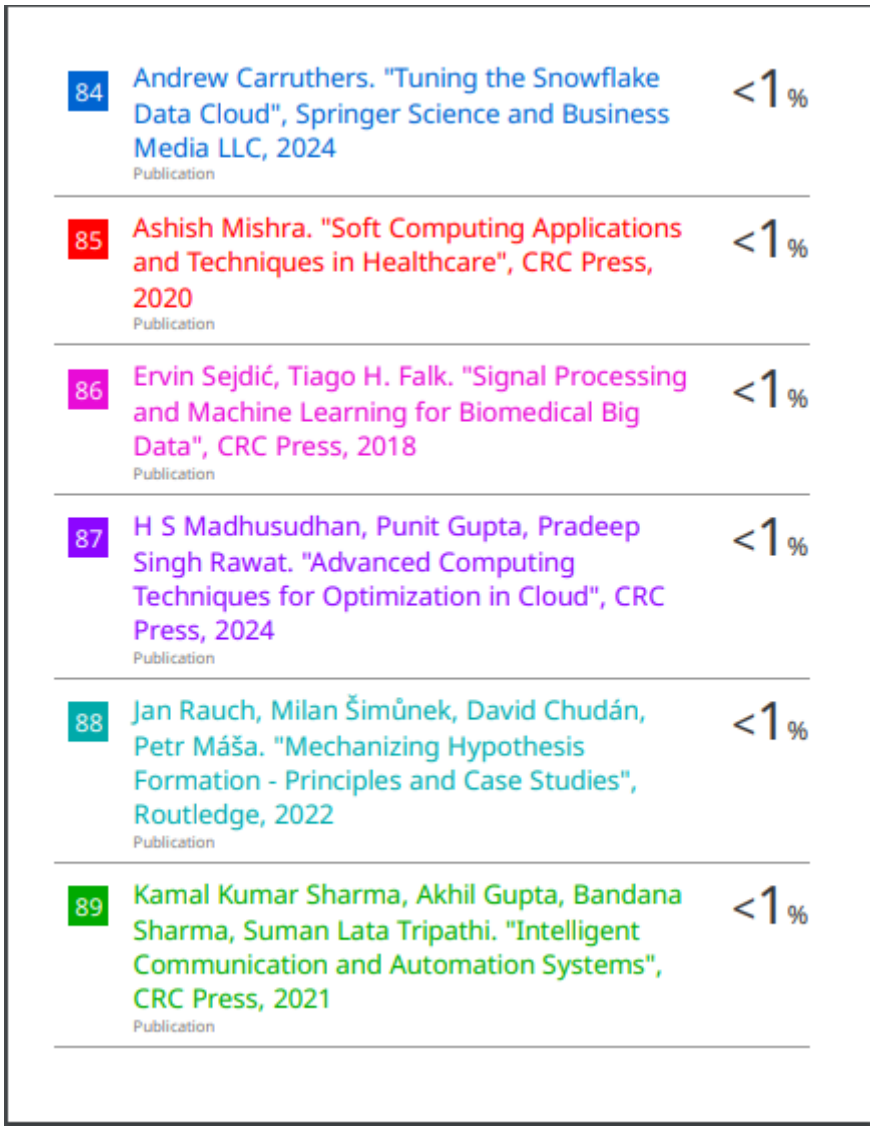


PLAGIARISM CHECK RESULT

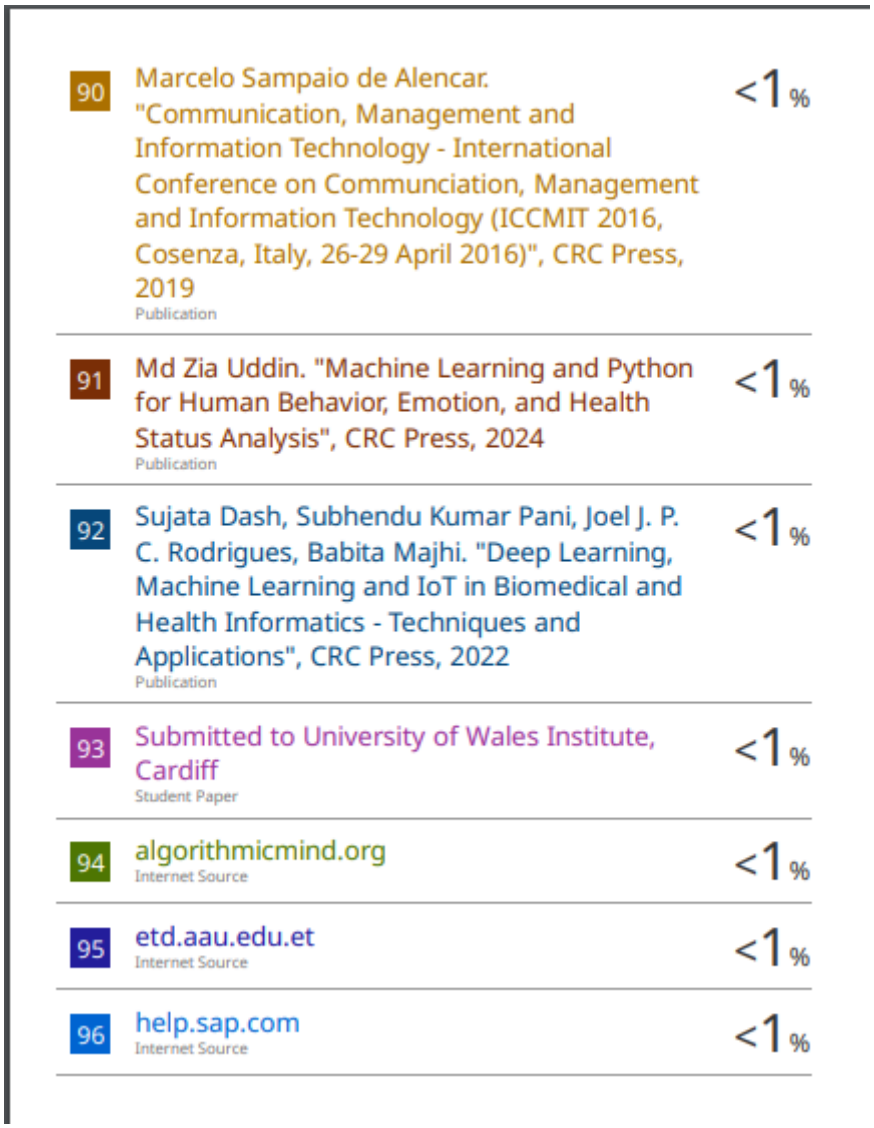


	Proceedings of the International Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023), Dec 1–2, 2023, Kolkata, India", CRC Press, 2024 Publication	
77	Maimuna Akhtar, Kazi Asif Ahmed, Ferdib-Al-Islam. "An Improved Prediction of Polycystic Ovary Syndrome Using SMOTE-based Oversampling and Stacking Classifier", 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023 Publication	<1 %
78	Salvatore, Dominick. "Managerial Economics in a Global Economy", Oxford University Press Publication	<1 %
79	Submitted to University of Leicester Student Paper	<1 %
80	dalspace.library.dal.ca Internet Source	<1 %
81	www.arabnewsservice.com Internet Source	<1 %
82	www.researchsquare.com Internet Source	<1 %
83	123dok.com Internet Source	<1 %

PLAGIARISM CHECK RESULT

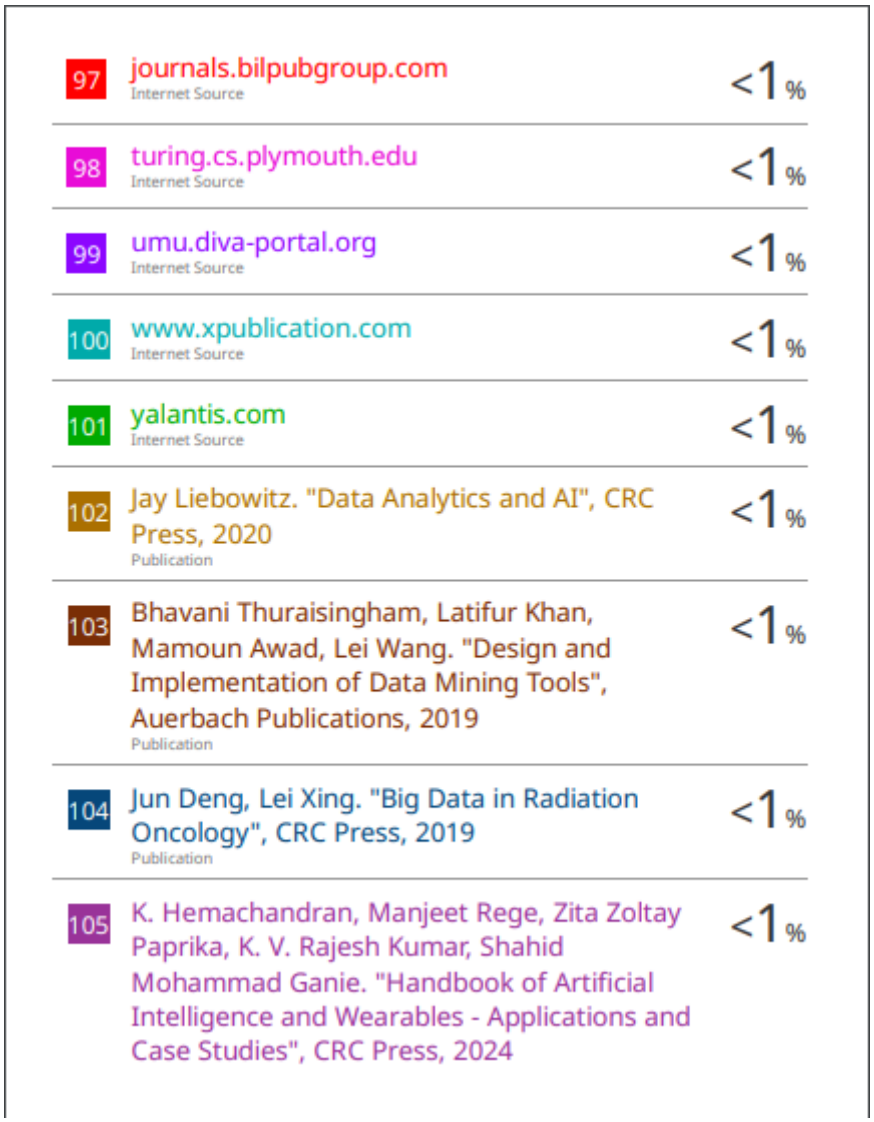


PLAGIARISM CHECK RESULT



90	Marcelo Sampaio de Alencar. "Communication, Management and Information Technology - International Conference on Communciation, Management and Information Technology (ICCMIT 2016, Cosenza, Italy, 26-29 April 2016)", CRC Press, 2019 Publication	<1 %
91	Md Zia Uddin. "Machine Learning and Python for Human Behavior, Emotion, and Health Status Analysis", CRC Press, 2024 Publication	<1 %
92	Sujata Dash, Subhendu Kumar Pani, Joel J. P. C. Rodrigues, Babita Majhi. "Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics - Techniques and Applications", CRC Press, 2022 Publication	<1 %
93	Submitted to University of Wales Institute, Cardiff Student Paper	<1 %
94	algorithmicmind.org Internet Source	<1 %
95	etd.aau.edu.et Internet Source	<1 %
96	help.sap.com Internet Source	<1 %

PLAGIARISM CHECK RESULT



PLAGIARISM CHECK RESULT

Publication

106 M.L. Jeremic. "Strata Mechanics in Coal Mining", A.A. BALKEMA, 2020 <1%
Publication

107 O. P. Verma, Seema Verma, Thinagaran Perumal. "Advancement of Intelligent Computational Methods and Technologies - Proceedings of I International Conference on Advancement of Intelligent Computational Methods and Technologies (AICMT2023)", CRC Press, 2024 <1%
Publication

108 R. N. V. Jagan Mohan, Vasamsetty Chandra Sekhar, V. M. N. S. S. V. K. R. Gupta. "Algorithms in Advanced Artificial Intelligence", CRC Press, 2024 <1%
Publication

109 Seong-Uk Baek, Jong-Uk Won, Jin-Ha Yoon. "Association between long working hours and the onset of problematic alcohol use in young workers: A population-based longitudinal analysis in South Korea", Journal of Affective Disorders, 2024 <1%
Publication

Exclude quotes Off Exclude matches Off

PLAGIARISM CHECK RESULT

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1of 1

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	ANG SENG CHUN
ID Number(s)	20ACB02022
Programme / Course	Bachelor of Information Systems (Honours) Information Systems Engineering - IA
Title of Final Year Project	A New Perspective on Income Earnings Using AI

Similarity	**Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)**
Overall similarity index: 12 % **Similarity by source** Internet Sources: 9 % Publications: 6 % Student Papers: 6 %	
Number of individual sources listed of more than 3% similarity: 0	

Parameters of originality required and limits approved by UTAR are as Follows:

(i) Overall similarity index is 20% and below, and

(ii) Matching of individual sources listed must be less than 3% each, and

(iii) Matching texts in continuous block must not exceed 8 words

Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Signature of Supervisor

Name: Dr Abdulkarim M. Jamal Kanaan Jebna

Date: 13 / 9 / 2024

Signature of Co-Supervisor

Name: _____

Date: _____

PLAGIARISM CHECK RESULT

FYP 2 CHECKLIST

UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY
(KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	20ACB02022
Student Name	Ang Seng Chun
Supervisor Name	Dr Abdulkarim Kanaan Jebna

TICK (√)	DOCUMENT ITEMS Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
√	Title Page
√	Signed Report Status Declaration Form
√	Signed FYP Thesis Submission Form
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
	List of Symbols (if applicable)
√	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
√	Appendices (if applicable)
√	Weekly Log
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)
√	I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report.

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

_____

(Signature of Student)
Date: 13 September 2024

PLAGIARISM CHECK RESULT