COUNTRY'S INCOME LEVEL CLASSIFICATION BY SUPERVISED LEARNING APPROACH BASED ON LEGAL GENDER EQUALITY

HO WENG TIM

ABSTRACT

A country's income and economic development are highly correlated with women's financial empowerment. Legislation is important to guarantee gender-equal access to social-economic advantages, which can help to develop gender equality and female empowerment. Gender equality is a form of human right and the phenomenon of giving equal opportunities and resources to both women and men. There is a need to push the bottom economic country's income category steps to a higher level by eliminating this legal barrier. There is a lack of an automation approach that can relate a country's economic growth with legal gender equality. Supervised learning approaches are proposed in this proposal as it is an effective technology that can aid in the classification and prediction of country income levels based on legal gender equality. Before model building, exploratory data analysis and data pre-processing are carried out. The models that were tested in this project included based and tuned support vector machine, random forest, gradient boosting and artificial neural network. The evaluation matrices that were used included sensitivity, specificity, accuracy and F1 score. Based on the output, it was found that tuned random forest and tuned gradient boosting achieve the highest performance. By utilizing this approach, interest institutions can make a virtual adjustment to gender equality laws, foresee the possibility of economic improvement and implement the best strategy. In this context, the fastest and key solution to boost economic growth and achieve women's financial empowerment can be recognized effectively.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Research Introduction

Economic development is the transformation of poor global markets into contemporary developed countries, and it refers to both qualitative and quantitative changes in a country's economy (Krueger, n.d.). It is a coordinated attempt by an accountable government entity to steer private sector participation into alternatives that will result in extensive economic growth. Economic growth can offer enough revenue for the regional workforce, viable business possibilities for businesses, and tax income to keep the systems in place to sustain this progress (California Association for Local Economic Development, 2022). It aims to raise total per capita earnings in order to boost material living conditions (Krueger, n.d.). The process of raising a nation's actual gross domestic product (GDP) is referred to as economic growth. The expansion of real GDP or gross national product (GNP) during a specific time period can be used to quantify growth. In order to understand economic growth more deeply, it is necessary to examine the factors that contribute to a consistent increase in production capacity (CFI, 2021).

Economic development is a crucial factor in economic expansion because it generates new employment options and promotes an enhanced standard of living for both current and future citizens by enabling expanded access to opportunities caused by economic growth. There are many advantages that economic development brings which included creating employment opportunities, industry diversification, the continuation and growth of businesses, economic bolstering, increase tax revenue, and an enhanced standard of living (Amanda, 2022). Economic growth contributes to the production of more affordable, higher-quality goods and services. It has led to an upsurge in both consumer and commercial goods manufacturing. It causes a rise in per capita revenue, which raises a country's overall national wealth. Economic growth encourages social equality so that revenue and wealth are distributed equally and everybody has access to some degree of material prosperity, status, and standard of living (SubjectQuery, 2019).

Gender equality is when individuals of both sexes should get the same respect, privileges, and obligations. It is both a basic human right and a prerequisite for economic development (Vic.Gov.Au, n.d.). Society thrives when all genders are given equal employment possibilities. Studies have shown that varied workplaces are more efficient, and gender diversity is part of this diversification. When all genders are provided equal employment opportunities, rates of poverty are lowered, societies are strengthened, and a country's GDP is greatly increased (Martinez, 2022). Nevertheless, despite advancements, women and girls around the world still do not completely enjoy equal rights, and their power as change-agents for the economy, society, and sustainable development is still unrealized (United Nations Global Compact, n.d.).

For men and women to have comparable legislative entitlements in all spheres of life and to prevent institutionalized prejudice against women, complete gender equality legislation is essential (Legistrationline, n.d.). Laws could be necessary to guarantee gender-equitable chances for economic advantages that will improve the economic independence of women (OECD, 2018). Gender equality must be achieved for a nation's economy to be sustainable and thrive (International Monetary Fund, 2022; UN Women, 2018). To improve a country's economic standing, it is essential to guarantee that each gender has equal access to work opportunities, management positions, and choices at all levels (United Nation, n.d.). There are several strategies utilized to reduce gender disparity and accelerate world economic development, but no successful method has been found to link legal gender equality to national income.

Hence, an automated system that can be used to discover the regulation obstacle in order to achieve gender equality is important. The method suggested in this project to achieve this goal is supervised learning, which uses labeled data to train models to identify inputs and predict outcomes (IBM, 2020). In order to effectively eliminate the gender pay gap, discover legislative obstacles to women's economic achievement, and boost worldwide financial development, this project is developing an automated method to define a country's income status based on gender equality laws. In this project there are few algorithms were used which included artificial neural network, gradient boosting, support vector machine and random forest. The results of these models are compared to select the best model.

1.2 Problem Statement

The main problem statement of this study is there is a lack of approaches that leverages the key variable for prediction because the independent factors of other supervised learning research are not the primary impediment to economic growth. Additionally, there is a lack of supervised learning approaches that can be applied across national borders because the majority of these approaches come from research that places a greater emphasis on the economic development of particular nations or areas, making them inapplicable to other nations. Besides, there is a shortage of an efficient supervised learning method that can be utilized as a benchmark to help determine the major legal obstacles that impede women's financial empowerment and global economic progress.

1.3 Research Question

For the purpose of this study, the following questions are addressed:

1. What are the effective supervised learning approaches to identify income categories of a country based on gender equality laws?
2. How can the models be optimized to be the best fit for accurately classifying the income categories of a country based on legal gender equality?
3. How the best supervised learning model for income categories classification of a country based on legal gender equality be determined?

1.4 Aim and Objective

The main aim of this research is to develop an effective supervised learning approach to identify the income categories of a country based on legal gender equality.

For the purpose of this study, the following objectives are addressed:

- To identify effective supervised learning approaches for income categories classification of a country based on legal gender equality.
- To optimize and fine-tune the models to be the best fit for accurately classifying the income categories of a country based on legal gender equality.
- To measure and estimate the performance of developed supervised learning models for income category classification of a country based on legal gender equality.

1.5 Research Scope

A classification model used for country income levels is developed through supervised learning algorithms in this research. The programming language used in this research is Python. The country's income levels are divided into high-income levels, upper-middle income, low-middle income, and low-income levels. The input dataset is obtained from the Data Catalog of the World Bank Group which is an international financial institution. The information of the dataset is collected by the Women, Business and the Laws, which is a project under the World Bank Group that aims to gather specific information on the legislation and policies that limit women's economic potential in order to promote gender equality. 50 years of data which is from the year 1971 to 2021, including 190 countries in 7 regions are used as the input label dataset. The regions involved included: Sub-Saharan Africa, South Asia, Middle East and North Africa, Latin America and Caribbean, Organization for Economic Co-operation and Development (OECD), Europe and Central Asia and East Asia and Pacific. The gender equality laws are grouped into 8 indicators: pension, assets, entrepreneurship, parenthood, marriage, pay, workplace, and mobility. The indicators focus on rules that may limit women's capacity to pursue career opportunities and start their own businesses. Each indicator contains four or five questions about whether laws and policies exist that support socioeconomic gender equality. 35 legal topics in total are examined in Women, Business, and the Law.

1.6 Significance of research

Numerous studies have demonstrated that gender equality and the financial empowerment of women have a substantial impact on national and even global economic growth. To help with the accomplishment of the Sustainable Development Goals (SDGs) before 2030, this recommended research is crucial. Since the most recent United Nations report revealed that the progress toward the SDGs is still agonizingly slow, there is a need for ways to expedite this progress. This suggested supervised learning is crucial since it is a method of automation that many nations can utilize. According to other studies, the majority of supervised learning techniques related to economic growth are only applicable to particular nations or regions. In light of the fact that it includes 190 countries across 7 regions, the suggested supervised learning approach helps to promote economic progress.

The variables included in this project are primarily legislation-related indicators that have an impact on women's financial possibilities, making the suggested supervised learning approach more trustworthy. Numerous studies have demonstrated the strong link between economic growth and women's financial empowerment. Legislation frequently promotes gender stereotypes that limit women's economic participation (IMFBlog, 2022). Legislation can affect women's decisions to enter and remain in the workforce (Wadhwa & Halim, 2020). Globally, income disparities between men and women may lead to a loss of $23,620 in capital per person, which would impede economic development (The World Bank, 2018). According to the International Monetary Fund, reducing gender inequality may increase GNP by 35%. (Malaza & Parekh, 2020). The financial empowerment of women is a prerequisite for the nation's economic progress. When contrasted to other methods of existing studies, the factors are more trustworthy to be utilized to classify and predict the income level of a country because they directly influence the financial empowerment of women.

The secret to advancing the world economy is gender equality. With the exception of 2030, there are fewer than 8 years left. It is difficult for a nation at the bottom of the economic scale that has maintained a culture of gender prejudice for generations to change all the laws pertaining to gender inequality in such a short amount of time. Therefore, it is crucial to pinpoint the main legislative impediment that delays the process of improving women's financial prospects and raising the nation's revenue. In order to identify the primary law obstacle and remove it, this project was created to serve as automation guidelines for the income level prediction of a country based on gender equality legislation. Interest organizations can utilize this approach to identify legislative obstacles to women's financial success, get rid of them, and put the best plan of action into place in order to promote economic progress.

Using this supervised learning approach, organizations can practically change gender equality legislation in the automated system to anticipate the potential of the country's income level development. Many nations have been stuck at the same income level for years. In this situation, organizations can successfully pinpoint the major challenges, deal with these problems, and address them in order to hasten the transition of the nation's revenue from the old level to a higher level. This planned execution not only promotes economic growth but also empowers women financially at the same time. Since studies and research have shown that the COVID-19 pandemic will worsen women's financial chances and cause more women than men to lose their jobs, this situation will ultimately endanger national or even global economic stability. Therefore, it is crucial to find efficient solutions in order to avoid reversing the progress made in women's financial empowerment and global economic development.

CHAPTER 2

LITERATURE REVIEW

2.1 Background

There are many studies or professions had pointed out that gender equality is highly correlated with economic development. In addition to reducing gender bias, assisting women in achieving their capacity in the field of finance has a substantial positive effect on the development, profitability, and future of economies around the world (Christopherson et al., 2022). By reducing the wage disparity between men and women, the global GDP might rise by 26%, benefiting both developed and developing nations when women entering the labor market at the same frequency as men, working a similar amount of time as men, and being engaged in the same positions as males across industries would be beneficial for both developed and emerging nations. By matching the pace at which the country in their region is advancing toward equality, even lacking complete equalization, countries can increase their GDP by $12 trillion (Growing Economies Through Gender Parity, n.d.). McKinsey Global Institute, an organization that provides true findings from economics and business studies to assist in policy and management choices reported the growth rate of GDP by 2025 with gender equity throughout all regions' workforce in Figure 2.1.

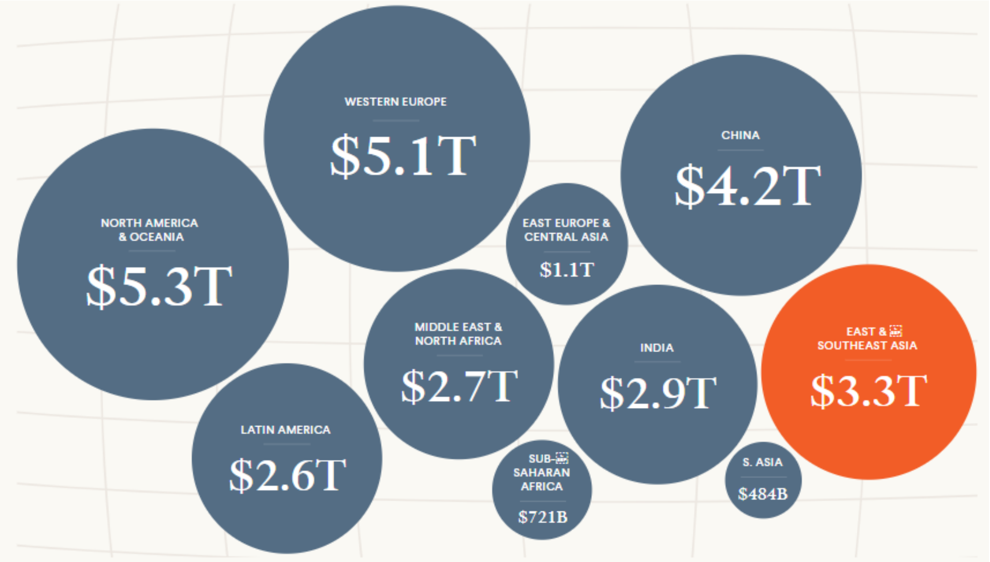


Figure 2.1 Growth rate of GDP by 2025 with gender equity throughout all regions' workforce

According to numerous studies, the attainment of women 's financial independence is said to have a substantial impact on global growth in the economy. International Monetary Fund (IMF), a global financial agency that promotes economic growth and sustainable stability is also concerned about female empowerment due to its impact on financial stability (International Monetary Fund, 2022). Furthermore, according to economists at BofA Securities, if gender equality is accomplished by 2025, the global GDP might reach USD 28 trillion, whereas lost assets as a result of gender inequality are anticipated to total USD 160.2 trillion (The Economic Times, 2021). Moreover, Bertay et al. (2020), a study that investigates the relationship between gender equality with economic growth reported that greater gender parity may promote economic expansion. The study by Akbulaev et al. (2020) also identified gender equality as the main factor that affects economic growth. The theoretical literature survey of Silva et al. (2021) pointed out that a huge percentage of theories under investigation contend that gender inequality hinders economic growth, especially in the long run. In this regard, the historical study demonstrates that positive results in the global economic expansion are linked to the economic progress of women and the removal of gender inequalities.

Gender equality has been a long-term-term objective for women's empowerment advocates. Nevertheless, gender disparities and legal restrictions remain in place today, and additional laws that bias against men and women in the workplace still exist (Christopherson et al., 2022; Wadhwa & Halim, 2020). Gender inequality is when one party or gender is frequently given more priority or advantage than the other due to prejudice based on gender. Discrimination between men and women was clearly visible just a quarter century ago. Nearly worldwide, women make less money and are less financially efficient than men. And compared to men, women have fewer opportunities to direct their life and make choices. Countries that forbid women from holding assets, become entrenched in laws and practices that refuse to uphold women's rights. (UNICEF, n.d.; Save the children, n.d.). Gender bias in the working place can take many different forms, such as disparities in pay, opportunities for promotion, incidents of sexual assault, and discrimination. It often shows up in subtle forms, like fewer opportunities for mothers and an increase in female stress (Woll, 2021).

The McKinsey Global Institute pointed out that legal protection is one of the most crucial factors that can be used to speed up gender equality. The majority of nations have laws that make it more difficult for females to find employment. For example, more than 100 nations have at least one female job limitation, while 59 have no lawful safeguards against workforce sexual harassment and 18 demand a husband's consent before a woman can be employed outside the home (Growing Economies Through Gender Parity, n.d.). The study of Bertay et al. (2020) which investigates the relationship between economic growth and gender equality also brings empirical evidence in defence of the claim that women's rights are fundamentally important and demand top priority on the agendas of policymakers. In addition to being a matter of human rights, equitable, and social equality, laws intended to guarantee a level playing ground for women, such as enhancing the law's rule and women's legal protections, in general, and especially, women's health, access to schooling, banking sectors, and technology, are also related policy enablers to enhance financial expansion.

In the past, there are many parties are putting effort to achieve economic development and remove gender inequality. In the 19th century, the battle against prejudice against women began. The American Congress was convinced to enact a statute ensuring equal wages for women working as federal employees in 1872 by lawyer Belva Ann Lockwood. A little over a century later, in 1963, the Equal Pay Act was established, declaring it illegal to pay men and women different rates in any industry. Women were given the same rights under the Civil Rights Act of 1964, which was revised in 1991 to enable sexual harassment complaints to be made by women against employers (Wooll, 2021). In the 20th century, the United Nations adopted 17 Sustainable Development Goals (SDGs) in 2015 with the aim of advancing the fight against all forms of poverty (BRD, n.d.). The goals established are unique, they call on all countries to contribute to fostering prosperity while protecting the environment in order to ensure that everyone experiences prosperity and peace by 2030 (Gavi, 2020).

Under the SDGs, gender equality and economic growth are included as the fifth and eighth targets, respectively (The Global Goals, n.d.; Sustainable Development Goals, n.d.). In 2021, UNICEF also introduced a new gender strategy (2021-2030) outlining goals for achieving gender equity in all of their programs, workspaces, and practices. The Gender Action Plan integrates a new plan for gender parity throughout internal laws, practices, and accountable measures, taking into consideration both operational and organizational objectives. One of their goals is to improve laws and procedures to promote workplaces and activities that are more gender-transformative (unicef, n.d.). Although the federal statute prohibits it, prejudice and gender disparity find sneaky ways to enter the workplace. Despite some advances, there is still a gender imbalance in society today (Wooll, 2021).

2.2 Related Work

The aim of this project is to develop an effective supervised learning approach to identify the income categories of a country based on legal gender equality. In the following section, related work that is related to economic development and current approach that relate to gender bias are reviewed under section 2.2.1 and section 2.2.2. Furthermore, the supervised learning algorithm that was used to predict income categories is reviewed to find out the suitable algorithm that can be utilized to build the prediction models.

2.2.1 Economic Development

When key ideas like recurrent neural network (RNNs), restricted Boltzmann machines (RBMs) and backpropagation were found in the 1980s, and fields like computer vision drew a lot of interest, supervied learning themes began to emerge in the economic literature on a bigger level. Supervised learning remains an applicable statistical technique now, however, because of its distinct methodology, it frequently surpasses traditional statistical methods. It rules the area in regard to prediction accuracy, particularly in forecasting (Korab, 2021). PWC UK, one of the top consulting firms in the globe, is one of the businesses that conduct research on how machine learning (ML) and artificial intelligence(AI) services could affect the economy. According to PWC, AI and ML could significantly assist in economic development in three key sectors, including increased efficiency, improved products, and the creation of new businesses (Addepto, 2019).

There are different studies using different indicators to predict economic development through a supervised learning model. For example, the study by Mele & Magazzino, (2020) used the Long Short-Term model to relate steel and iron production to economic growth in China. Besides, the study by Dutt & Tsetlin, (2020) used high-dimensional sparse (HDS) regression models and post-LASSO models to relate the poverty measure with economic development. Furthermore, the study of Storm et al. (2019) used a supervised learning approach such as neural network, decision tree and shrinkage methods to relate the agricultural sectors with economic growth. Moreover, the study by Ozden & Guleryuz, (2022) used Bayesian Tuned Gaussian Process Regression (BT-GPR) and Bayesian Tuned Support Vector Machine to develop economic development prediction model by using human capital as the indicator. Besides that, the study by Cogoljevic et al. (2018) analyzed the economic development prediction by using the energy resource indicator. In this study, the authors used energy resources mix as an indicator to predict the gross domestic product through the artificial neural network.

In the study by Farjallah & Sghaier, (2022), three types of supervised learning algorithms are used to predict economic growth by using the quality of governance as the indicator in the MENA region. The supervised learning algorithms that were used in this study are random forest, linear regression, boosted tree and support vector machine. Among these supervised learning algorithms, random forest is the most reliable model that can be used to predict economic growth based on the quality of governance. Besides, the study by Matsumoto & Samonte, (2022) also used macroeconomic indicators to predict the GDP development of the Philippines through gradient boosting, random forest and deep learning approaches. Moreover, the study by Cao et al. (2022) used multi-source open geospatial data as an indicator to build a multi-view graph neural network (MVGNN) model for economic development prediction in China.

The study by Hossain et al. (2021) investigates the GDP economic growth in Bangladesh by using supervised learning techniques through parameters such as unemployed rate, remittance, total investment, government debt and inflation rate. The aim of this study is to assist everyone in connecting to the economics area and assisting economists in proving their economic predictions. The supervised learning techniques they used included linear regression, random forest regressor, gradient boosting regressor, k-neighbor regressor, adaboost regressor, lasso regressor, and decision tree regressor. Their result showed that random forest regressor is the best model used to predict GDP economic growth. Another study by Richardson et al. (2018) used a supervised learning model compared with traditional multivariate and univariate statistical analysis to predict the real-time GDP economic data in New Zealand. The result reported that the supervised learning model has higher accuracy and higher performance than the traditional statistic method.

The study by Barhoumi et al. (2020) used supervised learning approaches to predict the GDP economic growth in Sub-Saharan Africa based on macroeconomic indicators. The supervised learning algorithms that were used are random walk, AR(1) processes, OLS regression, support vector machine, random forest and elastic net. Among these algorithms, the support vector machine, random forest and elastic net give high performance against other algorithms. Moreover, the study by Yoon, (2021) forecast the economic GDP growth in Japan by using gradient boosting and a random forest approach. This study compared their result with the prediction from the Bank of Japan and the International Monetary Fund. The result shows that the supervised learning approach outperformed the benchmark forecast. Between the two machine learning algorithms, gradient boosting has higher accuracy than random forest in this study.

Based on the above literature review, brings out information that there are many supervised learning approaches are built based on different indicators such as governance quality, energy resource, macroeconomic, geospatial data and poverty measures. However, these indicators are not the key indicators that can affect economic development. Based on the literature review in the background study which is section 2.1, there are many studies and trusted organizations that emphasized the achievement of gender equality and women’s financial empowerment as the primary and main factors that should be focused on economic development. Since the independent factors of existing supervised learning studies are not the main barrier to growth in the economy, this literature analysis uncovered the absence of supervised learning strategies that utilize the significant variables for predictions.

Furthermore, based on the literature review in this section, most studies are only focused on certain areas or regions only such as the studies of Mele & Magazzino, (2020) and Cao et al. (2022) which only focus on the economic development of China, Farjallah & Sghaier, (2022) only focus on MENA region, Hossain et al. (2021) only focus on Bangladesh, Richardson et al. (2018) only focus on New Zealand, Matsumoto & Samonte, (2022) only focus on the Philippines, Yoon, (2021) only focus on Japan and Barhoumi et al. (2020) which only focus on Sub-Saharan Africa. Hence, it can be proved that there is a lack of supervised learning approaches that can be applied across national borders because the majority of these approaches come from research that places a greater emphasis on the economic development of particular nations or areas, making them inapplicable to other nations. In the nutshell, there is a need for a supervised learning model that used key indicators for economic development prediction while the model is applicable to all countries worldwide instead of being limited to certain nations or areas only.

2.2.2 Gender Equality

In 90% of nations, there is at least one legislative restriction on women's ability to own assets, receive inheritances, or create bank accounts. The global workforce engagement of women in the labor force is 49 %, which is 27 % less than the proportion for males (Malaza & Parekh, 2020). However, even technology nowadays especially artificial intelligence or machine learning also found to have gender bias issues. The issue of model bias is becoming more and more obvious as artificial intelligence or machine learning gets more extensively used. Artificial intelligence has a growing impact on people's attitudes and actions in daily life. Nevertheless, years of progress toward gender equality may be silently undone due to the excessive representation of men in the development of these innovations. The constructivist approach was created by humans over years to guide judgments rather than relying exclusively on individual perspectives. However, supervised learning largely picks up new information by analyzing the data that is supplied to it. While a machine's capacity to handle massive amounts of data may partially alleviate this, if that data is laced with traditional gender conceptions, the technology's eventual application will continue to reinforce this bias (Leavy, 2018).

Bias errors occur because the AI's information will be incomplete if there are insufficient contributions from women. Naturally, people are in charge of machine learning, and the human bias is transmitted to AI, so the artificial intelligence machine will eventually contain human bias (International Women's Day, n.d.). Many organizations base their decisions on artificial intelligence (AI) systems that use machine learning (ML), in which a number of algorithms analyze vast amounts of data and train from it in order to detect the pattern and anticipate future events. These algorithms help determine how much credit financial institutions extend to certain clients, to who the medical sector gives COVID-19 vaccines first, and which job applicants are called in for interviews. However, gender prejudice is ubiquitous in these systems and has a significant negative influence on women's emotional, financial, and physical security in the long- and short-term. Additionally, it might accentuate and perpetuate undesirable gender prejudices and stereotypes that already exist (Smith & Rustagi, 2021).

Although the gender difference in data is not necessarily lethal, the development and application of artificial intelligence models in several sectors can seriously harm the well-being of women. Clinical studies that could indeed exclude sample sizes of females such as pregnant females, women in menopause, or women using birth control pills may produce health advice that is not always appropriate for the women's body, contributing to gender disparity in health that is not solely premised on biological and socioeconomic variations between women (Niethammer, 2020). The study by Chung et al. (2021) took into account an AI model for the prejudice analysis that makes preliminary severity predictions based on the medical records of COVID-19 patients. One model was developed using only the male community, and the second model was developed using only the female group. The authors then tested both models. The findings demonstrated that, in comparison to the unbiased model, the gender-dependent AI model offered lesser accuracy. Moreover, even using test data from the same gender used for training, the authors discovered that the accuracy of the bias model was also worse than that of the unbiased model.

Joy Buolamwini and Timnit Gebru have looked into the prejudices of AI face detection programmes in their book Gender Shades. Their research demonstrates that there can be overt gender discrimination in some implementations of AI facial recognition software. The incorrect or insufficient data sets used to train the systems are one cause of this (ARS Electronica, 2022). Besides, on social networking sites like Facebook and Twitter as well as online job boards like Indeed and LinkedIn, employees are rapidly discovering new chances. These platforms use algorithms to determine which job openings users see and how suitable they think they are for a given position. According to studies, however, individuals who choose "female" as their gender see fewer employment adverts for high-paying positions than those who choose "male." The paper claims that the algorithms can discriminate against females even if they have the same skills and experience as men and even if they are more experienced. Due to the fact that men were more frequently looking for new employment options, LinkedIn found that males were more frequently shown positions available than females (Huet, 2022).

Moreover, gender bias in search algorithms nowadays also has an effect on users. According to a recent study by a group of social psychologists, gender-neutral websites still give outcomes that are overwhelmingly male. Additionally, by encouraging bias against women and even affecting employment judgments, these search terms have an impact on visitors. Madalina Vlasceanu, a postdoctoral researcher in the Department of Psychology at New York University, claims there is growing worried that the algorithms utilized by contemporary artificial intelligence technologies generate biased results since they are trained on data that contains social prejudices. The researchers carried out research to ascertain whether prejudices in algorithm results are related to the level of inequalities in a community and, if so, whether access to such production could lead human decision-makers to act according to these prejudices. They investigated whether terms like "person," "student," or "human" should correspond to a male or a female with equal probability. However, the results show these terms are more frequently thought to be male. Here, they carried out Google image searches for "person" within each of the 37 nations. The findings demonstrated that computational sex discrimination correlates with social gender disparity and that the percentage of male images produced by these searches was higher in countries with larger gender discrimination. 3 months later, the research was performed by the researchers with a group of 52 countries, 31 of which were included in the initial study. The findings supported the initial findings of the study that gender inequities at the social level are mirrored in model output (NYU, 2022).

Furthermore, the study of Buolamwini, (2019), research revealed significant racial and gender prejudice in AI systems offered by well-known tech companies like Amazon and Microsoft. All companies fared noticeably better on man faces than on woman faces when asked to identify the gender of a face. For guys with a lighter complexion, the mistake percentages at the businesses the author studied were no higher than 1%. The mistakes increased to 35% for women with darker skin. The gender prejudices in the recruitment algorithm created by Amazon to select resumes for application forms were evident by devaluing resumes that included the term "female". According to studies, there is a direct correlation between the absence of females in development companies and the creation of AI systems that exhibit gender prejudices (Adams, 2019).

Equality is becoming increasingly important in machine learning generally, and it is crucial that female remain at the center of those who define the term. In addition to being a fundamental right, promoting women 's employment in the field of machine learning is also crucial to preserving the gains in women's rights that have been made attributable to years of feminist ideas (Leavy, 2018). One of the most obvious examples of gender inequality is the unfair mistreatment of females in the legal system. Legal restrictions that are frequently based on gender inhibit women and, by extension, economies from attaining their full capacity. As a result, certain nations have fallen behind in achievements in eliminating the legislation that serves as an obstacle to female empowerment participation. Additionally, cross-country disparities in gender-biased legislation have remained and even expanded over time. This research indicates that more women's rights in the legislation improve cross-country economic convergence over time, based on a worldwide dataset since the 1970s. The findings push for change and give cause for optimism for the future. Researchers suggest that legislation changes in favor of women's rights, which may be feasible in the near term, aid in bringing life quality in developing nations up to par with those in developed countries (Sever, 2022). Hence, there is a need to developed an automation machine which do not have gender bias and able to tackle the gender inequality problem in social.

2.2.3 Income Level Classification

To identify trends in huge data that result in usable findings, data scientists employ a wide range of machine learning methods. On a broad scale, these various algorithms can be divided into supervised and unsupervised learning categories based on how they "train" data to produce forecasting results. Supervised learning is the main method that is used in this capstone project. The input used to educate the system must already have the right results marked in order for supervised learning to work (GreeksforGreeks, 2022). The target variable of the dataset in this project is the income group which is a categorical variable. In this section, the objective of the literature review here is to find out the popular supervised learning algorithm that is used in income or financial status prediction. The most commonly used supervised learning algorithms are filtered out at the end of this section and planned into this capstone project model building.

The study by Singh, (2017) used the US census data to build a prediction model for individual income levels. In this case study, the dataset consists of 14 independent variables and 1 dependent variable. The dependent variable is ‘incomelevel’ which represents the level or group of income that is divided into 4 categories which are “greater than USD 50000”, “>50K”, “less than or equal to USD 50000” and “<=50K”. The author carries out exploratory data analysis for continuous and categorical variables respectively, cleans the data by carrying out missing value imputation, data normalization, and then builds the predictive model by using boosting algorithm. After, the model building, the author validates the result by accuracy, sensitivity and specificity parameters. The result shows that this machine learning algorithm brings good performance with 86% accuracy, 88% sensitivity and 78% specificity.

The study by Bekena & Menji, (2017) used a dataset obtained from the UCSI Machine Learning Repository which consists of 32561 observations and 13 attributes to predict the individual income level for 42 countries. There are a few supervised learning models are built in this study which included random forest, gaussian naïve bayes and decision tree classifier. In this study, the researcher carries out exploratory data analysis, data encoding, feature engineering and missing value for data preparation. The result of the predictive model shows that the random forest classifier and decision tree give the highest performance by bringing an accuracy of around 85% while the gaussian naïve bayes give 78% accuracy. However, the random forest classifier is chosen as the final best model even though it has the same accuracy as the decision tree. This is because the random forest classifier can better avoid overfitting problems than the decision tree. The study of Bramesh & Puttaswamy, (2019) also uses the same dataset to build the prediction model. In this study, the authors used gaussian naïve bayes, support vector machine, random forest, decision tree and gradient boosting algorithms to build the model. The results show that gradient boosting classifier outperformed other classifiers.

Another study by Chakrabarty & Biswas, (2018) used the dataset obtained from the University of California Irvine Machine Learning Repository which consists of 48,842 observations and 14 attributes for 42 nations. This study attempts to demonstrate how machine learning and data mining methods can be used to address the issue of income inequality. Based on a specific set of characteristics, classification has indeed been conducted to identify if an individual's annual earnings in the United States belong in the income category of more than 50,000 Dollars or less equal to 50,000 Dollars. Before model building, feature selection, data visualization, missing value imputation, data encoding, shuffling and splitting into 80% training and 20% testing set are done for data preparation. The predictive model built in this study is gradient boosting classifiers and ensemble learning. Hyperparameter tuning is carried out through Grid Search Algorithm to find out the best parameter. The parameter that was used to check results included accuracy, F1-Score, recall, precision and Receiver Operator Characteristic Curve (ROC Curve). The study conclude that gradient boosting classifier has higher performance.

The study of Srivastava et al. (2020) used a dataset with 31978 observations and 13 variables to build random forest, random tree, naïve bayes, and REPTree model that predicts the income status of individuals. Before model building, the researcher explores and cleans the data by using WEKA software. The result of this study shows that REPTree is the most accurate model in this study. The study of Khanna, (2018) used the census dataset obtained from the UCI repository adult dataset which consists of 14 attributes and 48842 observations. It is a comparative study that uses neural networks, support vector machines, random forests, naïve bayes and decision trees to build the models for individual income level prediction. Before building the model, the author carries out missing value imputation, removing redundant values, data encoding, and splitting into training and testing sets. The parameters that were used for result evaluation included the ROC curve, confusion matrix, F1-score, accuracy, precision, recall and Gini coefficient. The results of this study show that neural networks outperformed other supervised learning models with 0.77 AUC scores and 86.3% accuracy.

The study by Voleti & Jana, (2022) predicts employee income levels by using 50,000 observations and 15 attribute datasets obtained from Kaggle to build logistic regression and support vector machine models. The target variable of this dataset is the income level which is divided into 2 categories which are income above 50,000 USD and below 50,000 USD. First, the author carries out exploratory data analysis through data visualization techniques such as charts, graphs and histograms etc. Then, the pre-processing process that was done in this study includes feature engineering, missing value imputation, remove null or duplicate data. The performance measures are F1 score, accuracy, recall score and precision. The result of this study shows that the support vector machine has a higher performance than logistic regression with 89% accuracy. Based on the project of Nauli, (2021) that used adult census income data that is obtained from Kaggle to build a support vector machine, random forest, decision tree and naïve bayes models for the. Hyperparameter tuning by using Grid Search is also done in this project. The results of the models are compared to find out the best model. In this project, the result shows that random forest is the best model that has 81.58% accuracy and 88.26% sensitivity.

The study by Temraz, (2019) used a dataset with 48,842 observations and 15 variables to classify the annual income of individuals by using decision tree, random forests and artificial neural network algorithms. There are 27 models are built in this study with different parameter settings. In this study, the author carries out exploratory data analysis to understand the dataset and explore the missing value in the dataset. The pre-processing step that the author carried out in this study is data transformation, missing value imputation, feature engineering by selecting the important variable and resampling. In terms of sensitivity and specificity, the decision tree classifier scored the highest overall. As opposed to this, the random forest classifier received the highest ratings for precision and accuracy.

Based on the above literature review, most of the studies are investigates the income level in a different field, this is because there is a lack of income level classification studies in terms of countries by using machine learning approaches. Based on the review, we found that most of the studies used data visualization such as bar chart, histogram, plot to understand the data. Besides, authors of studies carry out data pre-processing such as imputing missing value, data encoding, dropping duplicate or null data and feature engineering for data preparation. Moreover, there are a few studies that also carry out hyperparameter tuning by using the Grid Search method. Further, the literature review shows that random forest, support vector machine, artificial neural network, and gradient boosting are the common machine learning algorithms that have good performance in predicting income levels. Hence, in this capstone project, the machine learning models that are used for country's income levels prediction are random forest, support vector machine, artificial neural network and gradient boosting models.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Flow Chart

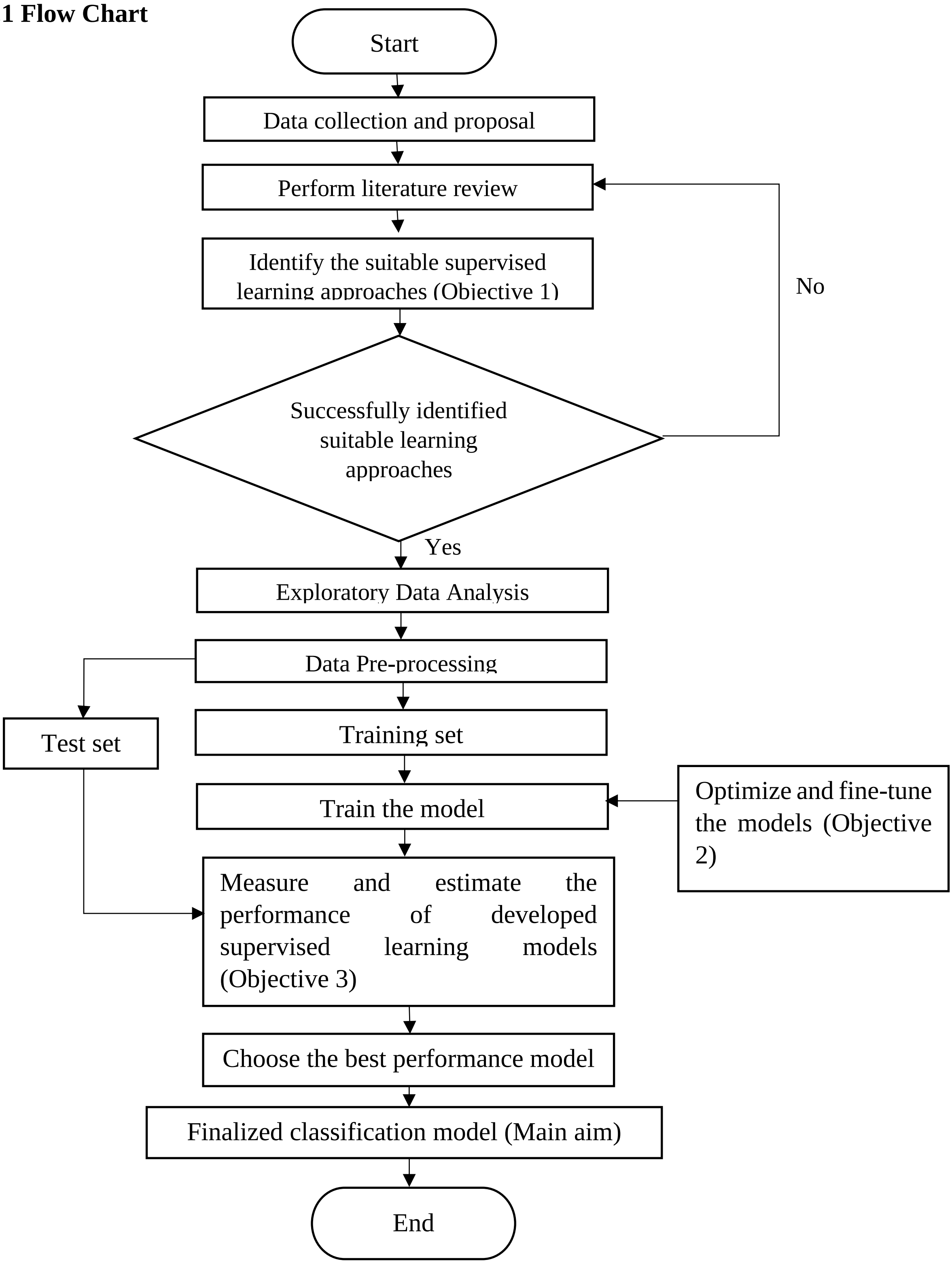

Start
Data collection and proposal
Perform literature review
Identify the suitable supervised learning approaches (Objective 1)
Successfully identified suitable learning approaches
No
Yes
Exploratory Data Analysis
Data Pre-processing
Test set
Training set
Train the model
Optimize and fine-tune the models (Objective 2)
Measure and estimate the performance of developed supervised learning models (Objective 3)
Choose the best performance model
Finalized classification model (Main aim)
End

Figure 3.1 Methodology Flow Chart

3.2 Methodology Background

A subfield of artificial intelligence (AI) and computer science called machine learning concentrates on using algorithms and data to simulate how people learned, gradually building up the accuracy of the system. Algorithms are developed using analytical techniques to produce classification or prediction and to find important discoveries in data mining tasks. The decisions made as a result of these findings ideally influence critical development indicators in programs and enterprises (IBM, 2020). It is suggested that supervised machine learning, a prominent categorization approach, be employed in this capstone project to categorize a country's income level depending on its legal gender equality. In this approach, input data is sent to the algorithms in order to obtain the best results. The process of learning continues until the highest level of effectiveness is reached (ODSC, 2020). In this section, the details of the methodology procedure are explained from data collection until model validation and comparison.

3.3 Methodology Procedure

3.3.1 Data Collection

The data is collected through the internet which is obtained from The World Bank Data Catalog, a platform that is intended to enable the World Bank development statistics simpler to access, acquire, analyze, and distribute. The dataset information is collected by a program under the World Bank Group which is the Women, Business and Law that intends to collect precise data on the legal restrictions on females' economic opportunities. The data is an excel file of 2.35 MB. In this dataset, there are a total of 8 indicators used as the independent variable, in each indicator, there are four to five questions concerning the existence or absence of regulations and legislation promoting socioeconomic gender equality. The indicator included: mobility, workplace, pay, marriage, parenthood, entrepreneurship, assets, and pension. For example, under the mobility indicator, questions such as "Can a woman apply for a passport in the same way as a man?" served as an attribute, and answer such as "Yes" or "No" are placed under the column. These indicators are related to the laws that can affect the economic opportunities of women. The dependent variable is the income group, which is divided into low, low-medium, upper-medium and high-income groups. This income group indicates countries' income levels.

3.3.2 Literature Review

A literature review is a study of academic resources on a certain topic. It provides an overview of current knowledge, related theories, methodology, and restrictions in existing investigations (McCombes, 2019). The next stage after the proposal was approved was to conduct a review of the literature about supervised machine learning methods for categorizing income levels. In this section, the literature review study different article, website and journal to give more understanding about the research background and justified the problem statement. Under supervised machine learning, there are many different types of algorithms, including support vector machines, logistic regression, random forests, gradient boosting, artificial neural networks, and others. Therefore, by examining the latest and existing works in the same field of classification, the literature review can help in screening the appropriate machine learning algorithm. The accomplishment of objective 1 in this project will be demonstrated by the effective identification of the supervised machine learning algorithms that will be used in this capstone project through the literature review. In this capstone project, there are few machine learning algorithms are identified, which are support vector machine, artificial neural network, random forest and gradient boosting.

3.3.3 Exploratory Data Analysis

First, exploration of the dataset is stated by using exploratory data analysis after the suitable supervised machine learning algorithms are identified. Exploratory data analysis is an important process that is used to do preliminary analyses of data in order to visualize the data, verify assumptions using statistical results, test hypotheses, identify patterns, and detect anomalies (Patil, 2018). This technique is generally used to identify animalities in the dataset and gain a deeper knowledge of the features in the dataset and their correlation (IBM, 2020). In this step, the dataset is explored to view the data types, count the number of observations and attributes, check the missing values, summary statistic, explore the dependent and independent variables, data balance conditions and identify outliers.

3.3.4 Data pre-processing

By preparing the original data for use in a machine learning algorithm, data pre-processing is a technique that can increase the model's precision and effectiveness. Original data often includes noise, is sparse, or has an unsuitable format that prevents it from being directly used in machine learning, therefore pre-processing is required to clean the data (GreeksforGreeks, 2021). Data pre-processing operations such as missing value imputation, data transformation, data encoding, feature engineering, feature selection, and splitting of the dataset into the training set and testing set for model building are performed in this step. The dataset eventually qualifies for model building and can be used with machine learning techniques.

3.3.5 Model Training

The pre-processed data is divided into a training set and a test set after processing. In the splitting, 30% went to the test set and 70% to the training set. Due to the necessity for enough data to train the machine learning model, the fraction of the training set is higher than the test set. The machine learning algorithms that were found in objective 1 are then given the training set so that they can be trained. Under the literature review, there are 4 machine learning algorithms identified which included support vector machine, random forest, artificial neural network and gradient boosting. The target variable, which is the income group, is present in the training set. The dataset's pattern is discovered by the machine learning algorithms, which also learn from the dataset and link the input data to the target variable. Then, 4 kinds of machine learning models are generated.

3.3.5.1 Support Vector Machine

A supervised machine learning model called a support vector machine (SVM) employs classification techniques to solve group classification tasks. An SVM algorithm can classify new text after being given sets of labelled training data for each class (Ray, 2021). Finding a hyperplane in an N-dimensional space that categorizes the data points clearly is the goal of the support vector machine algorithm. Data points that are nearer the hyperplane are known as support vectors, which have an impact on the hyperplane's orientation and position. Margins of the classifiers can be increased by using these support vectors (Gandhi, 2018). Figure 3.2 shows a sample of the support vector machines classifier from the study by Fernandes et al. (2020).

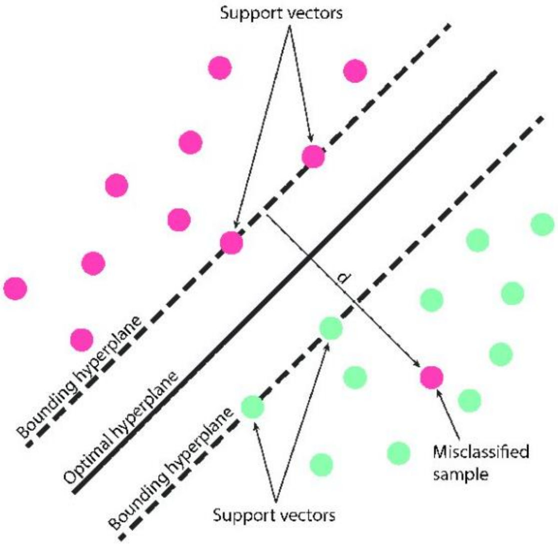


Figure 3.2 Support Vector Machine Classifier Sample

3.3.5.2 Random Forest

Leo Breiman and Adele Cutler are the creators of the widely used machine learning technique known as random forest, which mixes the result of various decision trees to produce a unique outcome. Its widespread use is motivated by its adaptability and usability because it can solve regression and classification issues. There are 3 key hyperparameters for random forest approaches that must be adjusted prior to training which are sampled feature count, tree count and node size (IBM, 2019). Every tree in the random forest gives out a classification forecast, and the classification that receives the highest voting represents the prediction made by the random forest model (Yiu, 2019). Figure 3.3 shows the sample of the random forest classifier.

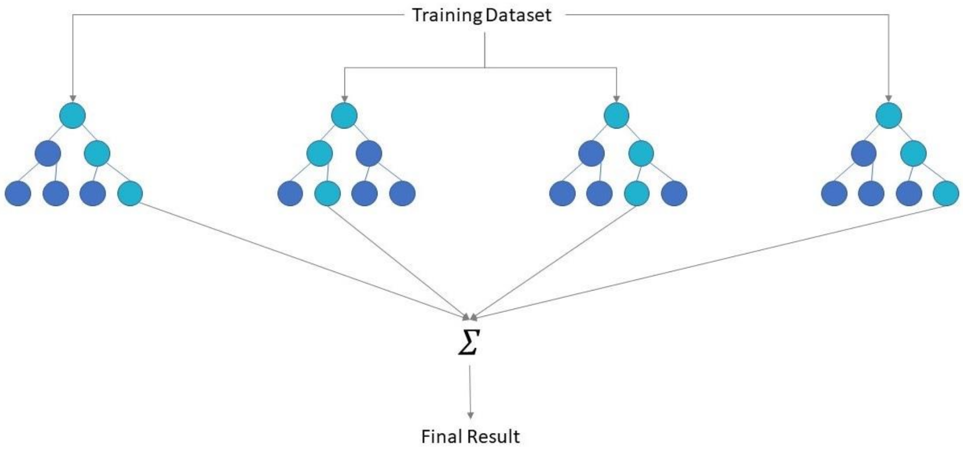


Figure 3.3 Random Forest Classifier Sample

3.3.5.3 Artificial Neural Network

They mirror the mechanism that natural neurons communicate with one another by taking their origin and architecture from the human brain. A node layer, which includes an input layer, one or more hidden layers, and an output layer, makes up artificial neural networks (ANNs). Each node, or artificial neuron, is interconnected and has a threshold and weight that go with it. A node gets activated and begins transferring data to the network's next layer if its output is greater than the predefined threshold value for that node. Instead, no data is transmitted to the network's next layer. Training data is essential for neural networks to develop and enhance their efficiency over time. Nevertheless, these algorithms become effective methods in artificial intelligence and computer science once they are adjusted for accuracy, enabling users to quickly categorize and group data (IBM, 2020). Figure 3.4 shows the sample of artificial neural network classifier.

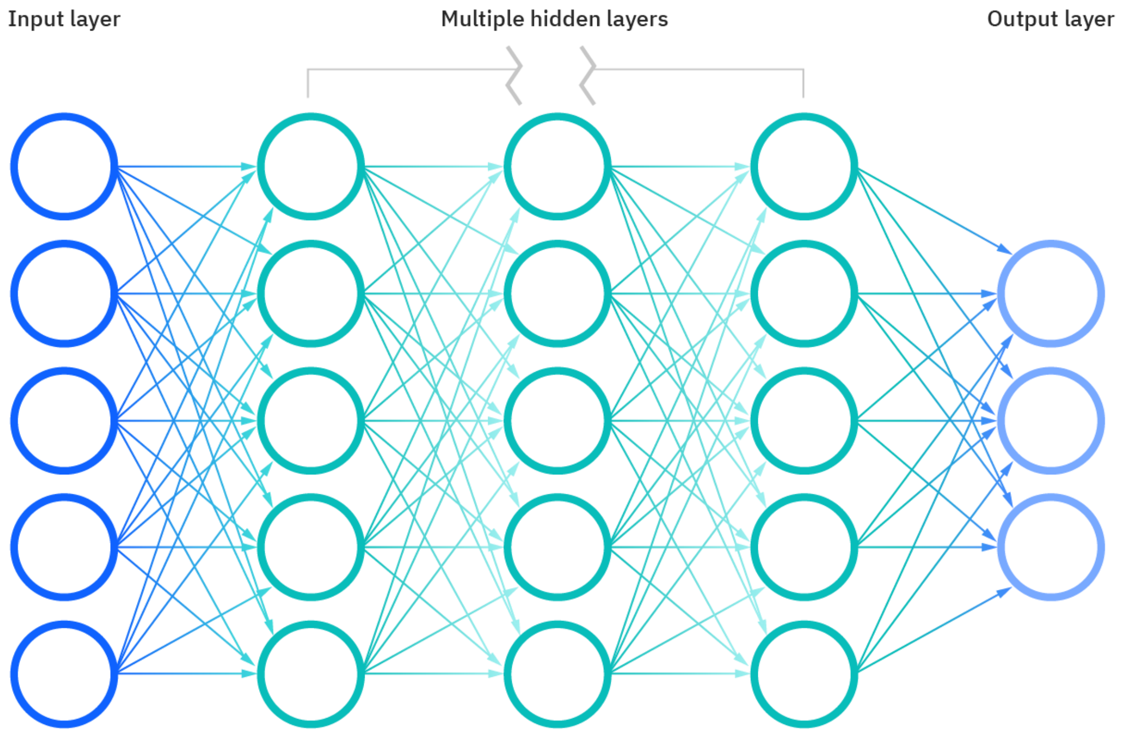


Figure 3.4 Artificial Neural Network Classifier Sample

3.3.5.4 Gradient Boosting

A decision tree for extensive and complicated data is represented by the machine learning boosting system known as gradient boosting. It is based on the assumption that, when merged with the prior set of models, the subsequent model will reduce the overall prediction error. This boosting technique consists of three key components: a loss function, a weak learner, and an additive model (Vaidya, n.d.). It operates under the premise that a predictor can be made more effective by combining several weak learners. Gradient boosting is a type of boosting strategy that builds a strong model by iteratively learning from each of the weak learners. The notion of boosting is based on the presumption that poor learners could be improved (Kurama, 2020).

3.3.6 Hyperparameter Tuning

There are a number of parameters in a machine learning model that are not learned from the training dataset. The model's performance is controlled by these variables. As part of the hyperparameter tuning phase, the parameters for each algorithm are optimized, and the hyperparameter space is searched for the best values to control how algorithms learn. Based on the literature review, most of the studies that carry out hyperparameter tuning are using the Grid Search method. Grid search is a tuning method that seeks to determine the ideal parameter settings. It is a thorough search that is done on a model's particular parameter values (Malik, 2020). Hence, this method is identified to be used in this capstone project for the model hyperparameter tuning part. To create an optimized machine learning model, the model is retrained after hyperparameter adjustment. Here, objective 2 of this capstone project is achieved.

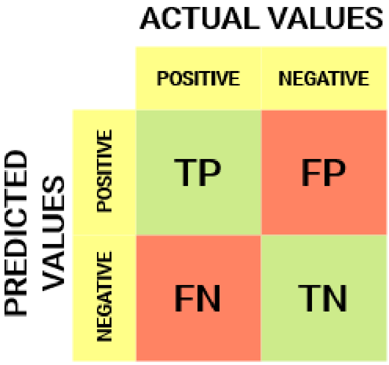
3.3.7 Model Performance Measure and Estimation

After training the model, the performance of the supervised machine learning model is evaluated utilizing the test set that was split during pre-processing. The area under the curve (AUC), receiver operating characteristic (ROC) curve, sensitivity, specificity, and accuracy are performance measures that are calculated throughout the model assessment process. In this case, objective 3 is accomplished. The result obtained to represent the dataset can then be found using this performance metric, as well as the future effectiveness of the chosen technique. The highest performance models can then be decided to serve as the model for categorizing the income level of the nation. Here, the capstone project's primary goal is accomplished.

3.3.7.1 Confusion Matrix

An N x N matrix called a confusion matrix is used to assess the effectiveness of a classification algorithm, where N is the total quantity of class labels. In the matrices, the real target variables are contrasted with those that the machine learning model anticipated. This gives a comprehensive understanding of the effectiveness of the classification model and the types of mistakes it is committing (Bhandari, 2022). Figure 3.5 below represents the example of a confusion matrix where:

- True positive (TP) - The model forecasted a positive result, and the real outcome was positive.
- True negative (TN) – The model forecasted a negative result, and the real outcome was negative.
- False positive (FP) – It is known as Type 1 error as the real outcome is negative but the model forecasted a positive result.
- False negative (FN) – It is known as Type 2 error as the real outcome is positive but the model forecasted a negative result.



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

Figure 3.5 Confusion Matrix

3.3.7.2 Accuracy

One approach to gauge how frequently a machine learning system labels a data source properly is to look at its accuracy. The percentage of accurately anticipated data points among all the data points is known as accuracy. Theoretically, it is calculated as the sum of true positives and true negatives divided by the sum of true positives, true negatives, false positives, and false negatives (Mishra, 2018). The formula for accuracy calculation is shown below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

3.3.7.3 Specificity

The parameter used to assess a model's capacity to forecast the true negatives of each given class is known as specificity in machine learning. This word is also known as a true negative rate in research, it is calculated as the true negatives divided by the sum of true negatives and false positives (deepchecks, n.d). The following formula can be used to determine it:

$$Specificity = \frac{TN}{TN + FP}$$

3.3.7.4 Sensitivity

The parameter used to assess a model's capacity to forecast the true positives of each given class is known as sensitivity in machine learning. This word is also known as a true positive rate in research, it is calculated as the true positive divided by the sum of true positives and false negatives (deepchecks, n.d). The following formula can be used to determine it:

$$Sensitivity = \frac{TP}{TP + FN}$$

3.3.7.5 Receiver Operating Characteristic (ROC)

The effectiveness of a classification model can be examined graphically using receiver operating characteristic (ROC) evaluation. The effectiveness of a classifier is assessed using two statistics: true positive rate and false positive rate. The false positive rate is drawn on the x-axis and the true positive rate is displayed on the y-axis in a two-dimensional graph of statistics. The generated graph can be utilized to assess whether a classifier outperforms random prediction and to evaluate the comparative performances of various models (Tan, n.d.). Figure 3.6 below shows an example of the ROC curve.

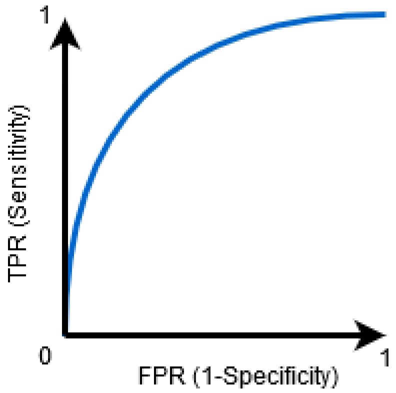


Figure 3.6 ROC Curve Sample

3.3.7.6 Area Under the Curve (AUC)

The capacity of a predictor to differentiate between categories is measured by the Area Under the Curve (AUC), which is used as a summation of the ROC curve. The model performs better at differentiating between the positive and negative classes, the higher the AUC value (Bhandari, 2022).

CHAPTER 4

IMPLEMENTATION

4.1 Data Collection

The data is collected online by The World Bank Data Catalog, which is a platform designed to make it easier to acquire, obtain, analyze, and share World Bank development statistics. Women, Business and Law under the World Bank Group, which aims to gather detailed data on the regulatory limitations on female empowerment possibilities, is responsible for compiling the dataset information. The dependent variable of the dataset is the ‘income group’ while the independent variables included 8 indicators. Under each indicator, there are almost 4 to 5 questions. In the session of 4.2, the variable description is explained in the table 4.1.

4.2 Description of the Dataset

Table 4.1: Description of the Dataset

Variable	Variable Description	Example
ID	The country ID code	AFG2021, AGO2021,…..
economy	The country	Afghanistan, Angola, …
wbcodev2	Shortform of country	AFG, AGO, ….
Region	Region	South Asia, Sub-Saharan Africa, ….
Income group	The income category of the country	Low income, high income, ….
reportyr	The report year for the income group and gender equality implementation status	2011, 2012,…
WBL index	This indicator draws on all eight areas of the Women Business and the Law (WBL)report including:Mobility,	20.25, 20.25, ….

	Workplace, Pay, Parenthood, Marriage, Entrepreneurship, Assets and Pensions.	
Mobility	Indicator that represents the women’s mobility ability in that country	100, 75, ...
Can a woman apply for a passport in the same way as a man?	A question that asks for women’s ability to apply for a passport in the same way as man	Yes, No..
Can a woman travel outside the country in the same way as a man?	A question that asks for women’s ability to travel outside the country in the same way as a man	Yes, No
Can a woman travel outside her home in the same way as a man?	A question that asks for women’s ability to travel outside her home in the same way as a man	Yes, No,…
Can a woman choose where to live in the same way as a man?	A question that asks for women’s ability to live in the same way as man	Yes, No, …
Workplace	Indicator that represents the women’s freedom in workplace	100, 75…
Can a woman get a job in the same way as a man?	A question that asks for women’s ability to get a job in the same way as man	Yes, No,…
Does the law prohibit discrimination in employment based on gender?	A question that asks for the ability of country to prohibit discrimination in employment based on gender	Yes, No,…
Is there legislation on sexual harassment in employment?	A question that asks for the ability of country to	Yes, No,…

	implement legislation on sexual harassment in employment	
Are there criminal penalties or civil remedies for sexual harassment in employment?	A question that asks for the ability of the country to implement criminal penalties or civil remedies for sexual harassment in employment	Yes, No,…
Pay	An indicator that represents the women's pay ability in the country	100,75,….
Does the law mandate equal remuneration for work of equal value?	A question that asks for the country to implement a law mandate equal remuneration for work of equal value	Yes, No…
Can a woman work at night in the same way as a man?	A question that asks about the ability of women to work at night in the same way as a man.	Yes,No,….
Can a woman work in a job deemed dangerous in the same way as a man?	This a question that asks about the ability of women to work in a job deemed dangerous in the same ways as a man.	Yes, No….
Can a woman work in an industrial job in the same way as a man?	A question that asks about the woman ability to work in an industrial job in the same way as a man	Yes, No….
Marriage	An indicator that represents women's freedom after marriage in the country	100, 75, …
Is there no legal provision that requires a married woman to obey her husband?	A question that asks about the present of legal provision that	Yes, No,…

	requires a married woman to obey her husband	
Can a woman be head of household in the same way as a man?	A question asks about the women’s ability to head of household in the same way as a man	Yes, No,…
Is there legislation specifically addressing domestic violence?	A question asks about the ability of the country to implement legislation specifically addressing domestic violence.	Yes, No,…
Can a woman obtain a judgment of divorce in the same way as a man?	A question asks about the ability of women to obtain a judgment of divorce in the same way as a man.	Yes, No,…
Does a woman have the same rights to remarry as a man?	A question asks about the ability of women to have the same rights to remarry as a man.	Yes, No,…
Parenthood	An indicator that presents the treatment for parents	100, 75,..
Is paid leave of at least 14 weeks available to mothers?	A question asks about the presence of paid leave of at least 14 weeks available to mothers	Yes, No,…
Length of paid maternity leave	The length of paid maternity leaves in days	14, 79,…
Does the government administer 100% of maternity leave benefits?	A question asks about the government administer 100% of maternity leave benefits.	Yes, No,..
Is there paid leave available to fathers?	A question ask for the presence of paid leave available to fathers.	Yes, No,..

Length of paid paternity leave	The length of paid paternity leaves in days	84,21,..
Is there paid parental leave?	The presence of paid parental leave	Yes, no,..
Shared days	The number of days that both father and mother to cliam paid leave	83,29,..
Days for the mother	Parental days for the mother	43,18,…
Days for the father	The parental days for the father	34,23,…
Is dismissal of pregnant workers prohibited?	A question asks about the presence of dismissal of pregnant workers prohibited	Yes, no,…
Entrepreneurship	An indicator that represents the ability of women in entrepreneurship	100, 75,…
Can a woman sign a contract in the same way as a man?	A question asks about the ability of women sign a contract in the same way as a man.	Yes, No,..
Can a woman register a business in the same way as a man?	A question asks about the ability of a woman register a business in the same way as a man	Yes, No,…
Can a woman open a bank account in the same way as a man?	A question asks about the ability of a woman open a bank account in the same way as a man	Yes, No,…
Does the law prohibit discrimination in access to credit based on gender?	A question asks about the presence of the law prohibit discrimination in access to credit based on gender	Yes, No,…

Assets	An indicator that represent the assets of women	100,75,…
Do men and women have equal ownership rights to immovable property?	A question asks about the ability of women to have equal ownership rights to immovable property as men.	Yes, No,…
Do sons and daughters have equal rights to inherit assets from their parents?	A question asks about the ability of daughters to have an equal right to inherit assets from parents as sons.	Yes, No,.
Do male and female surviving spouses have equal rights to inherit assets?	A question asks about the ability of female surviving spouses to have equal rights to inherit assets as males	Yes, No,…
Does the law grant spouse equal administrative authority over assets during the marriage?	A question asks about the presence of a law granting spouse equal administrative authority over assets during the marriage	Yes, No,..
Does the law provide for the valuation of nonmonetary contributions?	A question asks about the presence of a law that provides for the valuation of nonmonetary contributions.	Yes, No…
Pension	An indicator that repression the pension status of women	100, 75,…
Is the age at which men and women can retire with full pension benefits the same?	A question asks about the ability of women to retire with full pension benefits same as the men.	Yes, No,..
Is the age at which men and women can retire with partial pension benefits the same?	A question asks about the ability of women to retire with partial pension benefits same as man	Yes, No,..

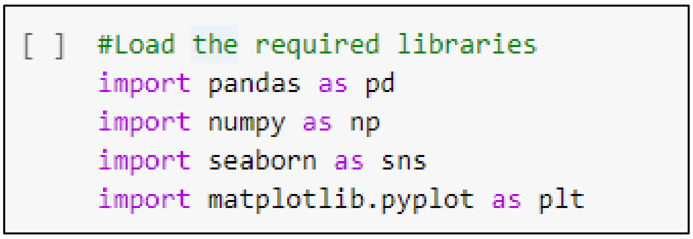
Is the mandatory retirement age for men and women the same?	A question asks about the mandatory retirement age for men and women having equal treatment.	Yes, No,….
Are periods of absence due to childcare accounted for in pension benefits?	A question asked about the availability of periods that absence due to childcare accounted for in pension benefits	Yes, No,…

4.3 Data Exploration

Exploratory data analysis is the critical stage of doing preliminary analyses of data in order to find trends, identify anomalies, test hypotheses, and double-check preconceptions with the aid of descriptive statistics and visualizations. Before exploratory data analysis, the needed libraries are imported. Next, exploratory data analysis is carried out to understand the dataset. The data structure, data types, number of attributes and observations, correlation, missing value and outlier detection are performed in this stage. The visualization techniques used in this stage included the bar chart, histograms, box plots and heatmap.

4.3.1 Import Libraries

Before importing the dataset, different kinds of libraries are imported for the following usage. The library importation code is shown below:



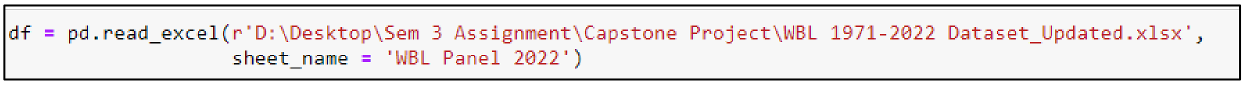
```
[ ]  #Load the required libraries
     import pandas as pd
     import numpy as np
     import seaborn as sns
     import matplotlib.pyplot as plt
```

Figure 4.1 Library Importation Code

The libraries that were imported include pandas, numpy, seaborn and matplotlib. For the purpose of manipulating and analyzing data, the Python programming language has a software package called pandas. It includes specific data structures and procedures for working with datasets and mathematical tables. Massive, multi-dimensional arrays and matrices are supported by NumPy, a toolkit for the Python language, together with a substantial number of high-level mathematical operations that may be performed on that array. One outstanding Python module for visualizing graphical quantitative graphing is Seaborn. To make the production of various statistical charts in Python more visually appealing, Seaborn offers a variety of color choices and elegant standard layouts. A more appealing visualization of the essential component of comprehending and analyzing data is what Seaborn Library tries to achieve. The Matplotlib package Pyplot offers a MATLAB-like user interface. Matplotlib is intended to be as user-friendly as MATLAB, with the added benefit of supporting Python and also being public and accessible.

4.3.2 Upload dataset

First, the dataset is uploaded into python environment by using the python code shown below:

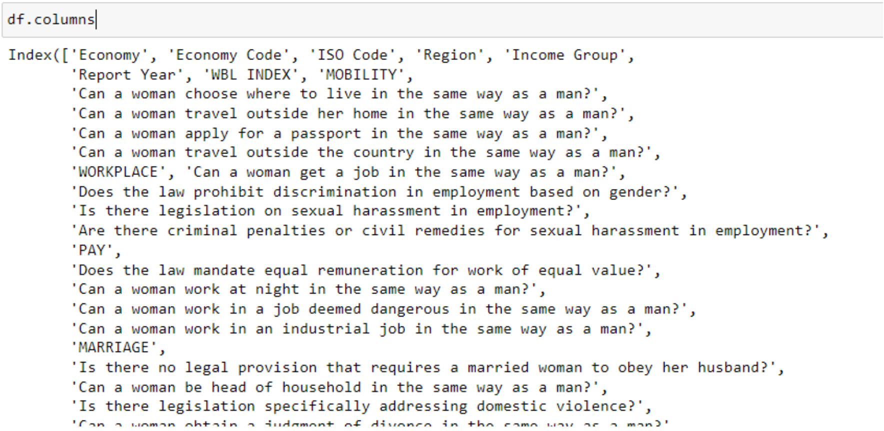


```
df = pd.read_excel(r'D:\Desktop\Sem 3 Assignment\Capstone Project\WBL 1971-2022 Dataset_Updated.xlsx',
                   sheet_name = 'WBL Panel 2022')
```

Figure 4.2 Upload Dataset Code

The excel dataset is uploaded by using the pandas library. The name of the excel file is ‘WBG.xlsx’ while the sheet that was selected is ‘1971-2021’ which contains 50 years of data.

4.3.3 Structure of the Dataset

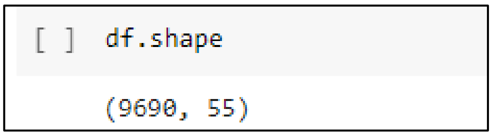


```
df.columns
```

```
Index(['Economy', 'Economy Code', 'ISO Code', 'Region', 'Income Group',
       'Report Year', 'WBL INDEX', 'MOBILITY',
       'Can a woman choose where to live in the same way as a man?',
       'Can a woman travel outside her home in the same way as a man?',
       'Can a woman apply for a passport in the same way as a man?',
       'Can a woman travel outside the country in the same way as a man?',
       'WORKPLACE', 'Can a woman get a job in the same way as a man?',
       'Does the law prohibit discrimination in employment based on gender?',
       'Is there legislation on sexual harassment in employment?',
       'Are there criminal penalties or civil remedies for sexual harassment in employment?',
       'PAY',
       'Does the law mandate equal remuneration for work of equal value?',
       'Can a woman work at night in the same way as a man?',
       'Can a woman work in a job deemed dangerous in the same way as a man?',
       'Can a woman work in an industrial job in the same way as a man?',
       'MARRIAGE',
       'Is there no legal provision that requires a married woman to obey her husband?',
       'Can a woman be head of household in the same way as a man?',
       'Is there legislation specifically addressing domestic violence?',
       'Can a woman obtain a judgment of divorce in the same way as a man?'
```

Figure 4.3 Code and Output for the Dataset Columns Exploration

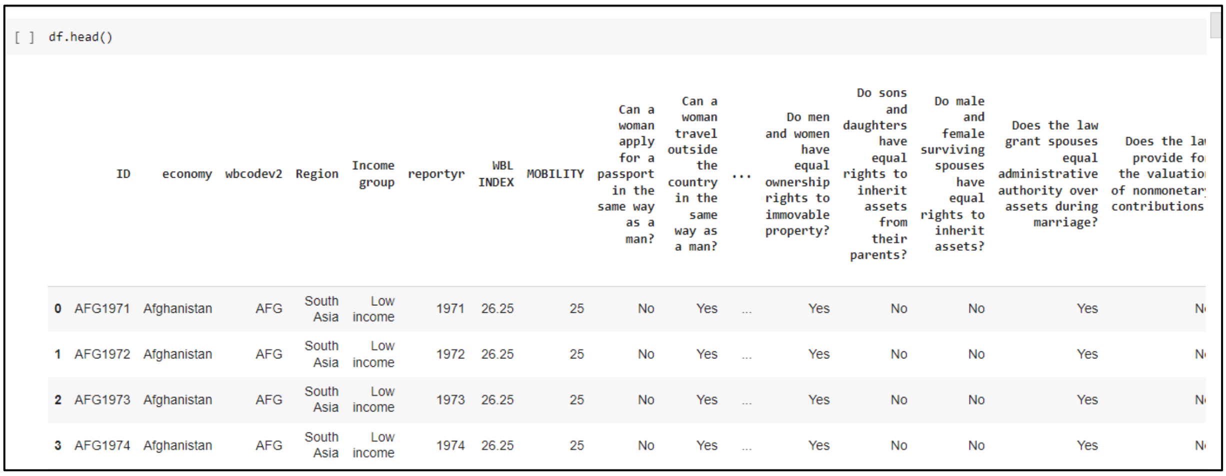
Based on figure 4.3, by using the python code above, the column in the dataset can be visualized in list form. Based on the output, we can understand that most of the variables are in a question form and under different categories.



```
[ ]  df.shape

     (9690, 55)
```

Figure 4.4 Code and Output for the Dataset Structure Information

Based on figure 4.4, by using the python code above, the shape of the dataset can be understood. The output showed that the dataset has 9690 observations and 55 variables.

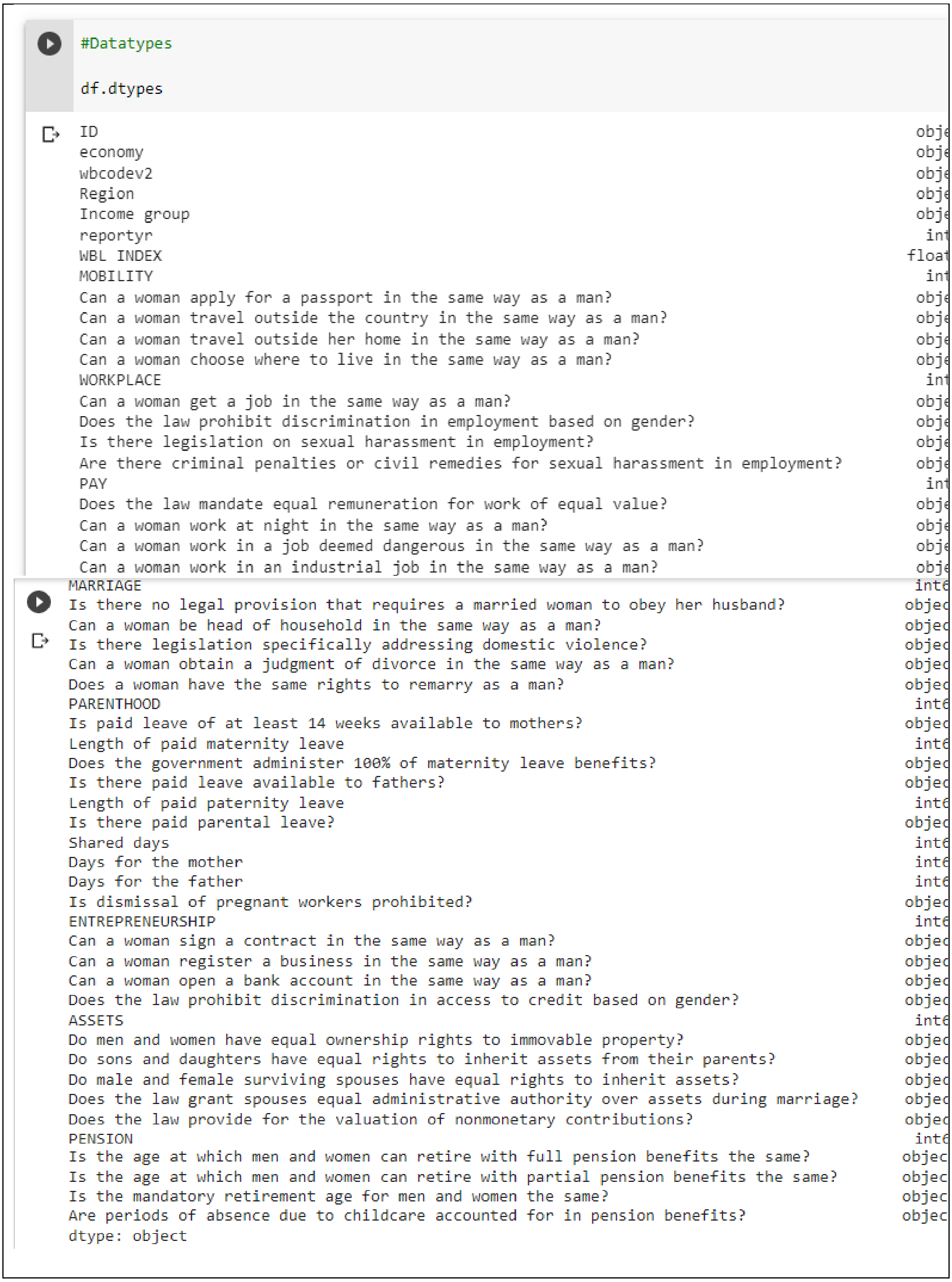


```
[ ]  df.head()
```

	ID	economy	wbcodev2	Region	Income group	reportyr	WBL INDEX	MOBILITY	Can a woman apply for a passport in the same way as a man?	Can a woman travel outside the country in the same way as a man?	...	Do men and women have equal ownership rights to immovable property?	Do sons and daughters have equal rights to inherit assets from their parents?	Do male and female surviving spouses have equal rights to inherit assets?	Does the law grant spouses equal administrative authority over assets during marriage?	Does the la provide fo the valuatio of nonmonetar contributions
0	AFG1971	Afghanistan	AFG	South Asia	Low income	1971	26.25	25	No	Yes	...	Yes	No	No	Yes	N
1	AFG1972	Afghanistan	AFG	South Asia	Low income	1972	26.25	25	No	Yes	...	Yes	No	No	Yes	N
2	AFG1973	Afghanistan	AFG	South Asia	Low income	1973	26.25	25	No	Yes	...	Yes	No	No	Yes	N
3	AFG1974	Afghanistan	AFG	South Asia	Low income	1974	26.25	25	No	Yes	...	Yes	No	No	Yes	N

Figure 4.5 Code and Output for the First Five Rows of the Dataset

Based on figure 4.5, by using the python code above, the first 5 rows of the dataset are shown. Based on the output, we can understand the content of each column. Based on this visualization, we can understand that most of the content is ‘Yes’ or ‘No’, which indicates that there is a higher probability that most of the variables are categorical form.



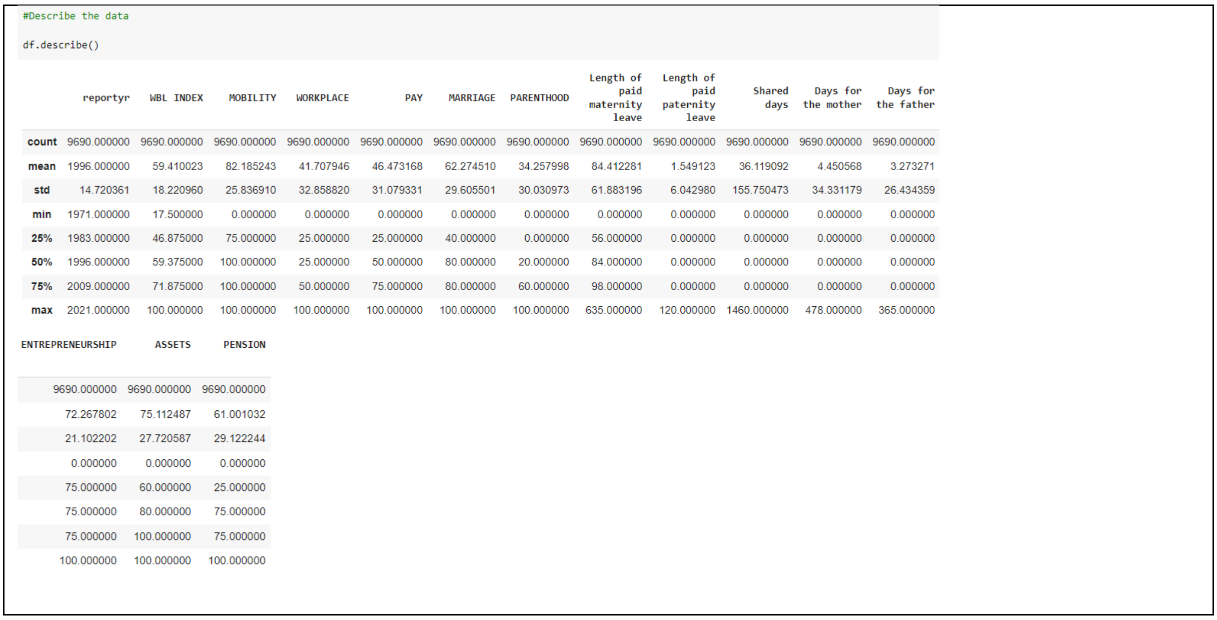
```
#Datatypes

df.dtypes
```

```
ID                                                                                      obj
economy                                                                                 obj
wbcodev2                                                                                obj
Region                                                                                  obj
Income group                                                                            obj
reportyr                                                                                 int
WBL INDEX                                                                              float
MOBILITY                                                                                 int
Can a woman apply for a passport in the same way as a man?                              obj
Can a woman travel outside the country in the same way as a man?                        obj
Can a woman travel outside her home in the same way as a man?                           obj
Can a woman choose where to live in the same way as a man?                              obj
WORKPLACE                                                                                int
Can a woman get a job in the same way as a man?                                         obj
Does the law prohibit discrimination in employment based on gender?                     obj
Is there legislation on sexual harassment in employment?                                obj
Are there criminal penalties or civil remedies for sexual harassment in employment?     obj
PAY                                                                                      int
Does the law mandate equal remuneration for work of equal value?                        obj
Can a woman work at night in the same way as a man?                                     obj
Can a woman work in a job deemed dangerous in the same way as a man?                    obj
Can a woman work in an industrial job in the same way as a man?                         obj
MARRIAGE                                                                                 inte
Is there no legal provision that requires a married woman to obey her husband?          objec
Can a woman be head of household in the same way as a man?                              objec
Is there legislation specifically addressing domestic violence?                         objec
Can a woman obtain a judgment of divorce in the same way as a man?                      objec
Does a woman have the same rights to remarry as a man?                                  objec
PARENTHOOD                                                                               inte
Is paid leave of at least 14 weeks available to mothers?                                objec
Length of paid maternity leave                                                           inte
Does the government administer 100% of maternity leave benefits?                        objec
Is there paid leave available to fathers?                                               objec
Length of paid paternity leave                                                           inte
Is there paid parental leave?                                                           objec
Shared days                                                                              inte
Days for the mother                                                                      inte
Days for the father                                                                      inte
Is dismissal of pregnant workers prohibited?                                            objec
ENTREPRENEURSHIP                                                                         inte
Can a woman sign a contract in the same way as a man?                                   objec
Can a woman register a business in the same way as a man?                              objec
Can a woman open a bank account in the same way as a man?                               objec
Does the law prohibit discrimination in access to credit based on gender?               objec
ASSETS                                                                                   inte
Do men and women have equal ownership rights to immovable property?                     objec
Do sons and daughters have equal rights to inherit assets from their parents?           objec
Do male and female surviving spouses have equal rights to inherit assets?               objec
Does the law grant spouses equal administrative authority over assets during marriage?  objec
Does the law provide for the valuation of nonmonetary contributions?                    objec
PENSION                                                                                  inte
Is the age at which men and women can retire with full pension benefits the same?       objec
Is the age at which men and women can retire with partial pension benefits the same?    objec
Is the mandatory retirement age for men and women the same?                             objec
Are periods of absence due to childcare accounted for in pension benefits?              objec
dtype: object
```

Figure 4.6 Code and Output for Data Types of Variables

Based on figure 4.6, by using the python code above, the data types of each variable are understood. Based on the output, it shows that most of the variables are categorical variables as expected previously. This data exploration provides information that label encoding or one-hot encoding is needed before the model building in the later pre-processing step.

4.3.4 Description Statistics

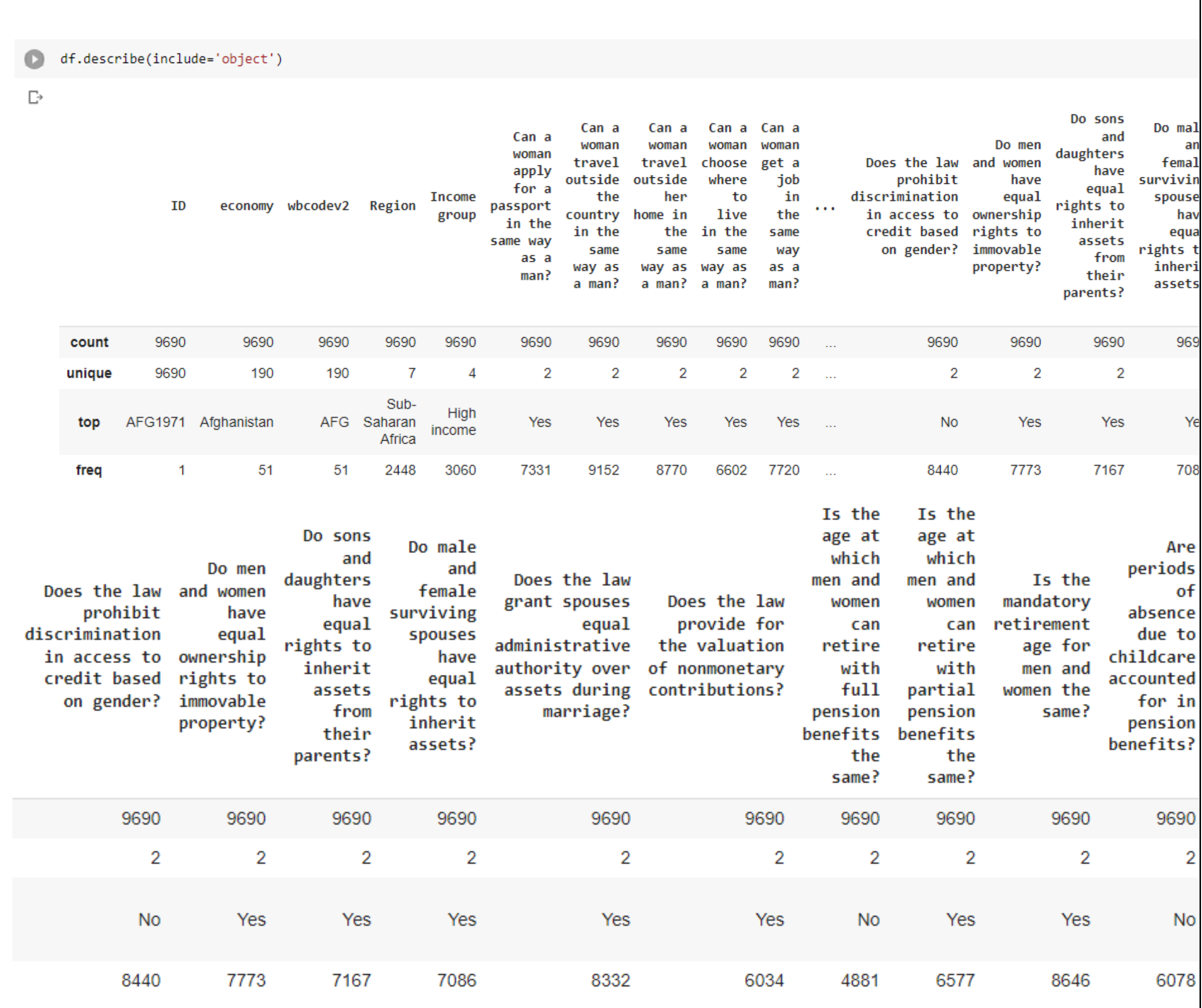


```
#Describe the data

df.describe()
```

```
                                                                                                          Length of    Length of
                                                                                                               paid         paid       Shared      Days for     Days for
          reportyr    WBL INDEX     MOBILITY    WORKPLACE          PAY     MARRIAGE   PARENTHOOD    maternity    paternity         days    the mother   the father
                                                                                                              leave        leave
count  9690.000000  9690.000000  9690.000000  9690.000000  9690.000000  9690.000000  9690.000000  9690.000000  9690.000000  9690.000000  9690.000000  9690.000000
mean   1996.000000    59.410023    82.185243    41.707946    46.473168    62.274510    34.257998    84.412281     1.549123    36.119092     4.450568     3.273271
std      14.720361    18.220960    25.836910    32.858820    31.079331    29.605501    30.030973    61.883196     6.042980   155.750473    34.331179    26.434359
min    1971.000000    17.500000     0.000000     0.000000     0.000000     0.000000     0.000000     0.000000     0.000000     0.000000     0.000000     0.000000
25%    1983.000000    46.875000    75.000000    25.000000    25.000000    40.000000     0.000000    56.000000     0.000000     0.000000     0.000000     0.000000
50%    1996.000000    59.375000   100.000000    25.000000    50.000000    80.000000    20.000000    84.000000     0.000000     0.000000     0.000000     0.000000
75%    2009.000000    71.875000   100.000000    50.000000    75.000000    80.000000    60.000000    98.000000     0.000000     0.000000     0.000000     0.000000
max    2021.000000   100.000000   100.000000   100.000000   100.000000   100.000000   100.000000   635.000000   120.000000  1460.000000   478.000000   365.000000

ENTREPRENEURSHIP       ASSETS      PENSION

     9690.000000  9690.000000  9690.000000
       72.267802    75.112487    61.001032
       21.102202    27.720587    29.122244
        0.000000     0.000000     0.000000
       75.000000    60.000000    25.000000
       75.000000    80.000000    75.000000
       75.000000   100.000000    75.000000
      100.000000   100.000000   100.000000
```

Figure 4.7 Code and Output for Data Description for Continuous Variable

Based on figure 4.7, by using the python code above, the data description for the continuous variables is shown. Based on the output, we can understand the count, mean, standard deviation, minimum value, quartile and maximum value of each continuous variable. Based on the minimum and maximum values, we can understand the possibility to present an outlier in the dataset.



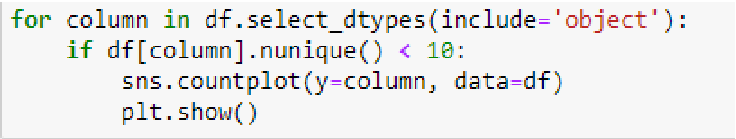
```
df.describe(include='object')
```

	ID	economy	wbcodev2	Region	Income group	Can a woman apply for a passport in the same way as a man?	Can a woman travel outside the country in the same way as a man?	Can a woman travel outside her home in the same way as a man?	Can a woman choose where to live in the same way as a man?	Can a woman get a job in the same way as a man?	...	Does the law prohibit discrimination in access to credit based on gender?	Do men and women have equal ownership rights to immovable property?	Do sons and daughters have equal rights to inherit assets from their parents?	Do mal an femal survivin spouse hav equa rights t inheri assets
count	9690	9690	9690	9690	9690	9690	9690	9690	9690	9690	...	9690	9690	9690	969
unique	9690	190	190	7	4	2	2	2	2	2	...	2	2	2	
top	AFG1971	Afghanistan	AFG	Sub-Saharan Africa	High income	Yes	Yes	Yes	Yes	Yes	...	No	Yes	Yes	Ye
freq	1	51	51	2448	3060	7331	9152	8770	6602	7720	...	8440	7773	7167	708

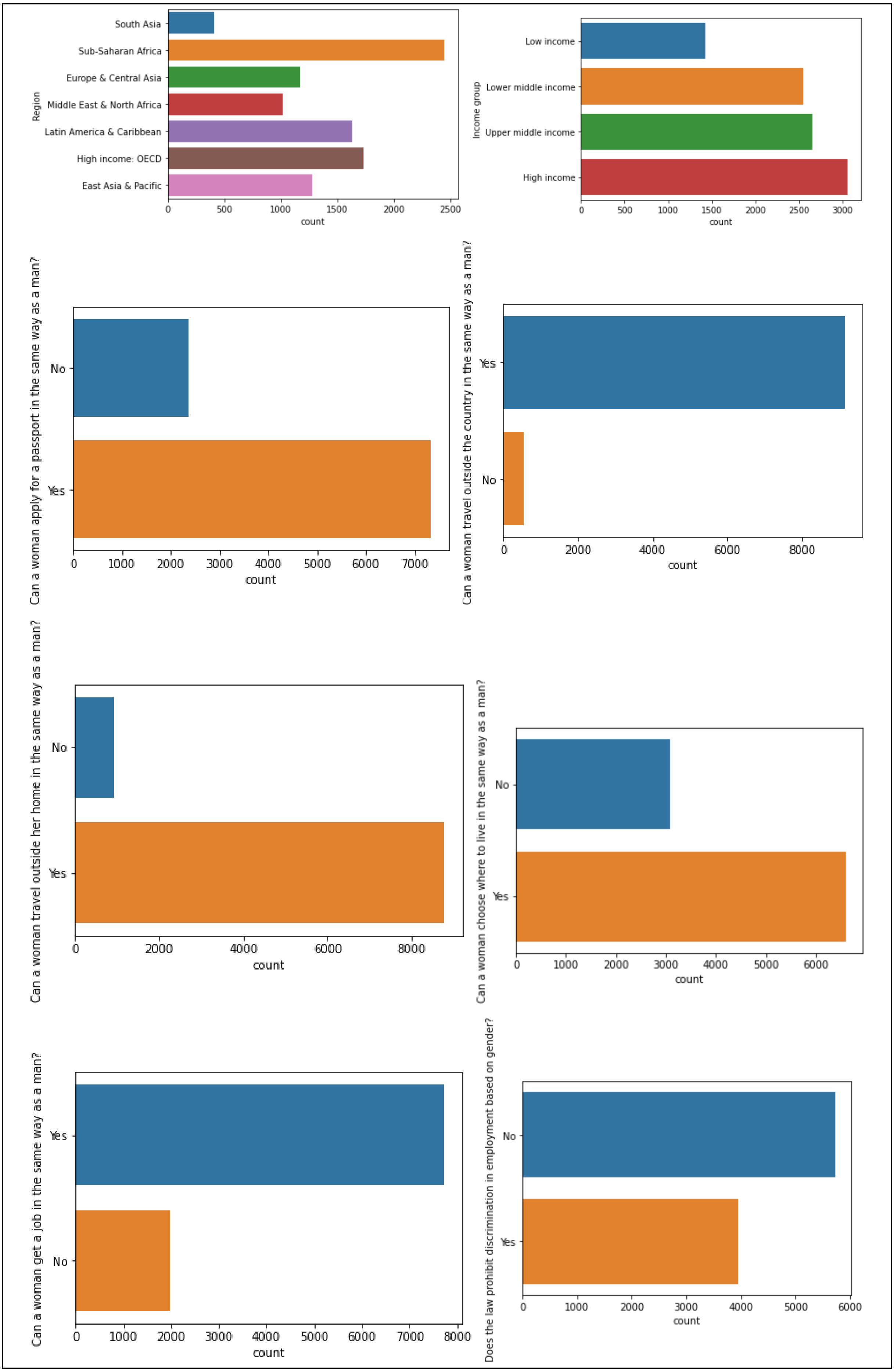
Does the law prohibit discrimination in access to credit based on gender?	Do men and women have equal ownership rights to immovable property?	Do sons and daughters have equal rights to inherit assets from their parents?	Do male and female surviving spouses have equal rights to inherit assets?	Does the law grant spouses equal administrative authority over assets during marriage?	Does the law provide for the valuation of nonmonetary contributions?	Is the age at which men and women can retire with full pension benefits the same?	Is the age at which men and women can retire with partial pension benefits the same?	Is the mandatory retirement age for men and women the same?	Are periods of absence due to childcare accounted for in pension benefits?
9690	9690	9690	9690	9690	9690	9690	9690	9690	9690
2	2	2	2	2	2	2	2	2	2
No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
8440	7773	7167	7086	8332	6034	4881	6577	8646	6078

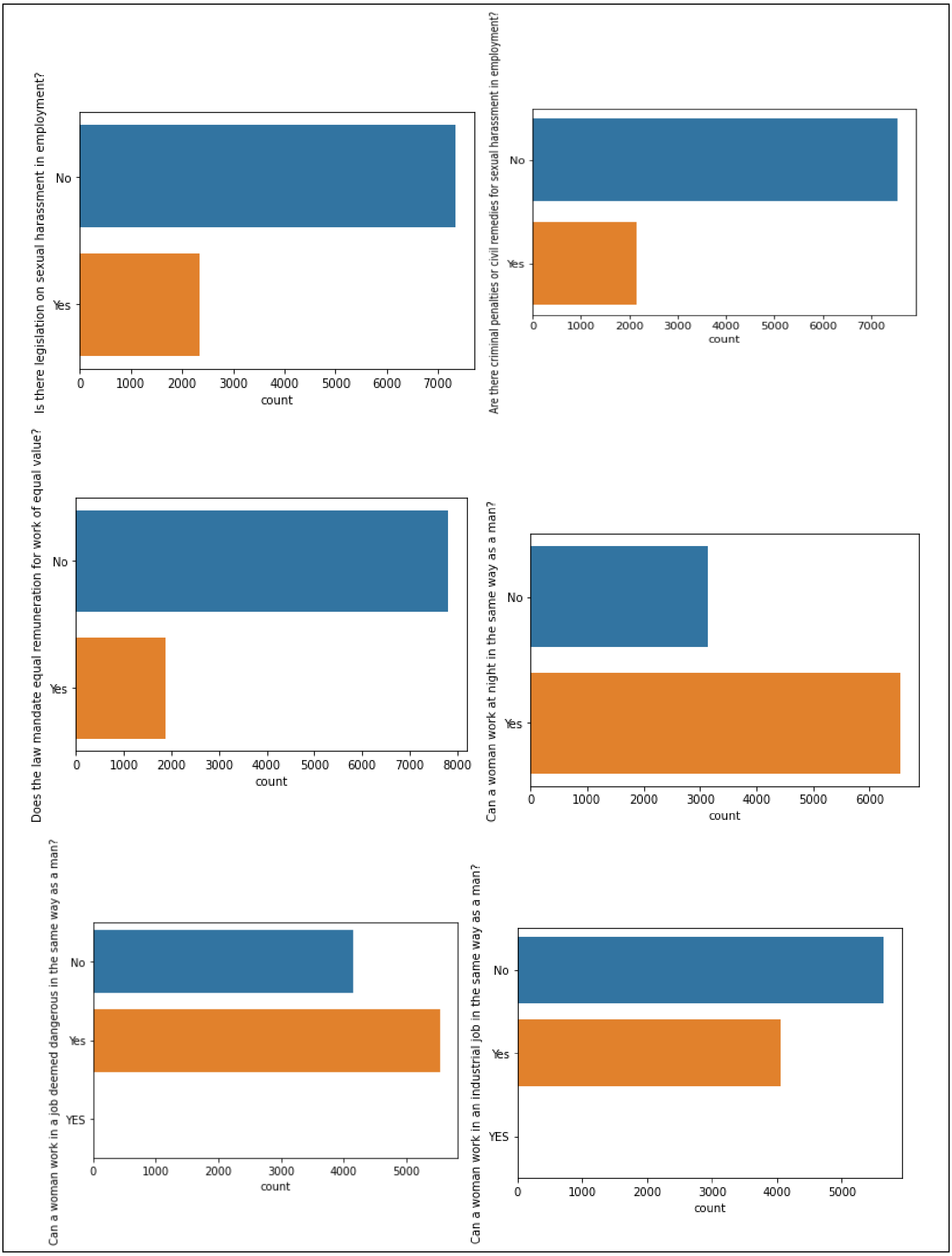
Figure 4.8 Code and Output for Data Description for Categorical Variable

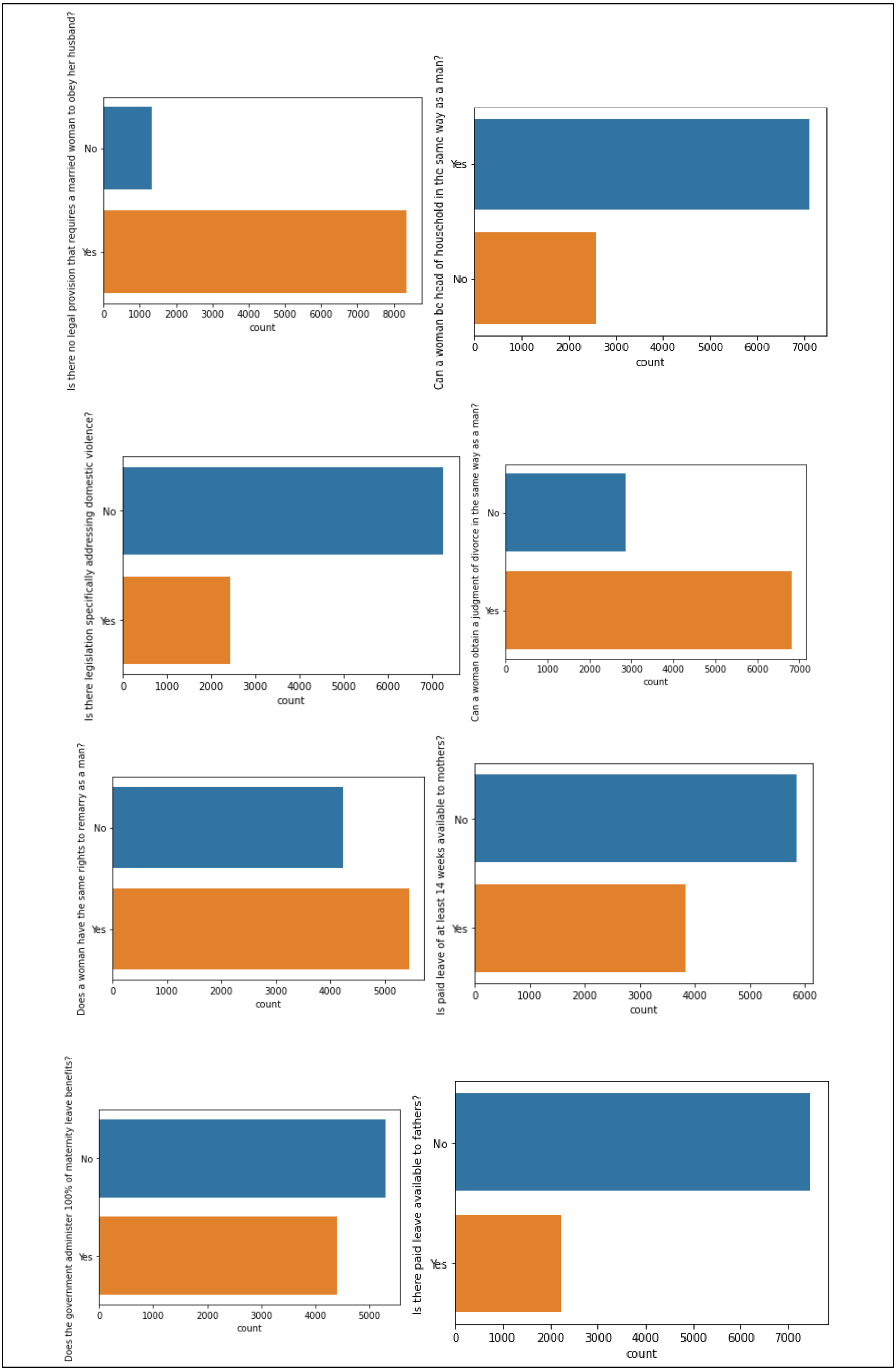
Based on figure 4.8, by using the python code above, the data description for the categorical variables is shown. Based on the output, we can understand the count, the number of categories under the variable and the frequency of the category.

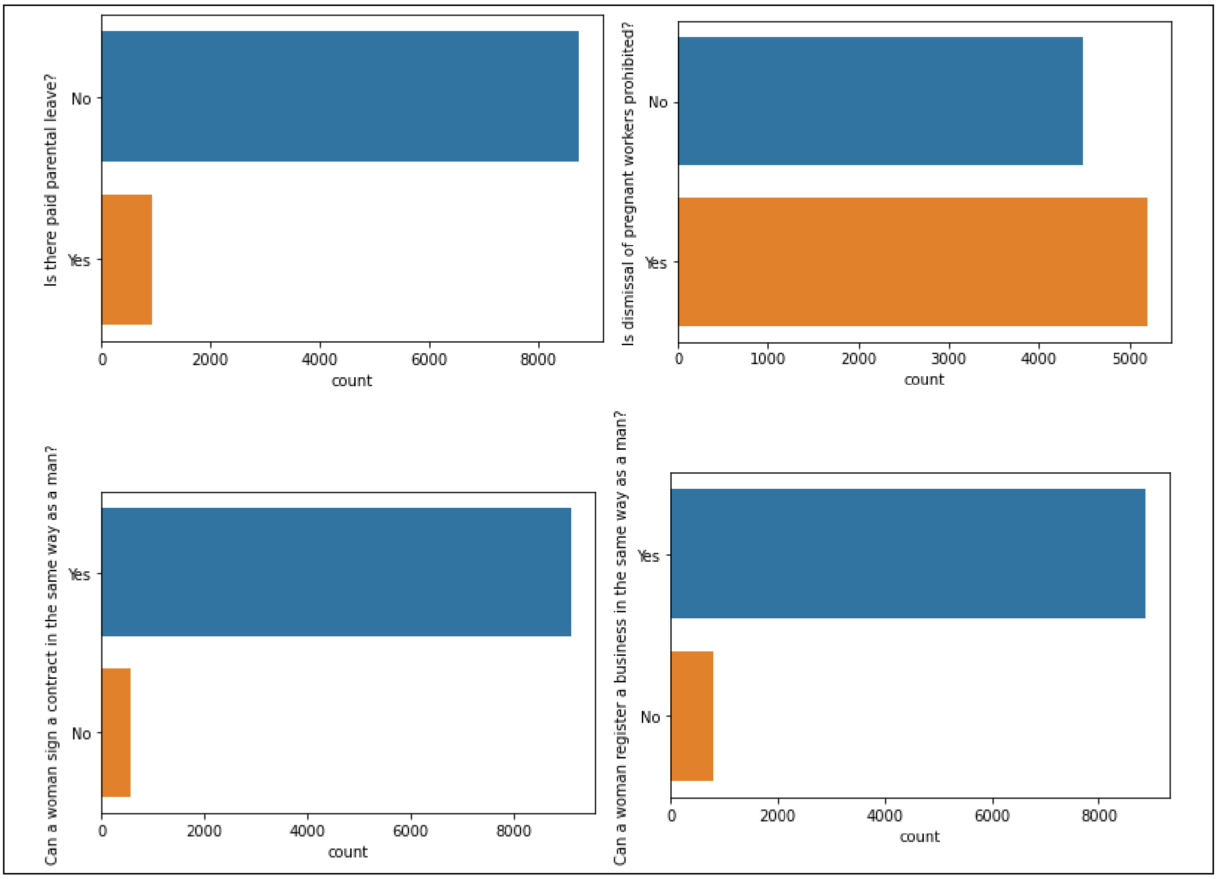


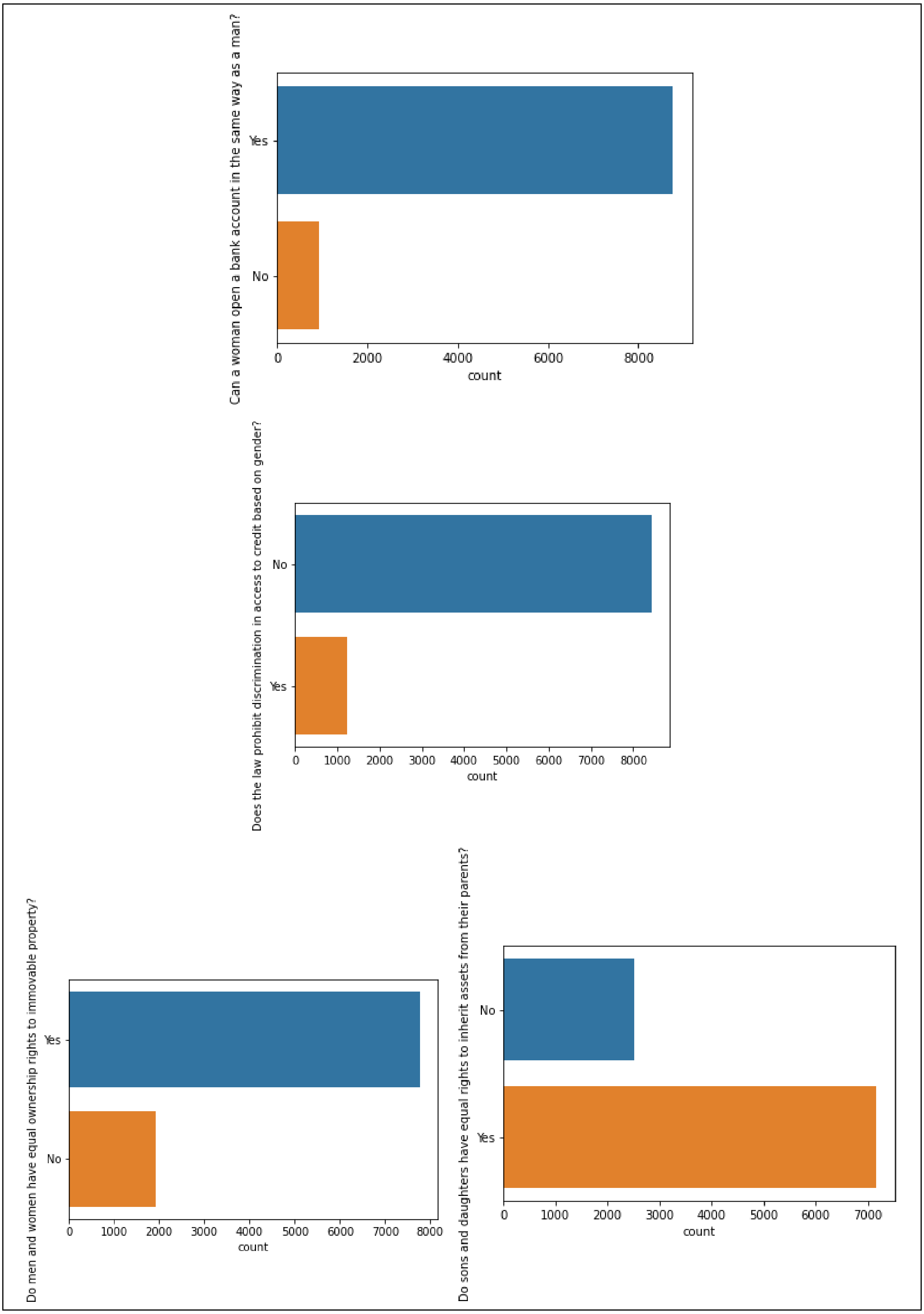
```
for column in df.select_dtypes(include='object'):
    if df[column].nunique() < 10:
        sns.countplot(y=column, data=df)
        plt.show()
```

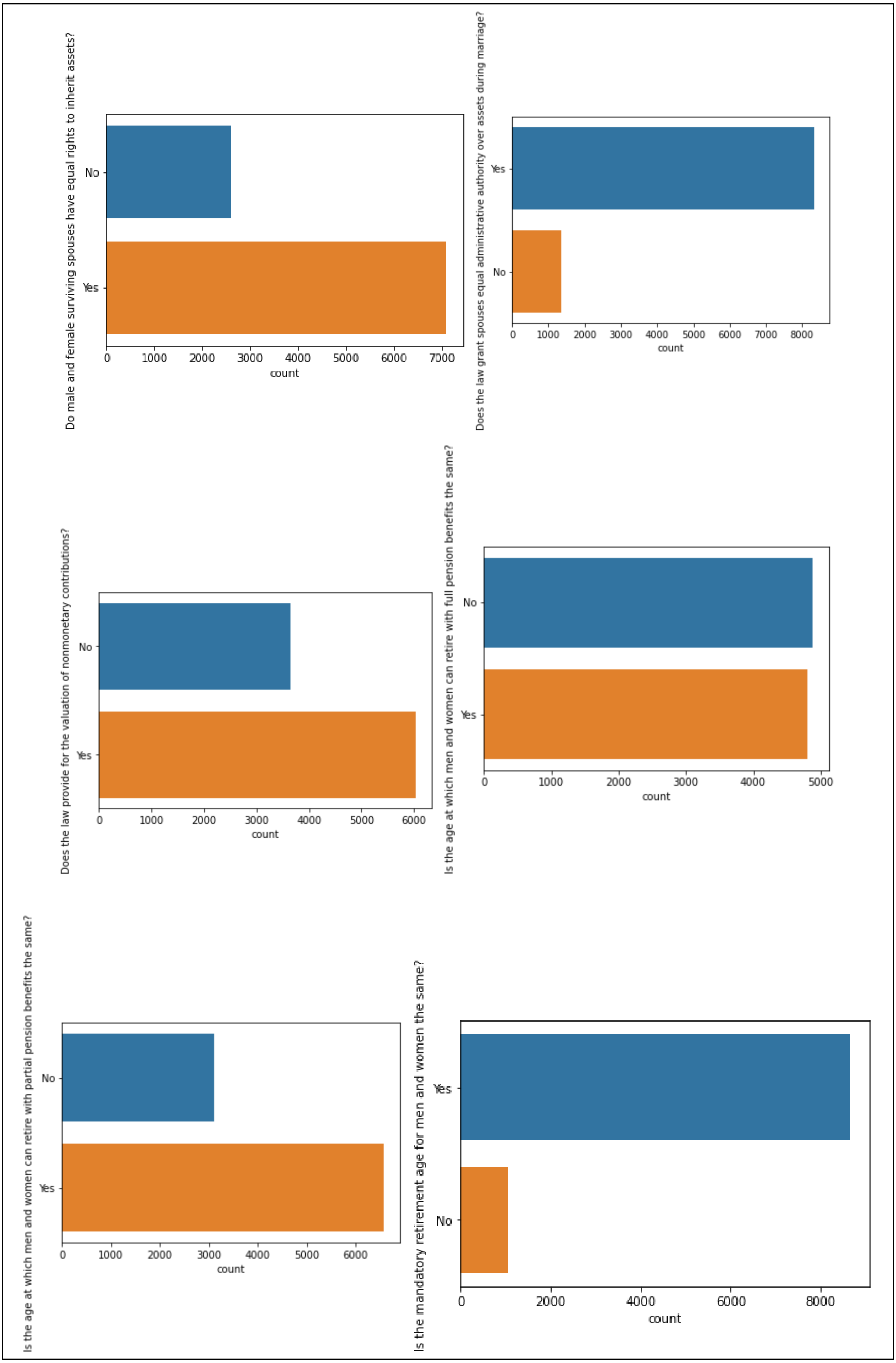

Region
South Asia
Sub-Saharan Africa
Europe & Central Asia
Middle East & North Africa
Latin America & Caribbean
High income: OECD
East Asia & Pacific
0
500
1000
1500
2000
2500
count
Income group
Low income
Lower middle income
Upper middle income
High income
0
500
1000
1500
2000
2500
3000
count
Can a woman apply for a passport in the same way as a man?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
count
Can a woman travel outside the country in the same way as a man?
Yes
No
0
2000
4000
6000
8000
count
Can a woman travel outside her home in the same way as a man?
No
Yes
0
2000
4000
6000
8000
count
Can a woman choose where to live in the same way as a man?
No
Yes
0
1000
2000
3000
4000
5000
6000
count
Can a woman get a job in the same way as a man?
Yes
No
0
1000
2000
3000
4000
5000
6000
7000
8000
count
Does the law prohibit discrimination in employment based on gender?
No
Yes
0
1000
2000
3000
4000
5000
6000
count


Is there legislation on sexual harassment in employment?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
count
Are there criminal penalties or civil remedies for sexual harassment in employment?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
count
Does the law mandate equal remuneration for work of equal value?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
8000
count
Can a woman work at night in the same way as a man?
No
Yes
0
1000
2000
3000
4000
5000
6000
count
Can a woman work in a job deemed dangerous in the same way as a man?
No
Yes
YES
0
1000
2000
3000
4000
5000
count
Can a woman work in an industrial job in the same way as a man?
No
Yes
YES
0
1000
2000
3000
4000
5000
count


Is there no legal provision that requires a married woman to obey her husband?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
8000
count
Can a woman be head of household in the same way as a man?
Yes
No
0
1000
2000
3000
4000
5000
6000
7000
count
Is there legislation specifically addressing domestic violence?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
count
Can a woman obtain a judgment of divorce in the same way as a man?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
count
Does a woman have the same rights to remarry as a man?
No
Yes
0
1000
2000
3000
4000
5000
count
Is paid leave of at least 14 weeks available to mothers?
No
Yes
0
1000
2000
3000
4000
5000
6000
count
Does the government administer 100% of maternity leave benefits?
No
Yes
0
1000
2000
3000
4000
5000
count
Is there paid leave available to fathers?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
count


Is there paid parental leave?
No
Yes
0
2000
4000
6000
8000
count
Is dismissal of pregnant workers prohibited?
No
Yes
0
1000
2000
3000
4000
5000
count
Can a woman sign a contract in the same way as a man?
Yes
No
0
2000
4000
6000
8000
count
Can a woman register a business in the same way as a man?
Yes
No
0
2000
4000
6000
8000
count


Can a woman open a bank account in the same way as a man?
Yes
No
0
2000
4000
6000
8000
count
Does the law prohibit discrimination in access to credit based on gender?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
8000
count
Do men and women have equal ownership rights to immovable property?
Yes
No
0
1000
2000
3000
4000
5000
6000
7000
8000
count
Do sons and daughters have equal rights to inherit assets from their parents?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
count


Do male and female surviving spouses have equal rights to inherit assets?
No
Yes
0
1000
2000
3000
4000
5000
6000
7000
count
Does the law grant spouses equal administrative authority over assets during marriage?
Yes
No
0
1000
2000
3000
4000
5000
6000
7000
8000
count
Does the law provide for the valuation of nonmonetary contributions?
No
Yes
0
1000
2000
3000
4000
5000
6000
count
Is the age at which men and women can retire with full pension benefits the same?
No
Yes
0
1000
2000
3000
4000
5000
count
Is the age at which men and women can retire with partial pension benefits the same?
No
Yes
0
1000
2000
3000
4000
5000
6000
count
Is the mandatory retirement age for men and women the same?
Yes
No
0
2000
4000
6000
8000
count

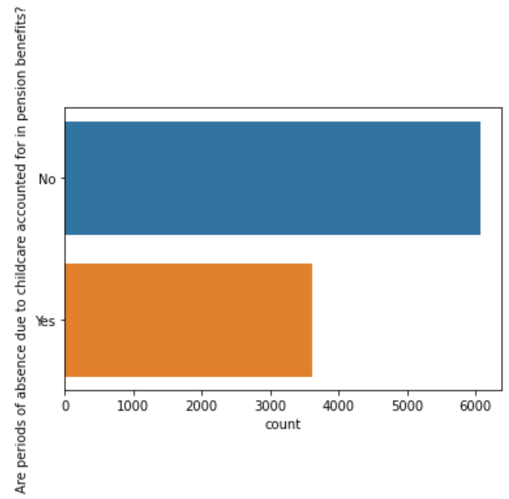
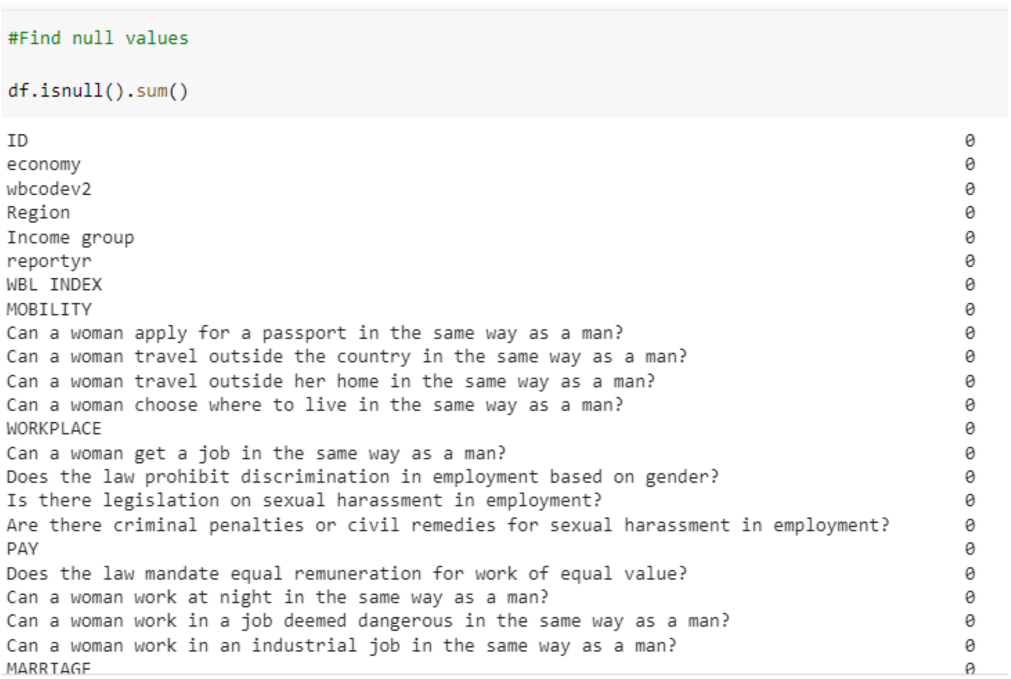


Figure 4.9 Code and the Bar Chart Visualization

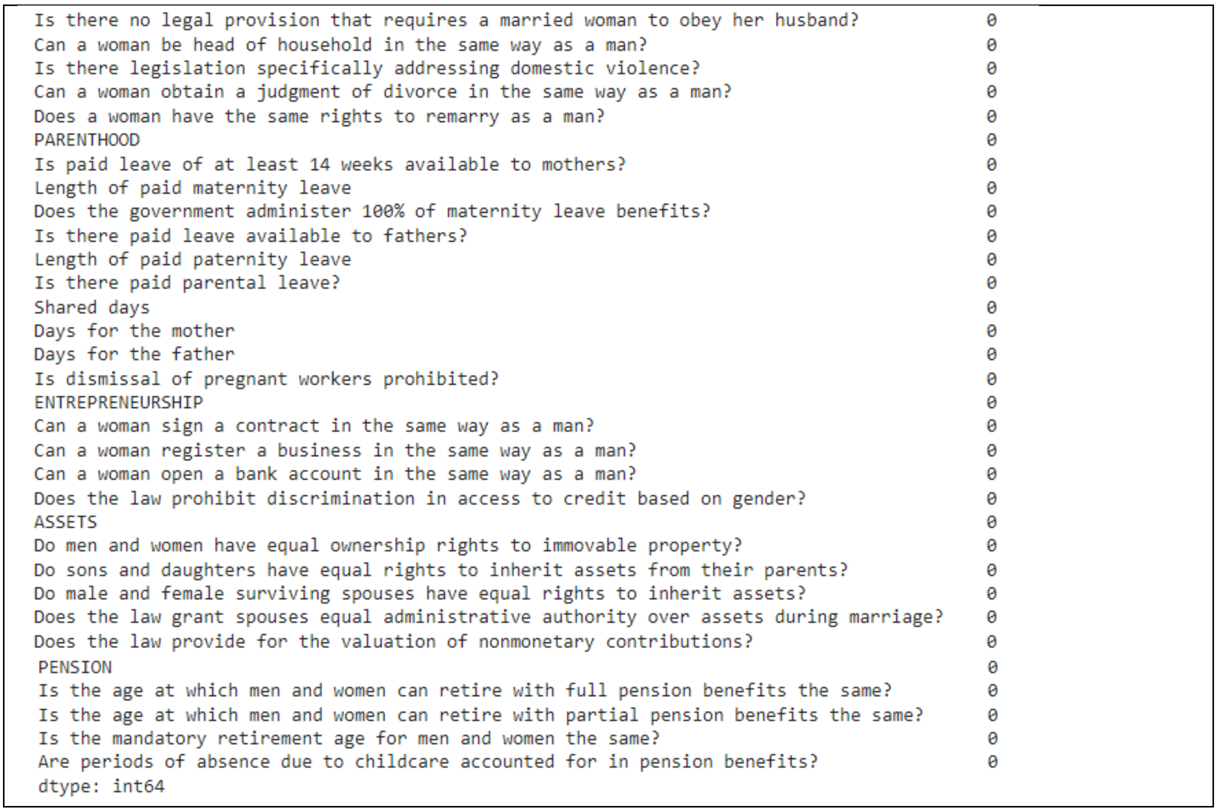
Based on figure 4.9, by using the python code above, the bar chart of each categorical variable is visualized. Based on the output, we can understand that most of the variables only have 2 variables, which are the ‘Yes’ or ‘No’ categories as expected previously. Only region and income groups are consisting more than 2 categories. Hence, here we can confirm that label encoding needs to be carried out in the later pre-processing step.

4.3.5 Missing Value Detection



```
#Find null values

df.isnull().sum()
```

```
ID                                                                                        0
economy                                                                                   0
wbcodev2                                                                                  0
Region                                                                                    0
Income group                                                                              0
reportyr                                                                                  0
WBL INDEX                                                                                 0
MOBILITY                                                                                  0
Can a woman apply for a passport in the same way as a man?                                0
Can a woman travel outside the country in the same way as a man?                          0
Can a woman travel outside her home in the same way as a man?                             0
Can a woman choose where to live in the same way as a man?                                0
WORKPLACE                                                                                 0
Can a woman get a job in the same way as a man?                                           0
Does the law prohibit discrimination in employment based on gender?                       0
Is there legislation on sexual harassment in employment?                                  0
Are there criminal penalties or civil remedies for sexual harassment in employment?       0
PAY                                                                                       0
Does the law mandate equal remuneration for work of equal value?                          0
Can a woman work at night in the same way as a man?                                       0
Can a woman work in a job deemed dangerous in the same way as a man?                      0
Can a woman work in an industrial job in the same way as a man?                           0
MARRIAGE                                                                                  0
```

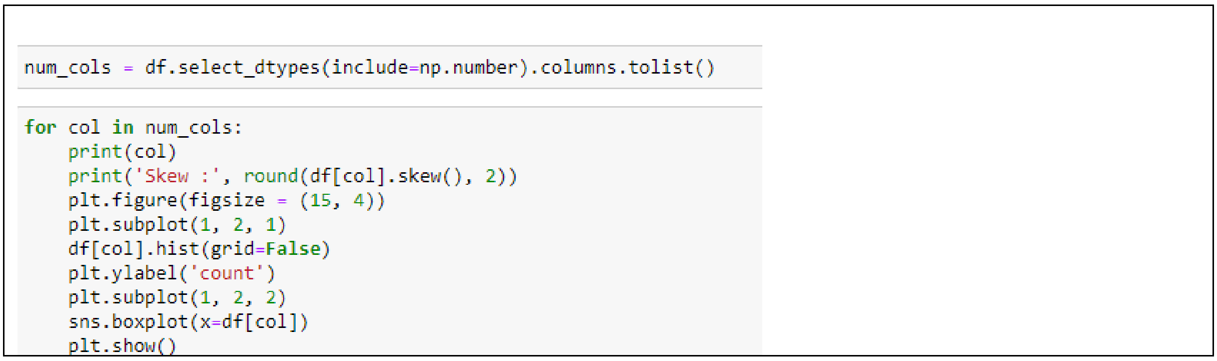


```
Is there no legal provision that requires a married woman to obey her husband?          0
Can a woman be head of household in the same way as a man?                              0
Is there legislation specifically addressing domestic violence?                         0
Can a woman obtain a judgment of divorce in the same way as a man?                      0
Does a woman have the same rights to remarry as a man?                                  0
PARENTHOOD                                                                              0
Is paid leave of at least 14 weeks available to mothers?                                0
Length of paid maternity leave                                                          0
Does the government administer 100% of maternity leave benefits?                        0
Is there paid leave available to fathers?                                               0
Length of paid paternity leave                                                          0
Is there paid parental leave?                                                           0
Shared days                                                                             0
Days for the mother                                                                     0
Days for the father                                                                     0
Is dismissal of pregnant workers prohibited?                                            0
ENTREPRENEURSHIP                                                                        0
Can a woman sign a contract in the same way as a man?                                   0
Can a woman register a business in the same way as a man?                               0
Can a woman open a bank account in the same way as a man?                               0
Does the law prohibit discrimination in access to credit based on gender?               0
ASSETS                                                                                  0
Do men and women have equal ownership rights to immovable property?                     0
Do sons and daughters have equal rights to inherit assets from their parents?           0
Do male and female surviving spouses have equal rights to inherit assets?               0
Does the law grant spouses equal administrative authority over assets during marriage?  0
Does the law provide for the valuation of nonmonetary contributions?                    0
 PENSION                                                                                0
 Is the age at which men and women can retire with full pension benefits the same?      0
 Is the age at which men and women can retire with partial pension benefits the same?   0
 Is the mandatory retirement age for men and women the same?                            0
 Are periods of absence due to childcare accounted for in pension benefits?             0
 dtype: int64
```

Figure 4.10 Code and Output Missing Value Detection

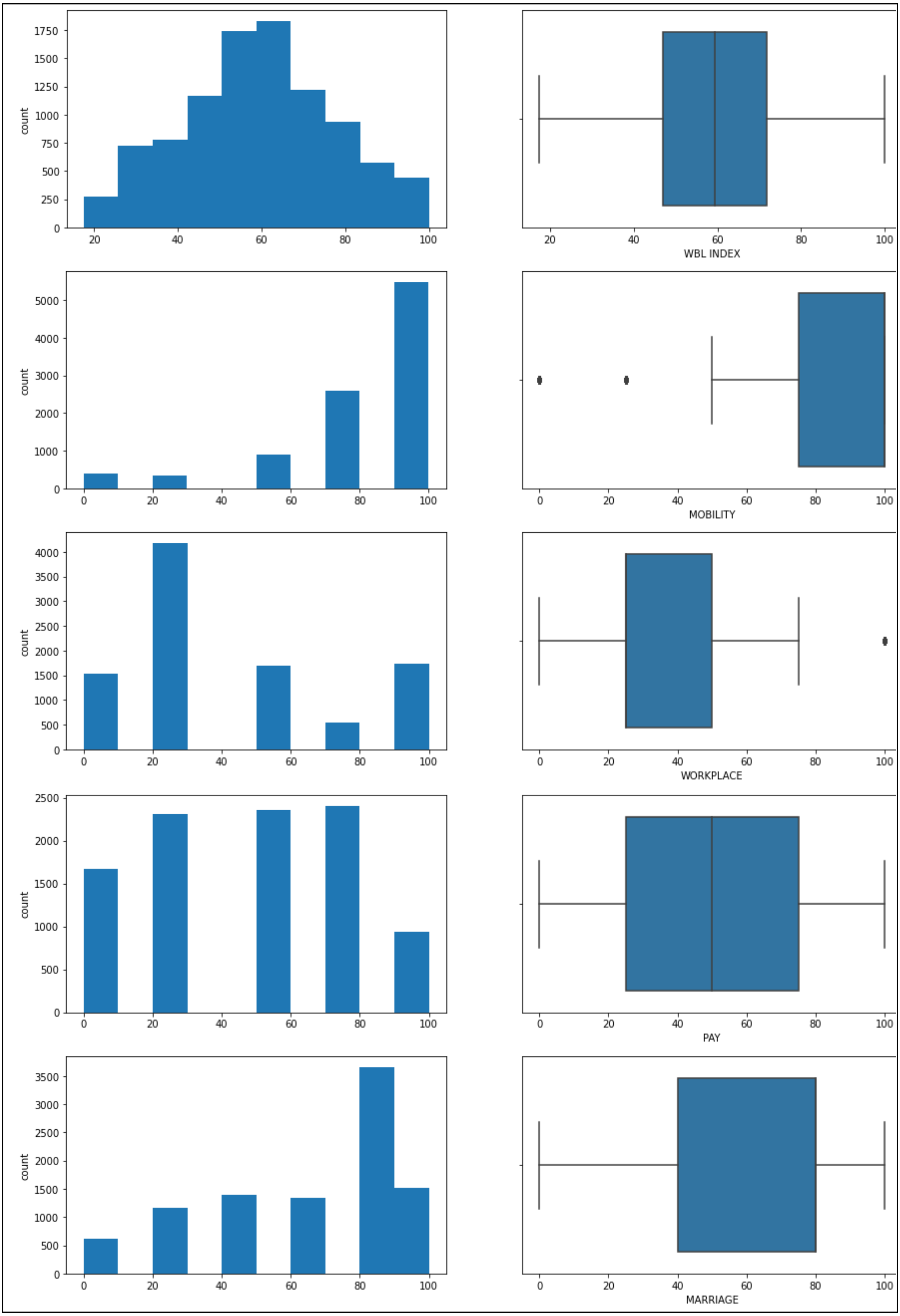
Based on figure 4.10, by using the python code above, the missing value of the dataset is explored. Based on the output, it was found that the dataset does not have any missing values. This finding indicates that there is no missing value imputation that needs to be done in the pre-processing step later.

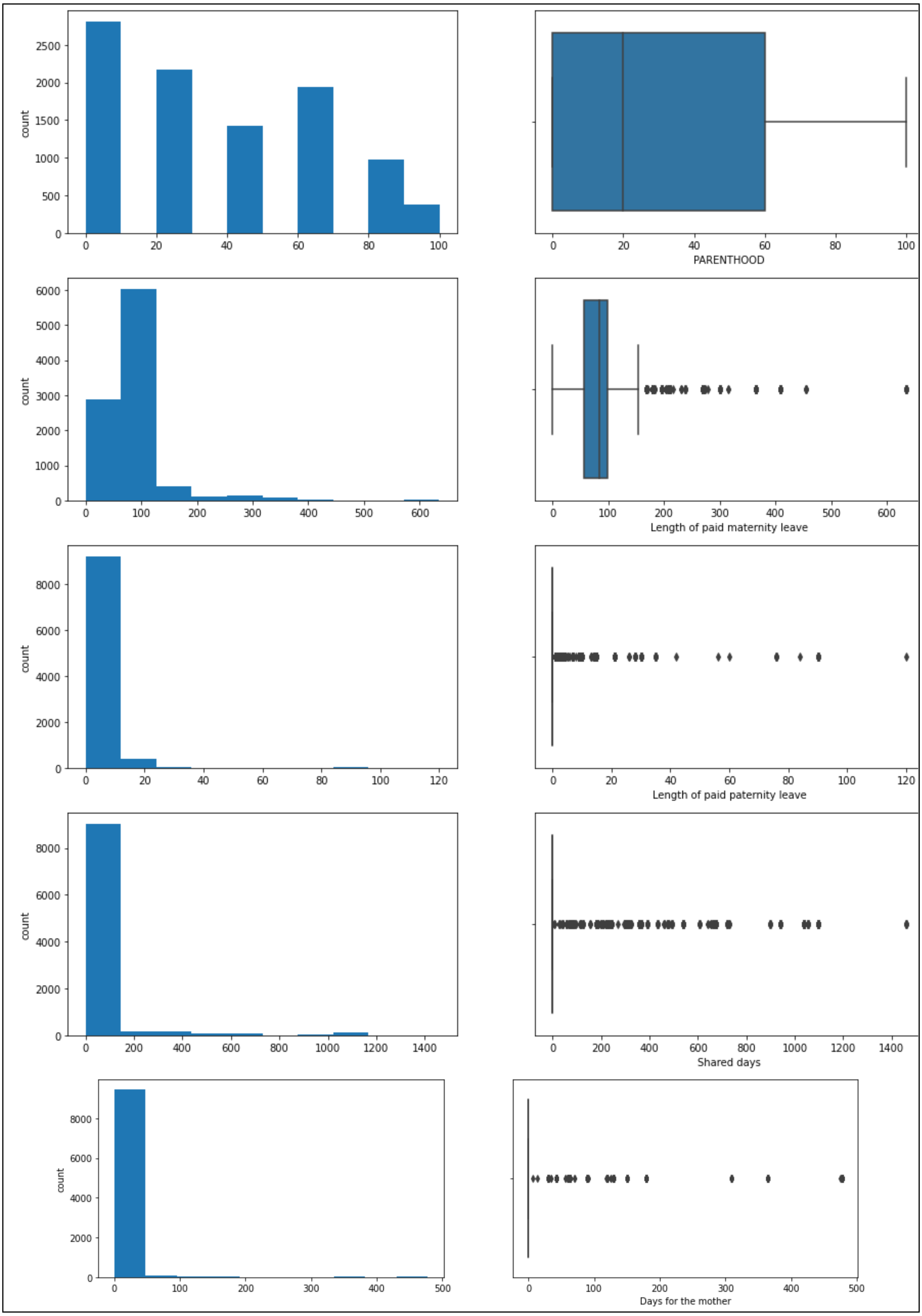
4.3.6 Distribution and Outlier Exploration



```
num_cols = df.select_dtypes(include=np.number).columns.tolist()
```

```
for col in num_cols:
    print(col)
    print('Skew :', round(df[col].skew(), 2))
    plt.figure(figsize = (15, 4))
    plt.subplot(1, 2, 1)
    df[col].hist(grid=False)
    plt.ylabel('count')
    plt.subplot(1, 2, 2)
    sns.boxplot(x=df[col])
    plt.show()
```


count
1750
1500
1250
1000
750
500
250
0
20
40
60
80
100
WBL INDEX
5000
4000
3000
2000
1000
MOBILITY
4000
3500
3000
2500
2000
1500
1000
500
WORKPLACE
2500
2000
1500
1000
500
PAY
3500
3000
2500
2000
1500
1000
500
MARRIAGE


count
2500
2000
1500
1000
500
0
0 20 40 60 80 100
0 20 40 60 80 100
PARENTHOOD
count
6000
5000
4000
3000
2000
1000
0
0 100 200 300 400 500 600
0 100 200 300 400 500 600
Length of paid maternity leave
count
8000
6000
4000
2000
0
0 20 40 60 80 100 120
0 20 40 60 80 100 120
Length of paid paternity leave
count
8000
6000
4000
2000
0
0 200 400 600 800 1000 1200 1400
0 200 400 600 800 1000 1200 1400
Shared days
count
8000
6000
4000
2000
0
0 100 200 300 400 500
0 100 200 300 400 500
Days for the mother

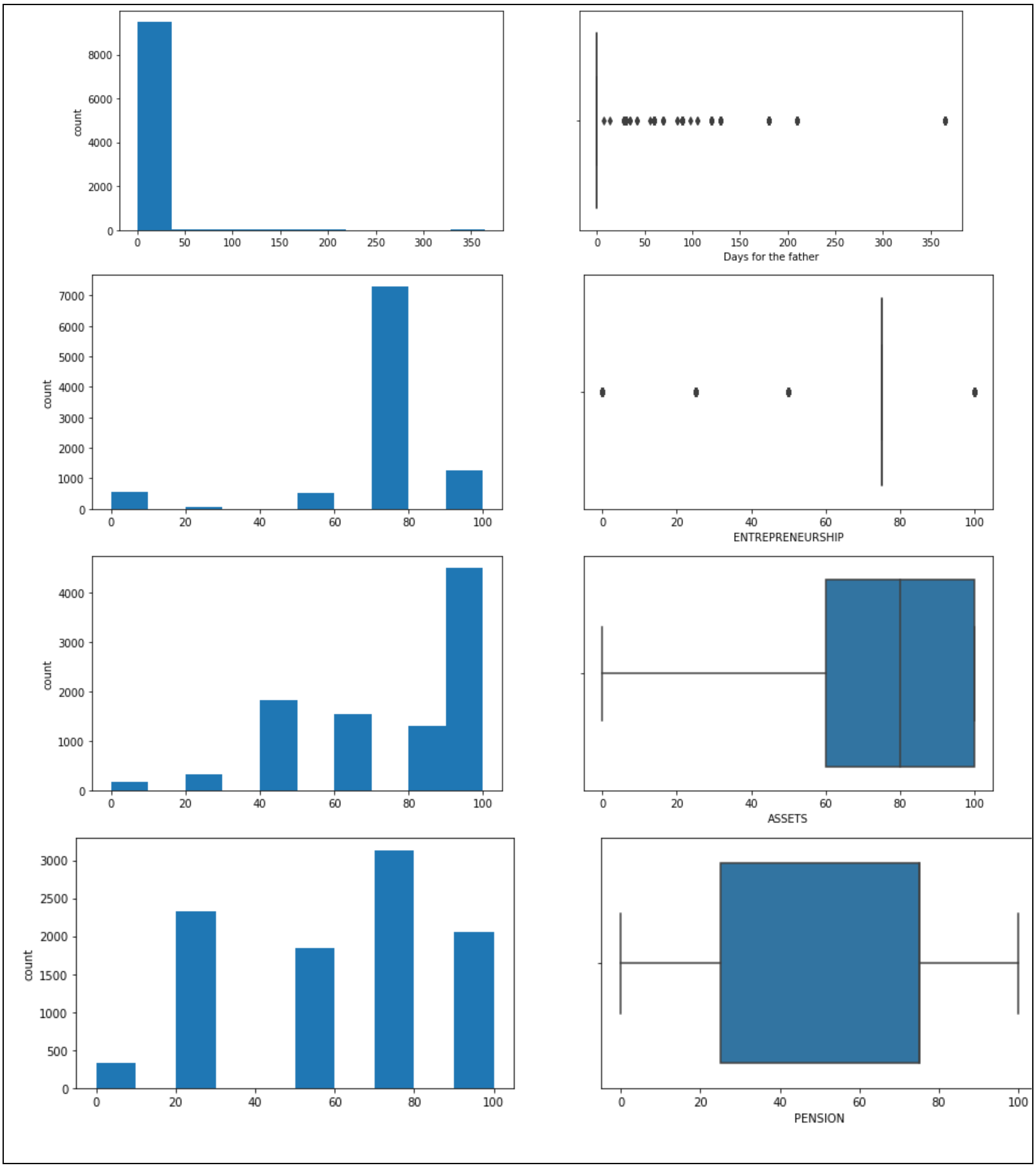
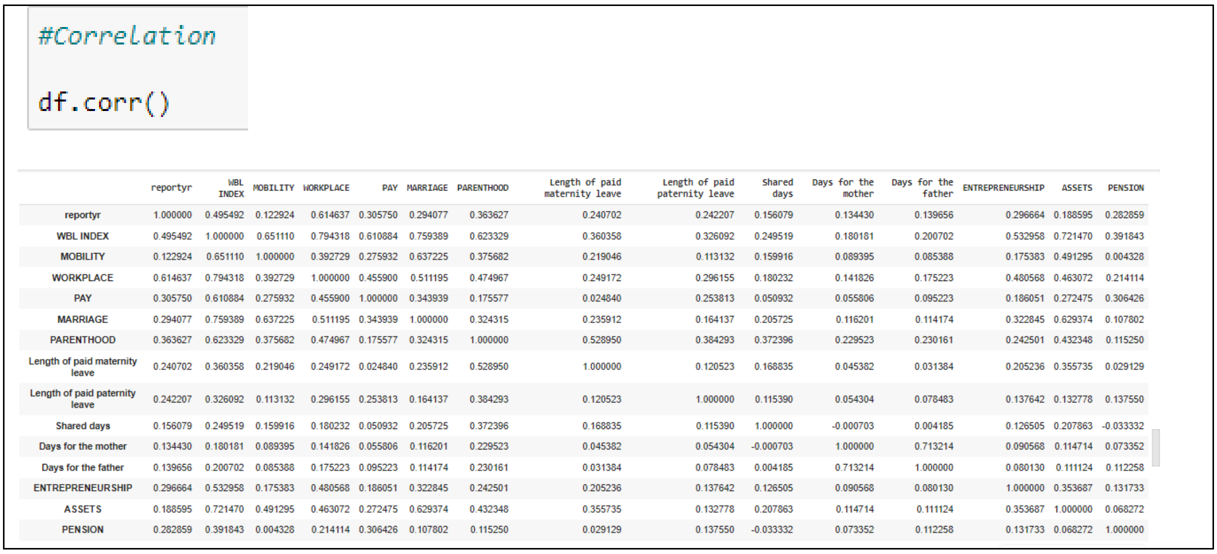


Figure 4.11 Code and the Histogram with Boxplot Output

Based on figure 4.11, by using the python code above, the histogram and box plot of the continuous variables are explored. Based on the output, it was found that the variables of ‘length of paid maternity leave’, ‘length of paid paternity leave’, ‘shared days’, ‘days for the mother’ and ‘days for the father’ are heavily right distributed and have many outliers. This brings information that we need to carry out outlier treatment before model building to prevent bias.

4.3.7 Correlation Exploration

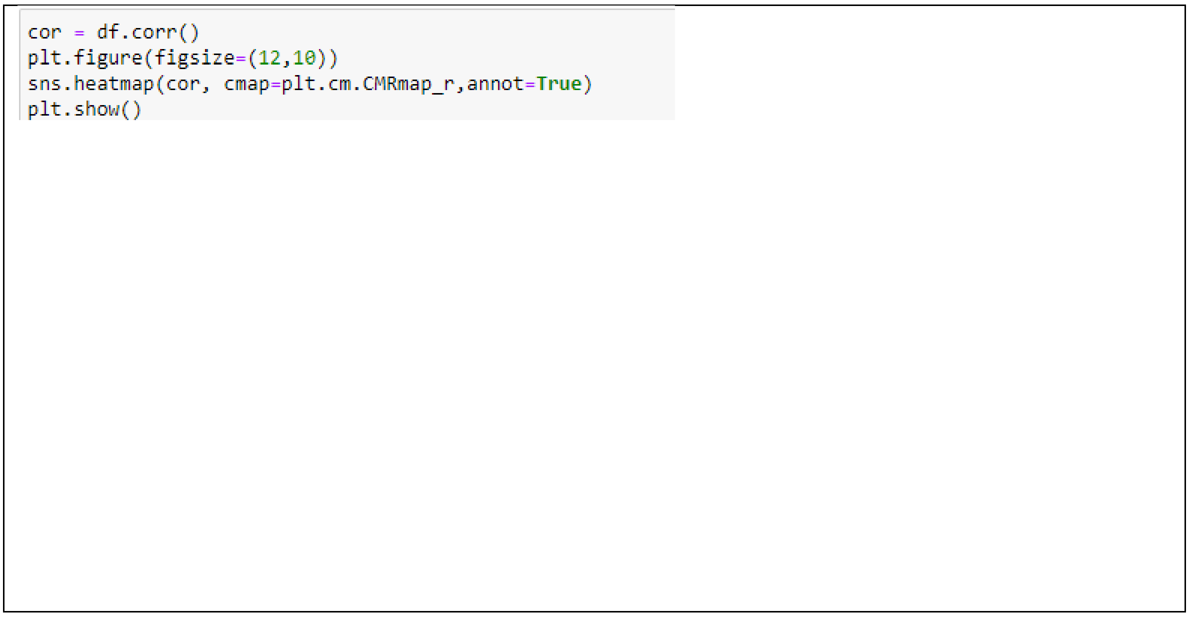


```
#Correlation

df.corr()
```

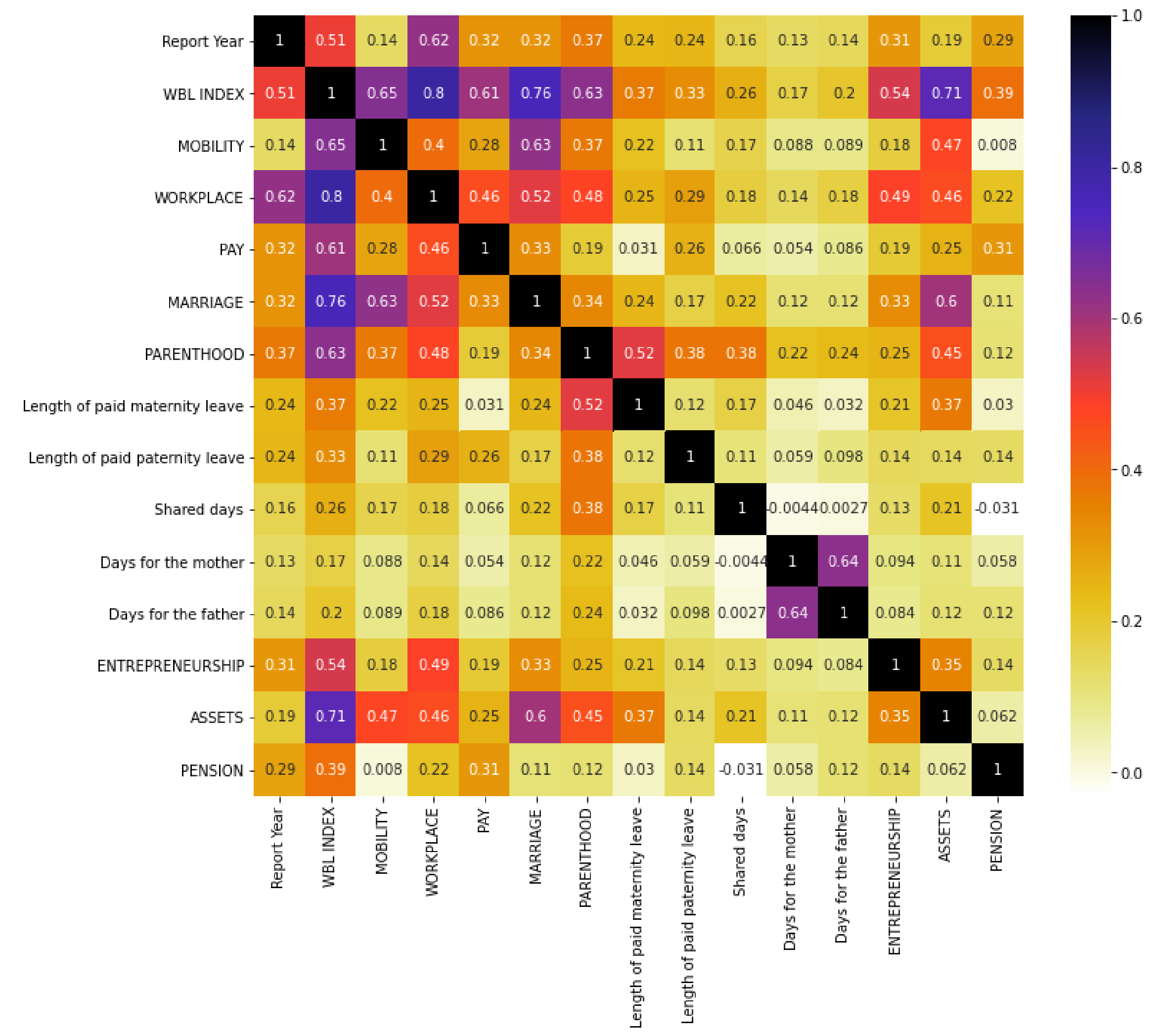
	reportyr	WBL INDEX	MOBILITY	WORKPLACE	PAY	MARRIAGE	PARENTHOOD	Length of paid maternity leave	Length of paid paternity leave	Shared days	Days for the mother	Days for the father	ENTREPRENEURSHIP	ASSETS	PENSION
reportyr	1.000000	0.495492	0.122924	0.614637	0.305750	0.294077	0.363627	0.240702	0.242207	0.156079	0.134430	0.139656	0.296664	0.188595	0.282859
WBL INDEX	0.495492	1.000000	0.651110	0.794318	0.610884	0.759389	0.623329	0.360358	0.326092	0.249519	0.180181	0.200702	0.532958	0.721470	0.391843
MOBILITY	0.122924	0.651110	1.000000	0.392729	0.275932	0.637225	0.375682	0.219046	0.113132	0.159916	0.089395	0.085388	0.175383	0.491295	0.004328
WORKPLACE	0.614637	0.794318	0.392729	1.000000	0.455900	0.511195	0.474967	0.249172	0.296155	0.180232	0.141826	0.175223	0.480568	0.463072	0.214114
PAY	0.305750	0.610884	0.275932	0.455900	1.000000	0.343939	0.175577	0.024840	0.253813	0.050932	0.055806	0.095223	0.186051	0.272475	0.306426
MARRIAGE	0.294077	0.759389	0.637225	0.511195	0.343939	1.000000	0.324315	0.235912	0.164137	0.205725	0.116201	0.114174	0.322845	0.629374	0.107802
PARENTHOOD	0.363627	0.623329	0.375682	0.474967	0.175577	0.324315	1.000000	0.528950	0.384293	0.372396	0.229523	0.230161	0.242501	0.432348	0.115250
Length of paid maternity leave	0.240702	0.360358	0.219046	0.249172	0.024840	0.235912	0.528950	1.000000	0.120523	0.168835	0.045382	0.031384	0.205236	0.355735	0.029129
Length of paid paternity leave	0.242207	0.326092	0.113132	0.296155	0.253813	0.164137	0.384293	0.120523	1.000000	0.115390	0.054304	0.078483	0.137642	0.132778	0.137550
Shared days	0.156079	0.249519	0.159916	0.180232	0.050932	0.205725	0.372396	0.168835	0.115390	1.000000	-0.000703	0.004185	0.126505	0.207863	-0.033332
Days for the mother	0.134430	0.180181	0.089395	0.141826	0.055806	0.116201	0.229523	0.045382	0.054304	-0.000703	1.000000	0.713214	0.090568	0.114714	0.073352
Days for the father	0.139656	0.200702	0.085388	0.175223	0.095223	0.114174	0.230161	0.031384	0.078483	0.004185	0.713214	1.000000	0.080130	0.111124	0.112258
ENTREPRENEURSHIP	0.296664	0.532958	0.175383	0.480568	0.186051	0.322845	0.242501	0.205236	0.137642	0.126505	0.090568	0.080130	1.000000	0.353687	0.131733
ASSETS	0.188595	0.721470	0.491295	0.463072	0.272475	0.629374	0.432348	0.355735	0.132778	0.207863	0.114714	0.111124	0.353687	1.000000	0.068272
PENSION	0.282859	0.391843	0.004328	0.214114	0.306426	0.107802	0.115250	0.029129	0.137550	-0.033332	0.073352	0.112258	0.131733	0.068272	1.000000

Figure 4.12 Code and the Correlation table

Based on figure 4.12, by using the python code above, the relationship between the continuous variables is explored. Based on the output, it was found that the WBL index has high correlations with more than 2 variables. This brings information that the variables might need to be dropped before model building since highly correlated variables have no meaning for model building. To better visualize and double-confirm the relationship, a heat map is plotted to visualize the relationship through color intensity.



```
cor = df.corr()
plt.figure(figsize=(12,10))
sns.heatmap(cor, cmap=plt.cm.CMRmap_r,annot=True)
plt.show()
```

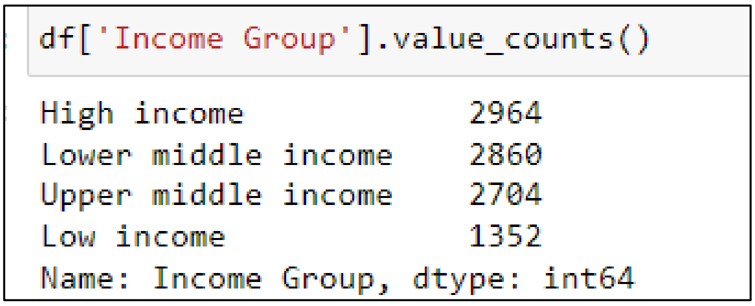


	Report Year	WBL INDEX	MOBILITY	WORKPLACE	PAY	MARRIAGE	PARENTHOOD	Length of paid maternity leave	Length of paid paternity leave	Shared days	Days for the mother	Days for the father	ENTREPRENEURSHIP	ASSETS	PENSION
Report Year	1	0.51	0.14	0.62	0.32	0.32	0.37	0.24	0.24	0.16	0.13	0.14	0.31	0.19	0.29
WBL INDEX	0.51	1	0.65	0.8	0.61	0.76	0.63	0.37	0.33	0.26	0.17	0.2	0.54	0.71	0.39
MOBILITY	0.14	0.65	1	0.4	0.28	0.63	0.37	0.22	0.11	0.17	0.088	0.089	0.18	0.47	0.008
WORKPLACE	0.62	0.8	0.4	1	0.46	0.52	0.48	0.25	0.29	0.18	0.14	0.18	0.49	0.46	0.22
PAY	0.32	0.61	0.28	0.46	1	0.33	0.19	0.031	0.26	0.066	0.054	0.086	0.19	0.25	0.31
MARRIAGE	0.32	0.76	0.63	0.52	0.33	1	0.34	0.24	0.17	0.22	0.12	0.12	0.33	0.6	0.11
PARENTHOOD	0.37	0.63	0.37	0.48	0.19	0.34	1	0.52	0.38	0.38	0.22	0.24	0.25	0.45	0.12
Length of paid maternity leave	0.24	0.37	0.22	0.25	0.031	0.24	0.52	1	0.12	0.17	0.046	0.032	0.21	0.37	0.03
Length of paid paternity leave	0.24	0.33	0.11	0.29	0.26	0.17	0.38	0.12	1	0.11	0.059	0.098	0.14	0.14	0.14
Shared days	0.16	0.26	0.17	0.18	0.066	0.22	0.38	0.17	0.11	1	-0.0044	0.0027	0.13	0.21	-0.031
Days for the mother	0.13	0.17	0.088	0.14	0.054	0.12	0.22	0.046	0.059	-0.0044	1	0.64	0.094	0.11	0.058
Days for the father	0.14	0.2	0.089	0.18	0.086	0.12	0.24	0.032	0.098	0.0027	0.64	1	0.084	0.12	0.12
ENTREPRENEURSHIP	0.31	0.54	0.18	0.49	0.19	0.33	0.25	0.21	0.14	0.13	0.094	0.084	1	0.35	0.14
ASSETS	0.19	0.71	0.47	0.46	0.25	0.6	0.45	0.37	0.14	0.21	0.11	0.12	0.35	1	0.062
PENSION	0.29	0.39	0.008	0.22	0.31	0.11	0.12	0.03	0.14	-0.031	0.058	0.12	0.14	0.062	1

Figure 4.13 Code and Heatmap

Based on figure 4.13, by using the python code above, the heatmap of the continuous variables is explored. Based on the output, it was found that the WBL index has high correlations with mobility, workplace, pay, marriage, parenthood, entrepreneurship and asset indicators. This brings information that these variables need to be dropped before model building since highly correlated variables have no meaning for model building, increasing the risk of errors and increasing the algorithm complexity.

4.3.8 Data Imbalance Exploration



```
df['Income Group'].value_counts()
```

```
High income             2964
Lower middle income     2860
Upper middle income     2704
Low income              1352
Name: Income Group, dtype: int64
```

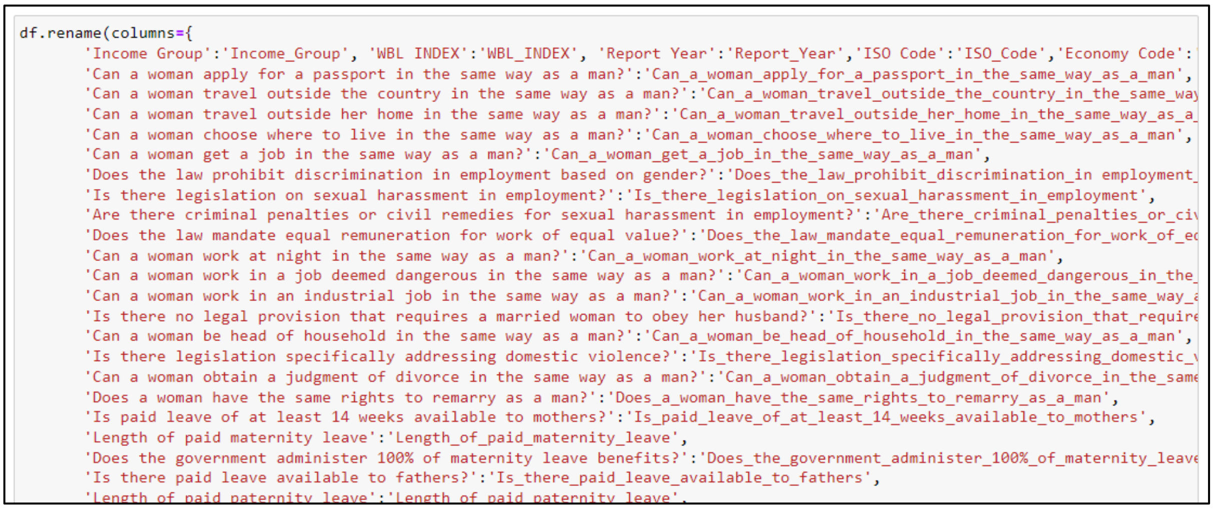
Figure 4.14 Code and Target Variable Category Exploration Output

Based on figure 4.14, by using the python code above, the category of the income group is explored to understand if there is an imbalance present in the dataset. Based on the output, it was found that there are 2964 high-income, 2860 lower-middle-income, 2704 upper-middle-income, and 1352 low-income. It is considered as no data imbalance issue happened, hence there is no data balancing technique needed to be implemented.

4.4 Data Pre-processing

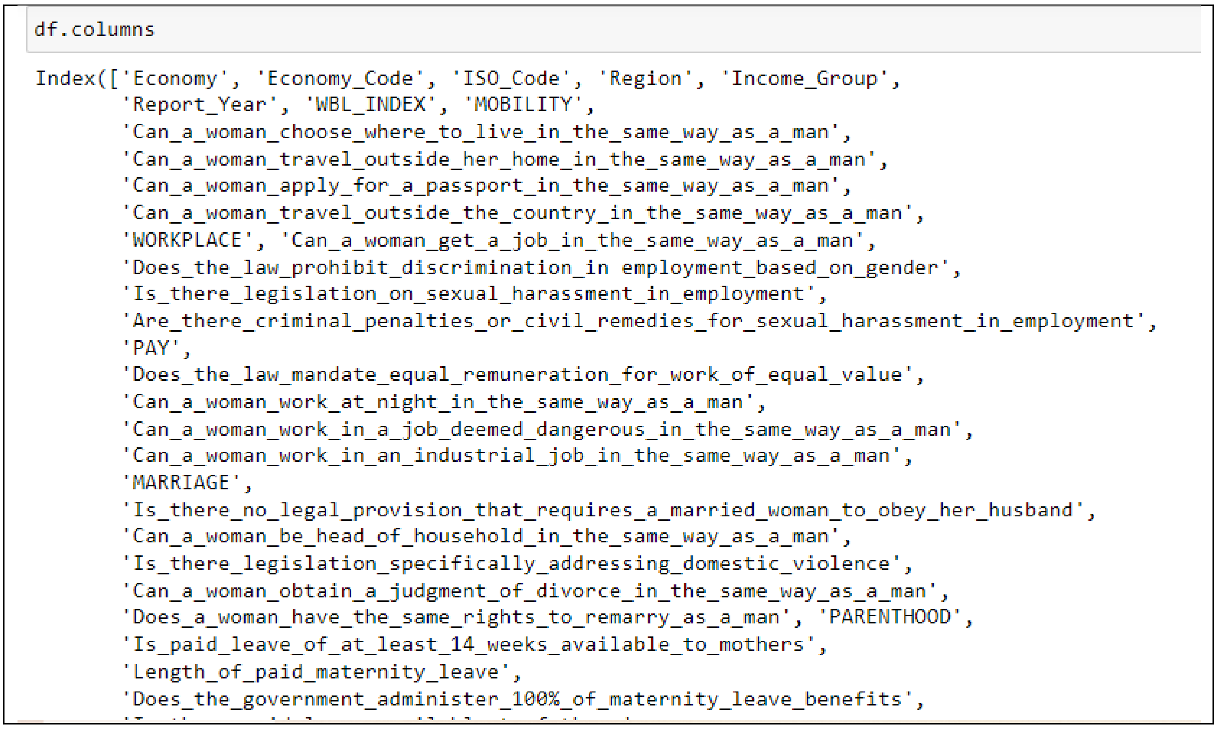
Data pre-processing is a method that can improve the accuracy and efficacy of the models by pre-processing the data to be used in an algorithm for computer system learning. Pre-processing is necessary to clean the actual information since it frequently contains noise, is sparse, or has an improper format that prohibits it from being utilized directly in learning algorithms. This stage involves performing data pre-processing tasks such as renaming columns, dropping useless columns, label encoding, removing outliers, and dataset splitting into the training set and testing set, dropping insignificant variables and saving the processed data into an excel file for model-building. Eventually, the dataset is ready for model creation and can be applied to supervised learning algorithms.

4.4.1 Rename Column



```
df.rename(columns={
        'Income Group':'Income_Group', 'WBL INDEX':'WBL_INDEX', 'Report Year':'Report_Year','ISO Code':'ISO_Code','Economy Code':
        'Can a woman apply for a passport in the same way as a man?':'Can_a_woman_apply_for_a_passport_in_the_same_way_as_a_man',
        'Can a woman travel outside the country in the same way as a man?':'Can_a_woman_travel_outside_the_country_in_the_same_way
        'Can a woman travel outside her home in the same way as a man?':'Can_a_woman_travel_outside_her_home_in_the_same_way_as_a_
        'Can a woman choose where to live in the same way as a man?':'Can_a_woman_choose_where_to_live_in_the_same_way_as_a_man',
        'Can a woman get a job in the same way as a man?':'Can_a_woman_get_a_job_in_the_same_way_as_a_man',
        'Does the law prohibit discrimination in employment based on gender?':'Does_the_law_prohibit_discrimination_in employment_
        'Is there legislation on sexual harassment in employment?':'Is_there_legislation_on_sexual_harassment_in_employment',
        'Are there criminal penalties or civil remedies for sexual harassment in employment?':'Are_there_criminal_penalties_or_civ
        'Does the law mandate equal remuneration for work of equal value?':'Does_the_law_mandate_equal_remuneration_for_work_of_eq
        'Can a woman work at night in the same way as a man?':'Can_a_woman_work_at_night_in_the_same_way_as_a_man',
        'Can a woman work in a job deemed dangerous in the same way as a man?':'Can_a_woman_work_in_a_job_deemed_dangerous_in_the_
        'Can a woman work in an industrial job in the same way as a man?':'Can_a_woman_work_in_an_industrial_job_in_the_same_way_a
        'Is there no legal provision that requires a married woman to obey her husband?':'Is_there_no_legal_provision_that_require
        'Can a woman be head of household in the same way as a man?':'Can_a_woman_be_head_of_household_in_the_same_way_as_a_man',
        'Is there legislation specifically addressing domestic violence?':'Is_there_legislation_specifically_addressing_domestic_v
        'Can a woman obtain a judgment of divorce in the same way as a man?':'Can_a_woman_obtain_a_judgment_of_divorce_in_the_same
        'Does a woman have the same rights to remarry as a man?':'Does_a_woman_have_the_same_rights_to_remarry_as_a_man',
        'Is paid leave of at least 14 weeks available to mothers?':'Is_paid_leave_of_at_least_14_weeks_available_to_mothers',
        'Length of paid maternity leave':'Length_of_paid_maternity_leave',
        'Does the government administer 100% of maternity leave benefits?':'Does_the_government_administer_100%_of_maternity_leave
        'Is there paid leave available to fathers?':'Is_there_paid_leave_available_to_fathers',
        'Length of paid paternity leave':'Length_of_paid_paternity_leave',
```

Figure 4.15 Rename the categorical variable



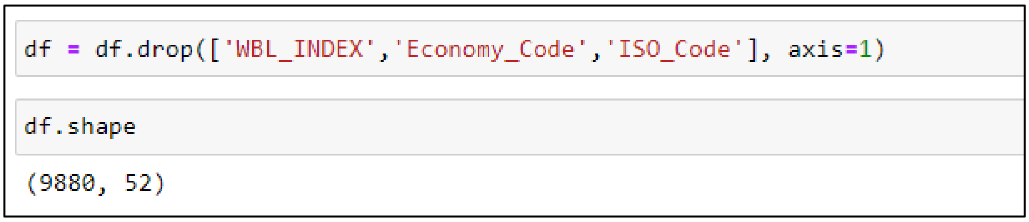
```
df.columns
```

```
Index(['Economy', 'Economy_Code', 'ISO_Code', 'Region', 'Income_Group',
       'Report_Year', 'WBL_INDEX', 'MOBILITY',
       'Can_a_woman_choose_where_to_live_in_the_same_way_as_a_man',
       'Can_a_woman_travel_outside_her_home_in_the_same_way_as_a_man',
       'Can_a_woman_apply_for_a_passport_in_the_same_way_as_a_man',
       'Can_a_woman_travel_outside_the_country_in_the_same_way_as_a_man',
       'WORKPLACE', 'Can_a_woman_get_a_job_in_the_same_way_as_a_man',
       'Does_the_law_prohibit_discrimination_in employment_based_on_gender',
       'Is_there_legislation_on_sexual_harassment_in_employment',
       'Are_there_criminal_penalties_or_civil_remedies_for_sexual_harassment_in_employment',
       'PAY',
       'Does_the_law_mandate_equal_remuneration_for_work_of_equal_value',
       'Can_a_woman_work_at_night_in_the_same_way_as_a_man',
       'Can_a_woman_work_in_a_job_deemed_dangerous_in_the_same_way_as_a_man',
       'Can_a_woman_work_in_an_industrial_job_in_the_same_way_as_a_man',
       'MARRIAGE',
       'Is_there_no_legal_provision_that_requires_a_married_woman_to_obey_her_husband',
       'Can_a_woman_be_head_of_household_in_the_same_way_as_a_man',
       'Is_there_legislation_specifically_addressing_domestic_violence',
       'Can_a_woman_obtain_a_judgment_of_divorce_in_the_same_way_as_a_man',
       'Does_a_woman_have_the_same_rights_to_remarry_as_a_man', 'PARENTHOOD',
       'Is_paid_leave_of_at_least_14_weeks_available_to_mothers',
       'Length_of_paid_maternity_leave',
       'Does_the_government_administer_100%_of_maternity_leave_benefits',
```

Figure 4.16 Check the renaming result

Firstly, the variable name is renamed, this is because when there is a blank between words in the variable name, it can easily cause errors and unable to call the variables. Hence, every blank between the word is replaced by a '_' to solve this problem as Figure 4.15. After renaming the variables, the variables are checked in list form. Based on the output in Figure 4.16, shows that the operation is successfully achieved.

4.4.2 Drop Useless Column



```
df = df.drop(['WBL_INDEX','Economy_Code','ISO_Code'], axis=1)
```

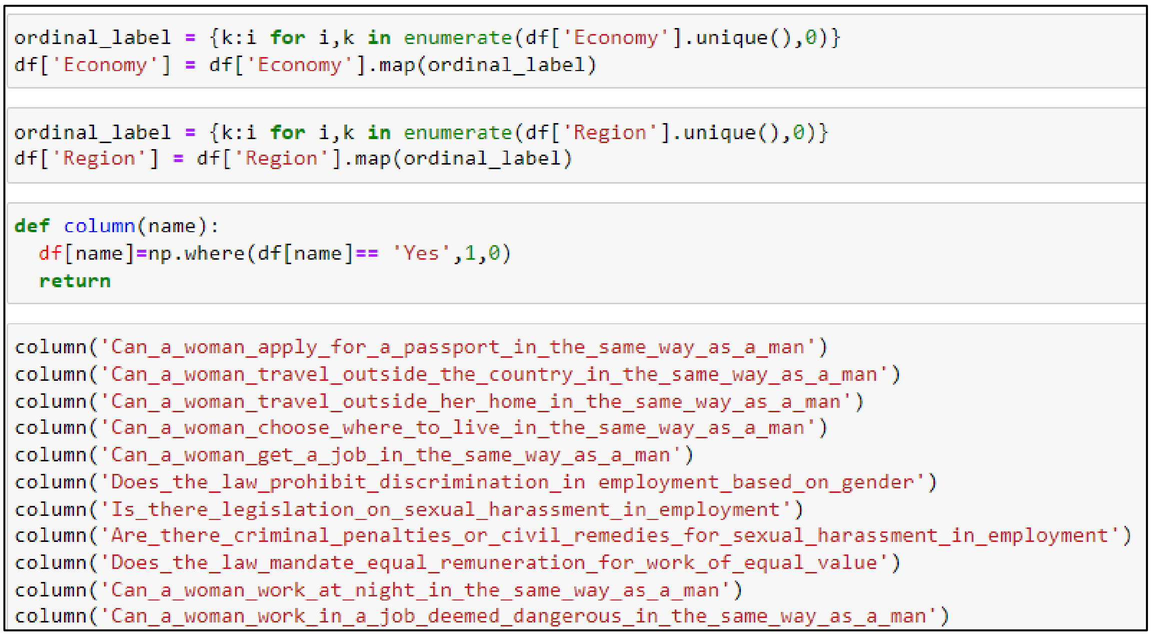
```
df.shape
```

```
(9880, 52)
```

Figure 4.17 Drop Useless Column

Based on the data exploratory analysis that was done previously, some of the useless variables are explored. For example, the ‘WBL index’ is highly correlated with other variables indicators during the correlation exploration. Besides, the ‘Economy Code’ and the ‘ISO_Code’ which has the same meaning as the ‘Economy’ variable. Hence, these variables needed to be dropped to prevent any bias and reduce the algorithm complexity.

4.4.3 Label Encoding



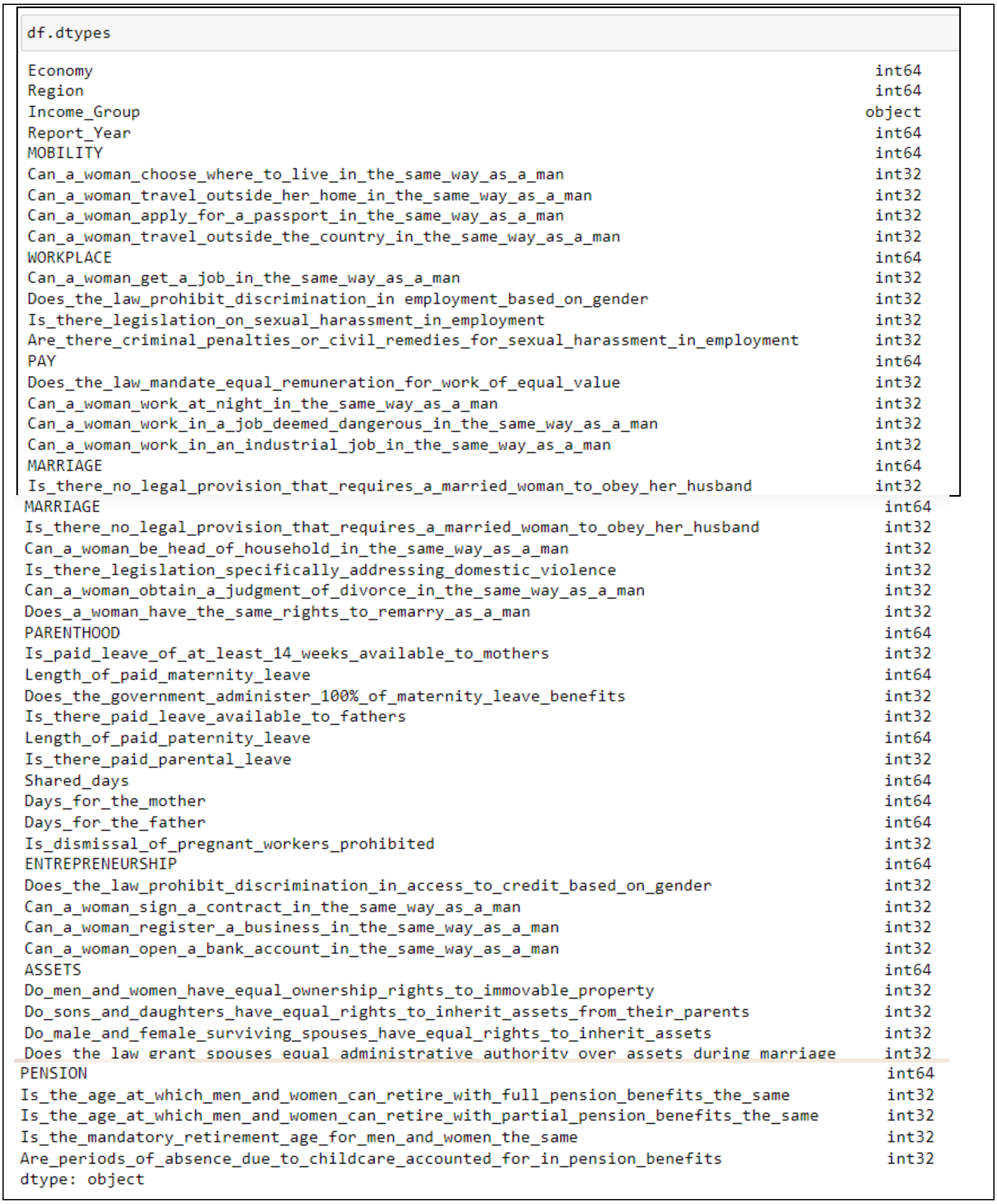
```
ordinal_label = {k:i for i,k in enumerate(df['Economy'].unique(),0)}
df['Economy'] = df['Economy'].map(ordinal_label)
```

```
ordinal_label = {k:i for i,k in enumerate(df['Region'].unique(),0)}
df['Region'] = df['Region'].map(ordinal_label)
```

```
def column(name):
  df[name]=np.where(df[name]== 'Yes',1,0)
  return
```

```
column('Can_a_woman_apply_for_a_passport_in_the_same_way_as_a_man')
column('Can_a_woman_travel_outside_the_country_in_the_same_way_as_a_man')
column('Can_a_woman_travel_outside_her_home_in_the_same_way_as_a_man')
column('Can_a_woman_choose_where_to_live_in_the_same_way_as_a_man')
column('Can_a_woman_get_a_job_in_the_same_way_as_a_man')
column('Does_the_law_prohibit_discrimination_in employment_based_on_gender')
column('Is_there_legislation_on_sexual_harassment_in_employment')
column('Are_there_criminal_penalties_or_civil_remedies_for_sexual_harassment_in_employment')
column('Does_the_law_mandate_equal_remuneration_for_work_of_equal_value')
column('Can_a_woman_work_at_night_in_the_same_way_as_a_man')
column('Can_a_woman_work_in_a_job_deemed_dangerous_in_the_same_way_as_a_man')
```

Figure 4.18 Label Encoding Operation Code



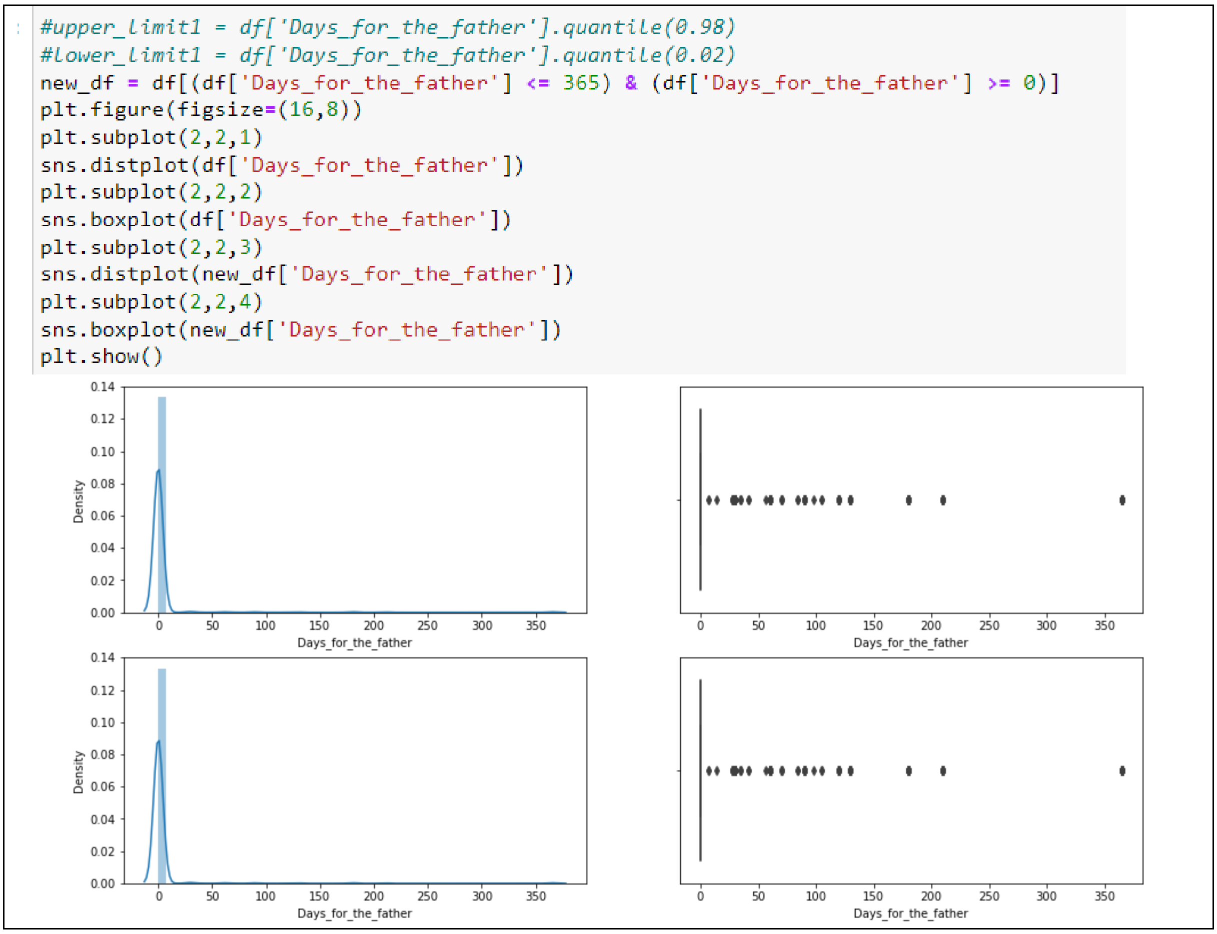
```
df.dtypes

Economy                                                                     int64
Region                                                                      int64
Income_Group                                                               object
Report_Year                                                                 int64
MOBILITY                                                                    int64
Can_a_woman_choose_where_to_live_in_the_same_way_as_a_man                   int32
Can_a_woman_travel_outside_her_home_in_the_same_way_as_a_man                int32
Can_a_woman_apply_for_a_passport_in_the_same_way_as_a_man                   int32
Can_a_woman_travel_outside_the_country_in_the_same_way_as_a_man             int32
WORKPLACE                                                                   int64
Can_a_woman_get_a_job_in_the_same_way_as_a_man                              int32
Does_the_law_prohibit_discrimination_in employment_based_on_gender          int32
Is_there_legislation_on_sexual_harassment_in_employment                     int32
Are_there_criminal_penalties_or_civil_remedies_for_sexual_harassment_in_employment    int32
PAY                                                                         int64
Does_the_law_mandate_equal_remuneration_for_work_of_equal_value             int32
Can_a_woman_work_at_night_in_the_same_way_as_a_man                          int32
Can_a_woman_work_in_a_job_deemed_dangerous_in_the_same_way_as_a_man         int32
Can_a_woman_work_in_an_industrial_job_in_the_same_way_as_a_man              int32
MARRIAGE                                                                    int64
Is_there_no_legal_provision_that_requires_a_married_woman_to_obey_her_husband    int32
MARRIAGE                                                                    int64
Is_there_no_legal_provision_that_requires_a_married_woman_to_obey_her_husband    int32
Can_a_woman_be_head_of_household_in_the_same_way_as_a_man                   int32
Is_there_legislation_specifically_addressing_domestic_violence              int32
Can_a_woman_obtain_a_judgment_of_divorce_in_the_same_way_as_a_man           int32
Does_a_woman_have_the_same_rights_to_remarry_as_a_man                       int32
PARENTHOOD                                                                  int64
Is_paid_leave_of_at_least_14_weeks_available_to_mothers                     int32
Length_of_paid_maternity_leave                                              int64
Does_the_government_administer_100%_of_maternity_leave_benefits             int32
Is_there_paid_leave_available_to_fathers                                    int32
Length_of_paid_paternity_leave                                              int64
Is_there_paid_parental_leave                                                int32
Shared_days                                                                 int64
Days_for_the_mother                                                         int64
Days_for_the_father                                                         int64
Is_dismissal_of_pregnant_workers_prohibited                                 int32
ENTREPRENEURSHIP                                                            int64
Does_the_law_prohibit_discrimination_in_access_to_credit_based_on_gender    int32
Can_a_woman_sign_a_contract_in_the_same_way_as_a_man                        int32
Can_a_woman_register_a_business_in_the_same_way_as_a_man                    int32
Can_a_woman_open_a_bank_account_in_the_same_way_as_a_man                    int32
ASSETS                                                                      int64
Do_men_and_women_have_equal_ownership_rights_to_immovable_property          int32
Do_sons_and_daughters_have_equal_rights_to_inherit_assets_from_their_parents    int32
Do_male_and_female_surviving_spouses_have_equal_rights_to_inherit_assets    int32
Does the law grant spouses equal administrative authority over assets during marriage    int32
PENSION                                                                     int64
Is_the_age_at_which_men_and_women_can_retire_with_full_pension_benefits_the_same    int32
Is_the_age_at_which_men_and_women_can_retire_with_partial_pension_benefits_the_same    int32
Is_the_mandatory_retirement_age_for_men_and_women_the_same                  int32
Are_periods_of_absence_due_to_childcare_accounted_for_in_pension_benefits   int32
dtype: object
```

Figure 4.19 Label Encoding Operation Result

Next, label encoding is done to convert the categorical variables into numerical variables. Label encoding is the process of transforming tags into numerical forms so that they may be handled by machines. The operation among those tags can then be accurately determined by machine learning techniques. It is a significant supervised learning pre-processing phase for the structured data. All the categorical variables are converted into numeric variables except the target variable since the model is able to determine categorical variables in the target variable.

4.4.4 Remove Outlier

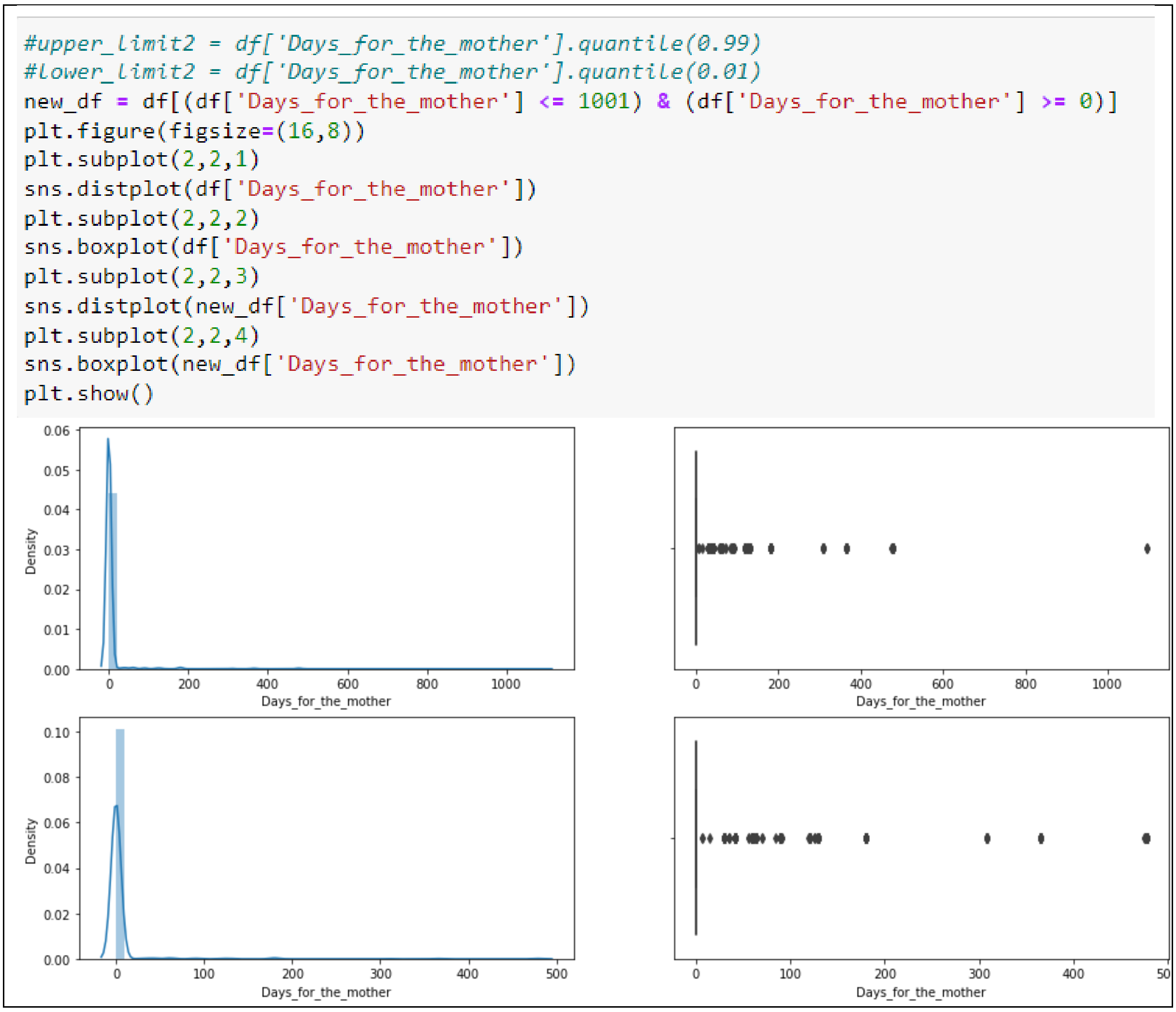
A finding that differs abnormally from other results in a population-based random selection is referred to as an outlier. Some outliers in the data reflect normal population variance and ought to be left free. They are referred to as real outliers. However, the false outlier needs to be removed because they indicate uncertainty, processing flaws, or inadequate sampling. Outliers make the data more variable, which reduces statistical significance. Therefore, eliminating outliers can make the findings provide accurate results. Based on the exploratory data analysis, there are a few variables that consist of heavy outliers and skewness. These variables included: Days_for_the_father, Days_for_the_mother, Shared_days, Length_of_paid_paternity_leave' and Length_of_paid_maternity_leave'.



```
#upper_limit1 = df['Days_for_the_father'].quantile(0.98)
#lower_limit1 = df['Days_for_the_father'].quantile(0.02)
new_df = df[(df['Days_for_the_father'] <= 365) & (df['Days_for_the_father'] >= 0)]
plt.figure(figsize=(16,8))
plt.subplot(2,2,1)
sns.distplot(df['Days_for_the_father'])
plt.subplot(2,2,2)
sns.boxplot(df['Days_for_the_father'])
plt.subplot(2,2,3)
sns.distplot(new_df['Days_for_the_father'])
plt.subplot(2,2,4)
sns.boxplot(new_df['Days_for_the_father'])
plt.show()
```

Figure 4.20 Outlier removal in the 'Days_for_the_father'

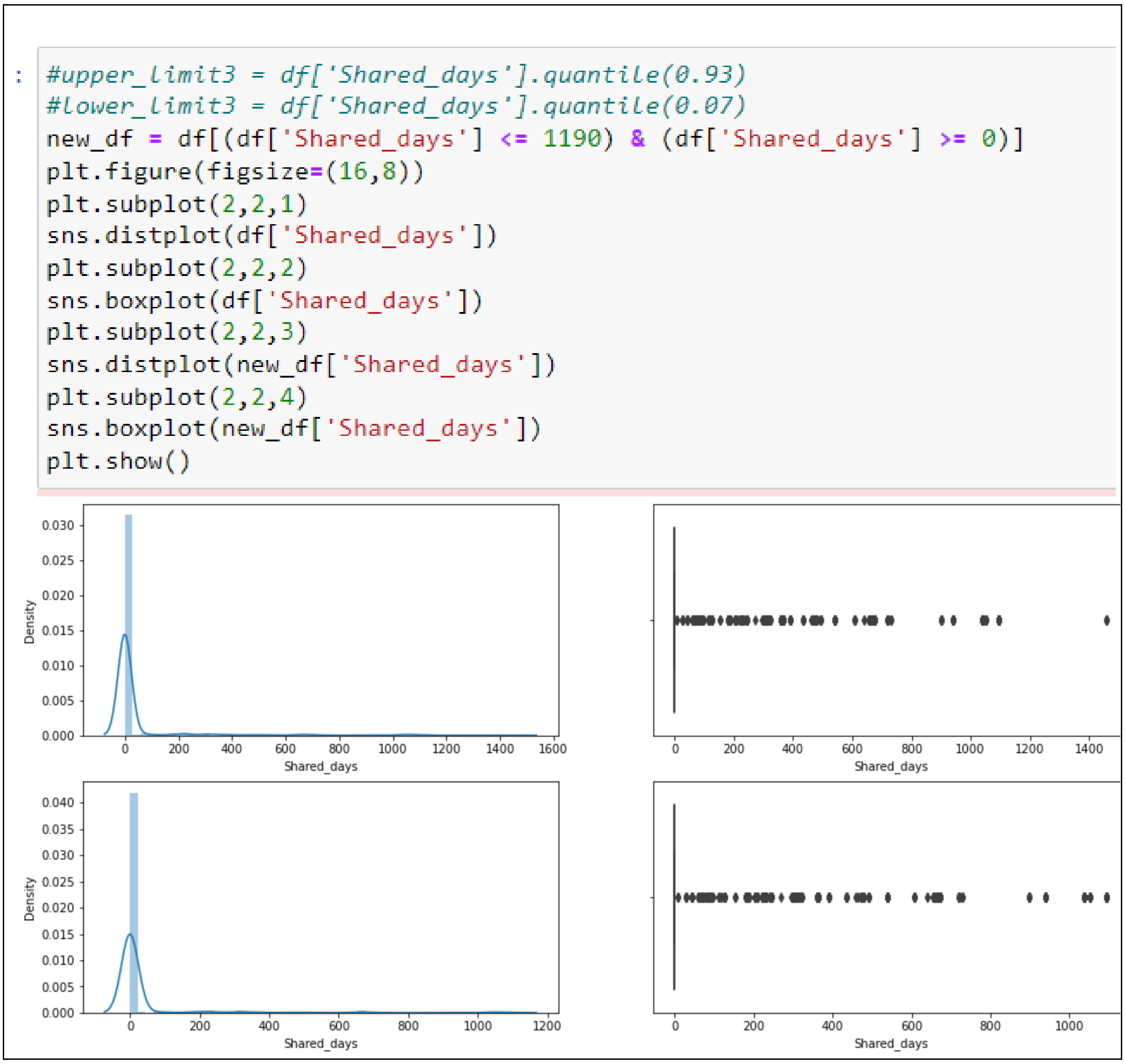
Figure 4.20 shows the coding for outlier removal and the output shows the before and after for ‘Days_for_the_father’. Although there are many outliers presented in the variable, most of them are reasonable since there are many different countries presented in the dataset, hence it is possible to have a big range between the value for parental leave. Based on the OECD family database, it was found that the country that gives the highest paid parental and home care leave reserved for fathers is Korea and Japan which is around 364 days while there are some companies that do not give any leave to father such as Australia, Columbia, Finland, Denmark, Estonia, Costa Rica, and Czech Republic. Hence, in the outlier removal, the upper limit is set as 365 days while the lower limit is set as 0 days. The result shows that it does not much difference between the before and after processing. This indicates that all the outliers are true and should not be removed.



```
#upper_limit2 = df['Days_for_the_mother'].quantile(0.99)
#lower_limit2 = df['Days_for_the_mother'].quantile(0.01)
new_df = df[(df['Days_for_the_mother'] <= 1001) & (df['Days_for_the_mother'] >= 0)]
plt.figure(figsize=(16,8))
plt.subplot(2,2,1)
sns.distplot(df['Days_for_the_mother'])
plt.subplot(2,2,2)
sns.boxplot(df['Days_for_the_mother'])
plt.subplot(2,2,3)
sns.distplot(new_df['Days_for_the_mother'])
plt.subplot(2,2,4)
sns.boxplot(new_df['Days_for_the_mother'])
plt.show()
```

Figure 4.21 Outlier removal in the ‘Days_for_the_mother’

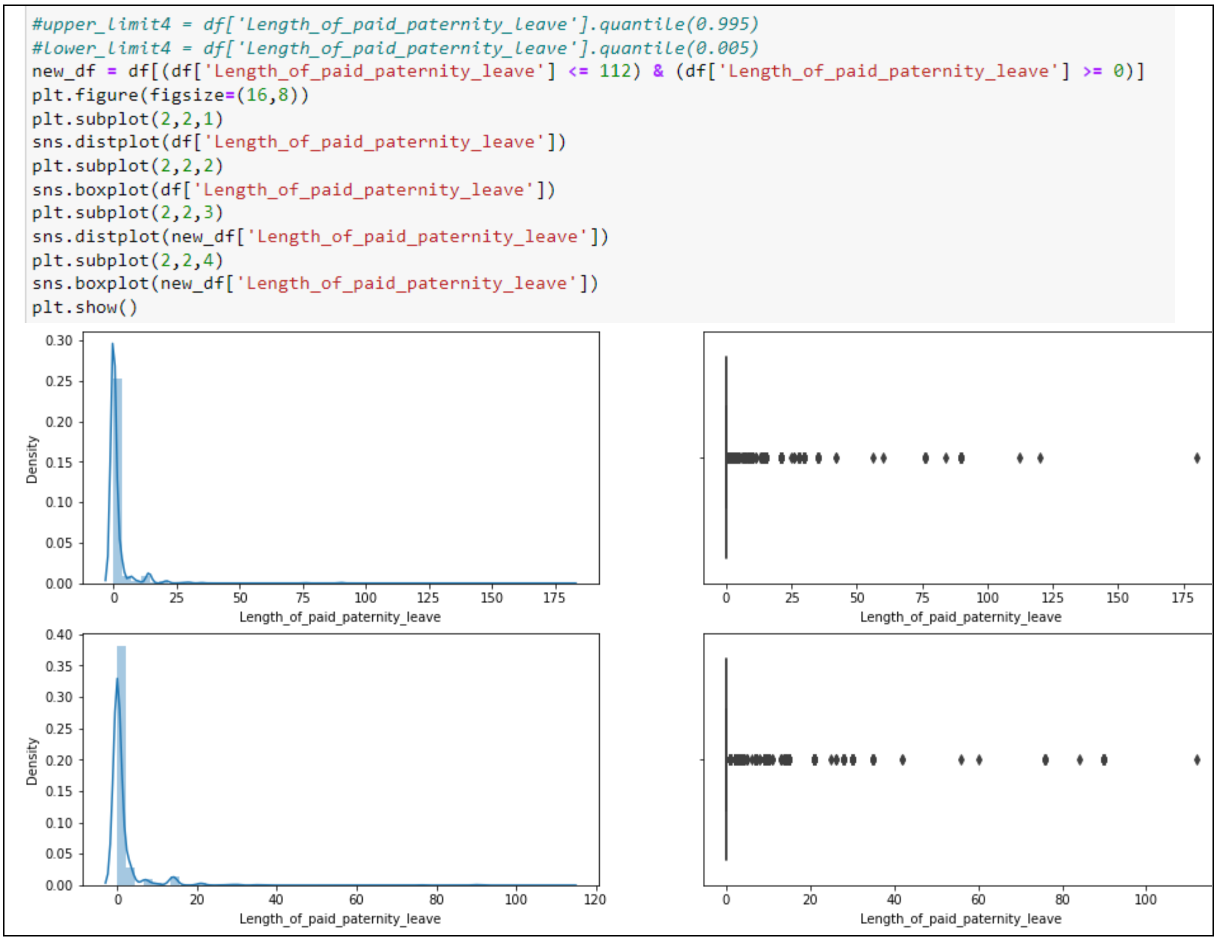
Figure 4.21 shows the coding for outlier removal and the output shows the before and after for ‘Days_for_the_mother’. Based on the OECD family database, it was found that the country that gives the highest paid parental and home care leave reserved for mothers is Finland which is around 1001 days while there are some companies that do not give any leave to fathers such as Spain, Malta, Cyprus, Greece and etc. Hence, in the outlier removal, the upper limit is set as 1001 days while the lower limit is set as 0 days. The result shows that the range is changed from more than 1400 to 1001. This indicates that all the false outliers are removed. Although there are still outlier exits, these are true outliers.



```
#upper_limit3 = df['Shared_days'].quantile(0.93)
#lower_limit3 = df['Shared_days'].quantile(0.07)
new_df = df[(df['Shared_days'] <= 1190) & (df['Shared_days'] >= 0)]
plt.figure(figsize=(16,8))
plt.subplot(2,2,1)
sns.distplot(df['Shared_days'])
plt.subplot(2,2,2)
sns.boxplot(df['Shared_days'])
plt.subplot(2,2,3)
sns.distplot(new_df['Shared_days'])
plt.subplot(2,2,4)
sns.boxplot(new_df['Shared_days'])
plt.show()
```

Figure 4.22 Outlier Removal in the ‘Shared_days’

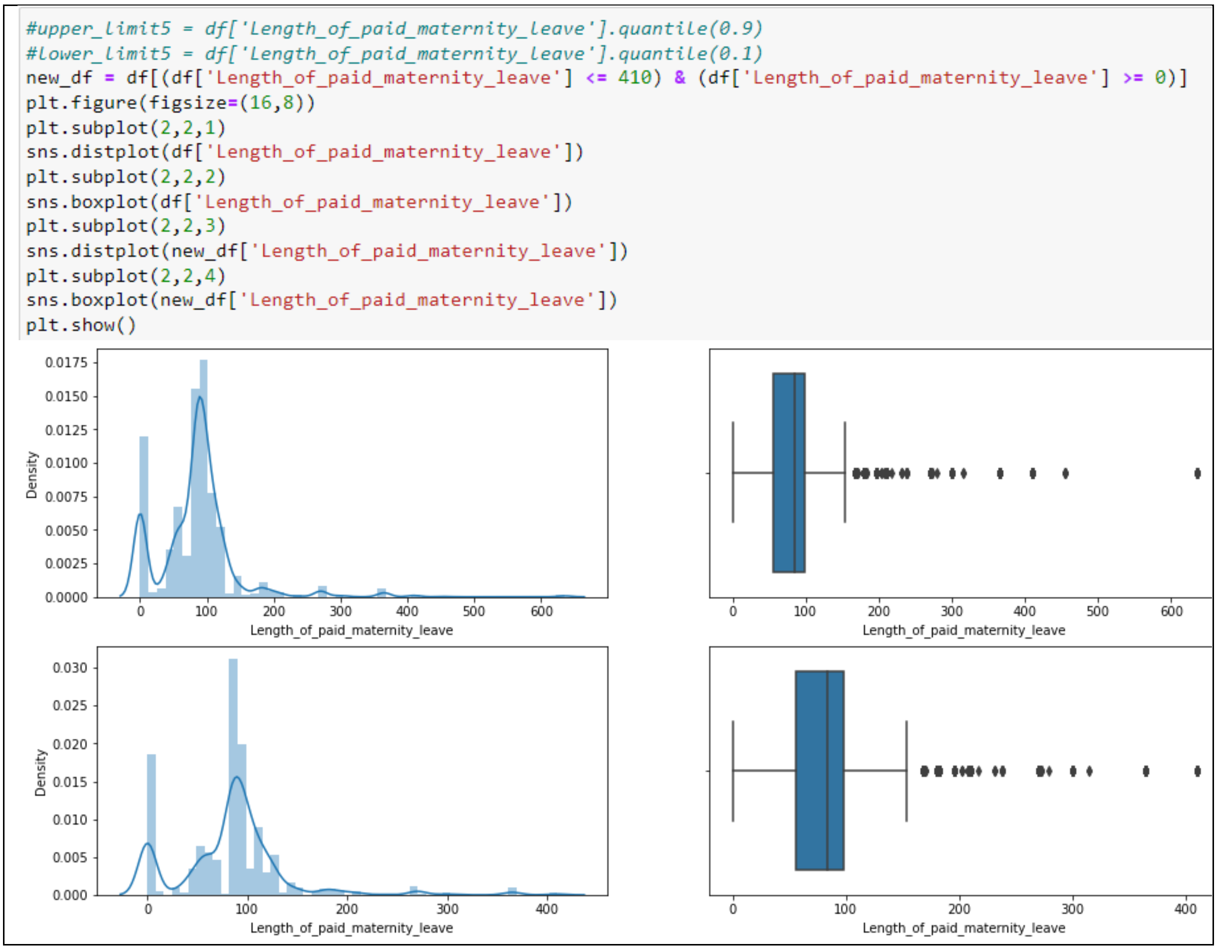
Figure 4.22 shows the coding for outlier removal and the output shows the before and after for ‘Shared days’. Based on the OECD family database, it was found that the highest number of shared days is 1190 days while there are some companies that do not give any leave. Hence, in the outlier removal, the upper limit is set as 1190 days while the lower limit is set as 0 days. The result shows that the range changed from more than 1400 to 1190. This indicates that all the false outliers are removed. Although there are still outlier exits, these are true outliers.



```
#upper_limit4 = df['Length_of_paid_paternity_leave'].quantile(0.995)
#lower_limit4 = df['Length_of_paid_paternity_leave'].quantile(0.005)
new_df = df[(df['Length_of_paid_paternity_leave'] <= 112) & (df['Length_of_paid_paternity_leave'] >= 0)]
plt.figure(figsize=(16,8))
plt.subplot(2,2,1)
sns.distplot(df['Length_of_paid_paternity_leave'])
plt.subplot(2,2,2)
sns.boxplot(df['Length_of_paid_paternity_leave'])
plt.subplot(2,2,3)
sns.distplot(new_df['Length_of_paid_paternity_leave'])
plt.subplot(2,2,4)
sns.boxplot(new_df['Length_of_paid_paternity_leave'])
plt.show()
```

Figure 4.23 Outlier Removal in the ‘Length_of_paid_paternity_leave’

Figure 4.23 shows the coding for outlier removal and the output shows the before and after for ‘Length_of_paid_paternity_leave’. Based on the OECD family database, it was found that the country that gives the highest length of paid paternity leave is Spain which is around 112 days while there are some companies that do not give any paternity leave to fathers such as Germany, Japan, United State and etc. Hence, in the outlier removal, the upper limit is set as 112 days while the lower limit is set as 0 days. The result shows that the range changed from more than 175 to 112. This indicates that all the false outliers are removed. Although there are still outlier exits, these are true outliers.

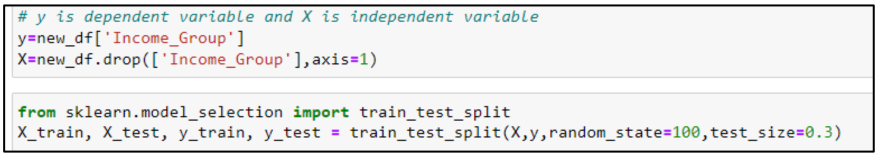


```
#upper_limit5 = df['Length_of_paid_maternity_leave'].quantile(0.9)
#lower_limit5 = df['Length_of_paid_maternity_leave'].quantile(0.1)
new_df = df[(df['Length_of_paid_maternity_leave'] <= 410) & (df['Length_of_paid_maternity_leave'] >= 0)]
plt.figure(figsize=(16,8))
plt.subplot(2,2,1)
sns.distplot(df['Length_of_paid_maternity_leave'])
plt.subplot(2,2,2)
sns.boxplot(df['Length_of_paid_maternity_leave'])
plt.subplot(2,2,3)
sns.distplot(new_df['Length_of_paid_maternity_leave'])
plt.subplot(2,2,4)
sns.boxplot(new_df['Length_of_paid_maternity_leave'])
plt.show()
```

Figure 4.24 Outlier removal in the ‘Length_of_paid_maternity_leave’

Figure 4.24 shows the coding for outlier removal and the output shows the before and after for ‘Length_of_paid_maternity_leave’. Based on the OECD family database, it was found that the country that gives the highest length of paid maternity leave is Bulgaria which is around 410 days while there are some companies that do not give any paid maternity leave to mothers such as the United States. Hence, in the outlier removal, the upper limit is set as 410 days while the lower limit is set as 0 days. The result shows that the range changed from more than 600 to 410. This indicates that all the false outliers are removed. Although there are still outlier exits, these are true outliers.

4.4.5 Train test splitting

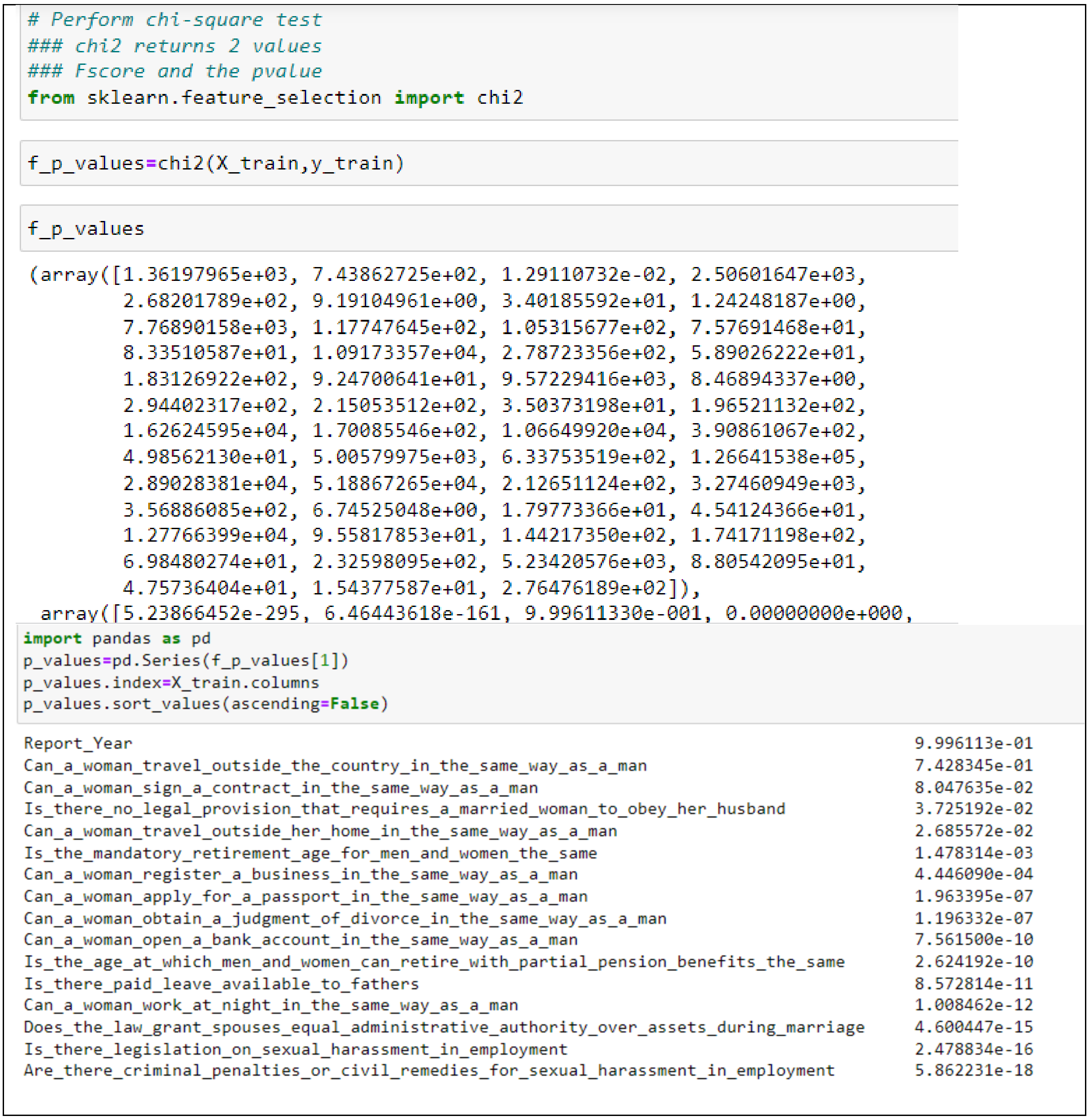


```
# y is dependent variable and X is independent variable
y=new_df['Income_Group']
X=new_df.drop(['Income_Group'],axis=1)

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=100,test_size=0.3)
```

Figure 4.25 Train Test Splitting

Further, the dataset is then split into training and testing sets. A model evaluation technique called train-test-split enables researchers to mimic how well a classifier could behave on untested data. The training set is split into 70% while the testing set is 30%. The training set is divided into X_train and y_train while the testing set is divided into X_test and y_test.

4.4.6 Chi-square Test



```
# Perform chi-square test
### chi2 returns 2 values
### Fscore and the pvalue
from sklearn.feature_selection import chi2
```

```
f_p_values=chi2(X_train,y_train)
```

```
f_p_values
```

```
(array([1.36197965e+03, 7.43862725e+02, 1.29110732e-02, 2.50601647e+03,
        2.68201789e+02, 9.19104961e+00, 3.40185592e+01, 1.24248187e+00,
        7.76890158e+03, 1.17747645e+02, 1.05315677e+02, 7.57691468e+01,
        8.33510587e+01, 1.09173357e+04, 2.78723356e+02, 5.89026222e+01,
        1.83126922e+02, 9.24700641e+01, 9.57229416e+03, 8.46894337e+00,
        2.94402317e+02, 2.15053512e+02, 3.50373198e+01, 1.96521132e+02,
        1.62624595e+04, 1.70085546e+02, 1.06649920e+04, 3.90861067e+02,
        4.98562130e+01, 5.00579975e+03, 6.33753519e+02, 1.26641538e+05,
        2.89028381e+04, 5.18867265e+04, 2.12651124e+02, 3.27460949e+03,
        3.56886085e+02, 6.74525048e+00, 1.79773366e+01, 4.54124366e+01,
        1.27766399e+04, 9.55817853e+01, 1.44217350e+02, 1.74171198e+02,
        6.98480274e+01, 2.32598095e+02, 5.23420576e+03, 8.80542095e+01,
        4.75736404e+01, 1.54377587e+01, 2.76476189e+02]),
 array([5.23866452e-295, 6.46443618e-161, 9.99611330e-001, 0.00000000e+000,
```

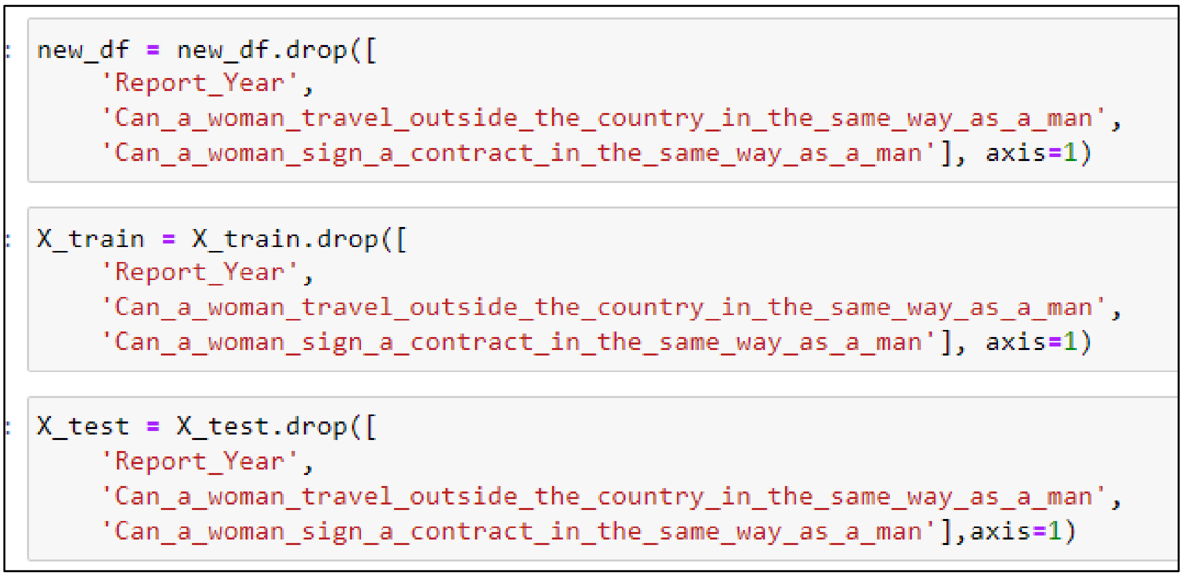
```
import pandas as pd
p_values=pd.Series(f_p_values[1])
p_values.index=X_train.columns
p_values.sort_values(ascending=False)
```

```
Report_Year                                                                9.996113e-01
Can_a_woman_travel_outside_the_country_in_the_same_way_as_a_man           7.428345e-01
Can_a_woman_sign_a_contract_in_the_same_way_as_a_man                      8.047635e-02
Is_there_no_legal_provision_that_requires_a_married_woman_to_obey_her_husband  3.725192e-02
Can_a_woman_travel_outside_her_home_in_the_same_way_as_a_man              2.685572e-02
Is_the_mandatory_retirement_age_for_men_and_women_the_same                1.478314e-03
Can_a_woman_register_a_business_in_the_same_way_as_a_man                  4.446090e-04
Can_a_woman_apply_for_a_passport_in_the_same_way_as_a_man                 1.963395e-07
Can_a_woman_obtain_a_judgment_of_divorce_in_the_same_way_as_a_man         1.196332e-07
Can_a_woman_open_a_bank_account_in_the_same_way_as_a_man                  7.561500e-10
Is_the_age_at_which_men_and_women_can_retire_with_partial_pension_benefits_the_same  2.624192e-10
Is_there_paid_leave_available_to_fathers                                  8.572814e-11
Can_a_woman_work_at_night_in_the_same_way_as_a_man                        1.008462e-12
Does_the_law_grant_spouses_equal_administrative_authority_over_assets_during_marriage  4.600447e-15
Is_there_legislation_on_sexual_harassment_in_employment                   2.478834e-16
Are_there_criminal_penalties_or_civil_remedies_for_sexual_harassment_in_employment  5.862231e-18
```

Figure 4.26 Chi-squared Test

Whenever the sampling weights are significant, chi-squared is a type of statistical method employed in the study of contingency tables. To put it another way, the main purpose of this test is to determine whether two category factors have independent effects on the test statistic. Based on figure 4.26, the chi-squared test is carried out to filter insignificant variables. The P-value of each variable is shown in the output. When the p-value is more than 0.05, it is considered an insignificant variable. Based on the output that is shown in figure 4.25, it shows that there are 3 categorical variables that are not significant, which included: ‘Report_Year’, ‘Can_a_woman_travel_outside_the_country_in_the_same_way_as_man’, ‘Can_a_woman_sign_a_contract_in_the_same_way_as_man’. The insignificant variable does not bring value to the model while increasing the algorithm complexity. Hence, to avoid these issues, the insignificant variables need to be dropped in the next step.

4.4.7 Drop insignificant categorical variable



```
new_df = new_df.drop([
    'Report_Year',
    'Can_a_woman_travel_outside_the_country_in_the_same_way_as_a_man',
    'Can_a_woman_sign_a_contract_in_the_same_way_as_a_man'], axis=1)
```

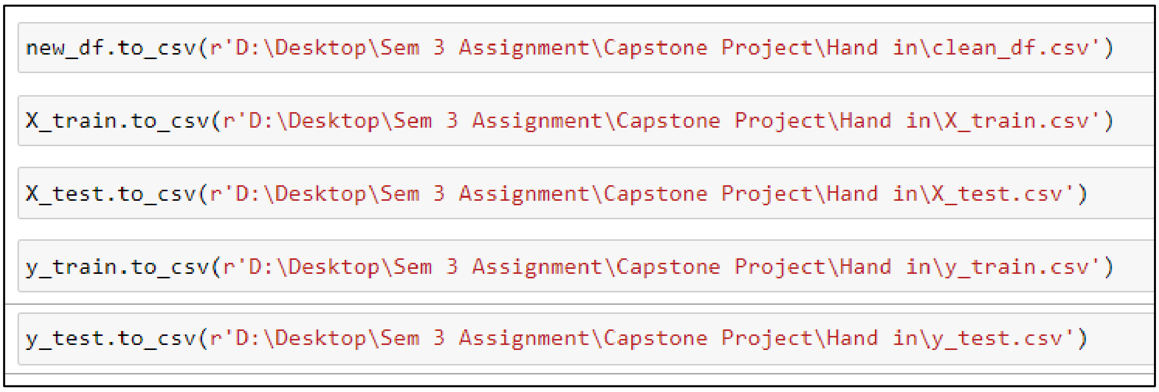
```
X_train = X_train.drop([
    'Report_Year',
    'Can_a_woman_travel_outside_the_country_in_the_same_way_as_a_man',
    'Can_a_woman_sign_a_contract_in_the_same_way_as_a_man'], axis=1)
```

```
X_test = X_test.drop([
    'Report_Year',
    'Can_a_woman_travel_outside_the_country_in_the_same_way_as_a_man',
    'Can_a_woman_sign_a_contract_in_the_same_way_as_a_man'],axis=1)
```

Figure 4.27 Dropped insignificant variables

After exploring the insignificant variables from the chi-square test, the insignificant variables that were explored are dropped in the new_df, X_train and X_test data frame.

4.4.8 Save data frame into csv file



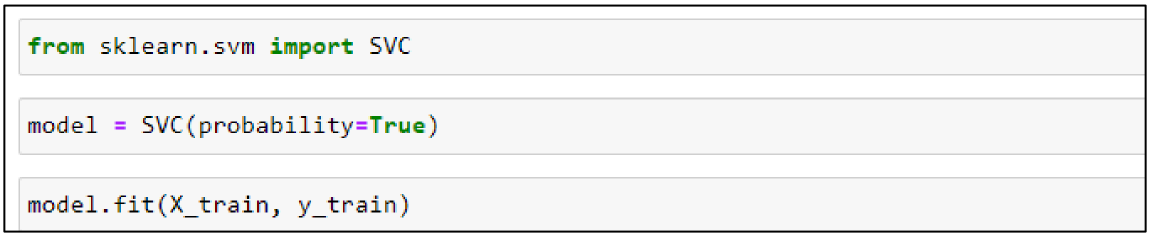
```
new_df.to_csv(r'D:\Desktop\Sem 3 Assignment\Capstone Project\Hand in\clean_df.csv')

X_train.to_csv(r'D:\Desktop\Sem 3 Assignment\Capstone Project\Hand in\X_train.csv')

X_test.to_csv(r'D:\Desktop\Sem 3 Assignment\Capstone Project\Hand in\X_test.csv')

y_train.to_csv(r'D:\Desktop\Sem 3 Assignment\Capstone Project\Hand in\y_train.csv')

y_test.to_csv(r'D:\Desktop\Sem 3 Assignment\Capstone Project\Hand in\y_test.csv')
```

Figure 4.28 Dropped Insignificant Variables

Last but not least, the cleaned data frame, X_train, y_train, X_test and y_test data frames are saved as a csv file for further usage in model building.

4.5 Model Building

4.5.1 Based Support Vector Machine (SVM)

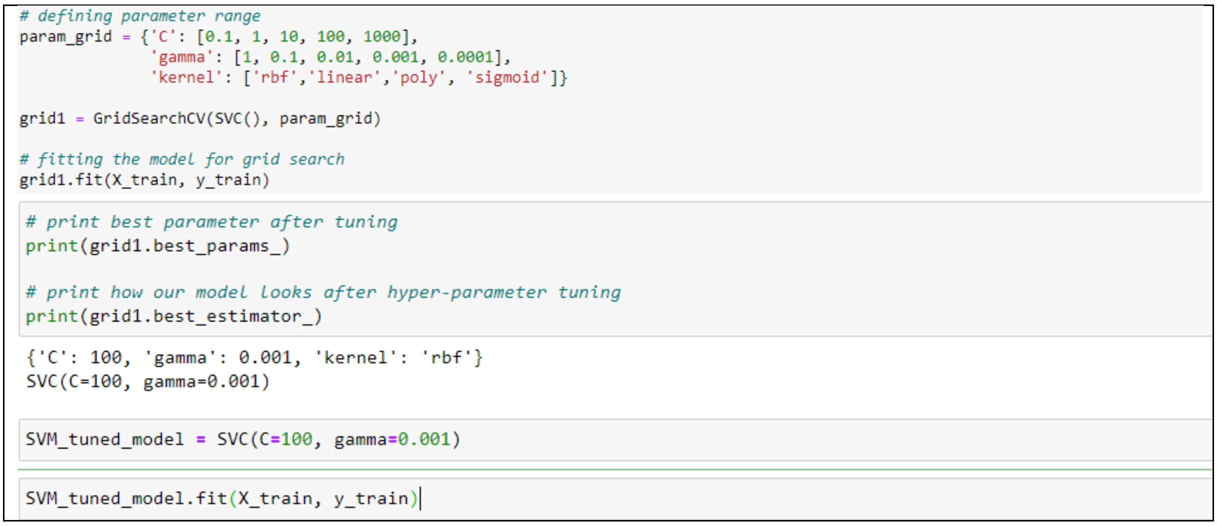


```
from sklearn.svm import SVC

model = SVC(probability=True)

model.fit(X_train, y_train)
```

Figure 4.29 Based Support Vector Machine Building

Before the support vector machine model building, the necessary library which is SVC is imported from the sklearn.svm to build the support vector machine model. The model is built and named a ‘model’ in this step. The model that was built is then fit with the training set which is X_train and y_train. In this based model, all the parameters are using the default setting without any modification.

4.5.2 Tuned Support Vector Machine



```
# defining parameter range
param_grid = {'C': [0.1, 1, 10, 100, 1000],
              'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
              'kernel': ['rbf','linear','poly', 'sigmoid']}

grid1 = GridSearchCV(SVC(), param_grid)

# fitting the model for grid search
grid1.fit(X_train, y_train)
```

```
# print best parameter after tuning
print(grid1.best_params_)

# print how our model looks after hyper-parameter tuning
print(grid1.best_estimator_)
```

```
{'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}
SVC(C=100, gamma=0.001)
```

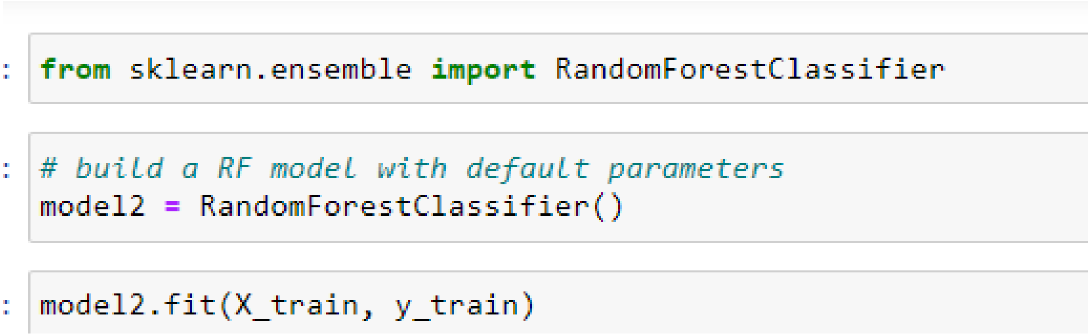
```
SVM_tuned_model = SVC(C=100, gamma=0.001)
```

```
SVM_tuned_model.fit(X_train, y_train)
```

Figure 4.30 Tuned Support Vector Machine Building

After the based support vector machine model building, the model is then tuned to obtain a better result. The hyperparameter tuning technique that was decided in this tuning process is Grid Search. This is because the optimum method for tuning the hyperparameters based on the training set will be discovered after an intensive search. Before the hyperparameter tuning process, the Grid Search library is imported from the sklearn.model_selection. The hyperparameter that was tuned in this model is ‘C’, ‘gamma’ and ‘kernel’. Based on the result, it shows that 100 C and 0.001 gamma with the rbf kernel achieved the highest performance compared with other values. Then, by using this information, a new tuned model is built and named ‘SVM_tuned_model’. The model is then fit into the training set.

4.5.3 Based Random Forest (RF)



```
from sklearn.ensemble import RandomForestClassifier
```

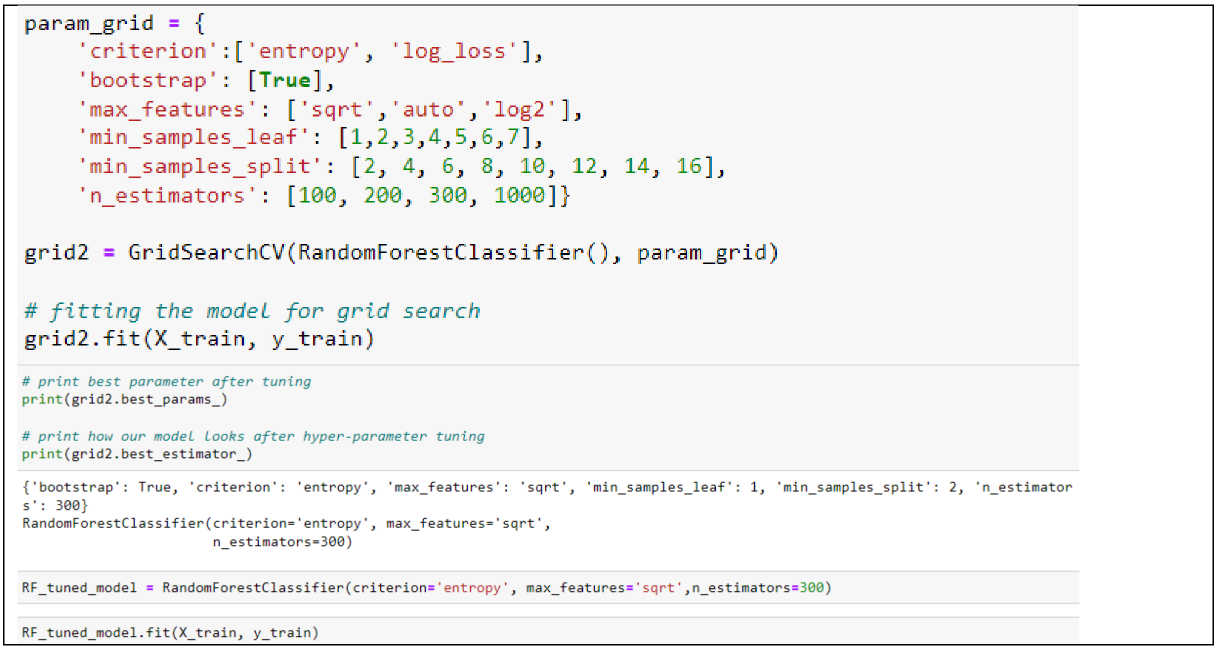
```
# build a RF model with default parameters
model2 = RandomForestClassifier()
```

```
model2.fit(X_train, y_train)
```

Figure 4.31 Based Random Forest

Before the random forest model building, the necessary library which is RandomForestClassifier is imported from the sklearn.ensemble to build the random forest model. The model is built and named ‘model2’ in this step. The model that was built is then fit with the training set which is X_train and y_train. In this based model, all the parameters are using the default setting.

4.5.4 Tuned Random Forest



```
param_grid = {
    'criterion':['entropy', 'log_loss'],
    'bootstrap': [True],
    'max_features': ['sqrt','auto','log2'],
    'min_samples_leaf': [1,2,3,4,5,6,7],
    'min_samples_split': [2, 4, 6, 8, 10, 12, 14, 16],
    'n_estimators': [100, 200, 300, 1000]}

grid2 = GridSearchCV(RandomForestClassifier(), param_grid)

# fitting the model for grid search
grid2.fit(X_train, y_train)
```

```
# print best parameter after tuning
print(grid2.best_params_)

# print how our model looks after hyper-parameter tuning
print(grid2.best_estimator_)
```

```
{'bootstrap': True, 'criterion': 'entropy', 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimator
s': 300}
RandomForestClassifier(criterion='entropy', max_features='sqrt',
                       n_estimators=300)
```

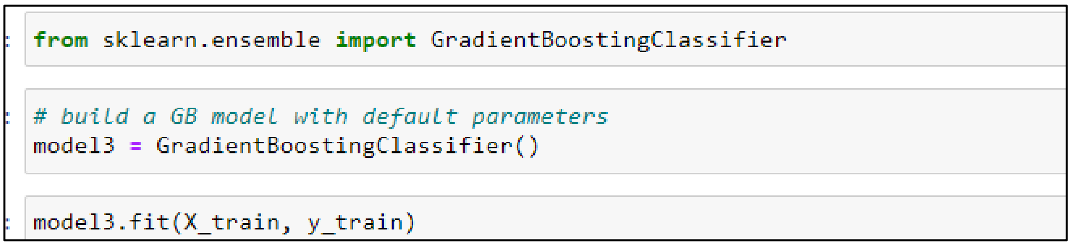
```
RF_tuned_model = RandomForestClassifier(criterion='entropy', max_features='sqrt',n_estimators=300)
```

```
RF_tuned_model.fit(X_train, y_train)
```

Figure 4.32 Tuned random forest

After the based random forest model building, the model is then tuned to obtain a better result. The hyperparameter tuning technique that was used in this tuning process is Grid Search. The hyperparameters that were tuned in this model are ‘criterion’, ’bootstrap’, ‘max_features’, ‘min_samples_leaf’, ‘minsamples_split’ and ‘n_estimators. Based on the result, it shows that entropy criterion, sqrt max features, and 300 n estimators with other default parameters setting are the best parameter grouping to build a better model. Then, by using this information, a new tuned model is built and named ‘RF_tuned_model’. The model is then fit into the training set.

4.5.5 Based Gradient Boosting (GB)



```
from sklearn.ensemble import GradientBoostingClassifier
```

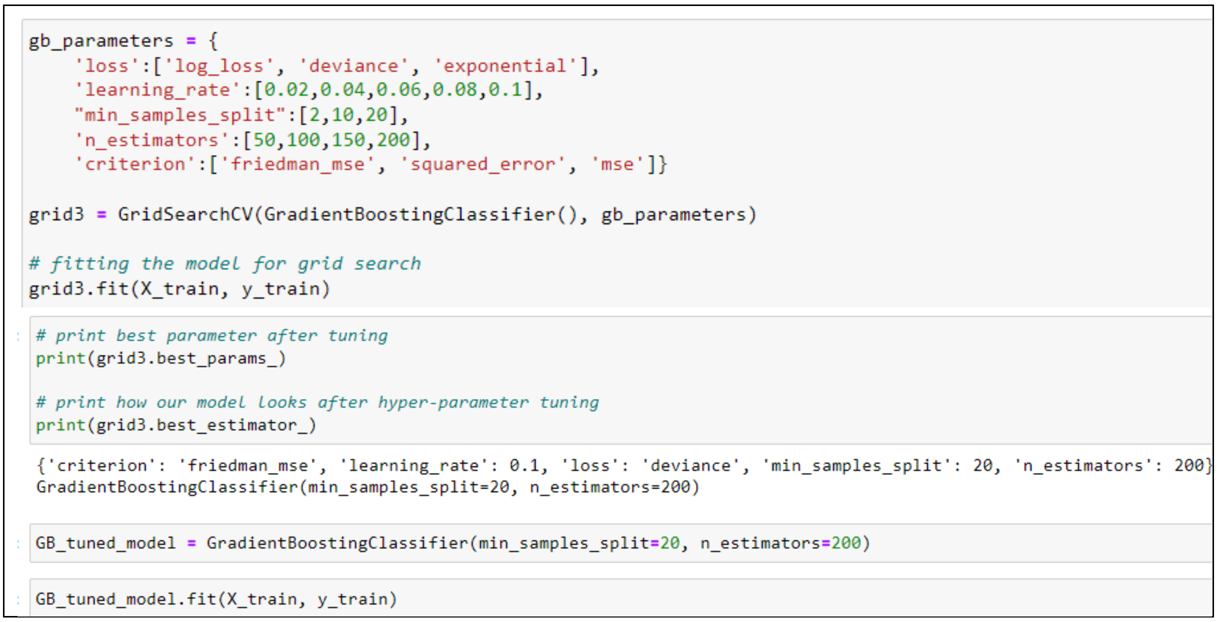
```
# build a GB model with default parameters
model3 = GradientBoostingClassifier()
```

```
model3.fit(X_train, y_train)
```

Figure 4.33 Based Gradient Boosting

Before the gradient boosting model building, the necessary library which is GradientBoostingClassifier is imported from the sklearn.ensemble to build the gradient boosting model. The model is built and named ‘model3’ in this step. The model that was built is then fit with the training set which is X_train and y_train. In this based model, all the parameters are using the default setting.

4.5.6 Tuned Gradient Boosting



```
gb_parameters = {
    'loss':['log_loss', 'deviance', 'exponential'],
    'learning_rate':[0.02,0.04,0.06,0.08,0.1],
    "min_samples_split":[2,10,20],
    'n_estimators':[50,100,150,200],
    'criterion':['friedman_mse', 'squared_error', 'mse']}

grid3 = GridSearchCV(GradientBoostingClassifier(), gb_parameters)

# fitting the model for grid search
grid3.fit(X_train, y_train)
```

```
# print best parameter after tuning
print(grid3.best_params_)

# print how our model looks after hyper-parameter tuning
print(grid3.best_estimator_)
```

```
{'criterion': 'friedman_mse', 'learning_rate': 0.1, 'loss': 'deviance', 'min_samples_split': 20, 'n_estimators': 200}
GradientBoostingClassifier(min_samples_split=20, n_estimators=200)
```

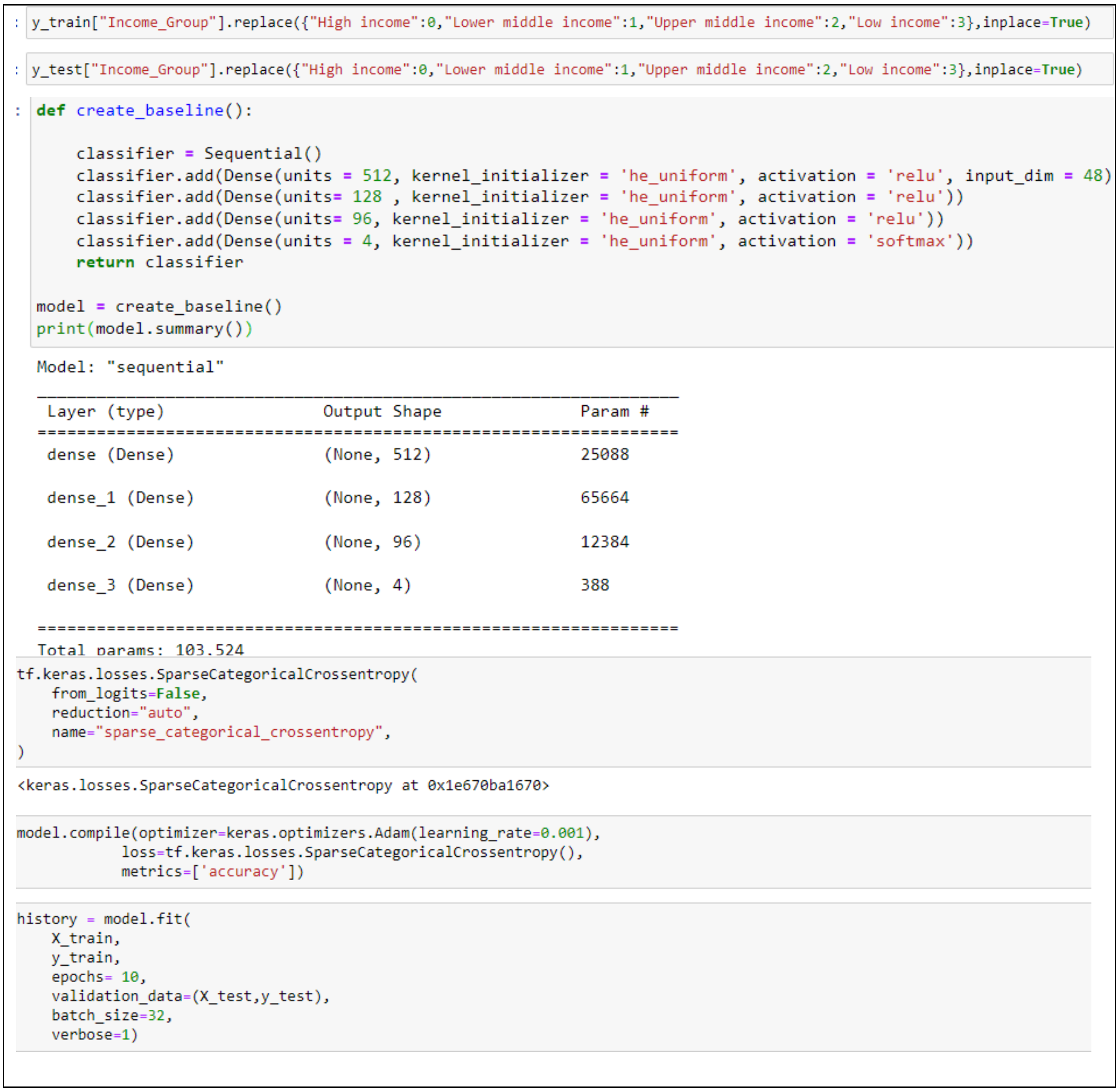
```
GB_tuned_model = GradientBoostingClassifier(min_samples_split=20, n_estimators=200)
```

```
GB_tuned_model.fit(X_train, y_train)
```

Figure 4.34 Tuned Gradient Boosting

After the based gradient boosting model building, the model is then tuned to obtain a better result. The hyperparameter tuning technique that was used in this tuning process is Grid Search. The hyperparameters that were tuned in this model are ‘loss’, ’learning_rate’, ‘min_samples_split’, ‘n_estimators’, and ‘criterion’. Based on the result, it shows 20 minimum samples split and 200 n estimators with other default parameter settings are the best parameter grouping to build a better model. Then, by using this information, a new tuned model is built and named ‘GB_tuned_model’. The model is then fit into the training set.

4.5.7 Based Artificial Neural Network (ANN)



```
y_train["Income_Group"].replace({"High income":0,"Lower middle income":1,"Upper middle income":2,"Low income":3},inplace=True)
```

```
y_test["Income_Group"].replace({"High income":0,"Lower middle income":1,"Upper middle income":2,"Low income":3},inplace=True)
```

```
def create_baseline():

    classifier = Sequential()
    classifier.add(Dense(units = 512, kernel_initializer = 'he_uniform', activation = 'relu', input_dim = 48))
    classifier.add(Dense(units= 128 , kernel_initializer = 'he_uniform', activation = 'relu'))
    classifier.add(Dense(units= 96, kernel_initializer = 'he_uniform', activation = 'relu'))
    classifier.add(Dense(units = 4, kernel_initializer = 'he_uniform', activation = 'softmax'))
    return classifier

model = create_baseline()
print(model.summary())
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 512)               25088

 dense_1 (Dense)             (None, 128)               65664

 dense_2 (Dense)             (None, 96)                12384

 dense_3 (Dense)             (None, 4)                 388

=================================================================
Total params: 103,524
```

```
tf.keras.losses.SparseCategoricalCrossentropy(
    from_logits=False,
    reduction="auto",
    name="sparse_categorical_crossentropy",
)
```

```
<keras.losses.SparseCategoricalCrossentropy at 0x1e670ba1670>
```

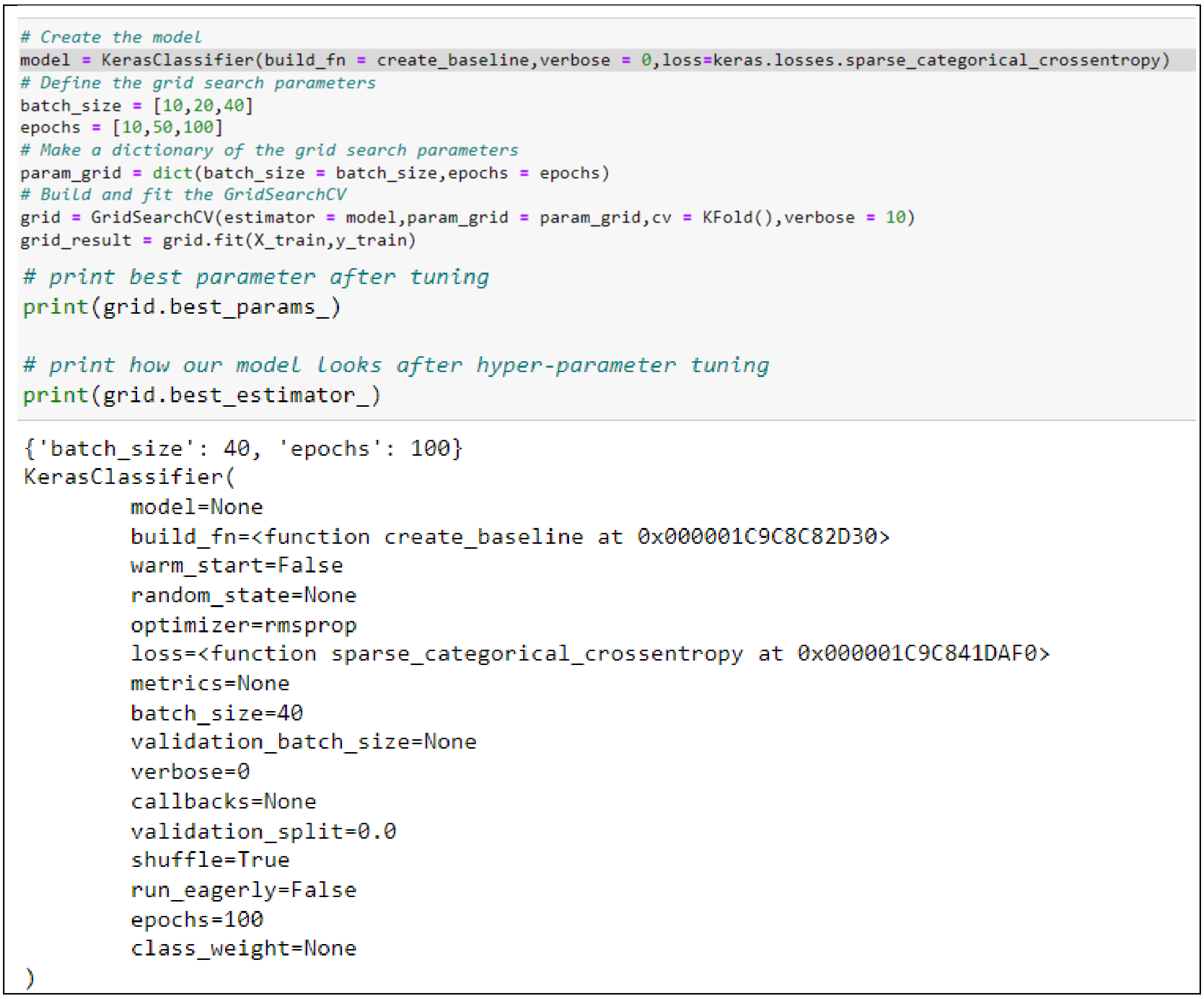
```
model.compile(optimizer=keras.optimizers.Adam(learning_rate=0.001),
            loss=tf.keras.losses.SparseCategoricalCrossentropy(),
            metrics=['accuracy'])
```

```
history = model.fit(
    X_train,
    y_train,
    epochs= 10,
    validation_data=(X_test,y_test),
    batch_size=32,
    verbose=1)
```

Figure 4.35 Based on Artificial Neural Network

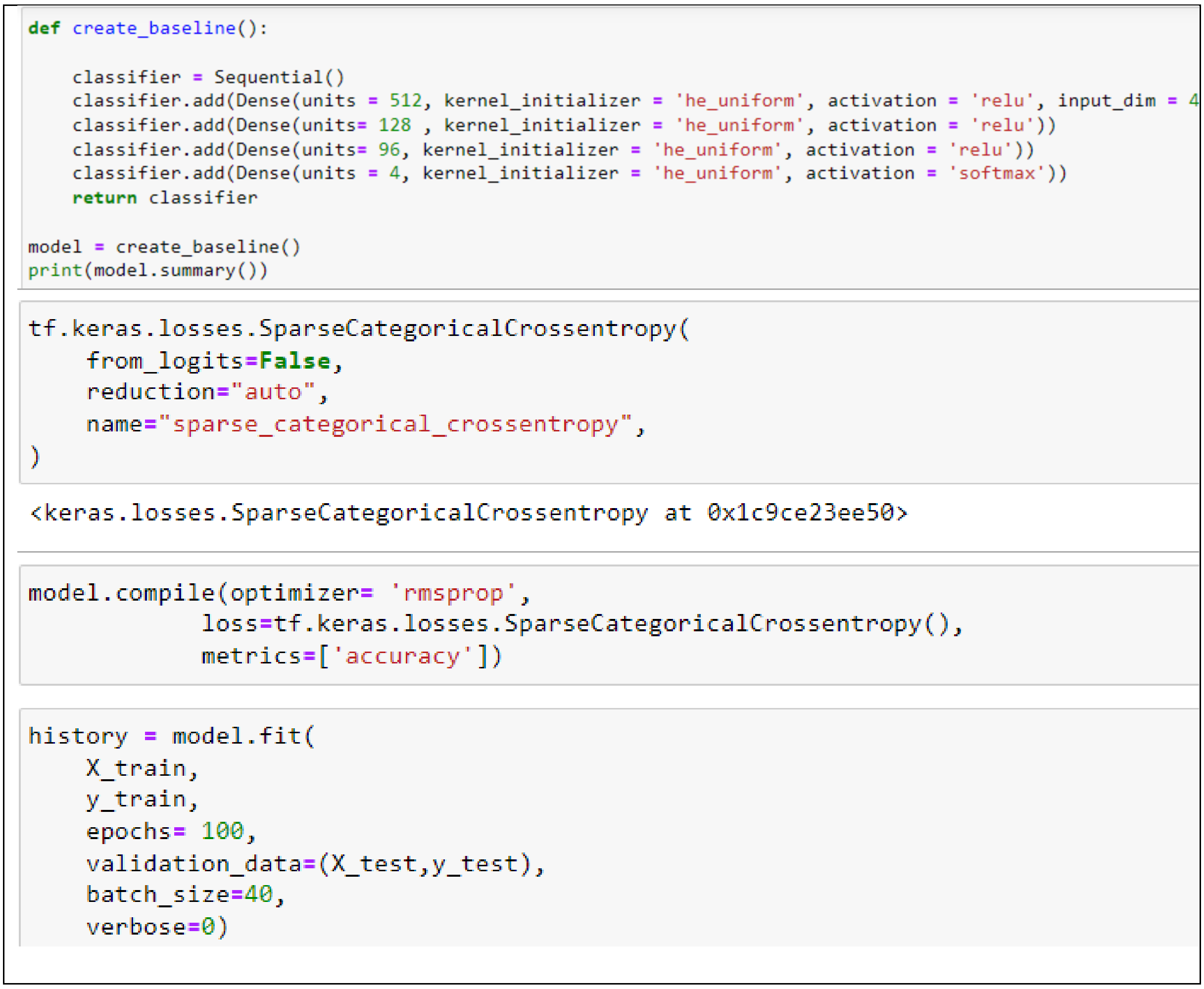
Before the artificial neural network building, the target variable in the training set is converted from the categorical variable into the numeric variable. This is because the artificial neural network model is unable to detect categorical target variables like other machine learning models. Then, a four-layer ANN classifier is built by using random units, ‘relu’ activation function in the input layer and the hidden layer and the ‘softmax’ activation function in the output layer. The ‘softmax’ activation function is normally used for categorical prediction. In the input layer, the input dimension is 48, this is because there are 48 variables fed into the model for training. In the output layer, 4 units were put because there are 4 categories under the target variable. The model summary is also shown in the figure above. Next, adam optimizer is used to optimize the result. The Adam optimizer produces outcomes that are better compared to those of conventional optimization methods, takes less effort to compute, and needs fewer tuning parameters. Adam is suggested as the standard optimizer for the majority of operations as a result of all of that. Moreover, the loss function used in this model is the sparse categorical cross-entropy. In situations where there are two or more label classes, use this crossentropy loss function. Since the target variable of the dataset has 4 label classes, hence sparse categorical is the most suitable loss function. Besides, the metric used in this model is ‘accuracy’ instead of ‘mse’ because it is a classification model. Last but not least, the model is fitted with 10 epochs and 32 batch sizes.

4.5.8 Tuned Artificial Neural Network



```
# Create the model
model = KerasClassifier(build_fn = create_baseline,verbose = 0,loss=keras.losses.sparse_categorical_crossentropy)
# Define the grid search parameters
batch_size = [10,20,40]
epochs = [10,50,100]
# Make a dictionary of the grid search parameters
param_grid = dict(batch_size = batch_size,epochs = epochs)
# Build and fit the GridSearchCV
grid = GridSearchCV(estimator = model,param_grid = param_grid,cv = KFold(),verbose = 10)
grid_result = grid.fit(X_train,y_train)
# print best parameter after tuning
print(grid.best_params_)

# print how our model looks after hyper-parameter tuning
print(grid.best_estimator_)
```

```
{'batch_size': 40, 'epochs': 100}
KerasClassifier(
	model=None
	build_fn=<function create_baseline at 0x000001C9C8C82D30>
	warm_start=False
	random_state=None
	optimizer=rmsprop
	loss=<function sparse_categorical_crossentropy at 0x000001C9C841DAF0>
	metrics=None
	batch_size=40
	validation_batch_size=None
	verbose=0
	callbacks=None
	validation_split=0.0
	shuffle=True
	run_eagerly=False
	epochs=100
	class_weight=None
)
```



```
def create_baseline():

    classifier = Sequential()
    classifier.add(Dense(units = 512, kernel_initializer = 'he_uniform', activation = 'relu', input_dim = 4
    classifier.add(Dense(units= 128 , kernel_initializer = 'he_uniform', activation = 'relu'))
    classifier.add(Dense(units= 96, kernel_initializer = 'he_uniform', activation = 'relu'))
    classifier.add(Dense(units = 4, kernel_initializer = 'he_uniform', activation = 'softmax'))
    return classifier

model = create_baseline()
print(model.summary())
```

```
tf.keras.losses.SparseCategoricalCrossentropy(
    from_logits=False,
    reduction="auto",
    name="sparse_categorical_crossentropy",
)
```

```
<keras.losses.SparseCategoricalCrossentropy at 0x1c9ce23ee50>
```

```
model.compile(optimizer= 'rmsprop',
            loss=tf.keras.losses.SparseCategoricalCrossentropy(),
            metrics=['accuracy'])
```

```
history = model.fit(
    X_train,
    y_train,
    epochs= 100,
    validation_data=(X_test,y_test),
    batch_size=40,
    verbose=0)
```

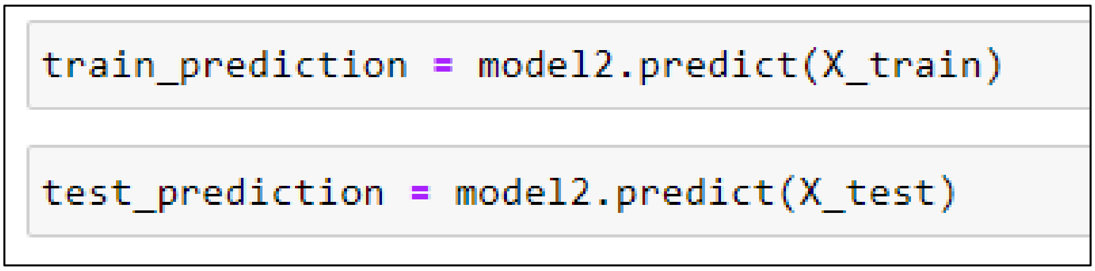
Figure 4.36 Tuned Artificial Neural Network

After the based artificial neural network model building, the model is then tuned to obtain a better result. The hyperparameter tuning technique that was used in this tuning process is also Grid Search. The hyperparameters that were tuned in this model are ‘epochs’ and ‘batch size’. Based on the result, it shows that 100 epochs and 40 batch sizes with other default parameter settings are the best parameter grouping to build a better model. Then, by using this information, a new tuned model is built and named ‘model’. The model is then fit into the training set.

4.6 Model evaluation

After the model building, the model is then used to carry out a prediction in the training and testing set. The result of the training and testing set are recorded to understand whether the model is overfitting, a good fit, or underfitting. The coding step of the model building is displayed in the session below. All the models are using the same steps to carry out the model evaluation.

Step 1: Train and test prediction



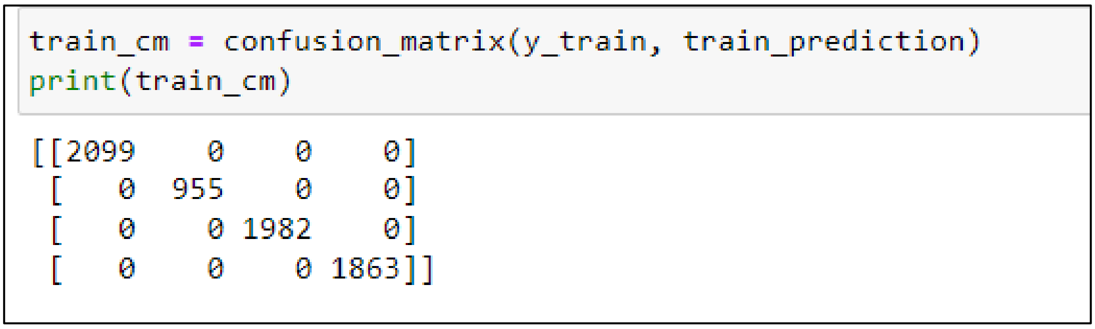
```
train_prediction = model2.predict(X_train)
```

```
test_prediction = model2.predict(X_test)
```

Figure 4.37 Train and test prediction

After the model building, the prediction is done on the training dataset (X_train) and the testing dataset (X_test).

Step 2: Confusion matrix



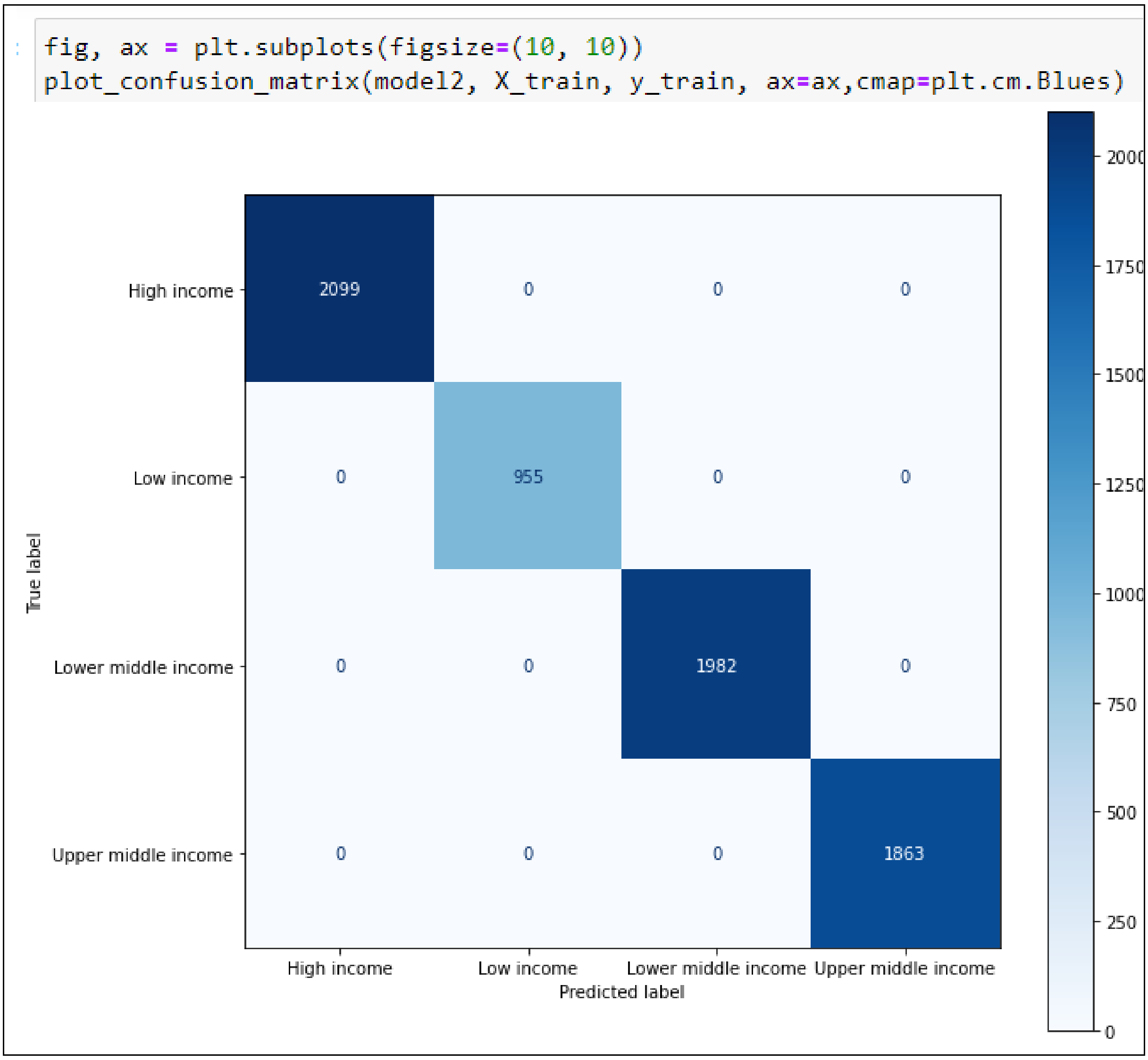
```
train_cm = confusion_matrix(y_train, train_prediction)
print(train_cm)
```

```
[[2099    0    0    0]
 [   0  955    0    0]
 [   0    0 1982    0]
 [   0    0    0 1863]]
```

Figure 4.38 Confusion Matrix

Next, the confusion matrix of the model is built to understand the accuracy of the model in classification prediction.

Step 3: Confusion matrix visualization

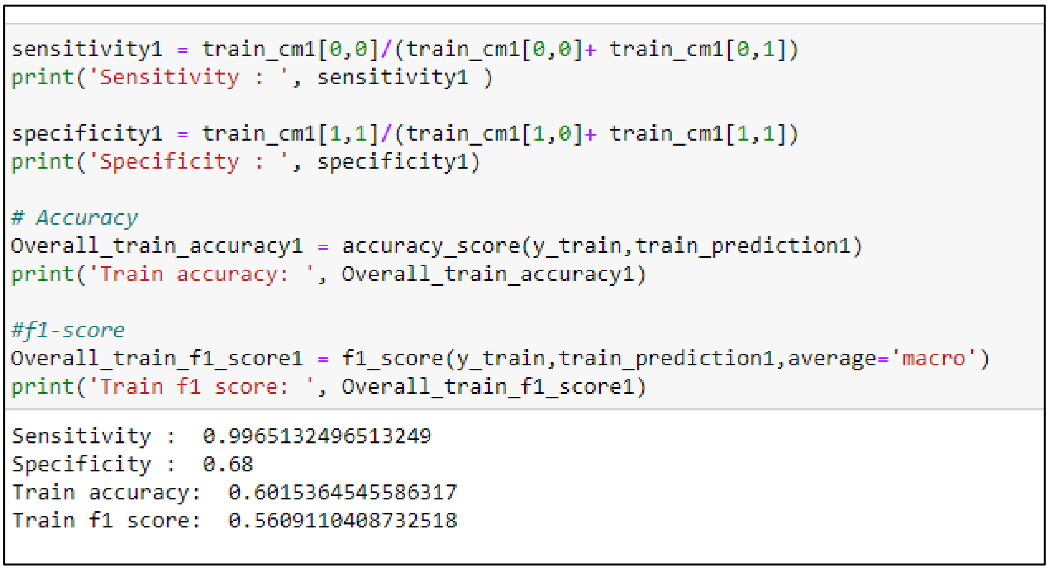


```
fig, ax = plt.subplots(figsize=(10, 10))
plot_confusion_matrix(model2, X_train, y_train, ax=ax,cmap=plt.cm.Blues)
```

Figure 4.39 Confusion Matrix Visualization

Further, the confusion matrix is further visualized by using the coding above. The output is shown in figure 4.39. The colour intensity is correlated with the count. The higher the count, the deeper the colour.

Step 4: Evaluate the model (Sensitivity, specificity, accuracy)



```
sensitivity1 = train_cm1[0,0]/(train_cm1[0,0]+ train_cm1[0,1])
print('Sensitivity : ', sensitivity1 )

specificity1 = train_cm1[1,1]/(train_cm1[1,0]+ train_cm1[1,1])
print('Specificity : ', specificity1)

# Accuracy
Overall_train_accuracy1 = accuracy_score(y_train,train_prediction1)
print('Train accuracy: ', Overall_train_accuracy1)

#f1-score
Overall_train_f1_score1 = f1_score(y_train,train_prediction1,average='macro')
print('Train f1 score: ', Overall_train_f1_score1)
```

```
Sensitivity :  0.9965132496513249
Specificity :  0.68
Train accuracy:  0.6015364545586317
Train f1 score:  0.5609110408732518
```

Figure 4.40 Model Evaluation

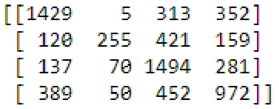
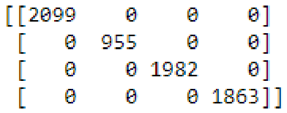
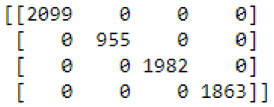
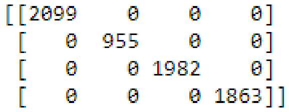
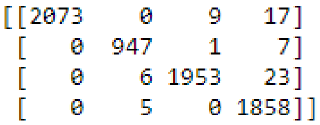
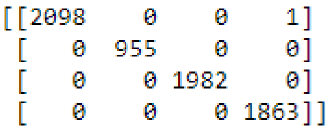
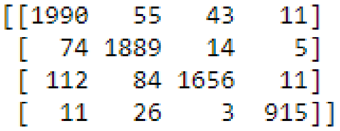
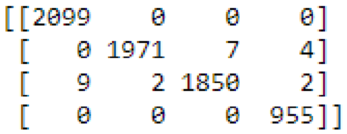
Moreover, the performance of the model is evaluated by using sensitivity, specificity, accuracy and f1 score.

CHAPTER V

RESULT AND DISCUSSION

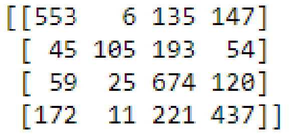
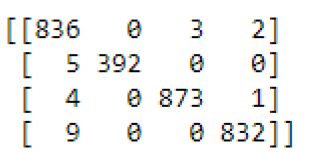
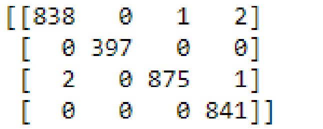
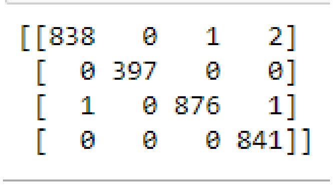
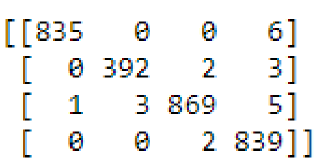
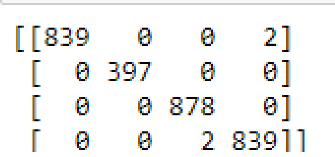
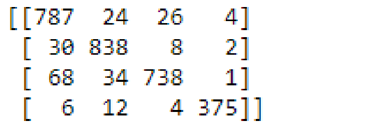
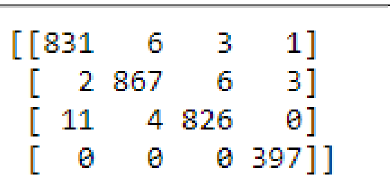
5.1 Training Result

Table 5.1 Result for Training Set

Model	Confusion Matrix	Sensitivity	Specificity	Accuracy	F1-score
Based SVM	 [[1429 5 313 352] [120 255 421 159] [137 70 1494 281] [389 50 452 972]]	0.9965132496513249	0.68	0.6015364545586317	0.5609110408732518
Tuned SVM	 [[2099 0 0 0] [0 955 0 0] [0 0 1982 0] [0 0 0 1863]]	1.0	1.0	1.0	1.0
Based RF	 [[2099 0 0 0] [0 955 0 0] [0 0 1982 0] [0 0 0 1863]]	1.0	1.0	1.0	1.0
Tuned RF	 [[2099 0 0 0] [0 955 0 0] [0 0 1982 0] [0 0 0 1863]]	1.0	1.0	1.0	1.0
Based GB	 [[2073 0 9 17] [0 947 1 7] [0 6 1953 23] [0 5 0 1858]]	1.0	1.0	0.9901434990578345	0.9900373942229949
Tuned GB	 [[2098 0 0 1] [0 955 0 0] [0 0 1982 0] [0 0 0 1863]]	1.0	1.0	0.9998550514567328	0.9998733555641945
Based ANN	 [[1990 55 43 11] [74 1889 14 5] [112 84 1656 11] [11 26 3 915]]	0.9731051344743277	0.9623025980641875	0.9349181040730541	0.938689828268858
Tuned ANN	 [[2099 0 0 0] [0 1971 7 4] [9 2 1850 2] [0 0 0 955]]	0.9966523194643712	1.0	0.9933323670097115	0.9940660355602297

5.2 Testing Result

Table 5.2 Result for Testing Set

Model	Confusion Matrix	Sensitivity	Specificity	Accuracy	F1-score
Based SVM	 [[553 6 135 147] [45 105 193 54] [59 25 674 120] [172 11 221 437]]	0.9892665474060823	0.7	0.5982414609401421	0.5591239297716828
Tuned SVM	 [[836 0 3 2] [5 392 0 0] [4 0 873 1] [9 0 0 832]]	1.0	0.9874055415617129	0.9918836658775786	0.99209315872404
Based RF	 [[838 0 1 2] [0 397 0 0] [2 0 875 1] [0 0 0 841]]	1.0	1.0	0.9979709164693946	0.9982411657178136
Tuned RF	 [[838 0 1 2] [0 397 0 0] [1 0 876 1] [0 0 0 841]]	1.0	1.0	0.9986472776462632	0.998823670430421
Based GB	 [[835 0 0 6] [0 392 2 3] [1 3 869 5] [0 0 2 839]]	1.0	1.0	0.9925600270544471	0.9922138600341814
Tuned GB	 [[839 0 0 2] [0 397 0 0] [0 0 878 0] [0 0 2 839]]	1.0	1.0	0.9986472776462632	0.9988234365243949
Based ANN	 [[787 24 26 4] [30 838 8 2] [68 34 738 1] [6 12 4 375]]	0.9704069050554871	0.9654377880184332	0.925938451132905	0.9306900264500475
Tuned ANN	 [[831 6 3 1] [2 867 6 3] [11 4 826 0] [0 0 0 397]]	0.9892729439809297	0.9953863898500577	0.983429151166723	0.9853044084097728

In the model evaluation step, there are a few modifications between the evaluation matrix in methodology. In our methodology, it was mentioned the receiver operating characteristic (ROC) and area under the ROC Curve (AUC) will be used in the model evaluation. However, after carrying out the experiment, it was found that this evaluation matrix support binary classification instead of multiclassification. While we apply our dataset in ROC coding, it keeps on showing errors since it is multiclassification data. Hence, after more literature review, the F1 score is used as the additional evaluation matrix to evaluate the model. The symmetrical average of a classifier's recall and precision is used to create the F1 score, which integrates both metrics into a single data point (educative, n.d). Instead of using precision and recall as the additional evaluation matrix, F1-score is chosen because it contributes to balancing the precision and recall. F1-score can assist in balancing the metric between positive and negative samples if users chose the positive class as the one with the fewest examples. F1-score values vary from 0 to 1. The model is more accurate when closer it is to 1. The formula for the F1 score is shown below:

$$F1\ score = \frac{2(P * R)}{P + R}$$

where:

P = precision

R= recall

Based on the literature review, the study of Chakrabarty & Biswas, (2018) aims to illustrate how technologies from machine learning and data mining can be applied to the problem of income disparity using f1-score as one of the evaluation matrices and found that gradient boosting is the best model. By employing 50,000 data and 15 feature datasets downloaded from Kaggle to create logistic regression and support vector machine models, a study from Voleti & Jana, (2022) forecasts worker wage levels and used the f1-score as one of the evaluation matrices. Moreover, the work by Srivastava et al. (2020) which built a random forest, random tree, naive Bayes, and REPTree model to predict people's income status using a dataset containing 31978 observations and 13 variables also utilized the f1-score as one of their evaluation matrices. Based on tables 5.1 and 5.2, objective 3 of this project is successfully achieved as it is able to evaluate and estimate the performance of the model based on different evaluation techniques.

The model result is divided into results for the training set and the testing set, this is because we want to understand whether the model is overfitting, underfitting, or a good fit. When the accuracy training set is significantly higher than the testing set, it indicates that there is overfitting happening in the model. On the other hand, when the accuracy of the testing set is significantly higher than the training set, it indicates that there is underfitting happening in the model. When the accuracy of the training set and testing set is nearly the same, it indicates that the model is a good fit. The value of accuracy, sensitivity, specificity and F1 score are listed in the table for all model that was built which included the based support vector machine, tuned support vector machine, based random forest, tuned random forest, based gradient boosting, tuned gradient boosting, based artificial neural network and tuned artificial neural network. Based on the result in table 5,1 and 5.2 above, it was found that all the models do not have any overfitting or underfitting problem since their training and testing accuracy has the nearly same value which indicates a good fit.

In the based support vector machine model, it was found that there is a high sensitivity score of 0.997 and a low specificity score of only 0.68 in the training set. Moreover, in the training set, the accuracy and F1 score in this model also give a lower performance with only 0.60 and 0.56. In the testing set, the sensitivity specificity, accuracy, and F1-score are 0.9989, 0.7, 0.598 and 0.559. Both the training and testing sets showed that there is lower specificity than sensitivity, this indicates that this model is more prone to give a false negative result. Since the accuracy and F1 score do not give a satisfactory result, hyperparameter tuning is necessary for this model. After hyperparameter tuning by using Grid Search, the performance of all the evaluation matrices in the support vector machine model is significantly improved to 1.0 in the training set. While in the testing set, the result is improved from 0.98 sensitivity to 1.0 sensitivity, 0.7 to 0.987 specificity, 0.598 to 0.99 accuracy and 0.56 to 0.99 F1 score. This result proved that the parameter that was chosen for tuning are playing a significant role in the model building. There is no overfitting or underfitting problem exists in this model.

Based on the result, the based random forest and based gradient boosting achieved a high performance with 1.0 in both sensitivity and specificity. Moreover, the based random forest achieved 0.998 accuracy and 0.998 F1-score while the based gradient boosting achieved 0.993 accuracy and 0.992 F1-score. In the based model, the random forest performed better than the gradient boosting. After hyperparameter tuning, the result of both models becomes nearly the same. The tuned random forest and tuned gradient boosting model achieved 1.0 specificity and sensitivity and 0.999 accuracy. The only difference value is the F1 score which is 0.998 for random forest and 0.999 for gradient boosting. There is not much difference between the result which indicates that both of the models are suitable to predict the income categories and give high performance. This finding is compatible with the literature review in Chapter 2 as it was mentioned majority of the models that achieved high performance are random forest and gradient boosting.

Moreover, other than machine learning, a deep learning approach which is an artificial neural network also used for country income level prediction. In the based artificial neural network, it achieves 0.97 sensitivity, 0.965 specificity, 0.926 accuracy and 0.926 F1 scores. This result is not performed as high as the based random forest and the based gradient boosting. After hyperparameter tuning, the tuned artificial neural network achieved 0.989 sensitivity, 0.995 specificity, 0.983 accuracy and 0.985 F1 scores. There is improvement after hyperparameter tuning which indicates that hyperparameter tuning is also effective in deep learning to improve outcomes. Hence, objective 2 in this project is achieved because the results of different models are successfully optimized after hyperparameter tuning. However, the result for the tuned artificial neural network is still not as good as the tuned random forest and gradient boosting. Hence, it can be concluded that tuned random forest and tuned gradient boosting are the most suitable model that can be used for the prediction of country income based on legal gender inequality. Here, the aim of this project is successfully achieved as it builds the best model for country income prediction.

In this study, there is a total of 8 models are built. It was found that every model gets improvement after hyperparameter tuning. This indicates that the based model still needs to be tuned even though it had achieved a good result because there is always room for improvement. A single training task is used to execute several experiments for hyperparameter tuning. Each trial represents the full performance of the training set with the settings of the selected hyperparameters specified inside the boundaries define (Kasture, 2020). In this project, the model that showed the most significant difference between the based model and the tuned model is the support vector machine. This indicates that the C and gamma parameters are playing a significant role in the support vector machine to determine the model’s effectiveness for country income classification. Just the c parameter needs to be optimized for a linear kernel. However, if researchers wish to employ an RBF kernel, researchers must concurrently optimize the c and gamma parameters. When gamma is big, the impact of c is minimal. If gamma is low, c has the same impact on the model as it would have on a linear one (Yidirim, 2020). Since our project is not using linear kernel for tuning only while the RBF kernel is also one of the hyperparameters in the support vector machine model, hence both the C and gamma hyperparameters are tuned and finally provide a satisfied result.

Moreover, in this study, the machine learning approach achieves higher performance than the deep learning approach which is the artificial neural network. This might be due to the number of hyperparameters that were tuned in the artificial neural network are being limited in Grid Search, which is only the number of epochs and batch size. In the future study, it is suggested that other hyperparameter tuning methods such as keras tuner which is specified for deep learning usage can be used to tune more hyperparameters such as the number of layers, the number of neurons, the value of learning rate and etc. Moreover, regularization methods such as the dropout method can be added to the artificial neural network to regularize the model in order to get a better result. Other than that, it is also suggested that instead of using the automated tuning method, the manual method also can be utilized to understand the effect of each hyperparameter on the model effectiveness during the income level classification.

The limitation of this project is choosing the right hyperparameter tuning method. There are many hyperparameter tuning methods are available such as grid search, random search, keras tuner, Bayes search and etc. Grid search is chosen for this project because it is the most common hyperparameter method found in the literature review. Grid search is a hyperparameter tuning method that creates a grid of the various possibilities of the hyperparameters. A model is trained and a rating is generated for every conceivable combo based on the testing set. Using this method, all feasible combinations of the hyperparameter settings are tested. Another limitation of this project is time consumption during the hyperparameter tuning process. Despite performing a thorough scan of all conceivable pairings, the strategy can be relatively wasteful in terms of learning duration. For the hyperparameter tuning process in this study, it takes more than hours to run the coding, which increases the risk of environmental corruption.

CHAPTER VI

CONCLUSION

Country income level prediction based on legal gender equality is important to improve the country’s economic and gender inequality that still exists nowadays. Random forest and gradient boosting are popular machine learning algorithms that always give high performance. Based on the literature review, the majority of the studies proved that these two algorithms are effective in income group classification. Based on this study, the finding is compatible with the literature review and contributes knowledge that tuned gradient boosting and tuned random forest are the best machine learning to predict country income level based on legal gender equality. In the future, the country’s income level prediction based on legal gender equality can use different hyperparameter tuning methods to find out the best tuning method with less time consumption and a lower risk of corruption. Moreover, manual parameter tuning also is suggested to understand the role of each hyperparameter in determining the effectiveness of the model. In the nutshell, the aim and objective of this project are successfully achieved and bring knowledge about the importance of hyperparameter tuning. Most importantly, this project contributes the best model to build supervised learning approaches in country income category prediction based on legal gender equality. Interest organizations can virtually change gender equality laws using this method, anticipate the likelihood of economic growth, and put the best plan into action. The most effective way to accelerate economic performance and increase women's financial empowerment can be identified in this framework.

REFERENCE

Addepto (2019, October 18). *Machine learning in Economics – How is it Used?* https://addepto.com/machine-learning-in-economics-how-is-it-used/

Amanda, R. (2022, April 21). *Top 6 Reasons that Economic Development is Important to a Region's Economy*. Orlando Economic Partnership. Retrieved August 12, 2022, from https://news.orlando.org/blog/top-6-reasons-that-economic-development-is-important-to-a-regions-economy-infographic/#:~:text=Economic%20development%20is%20a%20critical,for%20existing%20and%20future%20residents.

Amadeo, K. (2021). *What Is Economic Growth?* The Balance. https://www.thebalance.com/what-is-economic-growth-3306014

ARS Electronica. (2022, Jun 17). *Is this a man's world?* Acknowledging gender bias in AI. https://ars.electronica.art/aeblog/en/2022/06/17/gender-bias-in-ai/

Akbulaev, N., Aliyeva, B., & Sapena, J. (2020). *Gender and economic growth: Is there a correlation?* The example of Kyrgyzstan. Cogent Economics & Finance. https://doi.org/10.1080/23322039.2020.1758007

Bhandari, A. (2020). *Everything you Should Know about Confusion Matrix for Machine Learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

Barhoumi, K., Choi, S. M., Lyer, T, Li, J., Ouattara, F., Tiffin, A., & Yao, J. (2020). *A Machine-Learning Approach to Nowcast the GDP in Sub-Saharan Africa.* https://www.uneca.org/sites/default/files/AEC/2020/presentations/a_machine-learning_approach_to_nowcast_the_gdp_in_sub-saharan_africa_.pdf

Bertay, A. C., Dordevic, L. & Sever, C. (2020). *Gender Inequality and Economic Growth: Evidence from Industry-Level Data.* International Monetary Fund. file:///D:/Desktop/Sem%202%20Assignment/Capstone%20Project/Article/wpiea2020119-print-pdf.pdf

Bekena & Menji, S. (2017). Using decision tree classifier to predict income levels. *Munich Personal RePEc Archive.* https://mpra.ub.uni-muenchen.de/id/eprint/83406

Bhandari, A. (2022, June 16). *AUC-ROC Curve in Machine Learning Clearly Explained.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Bramesh, S. M. & Puttaswamy, B. S. (2019). Comparative Study of Machine Learning Algorithms on Census Income Data Set. *Journal of Engineering Research and Application.* http://dx.doi.org/10.9790/9622-0908017881

BRD. (n.d.). *Sustainable Development Goals (SDGs).* Retrieved August 19, 2022 from https://www.brd-org.se/blog/agenda-2030/?gclid=CjwKCAjw0a-SBhBkEiwApljU0tELPhOp7ZGmkBKddL4PZmoeFGVrTGtTSc_1i5ajuTCBIccRYAiZDxoCp8AQAvD_BwE

Buolamwini, J. (2019, February 7). *Artificial Intelligence Has a Problem With Gender and Racial Bias. Here's How to Solve It.* TIME. https://time.com/5520558/artificial-intelligence-racial-gender-bias/

California Association for Local Economic Development. (n.d.) *What is Economic Development?* Retrieved August 12, 2022, from https://caled.org/economic-development-basics/

Cao, R., Tu, W., Cai, J., Zhao, T., Xiao, J., Cao, J., Gao, Q., & Su, H. (2022). *Machine Learning-Based Economic Development Mapping from Multi-Source Open Geospatial Data.* ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences.https://ui.adsabs.harvard.edu/link_gateway/2022ISPAn..54..259C/doi:10.5194/isprs-annals-V-4-2022-259-2022

CFI. (2021, July 21). *Economic Growth.* https://corporatefinanceinstitute.com/resources/knowledge/economics/economic-growth/

Chakrabarty, N. & Biswas, S. (2018). *A Statistical Approach to Adult Census Income Level Prediction.* ResearchGate. https://www.researchgate.net/publication/328494313_A_Statistical_Approach_to_Adult_Census_Income_Level_Prediction/related

Christopherson, K., Yiadom, A., Johnson, J., Yazid, H., & Thiemann, C. (2022). *Tackling Legal Impediments to Women' s Economic Empowerment.* International Money Fund. https://www.imf.org/en/Publications/WP/Issues/2022/02/18/Tackling-Legal-Impediments-to-Womens-Economic-Empowerment-513392

Chung, H., Park, C., Kang, W. S. & Lee, J. (2021, November 29). Gender Bias in Artificial Intelligence: Severity Prediction at an Early Stage of COVID-19. *Frontliers.* https://doi.org/10.3389/fphys.2021.778720

Cogoljevic, D., Alizamir, M., Piljan, I., Prljic, K., & Zimonjic, S. (2017). A machine learning approach for predicting the relationship between energy resources and economic development. *Science Direct.* https://doi.org/10.1016/j.physa.2017.12.082

Deepchecks. (n.d). *Sensitivity and Specificity of Machine learning.* Retrieved August 17, 2022, from https://deepchecks.com/glossary/sensitivity-and-specificity-of-machine-learning/

Delgado, J. & Moura, Pedro. (2021). Electric Mobility: A Key Technology to Decarbonize the Economy and Improve Air Quality. *SpringerLink*. https://doi.org/10.1007/978-3-319-95864-4_127

Dutt, P. &_Tsetlin, I. (2020). Income distribution and economic development: Insights from machine learning. *Economic and Policy*. chrome-extension://dagcmkpagjlhakfdhnbomgmjdpkdklff/enhanced-reader.html?openApp&pdf=https%3A%2F%2Fonlinelibrary.wiley.com%2Fdoi%2Fpdfdirect%2F10.1111%2Fecpo.12157

Farjallah, N., & Sghaier, A. (2022). Machine Learning Based Modelling of Economic Growth and Quality of Governance: The MENA Region. *Research Square*. https://orcid.org/0000-0003-1106-8011

Fernandes, J. C., Teodoro, A. C., Lima, A.M.C., & Robles, E. R. (2020). Semi-Automatization of Support Vector Machines to Map Lithium (Li) Bearing Pegmatites. *Research Gate*. http://dx.doi.org/10.3390/rs12142319

Ganthi, R. (2018). *Support Vector Machine — Introduction to Machine Learning Algorithms.* Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Growing Economies Through Gender Parity. (n.d.). *Closing the gender gap in the workforce could add a staggering $28 trillion to the global GDP*. Retrieved August 12, 2022, from https://www.cfr.org/womens-participation-in-global-economy/

Gavi. (2020, February 18). *Sustainable Development Goals.* https://www.gavi.org/our-alliance/global-health-development/sustainable-development-goals?gclid=CjwKCAjw0a-SBhBkEiwApljU0o1yYZYKzZ-UqycH5n7K_zMR5ZGz1d-zzK_wDQhegDph3sSvI4tlHxoCDZ8QAvD_BwE

GreeksForGreeks. (2021, Jun 29). *Data Pre-processing in Data Mining.* https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

Hossain, A., Hossen, M., & Hasan, M. M. (2021). GDP Growth Prediction of Bangladesh using Machine Learning Algorithm. *Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks.* https://doi.org/10.1109/ICICV50876.2021.9388593

Huet, N. (2022, March 8). *Gender bias in recruitment: How AI hiring tools are hindering women's careers.* https://www.euronews.com/next/2022/03/08/gender-bias-in-recruitment-how-ai-hiring-tools-are-hindering-women-s-careers

International Monetary Fund. (2022, March 7). *How Empowering Women Supports Economic Growth.* https://blogs.imf.org/2022/03/07/how-empowering-women-supports-economic-growth/

IBM. (2020, August 19). *Supervised Learning*. https://www.ibm.com/cloud/learn/supervised-learning

IBM. (2020, December 7). *Random Forest*. https://www.ibm.com/cloud/learn/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

IBM. (2020, August 17). *Neural Networks*. https://www.ibm.com/cloud/learn/neural-networks#:~:text=Neural%20networks%2C%20also%20known%20as,neurons%20signal%20to%20one%20another.

IBM. (2020, July 15). *Machine Learning*. https://www.ibm.com/cloud/learn/machine-learning

International Women's Day. (n.d.). *Gender and AI: Addressing bias in artificial intelligence*. Retrieved August 18, 2022 from, https://www.internationalwomensday.com/Missions/14458/Gender-and-AI-Addressing-bias-in-artificial-intelligence#:~:text=Gender%20bias%20occurs%20during%20machine,incorporated%20within%20the%20AI%20system.

Kasture, N. (2020, November 16). *Why Hyper parameter tuning is important ?* Analytics Vidhya. https://medium.com/analytics-vidhya/why-hyper-parameter-tuning-is-important-for-your-model-1ff4c8f145d3

Khanna, R. (2018). *Comparative Study of Classifiers in predicting the Income Range of a person from a census data.* Towards Data Science. https://towardsdatascience.com/comparative-study-of-classifiers-in-predicting-the-income-range-of-a-person-from-a-census-data-96ce60ee5a10

Krueger, A. O. (n.d). *Economic development*. Encyclopedia Britannica. Retrieved August 12, 2022, from https://www.britannica.com/topic/economic-development

Korab, P. (2021, Jun 27). *Use of Machine Learning in Economic Research: What the Literature Tells Us*. https://towardsdatascience.com/use-of-machine-learning-in-economic-research-what-the-literature-tells-us-28b473f26043

Kurama, V. (2020). *Gradient Boosting In Classification: Not a Black Box Anymore!* PaperspaceBlog. https://blog.paperspace.com/gradient-boosting-for-classification/

Leavy, S. (2018). Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning. *Research Gate*. http://dx.doi.org/10.1145/3195570.3195580

Legislationline. (n.d). *Gender Equality.* Retrieved April 5, 2022 from https://www.legislationline.org/topics/topic/7

Malaza, T. & Perekh, N. (2020, December 8). *Can digital technology help create a more gender-equal society?* https://www.povertyactionlab.org/blog/12-8-20/can-digital-technology-help-create-more-gender-equal-society

Malik, F. (2020, Feb 19). *What Is Grid Search?* FinTechExplained. https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a

Martinez, H. (2022, Feb 7). *What is Gender Equality? Definition, Example.* https://unitedwaynca.org/blog/what-is-gender-equality/

Matsumoto, S. L., & Samonte, M. J. (2022). Philippine Economic Growth: GDP Prediction using Machine Learning Algorithms. *ICCBD 2021: 2021 4th International Conference on Computing and Big Data.* https://doi.org/10.1145/3507524.3507526

Mele, M. & Magazzino, C. (2020). A Machine Learning analysis of the relationship among iron and steel industries, air pollution, and economic growth in China. *Journal of Cleaner Production*. https://doi.org/10.1016/j.jclepro.2020.123293

McCombes. (2019, February 22). *What Is a Literature Review | Step-by-Step Guide & Examples*. Scribbr. https://www.scribbr.com/dissertation/literature-review/

Mishra, A. (2018, Feb 24). *Metrics to Evaluate your Machine Learning Algorithm*. Towards Data Science. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Nauli, M. E. (2021). *Income Prediction With Classification Models*. RPubs. https://rpubs.com/madeleinee/LBB-C2

Niethammer, C. (2020, May 2). *AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators*. Forbes. https://www.forbes.com/sites/carmenniethammer/2020/03/02/ai-bias-could-put-womens-lives-at-riska-challenge-for-regulators/?sh=27dc39b3534f

ODSC. (2020, November 5). *The A – Z of Supervised Learning, Use Cases, and Disadvantages*. https://opendatascience.com/the-a-z-of-supervised-learning-use-cases-and-disadvantages/

OECD. (2018, March 7). *The Impact of Legal Frameworks on Women's Economic Empowerment around the World: challenges and good practices.* https://www.oecd.org/mena/competitiveness/2107-March-on-Gender-Legal-Framework-Highlights.pdf

Ozden, E. & Guleryuz, D. (2022). Optimized Machine Learning Algorithms for Investigating the Relationship Between Economic Development and Human Capital. *SpringerLink*. https://link.springer.com/article/10.1007/s10614-021-10194-7

Patil, P. (2018, March 24). *What is Exploratory Data Analysis? Towards Data Science*. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15#:~:text=Exploratory%20Data%20Analysis%20refers%20to,summary%20statistics%20and%20graphical%20representations.

Ray, S. (2021, August 26). *Understanding Support Vector Machine(SVM) algorithm from examples (along with code)*. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Richardson, A., Mulder, T., & Vahbi, T. (2018). Nowcasting New Zealand GDP Using Machine

Learning Algorithms. *Centre for Applied Macroeconomic Analysis.* https://dx.doi.org/10.2139/ssrn.3256578

Roser, M. (n.d). *Economic Growth*. Retrieved April 10, 2022 from https://ourworldindata.org/economic-growth

Rubin, C., Hakspiel, J. & Gray, B. (2021, Feb 23). *Using Artificial Intelligence and Technology for Women's Economic Empowerment: How Can It Work?* Seep leading collaboration & learning. https://seepnetwork.org/Blog-Post/Using-Artificial-Intelligence-and-Technology-for-Women-s-Economic-Empowerment-Can-It-Work

Save the Children. (n.d). *Gender Discrimination: Inequality Starts in Childhood.* Retrieved August 13, 2022, from https://www.savethechildren.org/us/charity-stories/how-gender-discrimination-impacts-boys-and-girls#:~:text=Gender%20inequality%20is%20discrimination%20on,violated%20by%20gender%20based%20discrimination.

Sever, C. (2022, July 29). *Legal Gender Equality as a Catalyst for Convergence*. IMF Working Papers. https://www.imf.org/en/Publications/WP/Issues/2022/07/28/Legal-Gender-Equality-as-a-Catalyst-for-Convergence-521468

Silva, M. S. & Klasen, S. (2021). Gender inequality as a barrier to economic growth: a review of the theoretical literature. *Springer Link.* https://link.springer.com/article/10.1007/s11150-020-09535-6

Singh, A. (2017, September 13). *Predicting Income Level, An Analytics Casestudy in R*. Cloud x Lab. https://cloudxlab.com/blog/predicting-income-level-case-study-r/

Smith, G. & Rustagi, I. (2021, March 31). *When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity.* Stanford Social Innovation Review. https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity#

Srivastava, S., Sharma, D., & Sharma, P. (2020). *Comparing various Machine Learning Techniques for Predicting the Salary Status*. EasyChair Preprint. file:///D:/Desktop/Sem%202%20Assignment/Capstone%20Project/Article/EasyChair-Preprint-2625.pdf

Storm, H., Baylis, K., & Heckelei, T. (2019). Machine learning in agricultural and applied economics. *European Review of Agricultural Economics. 47* (3). 849 – 892. https://doi.org/10.1093/erae/jbz033

Tan, P.N. (n.d.). Receiver Operating Characteristic. *SpringerLink.* https://link.springer.com/referenceworkentry/10.1007/978-0-387-39940-9_569

Temraz, M. (2019). A Comparison of Supervised Learning Algorithms for the Income Classification. *International Journal of Computer Applications 182*(38), 19-25. http://dx.doi.org/10.5120/ijca2019918391

The Economic Times. (2021, March 5). *Gender inequality has cost world $70 tln since 1990: Report.* https://economictimes.indiatimes.com/news/economy/indicators/gender-inequality-has-cost-world-70-tln-since-1990-report/articleshow/81352405.cms?from=mdr

The Global Goals. (n.d.). *Gender Equality*. Retrieved April 6, 2022 from https://www.globalgoals.org/goals/5-gender-equality/

The World Bank. (2018). *Unrealized Potential: The High Cost of Gender Inequality in Earnings.* https://www.worldbank.org/en/topic/gender/publication/unrealized-potential-the-high-cost-of-gender-inequality-in-earnings#:~:text=Some%20key%20findings%3A&text=In%20low%2D%20and%20lower%2Dmiddle,of%20%2423%2C620%20per%20person%20globally.

The World Bank. (2022). *Women, Business and the Law 2022.* https://wbl.worldbank.org/en/wbl

UNDP. (n.d.). *The SDGs in Action.* Retrieved April 6, 2022 from https://www.undp.org/sustainable-development-goals?utm_source=EN&utm_medium=GSR&utm_content=US_UNDP_PaidSearch_Brand_English&utm_campaign=CENTRAL&c_src=CENTRAL&c_src2=GSR&gclid=CjwKCAjw0a-SBhBkEiwApljU0vXPK9TV5QcOXchnsd-c6Z-fjOUfyYiyhDd32THNd5Mo5GuVtBOnqRoCS-4QAvD_BwE

UN Women. (2018). *Facts and Figures: Economic Empowerment*. https://www.unwomen.org/en/what-we-do/economic-empowerment/facts-and-figures

United Nation Global Compact. (n.d.). *Gender Equality*. Retrieved August 12, 2022, from https://www.unglobalcompact.org/what-is-gc/our-work/social/gender-equality

Unicef. (n.d). *Gender equality.* Retrieved August 12, 2022, from https://www.unicef.org/gender-equality

Unido. (2019). *The Impact of New Digital Technologies on Gender Equality in Developing Countries.* https://www.unido.org/api/opentext/documents/download/16760725/unido-file-16760725

Vaidya, D. (n.d.). *Gradient Boosting.* Wall Street Mojo. Retrieved 18 August, 2022, from https://www.wallstreetmojo.com/gradient-boosting/#gradient-boosting-examples

Vic.Gov.Au. (n.d.). *Gender equality: what is it and why do we need it?* Retrieved August 12, 2022, from https://www.vic.gov.au/gender-equality-what-it-and-why-do-we-need-i

Voleti, R. & Jana, B. (2022). Predictive Analysis of HR Salary using Machine Learning Techniques. *International Journal of Engineering Research & Technology.* https://www.ijert.org/predictive-analysis-of-hr-salary-using-machine-learning-techniques

Wadhwa, D. & Halim, D. (2020). *Legal progress towards gender equality*. The World Bank, Atlas of the Sustainable Development Goals 2020: From World Development Indicators. https://datatopics.worldbank.org/sdgatlas/goal-5-gender-equality/

Wakefield. (n.d.). *A guide to the types of machine learning algorithms and their applications.* Saas. Retrieved April 9, 2022 from https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html#:~:text=As%20new%20data%20is%20fed,%2Dsupervised%2C%20unsupervised%20and%20reinforcement.

Wodon, Quentin, Onagoruwa, A., Malé, C., Montenegro, C., Nguyen, H., & Brière, B. (2020). *"How Large Is the Gender Dividend? Measuring Selected Impacts and Costs of Gender Inequality.*" World Bank Group. http://hdl.handle.net/10986/33396

Woll, M. (2021, October 15). *Gender inequality in the workplace: The fight against bias.* Retrieved August 13, 2022, from https://www.betterup.com/blog/gender-inequality-in-the-work-place

Wu, C., Huang, S.C., Chiou, C. C, Chang, T. & Chen, Y.C. (2021). The Relationship Between Economic Growth and Electricity Consumption: Bootstrap ARDL Test with a Fourier Function and Machine Learning Approach. *SpringerLink.* https://doi.org/10.1007/s10614-021-10097-7

Yiu, T. (2019, Jun 12). *Understanding Random Forest*. Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Yıldırım, S. (2020, May 31). Hyperparameter tuning for support vector machines — C and gamma parameters. Towards Data Science. https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines-c-and-gamma-parameters-6a5097416167

Yoon, (2021). Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. *Computational Economics. 57*(19), 1-19. https://link.springer.com/article/10.1007/s10614-020-10054-w