SOCIAL SECURITY ENROLLMENT

Statistical Analysis and Machine Learning Techniques

2024

Dr. Rania Mamdouh

Presented By

Saif Essam

Joy Ehab

Aya Elsherbiny

Mohamed Aly

Cairo University

Faculty of Economics and Political Science

Statistics Department

English Section

Determinants of Access to Social Security Coverage for the Egyptian Labor Force

Graduation Project

Presented by

Aya El Sherbiny (5200873)

Joy Ehab Zaky (5200172)

Mohammed Aly Abdelrahman (5201010)

Saif Essam Abdelmaboud (5200356)

Under The Supervision of

Dr. Rania Mamdouh

Acknowledgment

The Whole Gratitude is due to *ALLAH*

First and foremost, we would like to thank our supervisor Dr. Rania Mamdouh, Assistant Professor of Statistics, at the Faculty of Economics and Political Science, Cairo University. For her great guidance, support, and patience during this project. We also like to thank her for understanding our objectives and for her continuous support in organizing our thoughts and presenting them in the best possible way. Without her great assistance in developing this project, we will not be able to present it.

We wish also to express our sincere gratitude to Dr. Fatma El Zanaty Professor of Statistics, Faculty of Economics and Political Science, Cairo University, for her supervision of all graduation projects in the statistics department of the faculty, for her tireless efforts in answering all of our questions and providing us with the basis for creating a graduation project.

Finally, we dedicate this work and give our love, appreciation, respect, and gratitude to our families for their patience, encouragement, help, and support without whom we would never have enjoyed so many opportunities.

Abstract

Social security is a crucial component of any modern society, providing a safety net that ensures economic stability and security for individuals and families. The concept of social security continues to evolve, adapting to changing economic conditions, demographic shifts, and social needs, with the goal of improving the well-being of individuals and societies. The main objective of the paper is to identify the variables among all their categories that affect access to social security in Egypt the most. Additionally, using insights from the analysis for policymakers to form policy decisions that aim to enhance the effectiveness and inclusivity of Social Security in Egypt. The methodology adopted in our research includes explanatory analysis which involves descriptive analysis and measuring association. In addition to classification models using Binary Logistic Regression, Machine Learning techniques namely: Decision Trees and Random Forest, and Deep Learning using Convolution Neural Networks (CNN). Among the models, the Random Forest and CNN demonstrated the highest overall performance. Logistic Regression was notably effective in identifying individuals with social security coverage, achieving the highest sensitivity. Health insurance appeared as the most significant factor affecting social security coverage. The main demographic factor influencing an individual's decision to enroll in social security in Egypt is age, followed by marital status, then area of residence. As for the labor factors, the sector of employment is the most significant, followed by the individual's occupation and the industry. Finally, for the social factors, educational level is the most significant, followed by the father's presence at home. This comprehensive analysis aims to support policymakers in designing more effective and equitable social security systems for the Egyptian labor force.

Contents

List of Tables

List of Figures

List of Abbreviation

HIECS	Household Income, Expenditure, and Consumption Survey
MENA	Middle East and North Africa
SSI	Supplemental Security Income
CCT	Conditional Cash Transfer
ILO	International Labor Organization
LFS	Labor Force Survey
ERF	Economic Research Forum
CAPMAS	Central Agency for Public Mobilization and Statistics
VIF	Variance Inflation Factor
GVIF	Generalized Variance Inflation Factor
ROC	Receiving Operating Characteristic
AUC	Area Under the Curve
CNN	Convolution Neural Network

Chapter 1:

Introduction

This chapter overviews our research topic and highlights the study objectives. Section 1.1 provides a background of the study and importance of social security, followed by a historical context of social security in Egypt. A detailed review of the available literature on Social Security Coverage is presented in Section 1.2, identifying gaps and setting the stage for our research objectives. The research objectives are stated in Section 1.3. The research questions are formulated in Section 1.4. Finally, Section 1.5 presents the conceptual framework of the study. This chapter lays the groundwork for analyzing and understanding the dynamics that shape social security access in Egypt.

1.1 Background

“Should any political party attempt to abolish social security, unemployment insurance, and eliminate labor laws and farm programs, you would not hear of that party again in our political history.” is a quote said by Dwight David Eisenhower (the 34th US president). It highlights the importance of social security and how crucial it is to society. That being stated, we chose our research to be about the determinants of having access to social security coverage in the Egyptian labor market.

Social security by definition is a monetary assistance from the state to the people without adequate or no income. In the landscape of social welfare, the concept of social security stands as a cornerstone, playing a pivotal role in safeguarding the economic well-being of individuals and families, particularly in times of need or vulnerability. In the context of Egypt, where the intersection of tradition and modernity shapes the socio-economic fabric, understanding the determinants that influence an individual's decision to enroll in social security emerges as a matter of great importance. Egypt, a nation with a rich tapestry of history, culture, and diverse demographics, is undergoing a transformative period marked by economic reforms, urbanization, and demographic shifts. Against this backdrop, social security become not only essential mechanism for social protection but also an instrument for societal development and cohesion. The motive behind this research comes from a recognition of the gaps in our current understanding of the factors shaping social security enrollment in Egypt. While social security is designed to provide a safety net for all citizens, disparities, and challenges may exist, preventing the effective utilization of this enrollment by those who need them the most.

The "Household Income, Expenditure, and Consumption Survey (HIECS) 2019/2020" conducted in Egypt serves as a treasure, offering insights into the lives of individuals across diverse demographic, economic, and geographic spectra. By delving into this dataset, we seek to shed light on the differences in social security enrollment, deciphering the key determinants that influence individual choices in a society undergoing rapid transformation.

1.2 Historical Context of Social Security in Egypt

In the pre-20th century era, Egypt's historical social security landscape was characterized by traditional community support systems and familial care. Social protection was largely informal, relying on familial and community networks. As the mid-20th century unfolded globally, it witnessed the establishment of formal social security systems in many countries, responding to global changes and economic challenges. In Egypt, the post-war period saw initial efforts towards formalizing social security to address the needs of a growing population.

In the subsequent decades, particularly in the 1960s-1970s, Significant steps were taken to institutionalize social security in Egypt during this period. The government introduced social security to cover workers in specific sectors, industries, and public services.

The 1990s and 2000s saw discussions and reforms, influenced by structural adjustments, aiming to enhance the efficiency and sustainability of Egypt's social security systems. Moving into the 2010s, the landscape shifted further with the 2011 Egyptian Revolution, drawing attention to social justice concerns, including those related to social security. The government responded by addressing social protection challenges, striving to broaden coverage and improve the effectiveness of social security systems. Against this backdrop, the Household Income, Expenditure, and Consumption Survey (HIECS) 2019/2020 emerged, providing a contemporary snapshot of Egypt's social and economic conditions. Analyzing this dataset becomes instrumental in comprehending the present dynamics of social security in the country.

1.3 Literature Review

Understanding the determinants of individuals' decisions to enroll in social security is crucial for effective policy design and implementation. This section aims to synthesize existing research on social security determinants globally, with a focus on the Middle East and North Africa (MENA) region and Egypt specifically.

Rupp and Stapleton (1995) provided insights into the determinants of the growth in the Social Security Administration's disability programs. The article focused on trends in applications and awards for Disability Insurance and Supplemental Security Income (SSI). The research methodology encompassed state-level econometric analyses, utilizing a unique approach that "pools" time-series data for a cross-section of individual states. This method controls for changes in national factors, allowing an examination of state-specific dynamics. However, the approach has limitations, particularly in estimating factors with invariant changes across states. The study supplements econometric analyses with actuarial analyses and other evidence to offer a comprehensive understanding of the factors driving disability program growth. Case studies in five states, expert interviews, and a thorough literature review further validate the findings. Key factors explored include the business cycle, state budgetary pressures, the AIDS epidemic, immigration effects, changes in family structure, and programmatic factors. For instance, it is hypothesized that increases in the state unemployment rate positively affect the volume of applications and awards, while cutbacks in state-funded substitute programs have a positive impact. The study delves into specific variables for each program group, considering factors like the aging of the population, labor-force participation, economic restructuring, disabling work injuries, poverty rate, and the relative attractiveness of benefit programs.

Sieverding and Selwaness (2012) concentrated on addressing the shortcomings of the current social protection system, shedding light on the main determinants that lead to enrollment in social protection. Figuring out that social protection has two primary components, which are: social insurance and social assistance. Social insurance, often employment-based, is financed through contributions from workers and employers. It provides coverage for contingencies such as maternity, sickness, disability, and old age. Social assistance, on the other hand, consists of transfers to the poor or other vulnerable groups and is typically tax-financed. This can include pure income transfers or conditional transfers linked to labor supply or human capital investment, such

as Conditional Cash Transfer (CCT) programs (*Norton et al. 2002; Barrientos 2011*). They highlight variables such as age, household headship, and area of residence, noting that these factors significantly influence social protection coverage. For instance, younger individuals, non-heads of households, and rural residents are often less likely to be enrolled in social protection programs. This aligns with broader findings in the literature that demographic and geographic factors can impact access to social benefits.

Roushdy and Selwaness (2017) have done an empirical analysis exploring access to social insurance in the Egyptian labor market, addressing the critical question of who is covered and who underreports. the researchers delve into the intricacies of social insurance coverage in the context of the Egyptian labor market. The primary objective of the study is to empirically analyze the factors influencing access to social insurance among workers in Egypt, shedding light on the determinants of coverage and the phenomenon of underreporting. To achieve these objectives, the researchers utilized data from the Egyptian Labor Market Panel Survey of 2006 and 2012, employing a probit regression model to estimate the likelihood of social insurance enrollment for both wage and non-wage workers. The methodology involved addressing potential endogeneity issues between the type of work and social insurance access. The results of the study revealed that men, older individuals, married individuals, those with higher education, and white-collar, highly skilled workers were more likely to be socially insured. Interestingly, the analysis also highlighted correlations between the underreporting of insurable wages and various factors such as working outside the establishment, basic monthly wage, tenure, and occupational characteristics. The study provides valuable insights into the socio-demographic patterns of social insurance coverage in the Egyptian labor market, offering a nuanced understanding of the dynamics shaping access to social insurance programs.

Merouani et al. (2021) investigated social security enrollment as an indicator of state fragility and legitimacy through a field experiment in Maghreb countries. This study relied on the Sahwa dataset, originating from a comprehensive survey encompassing 10,000 households across Algeria, Egypt, Lebanon, Morocco, and Tunisia. Executed in 2016, the survey concentrated on youth empowerment, specifically analyzing individuals aged 15–29 from each household. The methodology employed involves a field experiment, allowing for a real-world exploration of individual choices regarding social security participation in the context of state fragility. The study

employed a weighted logit model to understand the factors influencing workers' choices regarding informality. This model encompassed socio-demographic variables to estimate the likelihood of opting for informality. The dependent variable, choosing informality, was measured based on responses to why individuals are not affiliated with the social security system. The logit model is particularly useful for this binary choice scenario. The ultimate goal of the study was to provide policy recommendations that enhance social security extension to all workers in the studied countries. Noteworthy findings include the association between education levels and insurance likelihood, highlighting the need for tailored incentives for highly educated workers. The study also emphasizes the importance of refining self-employed insurance schemes to attract broader participation. Job satisfaction, and individualistic tendencies emerge as significant determinants influencing social security participation. The study concluded with targeted policy recommendations, suggesting specific measures for categories such as contributing family workers, non-permanent employees, women, single individuals, poorly educated workers, low-income workers, and those in the building sector.

Kassem (2021) interviewed 12 policy makers and experts in relevant governmental, international and non-governmental bodies in depth, aiming to provide an analytical framework that identifies the governance conditions and challenges that are currently facing the Egyptian Social security system and to analyze the level of the social protection system as a whole and the program level, by examining both contributory and non-contributory cash transfer programs. The study adopted a framework that identifies three operational entry points of governance— Rules, Roles and Controls— as well as two spheres of analysis— program and sector levels. Snowballing sampling technique was used to reach the different study participants. The auther used qualitative research methods while conducting her research, where she divided the contributory and non-contributory schemes from the beginning of the analysis. She used an open-coding approach by segmenting and clustering data into themes and describing them using short sentences. Coding was then guided by the conceptual framework’s division of the meso- and micro-level analysis. The recommendations of this analysis center around the establishment of an institutional home for social protection, the establishment of a common vision, the development of the necessary institutional and administrative capacity, a rigorous monitoring and evaluation toolkit, undergoing regular governance examinations, among others.

Merouani and Lassassi (2021) addressed the challenges of social security coverage in developing countries, emphasizing it as a fundamental human right recognized by the International Labor Organization (ILO). Focusing on the case of informal workers in Algeria, the study employed the Contingent Valuation method through a face-to-face survey of 650 private sector workers in the Algiers governorate. The methodology involves understanding individuals' willingness to pay for social security, considering income and certainty about affording the premium. Techniques were employed to reduce hypothetical bias, including a double-bounded query for those unsure of their payment capacity. Results reveal an average willingness to pay of 2,900 Algerian dinars, slightly below the minimum required by the social security system. The study identified differences in willingness to pay based on demographics, income, age, and risk aversion.

Barsoum and Selwaness (2022) explored the potential impact of design changes on enrollment in Egypt's reformed social insurance system. This study utilized data from the Egyptian Labor Market Panel Survey of 2006 and 2012. The method involved estimating the likelihood of being enrolled in social insurance using a probit regression model for all workers, separately for wage and non-wage workers. The potential endogeneity between the type of work and social insurance access is addressed using an instrumental variables approach. Results show that men, older individuals, married individuals, those with higher education, and white-collar highly skilled workers are more likely to be socially insured. Under-reporting of insurable wages is positively correlated with working outside the establishment and basic monthly wage, while it is negatively correlated with tenure and white-collar occupations. High contribution rates and weak law enforcement encourage employers and employees to either not participate or contribute amounts lower than their actual wages. The study finds that wage workers are more likely to have social insurance coverage than non-wage workers. The results also showed that Experience increases the likelihood of having access to social insurance coverage among wage workers but does not seem to play a significant role among non-wage workers.

Assaad and Wahby (2023) analyzed the decline in social security coverage in Egypt from 2007 to 2021, aiming to determine whether this decline was due to compositional shifts in the structure of the economy or the workforce, or changes in coverage for specific types of jobs and workers, or the workforce, or changes in coverage for specific types of jobs and workers. Their

findings indicate that within-sector loss of coverage rather than compositional shifts in the economy or the workforce is responsible for the bulk of the decline in coverage, especially over the period of rapid decline from 2014 to 2017. However, to abstract from annual fluctuations and obtain a more accurate trend for each sub-period by sector, they smooth the data for sectoral employment and sectoral covered employment by using piecewise linear trendlines. They used multivariate analysis where they put into consideration if the individual has an additional job, the characteristics of the individual, and their effects on the likelihood of social security coverage. At the end, they concluded that the drop in social insurance coverage in Egypt, especially the drop that occurred in the period from 2014 to 2017 or even 2018 occurred because specific types of jobs were less likely to be covered rather than because the nature of jobs in the economy changed. This suggests that the decline is due to changes intrinsic to the social insurance system itself or in the proximate policy environment that affects the calculus of employers and workers to obtain coverage.

Loewe (2024) focused on social security concerning poverty and social inequality in Egypt. The analysis delved into the evaluation of implemented social protection systems and factors influencing social security effectiveness. In the study's exploration of poverty and social shocks, it was revealed that over a quarter of Egypt's population lives below the absolute poverty line, attributed to household vulnerability to social shocks. The poor include a significant percentage of women-headed households and the disabled, with social risks remaining a major contributor to poverty. In evaluating social protection systems, the study highlighted the forms of social insurance, social assistance, and state subsidies in Egypt. However, these systems suffer from inefficiencies and unequal distribution of benefits, slowing down their effectiveness in poverty prevention. The major problem identified is not the lack of resources but the unequal distribution and inefficiencies, limiting the ability to prevent poverty, especially for vulnerable populations with limited access to education, healthcare, and productive assets. The paper concluded that there was a pressing need for policy reform to enhance the effectiveness and equity of social security systems in Egypt. The focus of reform should include improving the allocation of resources, expanding coverage, and raising public awareness about social risks. Through comprehensive policy reform, Egypt can strengthen its social security systems and better protect citizens from the impacts of poverty and social shocks.

Selwaness and Barsoum (2024) declared that the rates of the participation of Egyptian workers in contributory social insurance have continued to decline, even within the years that the country has had positive annual growth rates. The main reason for this is that the cost of contribution has been rising, driven by a substantial increase in the minimum wage since 2016, which has made the scheme much more expensive. They also stated that the share of regular wage work in Egypt expanded from 24% in 2009 to 34% of total employment in 2021. Much of this 10-percentage point increase was in regular jobs with no coverage despite mandatory coverage. By 2021, only 35% of regular wage workers in the private sector were socially insured, down from 52% in 2009.

- **Gaps in the Literature**

1. **Lack of Individual-level Analysis:** Existing literature primarily offers a high-level view of social security challenges and policy recommendations, with limited focus on individual-level determinants. Insufficient exploration of how demographic and socio-economic variables, such as gender, age, marital status, and educational levels, directly influence social security status.
2. **Limited Focus on Specific Variables:** While some studies touch upon social security and its determinants, there is a need for a more detailed examination of the role played by specific demographic variables, chronic diseases, disability, and other relevant factors. The literature lacks comprehensive insights into how these variables interact and contribute to disparities in social security access.
3. **Incomplete Exploration of Policy Impacts:** While policy reform is highlighted as necessary, there is a lack of detailed exploration into the specific impact of policy changes on social security enrollment. Insufficient evidence-based insights into which policies are most effective and how they address disparities and challenges identified in the study.

1.4 Research Objectives

This paper focuses mainly on understanding and examining the factors that influence social security status for individuals in the labor market, this is further investigated using data from the Household Income, Expenditure, and Consumption Survey (HIECS) 2019/2020. To elaborate, the project focuses on the individuals who has the ability to work, and the chance to be enrolled in social security. To achieve this, the study will be divided into two main parts:

- ✓ **Part one: Exploratory Analysis**

  This part of the study will involve an exploratory analysis to investigate the relationship between various factors and social security status. This part will include a detailed descriptive analysis that summarizes the main characteristics of the dataset, visualizes the classification of the variables. This will be followed by measuring the association between social security enrollment and potential influencing factors, such as demographic characteristics (Age, Gender, Area), labor factors (Employment status, Sector of Employment, Occupation, Industry), health factors (Chronic Disease, Health Insurance, Disability Status) and social factors (Educational level, Father's presence at home, Marital Status). Techniques such as the chi-squared test of independence, or odds ratio measurements will be used to explore these relationships. This exploratory phase will help is understand the patters and correlations within the data, and it also provides a foundation for the second phase.

- ✓ **Part Two: Predictive/Classification Modeling**

  Based on the findings from the exploratory analysis, we can say that we can now classify the status of the individual based on their likelihood of enrolling in social security. In this part, we will apply binary logistic regression, and then evaluate the model using metrics such as accuracy, precision, and recall. After validating the model, it can be used to predict social security enrollment based on the characteristics. We will also apply machine learning techniques (Decision Tree, Random Forest) and deep learning techniques (CNN), to see alternative models.

Applying the analysis in our research is useful in real-life scenarios in several ways as the research can be helpful in policy Improvement as the insights reached at the end will be of great interest to policymakers to help them design and implement effective social security systems. This takes us to the next point of improvement which is resource allocation as inequalities can be identified thus helping policymakers prioritize resources and allocate them to those who genuinely need them so that resources are effectively utilized and provided to the target population. Additionally, the use of demographics in the response variables is of great help in identifying which groups may be facing barriers to accessing social security. This can help ensure people with disabilities have the needed opportunities to benefit from it.

1.5 Research Questions

1) What are the primary demographic factors influencing an individual's decision to enroll in Social Security in Egypt?
2) How do Labor factors, such as employment status and occupation, correlate with Social Security enrollment among individuals?
3) To what extent does the presence of chronic diseases or disabilities influence an individual's likelihood of Social Security enrollment?
4) Is there a discernible impact of family structure, as indicated by the presence of the father at home, on an individual's decision to enroll in Social Security?
5) What is the most appropriate method to give the best results based on classifying the social security enrollment among the labor force?
6) What insights can be drawn from the analysis to inform policy-makers for enhancing the effectiveness and inclusivity of Social Security in Egypt?

1.6 Conceptual Framework

The conceptual framework carefully outlines the multifaceted factors influencing social security enrollment. It categorizes the factors, as shown in (Figure 1.1), into four main groups: demographic factors, health factors, labor factors, and social factors.

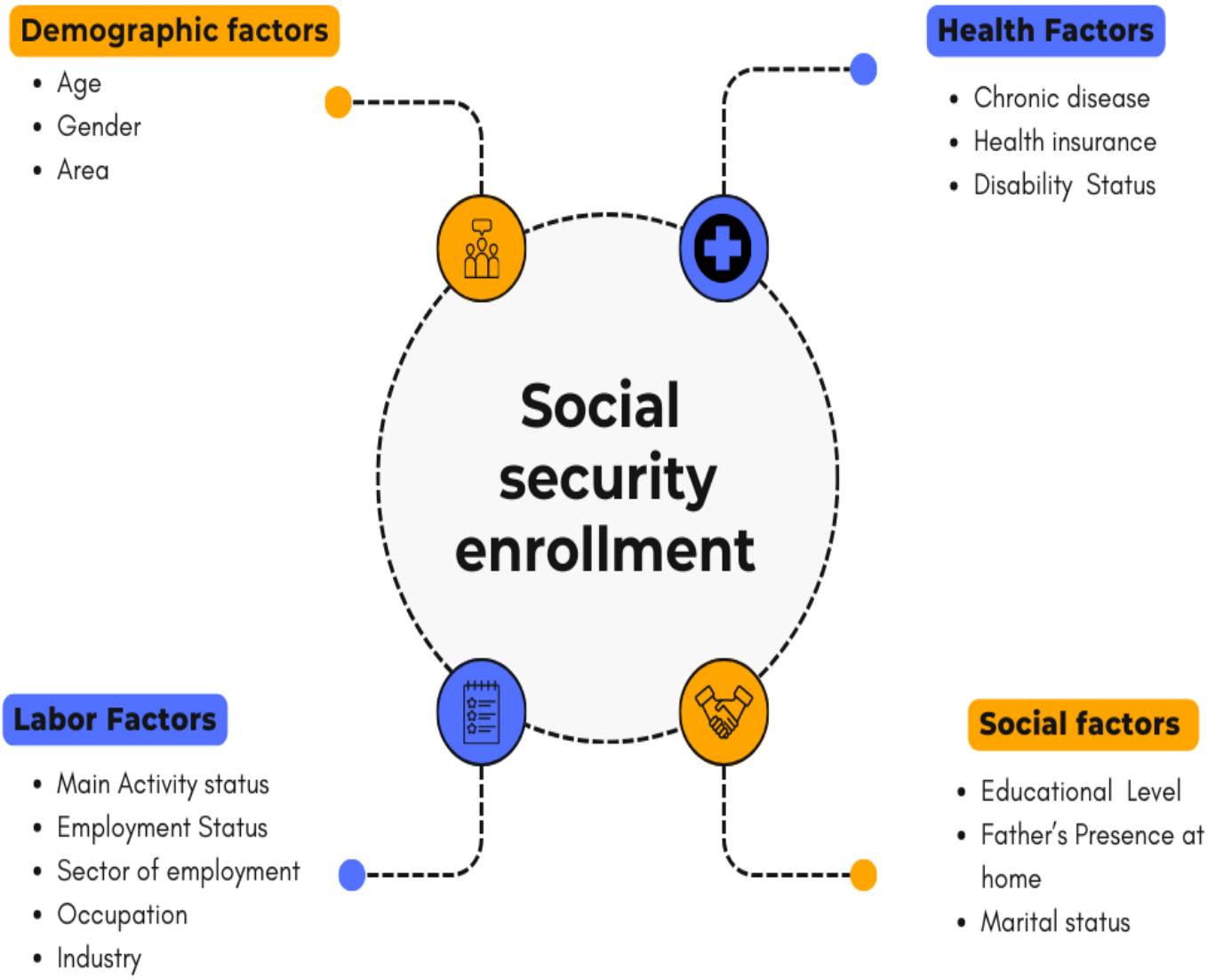

Figure 1.1: Conceptual Framework for Social Security Enrollment

Chapter 2:

Sample Design and Research Methodology

This chapter explains the sample design and the research methodology. In section 2.1, we provide a detailed description of the dataset, sampling technique, and the steps of data preparation and cleaning. Section 2.2 explains each variable in the dataset and its categories. Section 2.3 addresses the limitations of the dataset. Section 2.4 describes the methodology used in this research to ensure transparency, reproducibility, and robustness in our findings. And finally, section 2.5 visualizes the scaling of the quantitative variable (Age).

2.1 Data Description

The data for this project was collected during the "Household Income, Expenditure, and Consumption Survey (HIECS) 2019/2020" conducted in Egypt, making it a cross-sectional analysis capturing a snapshot of social security status and its determinants during that period. The fieldwork spanned from October 1, 2019, to September 30, 2020, with the dataset subsequently published by the Economic Research Forum (ERF) on June 12, 2023. The survey, which covered Egypt, Morocco, and Jordan, focused on individual and household information, including demographic, social, labor, and health characteristics. The dataset encompasses approximately 45200 thousand individuals, providing a comprehensive and representative source for the examination of social security enrollment and its determinants among individuals in Egypt. The data that we are working on represents only 50% of the data collected by the Central Agency for Public Mobilization and Statistics (CAPMAS), as this is the only accessible portion shared by ERF and CAPMAS.

As our analysis focuses on the active labor force, we started by removing anyone under the legal working age, which the survey set to be 6 years based on the International Labor Organization (ILO). Therefore, we removed anyone below 6 years old. After reviewing the remaining observations, we found that many people were not actively working or had never been actively working, as they were still studying or undergraduates. So, we decided to keep the age at 18, for which the data became balanced and more reliable. Thus, we applied our research to 13,472 observations.

- Sampling Frame: The Individual data from HIECS 2019/2020, focusing on demographic, social, labor, and health characteristics.

- Sampling Technique: Stratified Random Sampling, ensuring representation across key demographics and regions.
- Data Source: Household Income, Expenditure, and Consumption Survey (HIECS) 2019/2020 from the Economic Research Forum (ERF) Data Portal.
- Data Collection Method: Survey data collected through interviews and self-reporting.

2.2 Description of Variables

- **Response Variable**

Social Security: This variable indicates whether an individual has access to social security. It is a categorical variable with two possible values: "Yes" (indicating the individual has social security coverage) and "No" (indicating no coverage).

- **Explanatory Variables**

➢ **Demographic Factors:**

Age: A continuous variable that determines the individual's age.

Gender: This is a Binary variable that explains the individual's gender (Male/ Female).

Area: If an individual lives in a Frontier governorate area, urban area, or rural area.

➢ **Labor Factors:**

Employment Status: This is a categorical variable representing the employment status of the individual, with values such as "Employee", "Employer", "Own-account, self-employed", and "Family worker".

Sector of Employment: This is a categorical variable indicating the sector in which the individual is employed, with values include "Public sector", "Private sector", "Government", "Joint/Cooperative", and "Other sectors".

Occupation: This is a categorical variable representing the individual's occupation. Categories include "Managers", "Professionals", "Technicians and associate professionals", "Clerical support workers", "Service workers and shop and market sales workers", "Skilled agricultural, forestry and

fishery workers", "Craft and related trades workers", "Plant and machine operators, and assemblers", and "Elementary occupations".

Industry: The industry of the economy in which an individual is employed. This variable includes the following categories: “Mining”, “Insurance and Real Estate”, “Electricity and Utilities”, “Public Administration”, “Transportation, Storage and Communication”, “Construction”, “Manufacturing”, “Commerce”, “Agriculture and fishing”, and “Other services”.

- **Health Factors:**

Disability Status: This is a categorical variable indicating whether the individual has a disability, with possible values "Yes" and "No".

Chronic Disease: This is a categorical variable indicating whether the individual has a chronic disease, with possible values "Yes" and "No".

Health Insurance: This is a categorical variable indicating whether the individual has health insurance, with possible values "Yes" and "No".

- **Social Factors:**

Father's Presence at Home: This is a categorical variable indicating whether the father is a household member, with possible values such as "Father is not a household member", "Father is a household member", or “Father is dead”.

Educational Level: This is a categorical variable representing the highest level of education attained by the individual. Categories include "none", "primary", "secondary", "post-secondary", "post-graduate", and "university".

Marital Status: The individual’s current marital condition, and it includes 5 categories: Married monogamous, a person who is married to only one person. Married Polygamous, married to more than one person**.** Widowed, their spouse died. Divorced/Separated, used to be married but now they are not. Lastly, Never Married

2.3 Limitations

Even though the data was very detailed and beneficial, and the survey covered a wide range of topics, it still missed some critical points that could have significantly enhanced our research. Firstly, the data does not include the individual's income, which is a key variable. Income levels are likely to be closely related to social security enrollment, as financial stability often influences an individual's ability and willingness to participate in social security. The absence of this data limits our ability to fully understand the economic factors influencing enrollment. Secondly, the data includes many missing values for observations below the age of 18. This issue arises from the data collection process, where questions directed at individuals under 18 were either unanswerable by this age group or frequently skipped. The dataset also may not fully represent all segments of the Egyptian labor force. Specific groups or regions might be underrepresented. Moreover, advanced models such as Random Forest and CNN, while highly accurate, lack interpretability, making it challenging to draw clear policy implications from the results. This limitation restricts our ability to analyze the direct impact of some determinants on individual social security participation. Consequently, these data gaps hinder our ability to perform a more comprehensive analysis and to draw precise conclusions about the factors influencing social security enrollment.

2.4 Methodology

We began with a detailed descriptive analysis using Python, R, and STATA followed by measuring the association between the variables. We used the appropriate method based on the types of variables under examination:

For any two binary variables, we used the odds ratio to measure the association between the variables.

Table 2.1: 2x2 Contingency Table

	Category 1	Category 2
Category 1	a	b
Category 2	c	d

$$Odds\ Ratio = {}^{a.d}/_{b.c}$$

$$CI\ of\ Ln\ (Odds\ Ratio) = \ln(OR) \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}})$$

$$CI\ of\ Odds\ Ratio = (e^{Ln(OR1)}, e^{Ln(OR2)})$$

For multi-categorical variables with social security, we used the chi-squared test of independence, followed by Cramer's V to measure the strength of the association.

For the Chi-Squared test of independence:

$\boldsymbol{H_0}$: There is no association between the two variables.

$\boldsymbol{H}_1$: There is an association of any kind.

$$x^2 = \sum \frac{(Observed-Expected)^2}{Expected}$$

For Cramer's V:

$$V = \sqrt{\frac{\frac{X^2}{n}}{\min\,(I-1,J-1)}}$$

$x^2 = chi\ squared\ calculated\ in\ the\ previous\ tes$t

$n = sample\ size$

$\min(I-1, J-1)$

$=$ the smallest number of degrees of freedom between the two categorical variables.

Where; V must lie between 0 and 1, where 0 indicates complete independence, and 1 indicates complete independence dependence or association between the variables.

For measuring the association between an ordinal variable and social security, we used the Goodman and Kruskal's gamma test.

The gamma Coefficient ranges between (-1,1):

1 = perfect positive correlation: if one value goes up, so does the other.

-1 = perfect inverse correlation: as one value goes up, the other goes down.

0 = there is no association between the variables

Gamma is calculated as:

$$\gamma = \frac{N_C - N_D}{N_C + N_D}$$

Where;

- N_c is the total number of pairs that rank the same (concordant pairs)
- N_d is the number of pairs that don't rank the same (discordant pairs).

As for the method of analysis, it involves Binary Logistic Regression, chosen due to its suitability for binary response variables, in this case, the social security status (Yes/No). This statistical technique allows us to model the relationship between social security enrollment and various explanatory variables, including demographic factors, socio-economic indicators, health variables, and occupational details.

We also applied the Decision Tree technique, where health insurance was implemented as the root node. To get better results, we then applied the Random Forest method, which enhances predictive accuracy by aggregating multiple decision trees. Finally, we used the Convolutional Neural Network (CNN) model, a deep learning technique that can provide improved results by capturing complex patterns in the data.

Finally, model evaluation is applied, including performance metrics like accuracy and precision, with cross-validation ensuring robustness across different subsets of the data. Ethical considerations, bias minimization, and explainability are integral components of our methodology to ensure fair, transparent, and trustworthy results. The goal is to derive actionable insights for policymakers based on a thorough analysis of the determinants of social security enrollment among workers in Egypt.

2.5 Data Screening

First, we examined the age structure within the dataset. Initially, we identified outliers within the age variable using the boxplot visualization technique as shown in Figure 2.1. The outliers were determined to be individuals with ages ranging from 80 to 89 years old. We found 21 outliers detected, representing only 0.1556% of the observations. This result is also supported by Rosner's test.

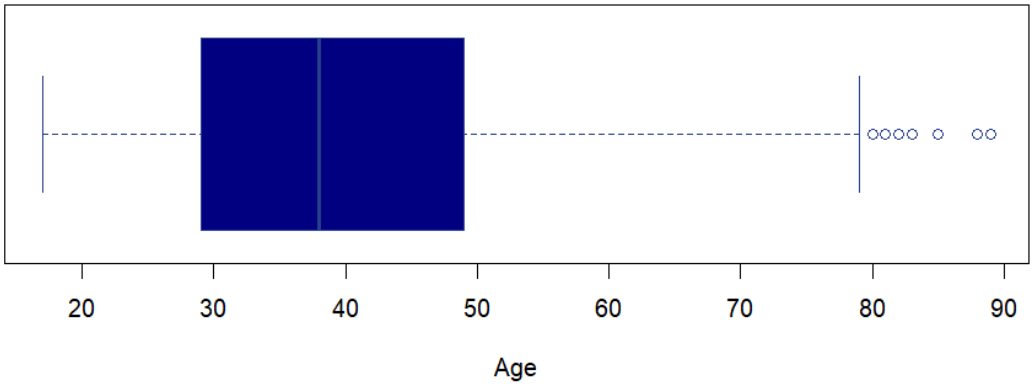

Figure 2.1: Boxplot for Age Structure

Despite their classification as outliers, it was decided not to remove these data points from the analysis. This decision was made in alignment with the study's objective, as it is essential to capture the full scope of the age variable to inform policy and understand the age dynamics within social security systems. Retaining the outliers allows for a comprehensive examination of the dataset, ensuring that any potential disparities or inequalities identified in the study are adequately addressed. By including individuals across a wide age range, ranging from younger to older adults, the analysis can capture the full spectrum of demographic diversity within the population under study. This approach enhances the robustness and inclusivity of the analysis, providing valuable insights into the factors influencing Social Security Status across different age groups. Fortunately, there was no missing values in the dataset.

Chapter 3:

Descriptive Analysis

This chapter provides a descriptive analysis of the dataset, in an attempt to visualize and explore the main features and relationships among the different variables. By analyzing demographic and socio-economic factors, we aim to identify significant patterns and correlations. These will help us understand the determinants of social security coverage and highlight policy interventions to enhance social protection for the Egyptian labor force.

Social Security

From Figure 3.1, we can observe that 60.4% of individuals do not have Social Security coverage, indicating that a significant majority of the surveyed population is not covered by Social Security systems. This lack of coverage could imply a gap in social protection and financial security for a large portion of the population. While 39.6% of individuals have social security coverage, a little over a third of the surveyed population is covered by social security. This proportion represents those who potentially have access to benefits such as pensions and other social security-related services. The chart highlights the need for policy interventions to increase social security coverage, ensuring that a larger proportion of the labor force is protected and has access to necessary social security benefits.

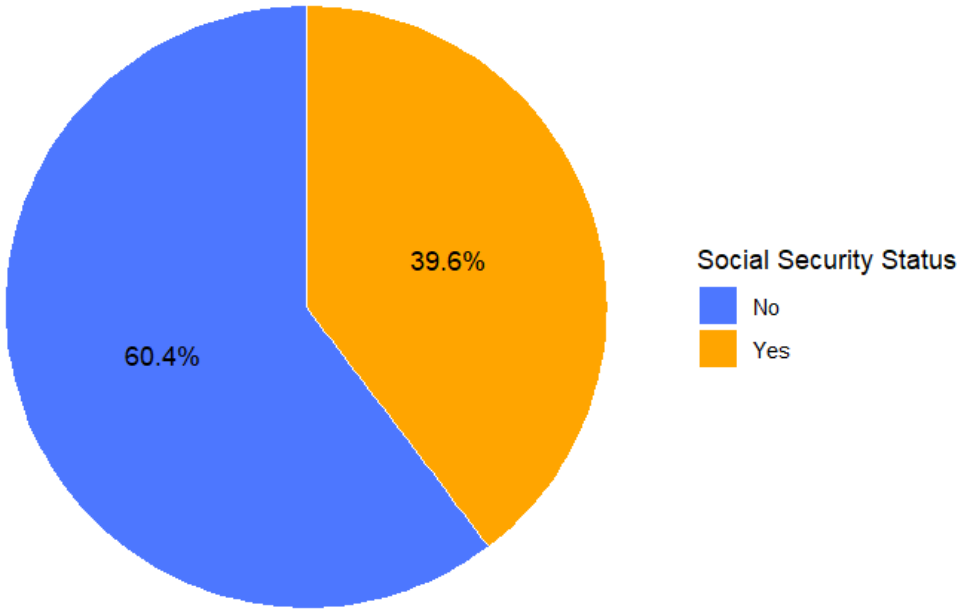

Figure 3.1: Pie Chart for Social Security Status

Age Structure

Here, the age distribution ranges from 18 to 89 years. As shown in Figure 3.2, it appears to be a little positively skewed and roughly symmetrical with a peak around the 30-35 age range. This suggests a notable concentration of individuals in their late 20s to early 40s, reflecting a predominantly working-age population. The average age of individuals in the dataset is approximately 39.38 years, indicated by the red dashed line.

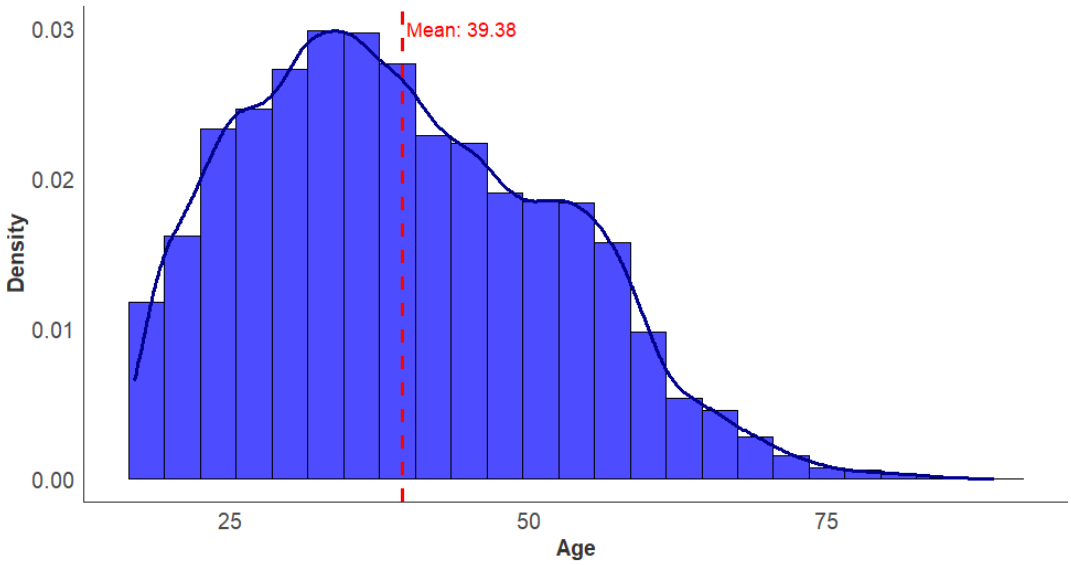

Figure 3.2: Histogram for the Age Distribution

Gender

As we can see from Figure 3.3, the dataset has a much higher proportion of males compared to females, with males making up 77.3% and females 22.7% of the sample. This gender imbalance could have implications for analyses related to gender-specific determinants of social security coverage and labor market outcomes.

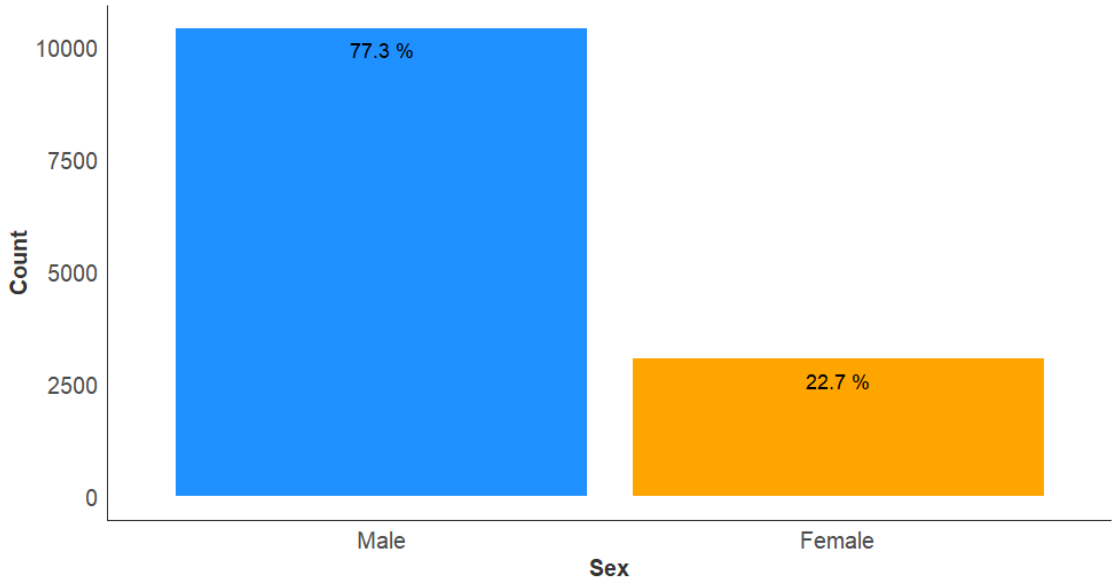

Figure 3.3: Bar Chart for Gender Distribution

From Figure 3.4 we can notice that a higher percentage of females (45.7%) receive social security benefits compared to males (37.9%).

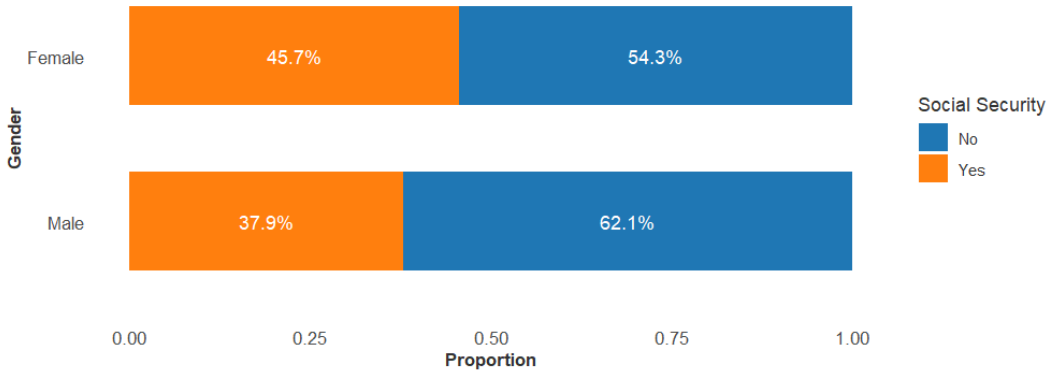

Figure 3.4: Stacked Bar Chart for Gender by Social Security

The odds ratio is used to assess the nominal association between benefiting from social security and the gender of the respondent. The odds ratio of 2.06543582 indicates that the odds of females being enrolled in social security is approximately 2 times higher than the odds of males

being enrolled in social security. As for the 95% confidence interval of the odds ratio (1.951502912, 2.1965723), it suggests that the difference in odds is statistically significant. To measure the association between them, we decided to use Pearson's chi-squared test which shows that there is a strong association between sex and social security (p-value = 1.151e-14).

Marital Status

From Figure 3.5, we can notice a general trend that individuals who are never married have a notably higher proportion of no social security coverage compared to other marital statuses. Married monogamous, married polygamous, and divorced/separated groups have a more balanced distribution, though there is still a slight tendency toward lacking coverage. On the other hand, widowed individuals have a significant majority of coverage of social security. This chart highlights disparities in social security coverage across different marital statuses, which could inform targeted interventions to improve social security access for groups who are more in danger.

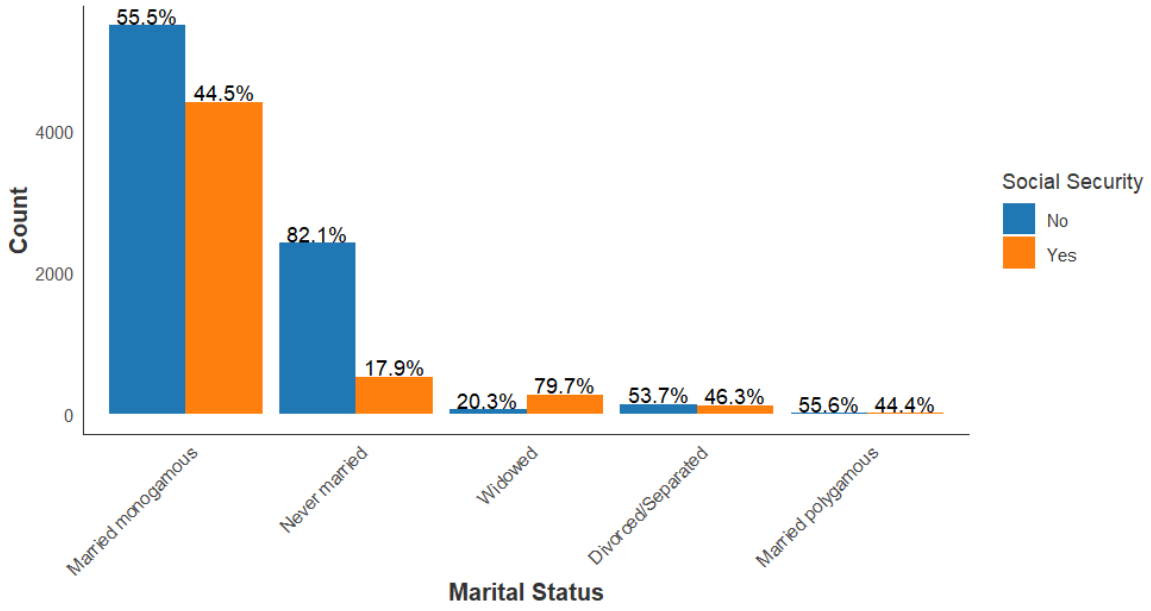

Figure 3.5: Clustered Bar Chart for the Marital Status by Social Security

Chi-squared test is used to check the independence between marital status and social security enrollment (χ^2 = 914.65, p-value < 2.2e-16), which shows that with 95% confidence, they

are not independent. To measure the strength of the association between them, we used Cramer's V (0.261), which indicates a moderate association between Marital Status and Social Security.

Area

From Figure 3.6, we can notice that Individuals in rural areas are more likely to lack social security coverage compared to those in urban areas and frontier governorates. Urban areas have a more balanced distribution of social security coverage, while frontier governorates show a majority with coverage.

The Chi-squared test is used to measure the independence between area and Social Security enrollment ($\chi 2 = 317.26$, $p < 2.2e{-}16$), indicating with 95% confidence that there is an association between them. Additionally, Cramer's V (0.1534579) is employed to assess the strength of association, suggesting a weak association between area and Social Security.

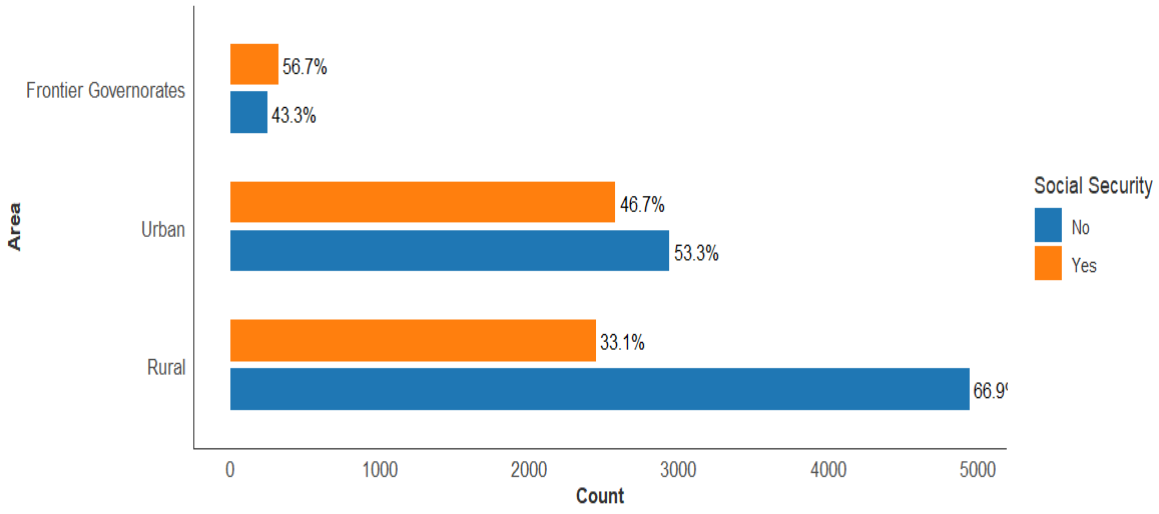

Figure 3.6: Clustered Bar Chart for Area by Social Security

Father's Presence at Home

As we can see from Figure 3.7, individuals with fathers present at home are the most likely to lack social security coverage. The distribution for individuals whose fathers is not household

members shows a more balanced coverage compared to other categories. While individuals whose fathers are deceased have a higher proportion of social security coverage compared to other groups.

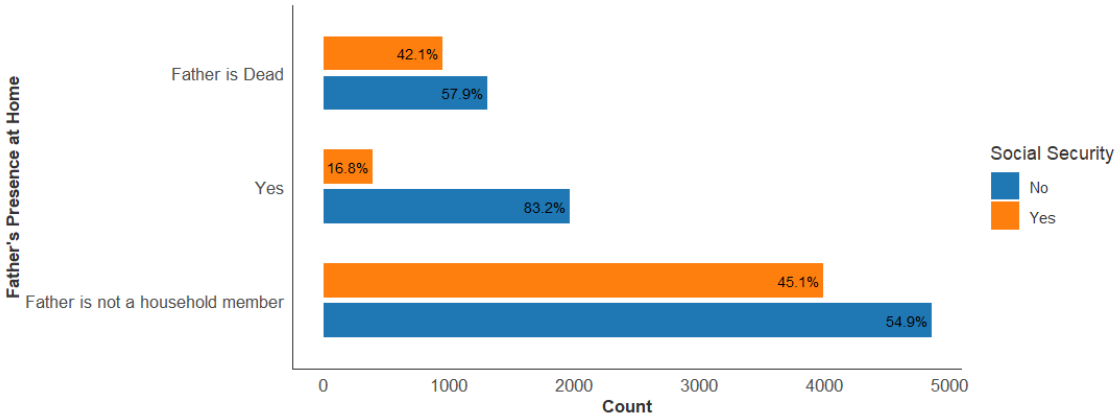

Figure 3.7: Clustered Bar Chart for Father's Presence at Home by Social Security

Chi-squared test is used to measure the independence between the father's presence at home and social security enrollment(χ^2 = 631.56, p-value < 2.2e-16), which shows that with 95% confidence, there is an association between them. To measure the strength of association we used Cramer's V (0.2165163), which indicates a weak association between the father's presence at home and Social Security.

Educational Level

We can notice from Figure 3.8 that higher levels of education (university, post-secondary, and postgraduate) are associated with higher proportions of social security coverage. While individuals with no education or lower levels of education (secondary, primary/lower secondary) are more likely to lack social security coverage. Hence, efforts to improve social security access could focus on individuals with lower educational attainment to reduce the coverage gap.

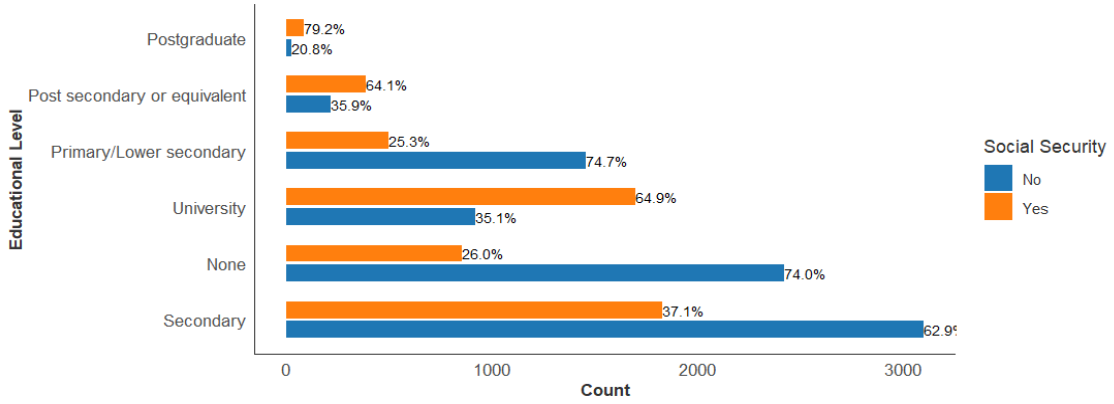

Figure 3.8: Clustered Bar Chart for the Educational Level by Social Security

Goodman and Kruskal's Gamma coefficient is used to study the association between social security enrollment and the educational level of individuals. The gamma value ($\gamma = 0.334$, $p < 2.2e\text{-}16$) suggests a positive moderate monotone association between social security enrollment and educational level. This indicates that individuals with higher educational levels are moderately more likely to be enrolled in social security.

Disability Status

As we can see from Figure 3.9, both disabled and non-disabled groups have a majority lacking social security coverage, but the proportion is higher among non-disabled individuals. The disabled group has a higher proportion of individuals with social security coverage compared to the non-disabled group, indicating better access for this group.

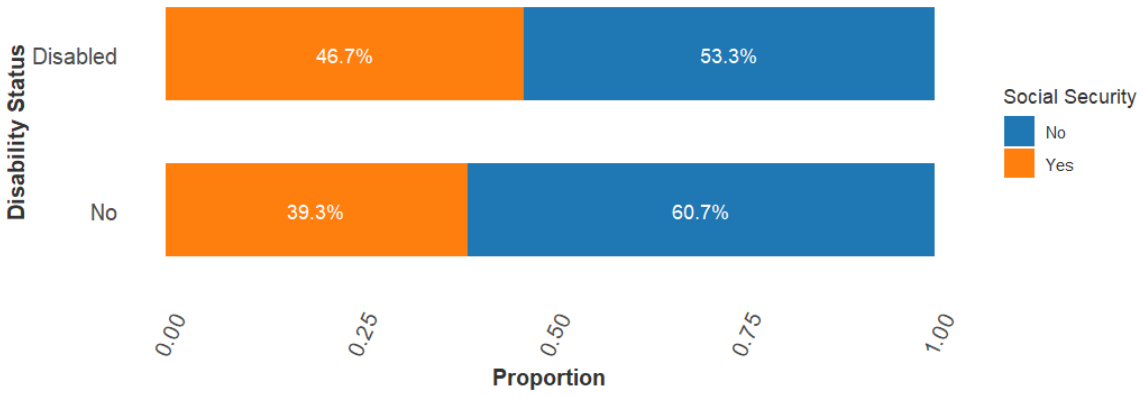

Figure 3.9: Stacked Bar Chart for Disability Status by Social Security

The nominal association between enrollment and benefiting from social security disability status is assessed using the odds ratio. The corresponding 95% confidence interval is (1.88165122, 2.366271477) indicating that the difference in odds is statistically significant. Where the odds ratio is 2.091515871 indicating that the odds of disabled individuals being enrolled in social security is about 2 times higher than the odds of non-disabled individuals being enrolled, and validating these results by chi-squared test, we can see that with 95% confidence, there is a significant association between disabilities and social security status.

Chronic Disease

From Figure 3.10, we can notice that Individuals with chronic diseases are more likely to have social security coverage than those without chronic diseases. Where there is a notable disparity in social security coverage between individuals with and without chronic diseases, with those having chronic diseases showing better coverage.

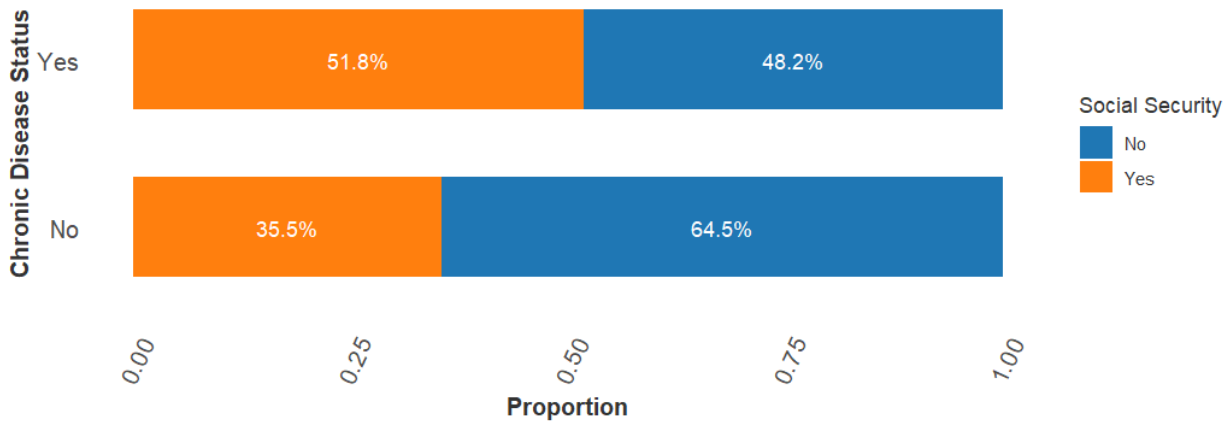

Figure 3.10: Stacked Bar Chart for Chronic Disease Status by Social Security

The odds ratio is used to assess the nominal association between benefiting from social security and if the individual has chronic disease or not. The 95% confidence interval of the odds ratio (1.807068, 2.114694) supports a significant relationship between enrollment in social security

and having a chronic disease, where having a chronic disease increases the odds of being enrolled by at least 80% and at most 111% at the 95% confidence level.

Health Insurance

Figure 3.11 shows that individuals with health insurance are much more likely to have social security coverage than those without health insurance. There is a notable disparity in social security coverage between individuals with and without health insurance, with those having health insurance showing significantly better coverage.

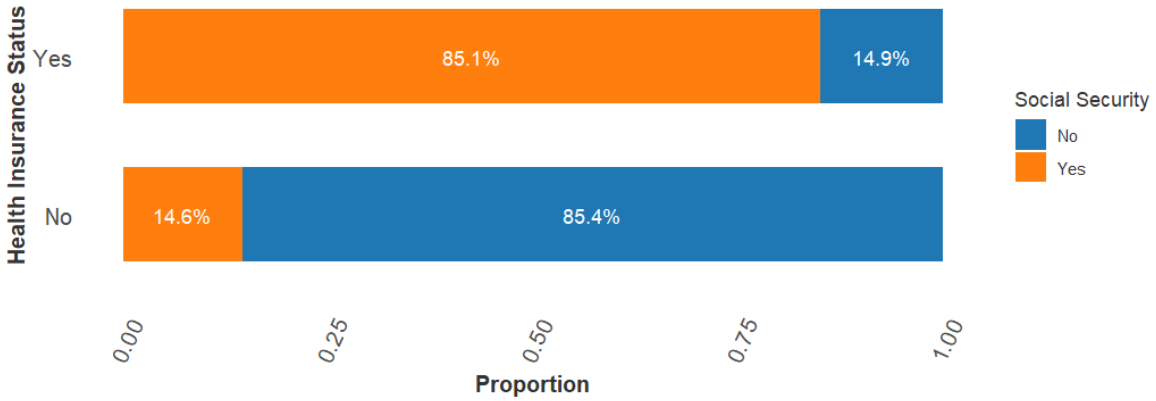

Figure 3.11: Stacked Bar Chart for Health Insurance Status by Social Security

The odds ratio indicates that the odds of individuals with health insurance being enrolled in Social Security are approximately 33.4 times higher than the odds of individuals without health insurance being enrolled. The confidence interval suggests that with 95% confident that the true odds ratio is between 30.25182 and 36.90782. applying chi squared test was giving the same results, that there is a statistically significant association between Health insurance status and social security status.

Employment Status

Here, from Figure 3.12, the vast majority of the surveyed individuals are employees, followed by employers and self-employed individuals. Contributing family workers make up the smallest segment.

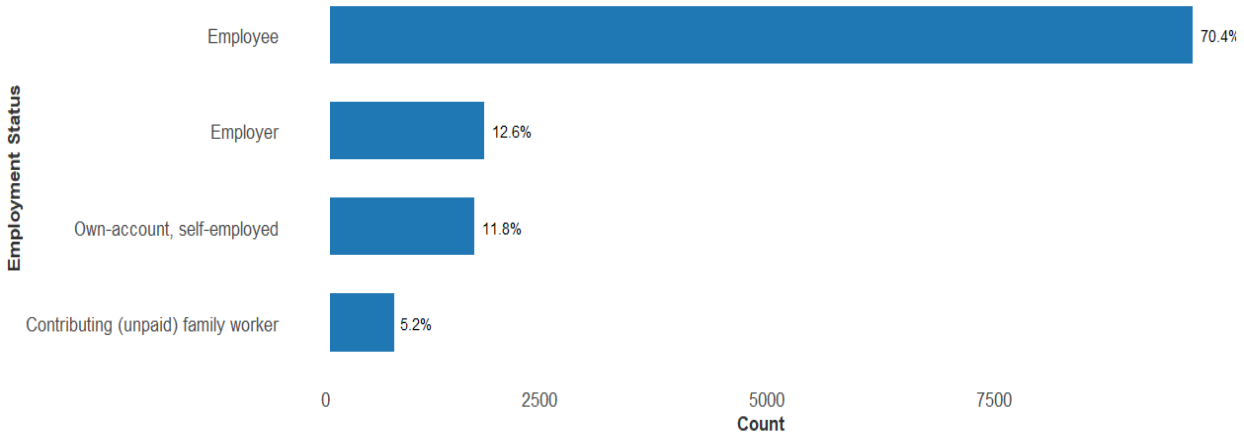

Figure 3.12: Bar Chart for the Employment Status

As shown in Figure 3.13, there is a general trend that all employment status categories show a majority lacking social security coverage, with the highest vulnerability observed among contributing (unpaid) family workers. Here, employees show a more balanced distribution compared to other categories, though still with a majority lacking coverage. Contributing family workers and self-employed individuals are particularly vulnerable, with the highest proportions lacking social security coverage.

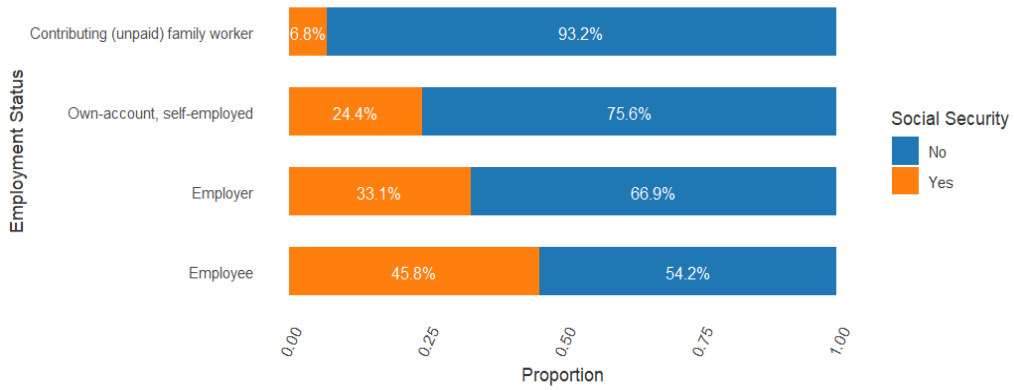

Figure 3.13: Stacked Bar Chart for the Employment Status by Social Security

The Chi-squared test is used to measure the independence between employment status and Social Security enrollment (χ^2 = 631.56, p < 2.2e−16), indicating with 95% confidence that there is an association between them. Additionally, Cramer's V (0.2165) is employed to assess the strength of the association, suggesting a weak association between employment status and Social Security.

Sector of Employment

From Figure 3.14, it is noticeable that the majority of the surveyed individuals are employed in the private sector, followed by the government sector. The public sector, joint/cooperative, and other sectors make up much smaller portions of the population. The chart highlights the dominance of the private sector in employment, with significant but smaller proportions in the government sector. This contributes to our objective and will help us achieve the desired information.

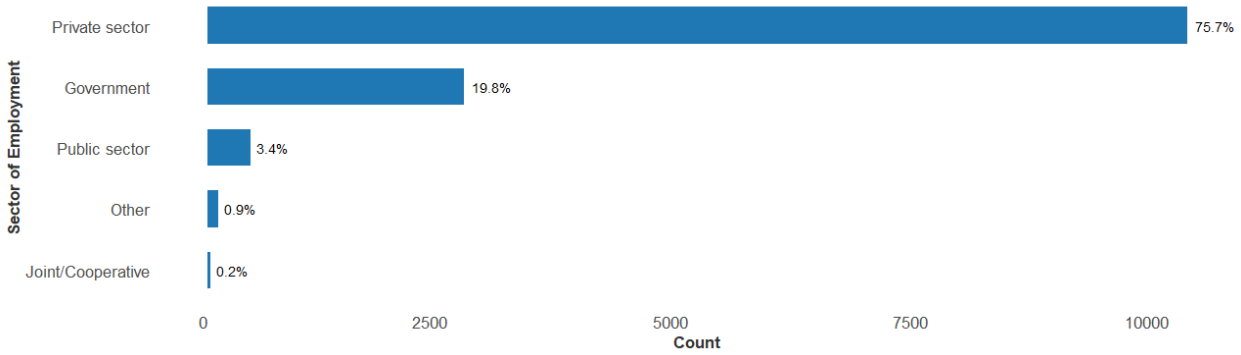

Figure 3.14: Bar Chart for the Sector of Employment

Figure 3.15 shows that the government and public sectors show high levels, providing the best social security coverage, while the private and "Other" sectors show significantly lower coverage. The private sector, in particular, shows a substantial gap in social security coverage, indicating the need for targeted interventions to improve coverage.

The Chi-squared test is used to measure the independence between the employment sector and Social Security enrollment ($\chi^2 = 4680.2$, p < 2.2e−16), indicating with 95% confidence that there is an association between them. Additionally, Cramer's V (0.5894) is employed to assess the strength of association, suggesting a strong association between the employment sector and Social Security.

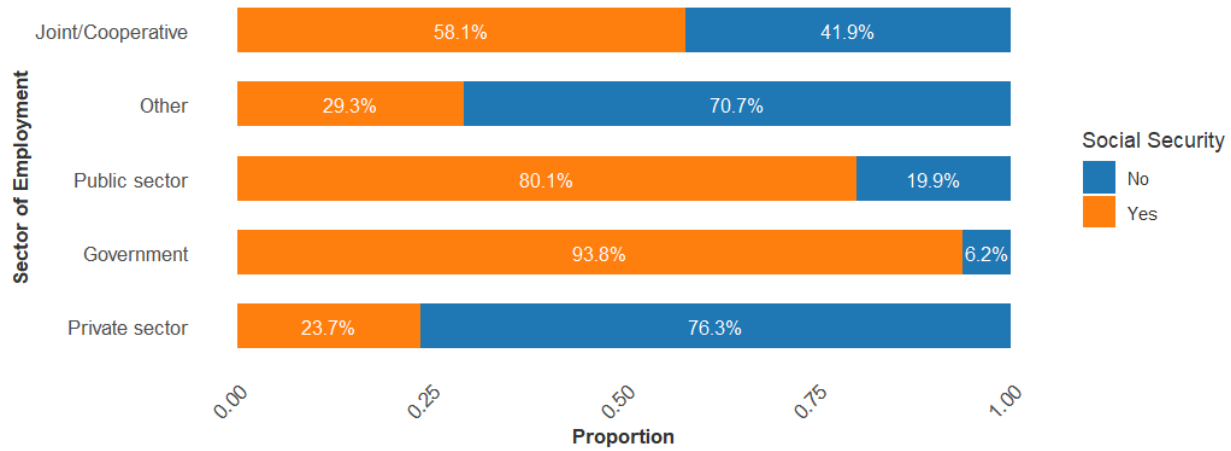

Figure 3.15: Stacked Bar Chart Sector of Employment by Social Security

Occupation

From Figure 3.16, we can see that the largest groups within the workforce are service workers and shop and market sales workers, skilled agricultural and fishery workers, and craft and related trades workers. Occupations such as legislators, senior officials, managers, and elementary occupations make up the smallest proportions of the workforce.

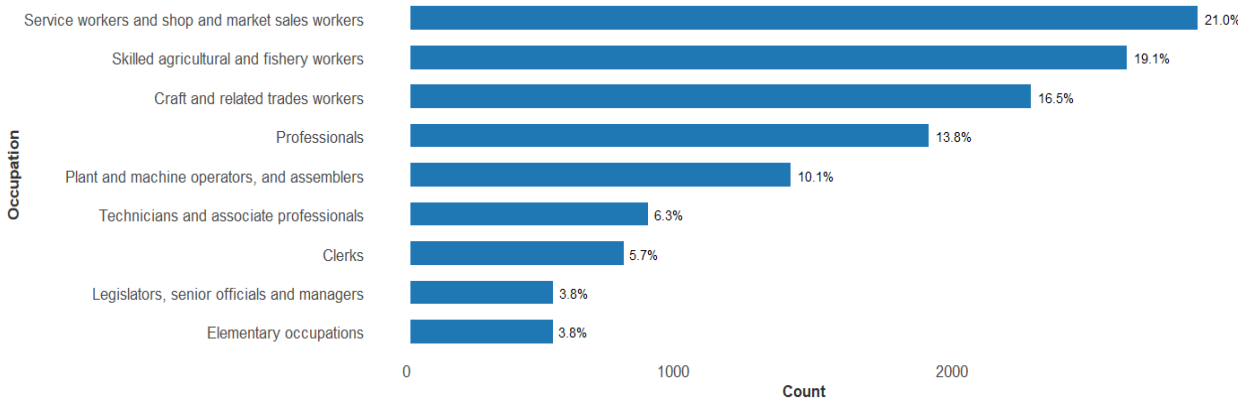

Figure 3.16: Bar Chart for the Distribution of Occupation

What we can notice from Figure 3.17 is that occupations such as clerks, professionals, and senior officials/managers have higher proportions of individuals with social security coverage. In contrast, the most exposed to danger groups, with the highest proportions lacking social security coverage, include elementary occupations, craft and related trades workers, and skilled agricultural and fishery workers.

The Chi-squared test is used to measure the independence between occupation and Social Security enrollment ($\chi^2 = 3422.8$, $p < 2.2e{-}16$), indicating with 95% confidence that there is an association between them. Additionally, Cramer's V (0.5040529) is employed to assess the strength of the association, suggesting a moderate association between occupation and Social Security.

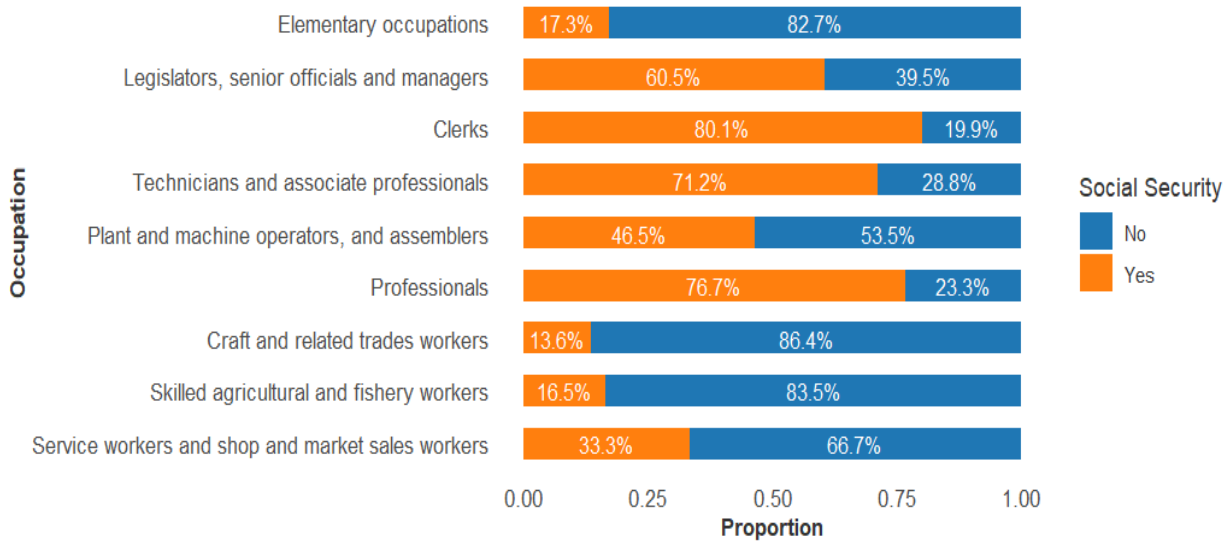

Figure 3.17: Stacked Bar Chart for the Occupation by Social Security

Industry

It is clear from Figure 3.18 that the largest groups within the workforce are in other services, agriculture, fishing, and commerce. Industries such as mining, finance, insurance, real estate, electricity, and utilities make up the smallest proportions of the workforce.

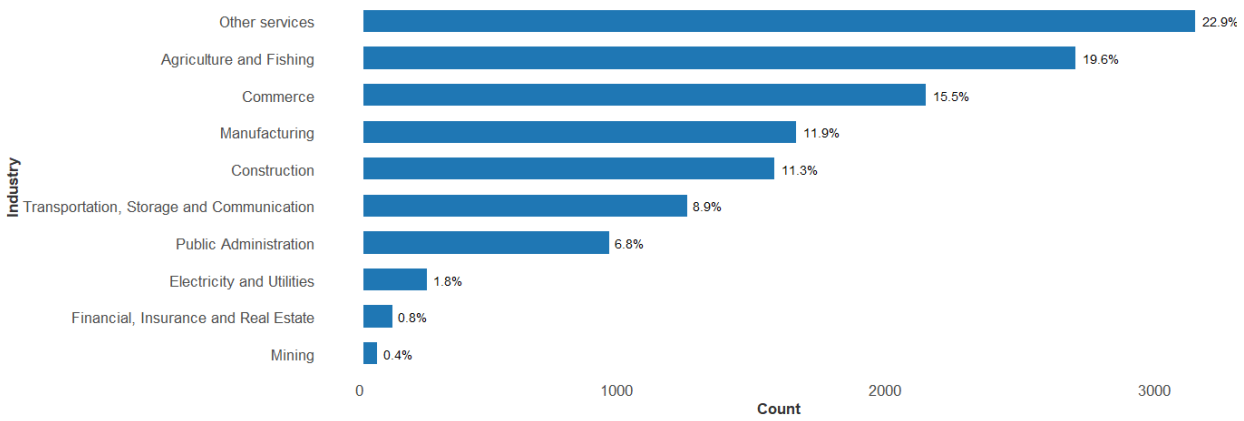

Figure 3.18: Bar Chart for Industry Distribution

As we can see from Figure 3.19, the public administration and electricity/utilities sectors show high levels of social security, providing the best social security coverage. In contrast, the construction, commerce, and agriculture/fishing sectors show significantly the lowest coverage. Other services, mining, finance, insurance, and real estate show approximately equally likely distribution of having social security coverage.

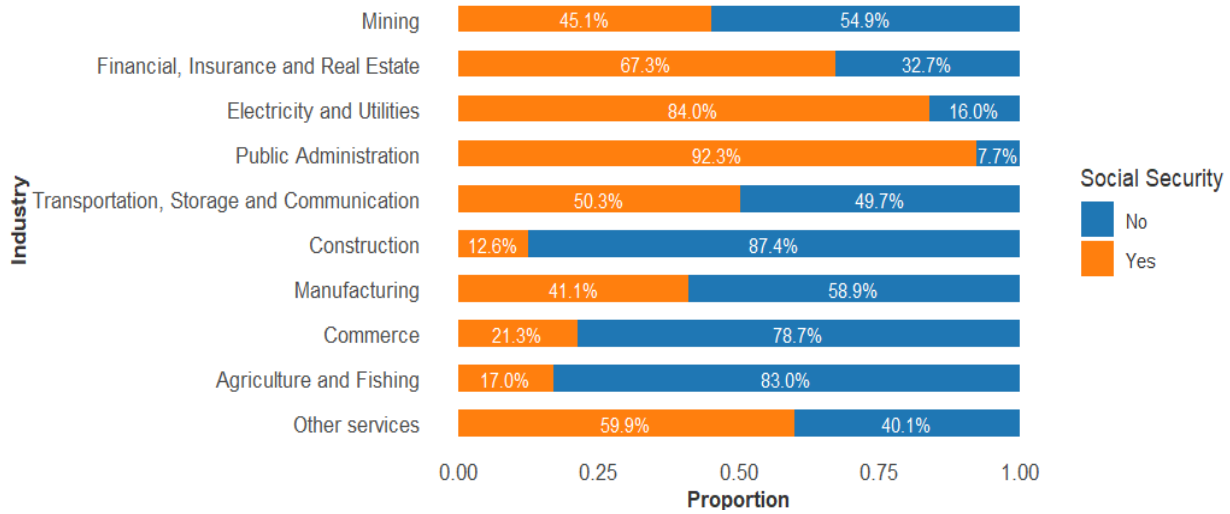

Figure 3.19: Stacked Bar Chart for Industry by Social Security

The Chi-squared test is used to measure the independence between Industry and Social Security enrollment ($\chi^2 = 9460.8$, p < 2.2e−16p), indicating with 95% confidence that there is an association between them. Additionally, Cramer's V (0.458) is employed to assess the strength of the association, suggesting a moderate association between employment status and Social Security.

Chapter 4:

Statistical Modeling: Binary Logistic Regression

In this chapter, a statistical modeling approach is employed to analyze the determinants of access to social security coverage. Binary logistic regression was chosen due to the binary nature of the dependent variable, whether an individual has social security coverage or not. The objective of this chapter is to utilize binary logistic regression to identify and quantify the impact of various socio-demographic and employment-related factors on the likelihood of an individual having social security coverage. By examining these determinants, we aim to provide a comprehensive understanding of the key variables influencing social security enrollment, thereby offering insights that can inform policy interventions to enhance social security coverage in Egypt.

4.1 Key Features of Binary Logistic Regression

1. Binary Outcome Variable: The dependent variable in binary logistic regression is binary, coded as 0 and 1. In simpler terms, it predicts the odds of an event occurring vs. not occurring based on the values of input variables.
2. Logit Transformation: The logistic regression model estimates the probability of the dependent variable successes (1) using the logit function. The logit function is the natural logarithm of the odds of the event occurring:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Where;

p: is the probability of success.

β_0: is the intercept (constant term).

β_i: are the coefficients of the independent variables X_i.

The coefficients in a logistic regression model are estimated using maximum likelihood estimation (MLE), which finds the parameter values that maximize the likelihood of observing the given sample data. Moreover, the logistic regression model is a linear probability model but uses the canonical link function to make the predicted probabilities range from 0 to 1 regardless of any input value for independent variables.

4.2 Variables Included in the Analysis

- Dependent Variable: Social Security Coverage (0 = No, 1 = Yes)
- Independent Variables:
    1. Age
    2. Gender (“Female” is the reference category).
    3. Disability Status (“Disabled” is the reference category)
    4. Chronic Disease (“No” is the reference category)
    5. Father’s Presence at Home (“Father is dead” is the reference category)
    6. Health Insurance (“No Health Insurance” is the reference category)
    7. Area (“Frontier Governorates” is the reference category)
    8. Marital Status (“Divorced/Separated” is the reference category)
    9. Educational Level (“None Educated” is the reference category)
    10. Employment Status (“Contributing (unpaid) family worker” is the reference category)
    11. Sector of Employment (“Government” is the reference category)
    12. Occupation (“Clerks” is the reference category)
    13. Industry (“Agriculture and Fishing” is the reference category)

4.3 Checking the Assumptions

Before fitting the model, the following assumptions were checked:

1. Large Sample Size.
2. Multicollinearity: Adjusted Generalized Variance Inflation Factor (GVIF) values for each predictor variable to ensure there was no correlation among them.
3. Linearity: The linearity of the continuous variable (Age) with the logit of the dependent variable using the Box-Tidwell test.

Checking Multicollinearity

Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can lead to unreliable estimates of coefficients and make it difficult to interpret the individual effect of the predictors. To detect multicollinearity, we use the Generalized Variance Inflation Factor (GVIF) as most of the explanatory variables are categorical. According to Fox and Monette (1992), the GVIF value of an explanatory variable ranges from 1 (indicating no multicollinearity) to infinity (indicating perfect multicollinearity). Moreover, as the number of explanatory variables in the model increases, the degrees of freedom decrease and the GVIF value gets larger even in the absence of serious multicollinearity. Consequently, the adjusted GVIF is used as it provides a more conservative estimate of the severity of multicollinearity in the model. The adjusted GVIF is given by $GVIF^{\frac{1}{2*Df}}$, accounting for the degrees of freedom and providing a measure to assess multicollinearity. A common rule of thumb is that an adjusted GVIF value greater than 4 indicates potential multicollinearity. However, values close to 1 suggest that there is minimal correlation between variables.

Table 4.1: GVIF and Adjusted GVIF for the Explanatory Variables

Feature	GVIF	Df	$(GVIF)^{\frac{1}{2*Df}}$
Age	2.110432	1	1.452732
Disability Status	1.039635	1	1.019625
Father's Present at Home	2.763946	2	1.289384
Health Insurance	1.280787	1	1.131718
Area	1.216000	2	1.050107
Marital Status	2.590295	4	1.126338
Educational Level	2.864530	5	1.110978
Employment Status	2.218195	3	1.142001
Sector of Employment	2.186205	4	1.102710
Occupation	140.161362	8	1.361960
Industry	129.393737	9	1.310173

Based on the adjusted GVIF results, as shown in Table 4.1, there is no indication of significant multicollinearity among the predictor variables in the final logistic regression model. The adjusted GVIF values for all the explanatory variables are between 1 and 1.5, suggesting that the estimates of the coefficients are reliable and not inflated due to multicollinearity. This confirms that the final model is appropriately specified, with each variable contributing uniquely to explaining the variance in access to social security coverage.

Checking Linearity

The Box-Tidwell test is used to check the linearity assumption between a continuous independent variable and the logit of the dependent variable in logistic regression. The test assesses whether the relationship between the continuous predictor and the logit transformation of the outcome variable is linear. The test was done having a p-value of (0.181), thus the assumption of linearity between the continuous variable "Age" and the logit of having social security coverage is satisfied.

4.4 Goodness of Fit

The Likelihood Ratio Test (LRT) is used to compare the goodness-of-fit between two nested models. In this case, we are comparing a full model that includes all the predictor variables with a reduced model. The test evaluates whether the fitted model is significantly worse than the full model in terms of explaining the variance in the dependent variable (Social Security coverage). A p-value of (0.4347), supports the conclusion that the fitted model is adequate and does not significantly differ in fit compared to the full model with 5% significance level, thereby simplifying the model without sacrificing explanatory power.

Moreover, the Hosmer and Lemeshow test is conducted. Based on the p-value (0.1275), the test indicates that the model fits the data well at the 5% level of significance.

4.5 Predictive Power of the Model

The classification table (Confusion Matrix) and the Receiving Operating Characteristic (ROC) curve are used to assess the predictive power of the model.

Confusion Matrix

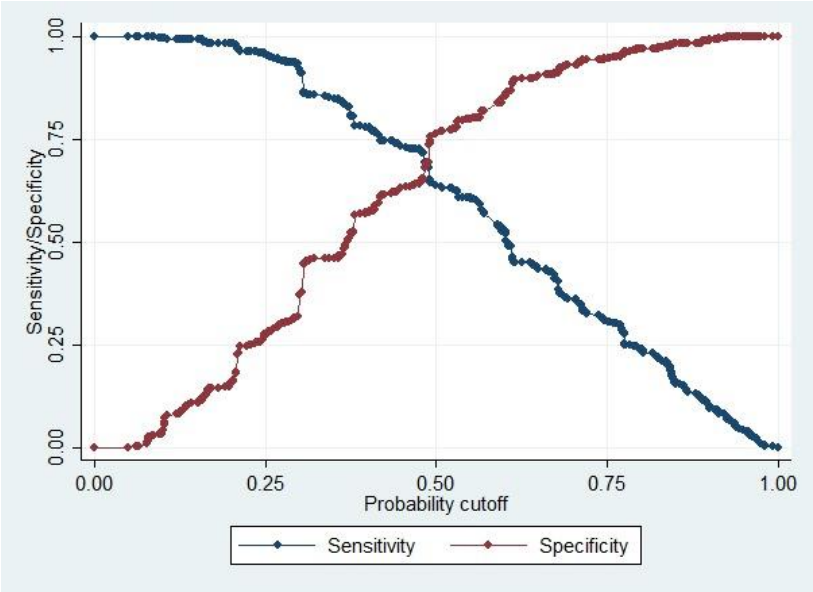

Figure 4.1: Sensitivity and Specificity Vs. Cut-off Points

From Figure 4.1, and after doing a for loop that iterates to reach the optimal cutoff point that has the least misclassification error. An optimal cutoff point of 0.47 was reached. Thus, the predictive performance of the binary logistic regression model was assessed. According to Table 4.2, The classification table shows the number of correct and incorrect predictions made by the model.

Table 4. 2: Confusion Matrix of the Binary Logistic Model

	Predicted Class	
Actual Class	Yes	No
Yes	1314	288
No	187	2252

The model, with a specificity of 92.33%, correctly identifies 92.33% of the individuals who do not have social security coverage. With a sensitivity of 82.02%, the model correctly identifies 82.02% of the individuals who have social security coverage. Furthermore, 88.25% of all cases are correctly classified, providing sufficient evidence of the good predictive power of the model as the model makes accurate predictions for a significant proportion of cases.

The Mcnemar's test has a p-value of ($4.468e^{-06}$) indicates that there is a significant difference between the model's performance on "No" and "Yes" predictions, suggesting that the model's performance is not due to random chance.

Receiving Operating Characteristics (ROC) Curve

The ROC curve is a graphical representation tool for evaluating the performance of the binary logistic regression model used to predict social security coverage among Egyptian workers. The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) across various threshold levels, providing a comprehensive view of the model's classification capabilities.

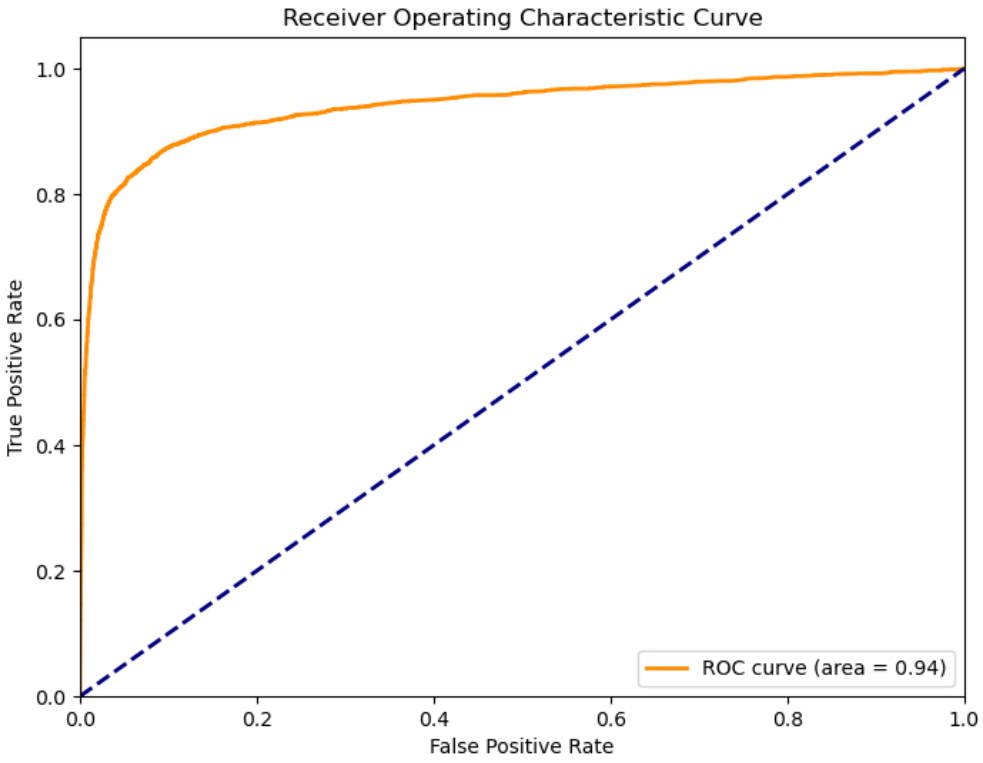

Figure 4.2: Logistic Regression ROC Curve

As can be seen from Figure 4.2, the ROC curve bows towards the upper left corner, showing an excellent balance between the true positive rate and the false positive rate, with an area under the curve (AUC) of approximately 0.94, which means that there is a 94% chance that the model will correctly distinguish between a randomly selected individual without social security coverage and a randomly selected individual with social security coverage. This confirms that the model has excellent discriminatory ability which performs well in distinguishing between individuals enrolled and not enrolled in Social Security. Thus, the model is highly reliable in its predictions, giving confidence that the probabilities computed by the model meaningfully rank examples from least to most likely to be positive.

4.6 The Fitted Model

After the stepwise backward elimination method was applied according to the AIC of each fitted model, picking the model with the minimum AIC and considering satisfying the assumptions, here is the final fitted model shown in Table 4.3, reporting the estimates and significance of these significant variables. Eleven variables significantly affect social security coverage, namely, age, disability status, father's presence at home, health insurance, area, marital status, educational level, employment status, sector of employment, occupation, and industry.

Table 4.3: Binary Logistic Regression Model Coefficients and Their Significance

Variables	Estimated Coefficient β	Estimated Odds Ratio e^{β}	P-value
Age	0.064082	1.0661803	$< \mathbf{2e^{-16}}$
Disability Status			
Not Disabled	-0.471150	0.6242841	**0.000858**
Father's Presence at Home			
Father is not a household member	0.536897	1.7106903	$\mathbf{9.31e^{-06}}$

Yes	0.183214	1.2010715	0.296505
Health Insurance			
Yes	2.553895	12.8570896	$< \mathbf{2e^{-16}}$
Area			
Rural	-0.757998	0.4686037	$\mathbf{8.74e^{-06}}$
Urban	-0.940040	0.3906123	$\mathbf{5.46e^{-08}}$
Marital Status			
Married Monogamous	-0.207951	0.8122472	0.363664
Married Polygamous	-0.123967	0.8834089	0.846957
Never Married	-0.575423	0.5624668	**0.018506**
Widowed	1.875013	6.5209019	$\mathbf{5.35e^{-10}}$
Educational Level			
Post Secondary or Equivalent	0.802958	2.2321335	$\mathbf{3.53e^{-05}}$
Post Graduate	0.889416	2.4337068	0.081190
Primary/Lower Secondary	0.167917	1.1828379	0.157349
Secondary	0.368337	1.4453290	**0.000312**
University	0.899642	2.4587235	$\mathbf{1.89e^{-10}}$
Employment Status			
Employee	0.309240	1.3623894	0.191128
Employer	0.673144	1.9603917	**0.005185**
Own-account/ Self-employed	0.076581	1.0795892	0.755243

Sector of Employment			
Joint/Cooperative	-0.585089	0.5570564	0.349227
Other	-2.075773	0.1254594	$\mathbf{1.55e^{-09}}$
Private Sector	-1.946782	0.1427327	$\mathbf{< 2e^{-16}}$
Public Sector	-0.834670	0.4340179	**0.000785**
Occupation			
Craft and related trade workers	-0.803678	0.4476795	**0.000272**
Elementary occupations	-0.784157	0.4565042	**0.002552**
Legislators, Senior Officials, and Managers	-0.339996	0.7117733	0.172530
Plant and Machine Operators, and Assemblers	0.317803	1.3741060	0.149308
Professionals	0.228517	1.2567353	0.279907
Service Workers and Shop and Marker Sales Workers	-0.181234	0.8342405	0.363926
Agriculture and Fishery Workers	-0.727968	0.4828893	**0.049983**
Technicians and Associate Professionals	0.148354	1.1599231	0.505079
Industry			
Commerce	0.048977	1.0501958	0.883432

Construction	-0.127607	0.8801990	0.711934
Electricity and Utilities	0.919537	2.5081299	**0.043569**
Financial, Insurance, and Real Estate	0.526121	1.6923555	0.249427
Manufacturing	0.897428	2.4532861	**0.006388**
Mining	0.041220	1.0420818	0.944466
Other Services	0.476304	1.6101119	0.147554
Public Administration	0.238730	1.2696355	0.529032
Transportation, Storage, and Communication	1.012842	2.7534162	**0.002856**

Interpretation of the Significant Parameters

- **Age**

At a 5% significance level, for each additional year of age for an Egyptian worker, the estimated odds of having social security coverage increases by approximately 6.6% for that worker, holding other variables constant.

- **Disability Status**
  - ✓ Not Disabled

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker without disability is approximately 37.5% lower than that of an Egyptian worker with disability, holding other variables constant.

- **Father's Presence at Home**
  - ✓ Father is not a household member

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker whose father is not a household member is approximately 71% higher than that of an Egyptian worker whose father is dead, holding other variables constant.

- **Health Insurance**
  - ✓ Has Health Insurance

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker having health insurance is approximately 12.86 times more likely to get social security compared to that of an Egyptian worker without health insurance, holding other variables constant.

- **Area**
  - ✓ Rural

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker who lives in rural areas is approximately 53% lower than that of an Egyptian worker who lives in frontier governorates areas, holding other variables constant.

  - ✓ Urban

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker living in urban areas is approximately 61% lower than that of an Egyptian worker living in frontier governorates, holding other variables constant.

- **Marital Status**
  - ✓ Never Married

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker who never married is approximately 43.7% lower than that of an Egyptian worker who is divorced or separated, holding other variables constant.

  - ✓ Widowed

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker who is widowed is approximately 6.52 times more likely to get social security compared to that of an Egyptian worker who is divorced or separated, holding other variables constant.

- **Educational Level**
  - ✓ Post Secondary or Equivalent

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker with post-secondary education or equivalent is approximately 2.23 times more likely to get social security compared to that of an Egyptian worker with no education, holding other variables constant.

  - ✓ Secondary

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker with secondary education is approximately 44.5% higher than that of an Egyptian worker with no education, holding other variables constant.

  - ✓ University

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker with a university degree is approximately 2.46 times more likely to get social security compared to that of an Egyptian worker with no education, holding other variables constant.

- **Employment Status**
  - ✓ Employer

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian employer is approximately 96% higher than that of an Egyptian contributing (unpaid) family worker, holding other variables constant.

- **Sector of Employment**
  - ✓ Private Sector

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker in the private sector is approximately 85.7% lower than that of an Egyptian worker in the government sector, holding other variables constant.

  - ✓ Public Sector

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker in the public sector is approximately 56.6% lower than that of an Egyptian worker in the government sector, holding other variables constant.

  - ✓ Other Sectors

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker in other sectors (national NGOs and private households) is approximately 87.5% lower than that of an Egyptian worker in the government sector, holding other variables constant.

- **Occupation**
  - ✓ Craft and Related Trade Workers

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian craft and related trade worker is approximately 55.2% lower than that of an Egyptian worker in clerical occupations, holding other variables constant.

✓ Elementary Occupations

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker in elementary occupations is approximately 54.35% lower than that of an Egyptian worker in clerical occupations, holding other variables constant.

✓ Agriculture and Fishery Workers

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker in agriculture and fishery occupations is approximately 51.7% lower than that of an Egyptian worker in clerical occupations, holding other variables constant.

- **Industry**

✓ Electricity and Utilities

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker in the electricity and utilities industry is approximately 2.51 times more likely to get social security compared to that of an Egyptian worker in the agriculture and fishing industry, holding other variables constant.

✓ Manufacturing

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker in manufacturing is approximately 2.45 times more likely to get social security compared to that of an Egyptian worker in the agriculture and fishing industry, holding other variables constant.

✓ Transportation, Storage, and Communication

At a 5% significance level, the estimated odds of having social security coverage for an Egyptian worker in transportation, storage, and communication is approximately 2.75 times more likely to get social security compared to that of an Egyptian worker in the agriculture and fishing industry, holding other variables constant.

4.7 Main Findings of the Logistic Regression

Our binary logistic regression analysis identified several key variables that significantly affect the likelihood of enrolling in social security coverage among the Egyptian labor force. These variables include age, with older individuals having higher odds of coverage; disability status, with disabled individuals more likely to have coverage; father's presence at home, where workers whose father is not a household member are more likely to have coverage; health insurance, where having health insurance greatly increases the likelihood of coverage; and area, with individuals in rural and urban areas less likely to have coverage compared to those in frontier governorates. Marital status also plays a role, with never-married and widowed individuals showing significant differences in coverage compared to divorced or separated individuals. Higher educational levels are associated with increased odds of coverage, while employment status shows that employers have significantly higher odds of coverage compared to unpaid family workers. Employment in private or other sectors significantly reduces the odds of coverage compared to government employment. Workers in craft, elementary, and agricultural occupations have lower odds of coverage compared to clerical workers, and workers in the electricity and utilities, manufacturing, and transportation sectors have higher odds of coverage compared to those in agriculture and fishing.

Chapter 5:

Machine Learning Algorithms

In this chapter, advanced machine learning techniques were applied, specifically the Decision Tree and Random Forest algorithms, to classify social security coverage among the Egyptian labor force. The objective of this chapter is to implement and compare these algorithms, analyze their predictive power, and identify the most influential factors affecting social security coverage. We aim to provide more information about the determinants of Social Security enrollment and suggest data-driven policy recommendations to improve coverage.

5.1 Decision Tree Algorithm

In this section, we will introduce the decision tree algorithm, how it works, its problems, and how to overcome these problems. A decision tree is implemented, and its predictive power is investigated. Finally, the important features according to the algorithm are detected.

5.1.1 Introduction to Decision Tree Algorithm

A Decision Tree is a supervised machine-learning algorithm used for both classification and regression tasks. It models decisions and their possible consequences, including chance event outcomes, resource costs, and utility. The structure of a decision tree, as we can see from Figure 5.1, includes nodes representing decisions or tests, branches representing the outcomes of these decisions, and leaf nodes representing the final outcomes or class labels. Each node splits the data into two or more branches. The first node, at the top of the decision tree, is called the root node. Leaf nodes or leaves are terminal nodes at the end of the decision tree diagram that do not split further into more nodes. A Leaf has a single branch leading to it and no branches coming off it.

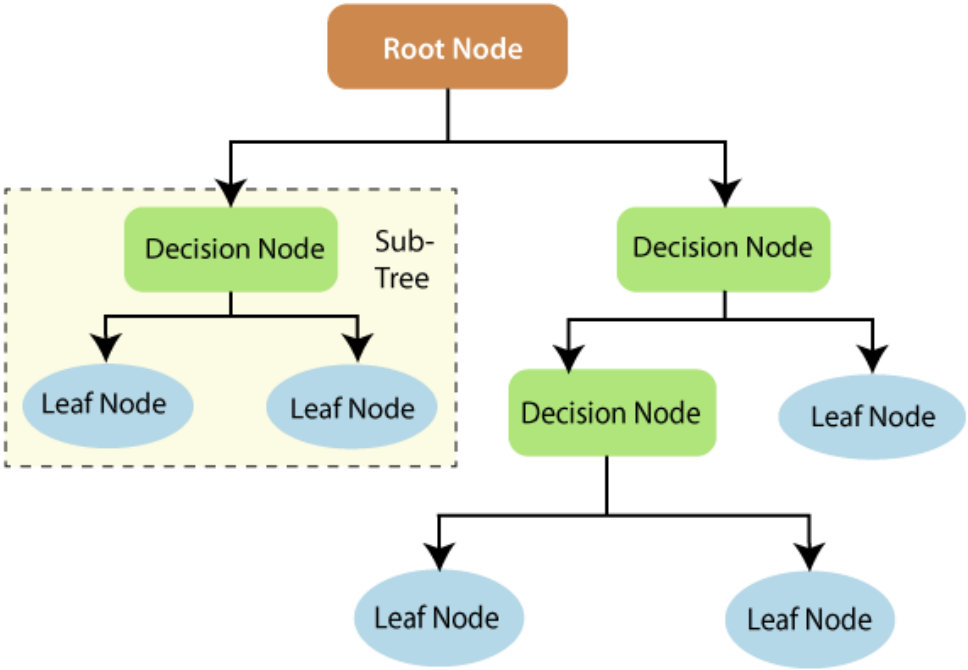

Figure 5.1: Decision Tree Structure

Source: ***Charbuty et al. (2021)***

5.1.2 How Decision Trees Work

The decision tree algorithm works by recursively splitting the dataset into subsets based on the feature values that result in the most significant information gain. The process involves:

1. **Selecting the Best Split:** At each node, the algorithm evaluates all possible splits based on each feature and selects the one that results in the most homogeneous subgroups. The "best" split is determined using a criterion such as the Gini index, Entropy, or Information Gain.

Gini Index:

The Gini index measures the impurity or diversity of a dataset. It is used to select the feature and threshold for splitting the data at each node. For a binary classification, the Gini index for a split is calculated as:

$$Gini(D) = 1 - \sum_{i=1}^{c} (Pi)^2$$

Where;

P_i is the proportion of instances in class i.

Gini Impurity computes the probability that a specific variable is being incorrectly classified when it is chosen at random. Its values range from 0 to 0.5. The lower the value of Gini, the lower the likelihood of misclassification and the better the split. Thus, a split that results in subsets with lower Gini index values is preferred, as it indicates more homogeneous subsets.

2. **Splitting the Data:** The dataset is split into two or more subsets based on the selected feature and threshold. This process continues recursively for each subset, creating a tree structure.
3. **Stopping Criteria:** The recursion stops when a stopping criterion is met. Common stopping criteria include:
   - A maximum tree depth is reached.
   - A minimum number of samples per node is reached.
   - No further information gain can be achieved.

5.1.3 Problems with Decision Trees

A decision tree follows a greedy algorithm to build the tree, which means it makes a series of locally optimal choices at each step with the aim of finding a global optimum. This involves selecting the best feature and threshold for splitting the data at each node, based on a criterion like the Gini index. Because the algorithm makes decisions based solely on local information, it may lead to suboptimal splits that do not result in the best overall tree structure. The locally optimal splits may not always combine to form the globally optimal decision tree. There are two main problems in this context:

1. **Overfitting:** The greedy approach can lead to overfitting, especially when they are allowed to grow too deep. Overfitting occurs when the model becomes too complex and captures noise in the training data, leading to poor generalization of new data.
2. **Bias and Variance Trade-off:** As shown in Figure 5.2, decision trees can have high variance because small changes in the data can result in different splits and trees. This makes them sensitive to the specific data they are trained on. Bias refers to the error introduced by approximating a real-world problem, which may be complex, by a simpler

model. As *Rhys, H. (2020)* mentioned, models with high bias are typically too simple and fail to capture the underlying patterns in the data (underfitting). While variance refers to the model's sensitivity to small fluctuations in the training data. Models with high variance are typically too complex and fit the training data very closely, but perform poorly on new data (overfitting). As model complexity increases, bias decreases, however, variance increases in return. The goal is to find a balance where both bias and variance are minimized.

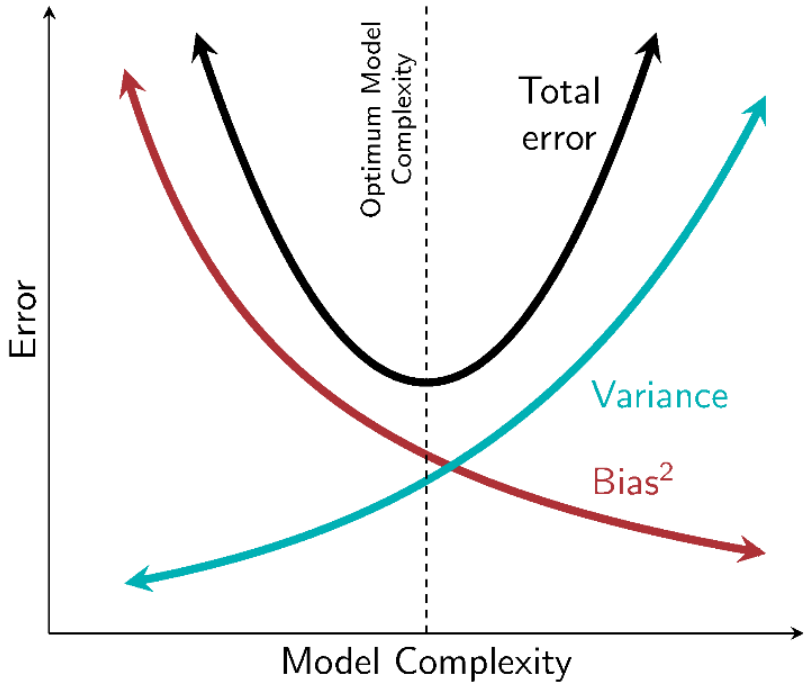

Figure 5.2: Bias-Variance Trade-off

Source: ***Rhys (2020)***

5.1.4 Overcoming the Problem of Decision Trees

1. **Pruning:** Pruning is a technique used to reduce the size of the decision tree by removing sections that provide little power to classify instances. This helps to reduce overfitting. Two common types of pruning are:
    - Pre-Pruning (early stopping): Introduce stopping criteria to prevent the tree from growing too deep. This can include setting a maximum depth, requiring a minimum

number of samples per leaf, or stopping when the information gain is below a certain threshold.

- Post-Pruning (removing branches after the tree is created): After the tree is fully grown, remove branches that have little importance or do not improve the model's performance on a validation set. Techniques like cost-complexity pruning or reduced error pruning can be used.

2. **Ensemble Methods:** Ensemble methods like Random Forests combine multiple decision trees to produce a more robust and accurate model. These methods reduce the variance and improve the generalization ability of decision trees.

5.1.5 The Implemented Decision Tree

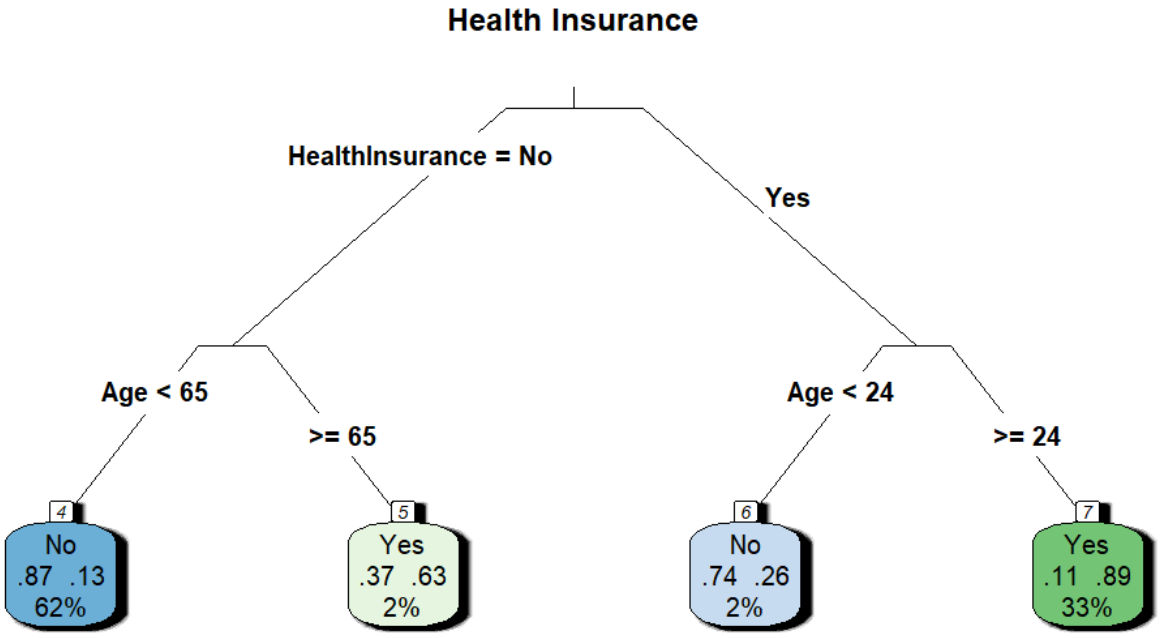

Figure 5.3: Decision Tree for the Social Security Enrollment

The decision tree, in Figure 5.3, clearly shows how different health insurance statuses and age groups affect the likelihood of having social security coverage. The root node of the tree is "Health Insurance", indicating that having health insurance is the most critical factor in determining social security coverage. The decision tree shows that:

For individuals without health insurance:

- Those under 65 years old are likely not to have social security coverage with a probability of 87%.
- Those aged 65 or older are more likely to have social security coverage with a probability of 63%.

While for individuals with health insurance:

- Those under 24 years old are less likely to have social security coverage with a probability of 74%.
- Those aged 24 or older are highly likely to have social security coverage with a probability of 89% probability.

Confusion Matrix

The confusion matrix heatmap provides a visual representation of the performance of the decision tree model in predicting social security coverage. The matrix compares the predicted classifications to the actual classifications.

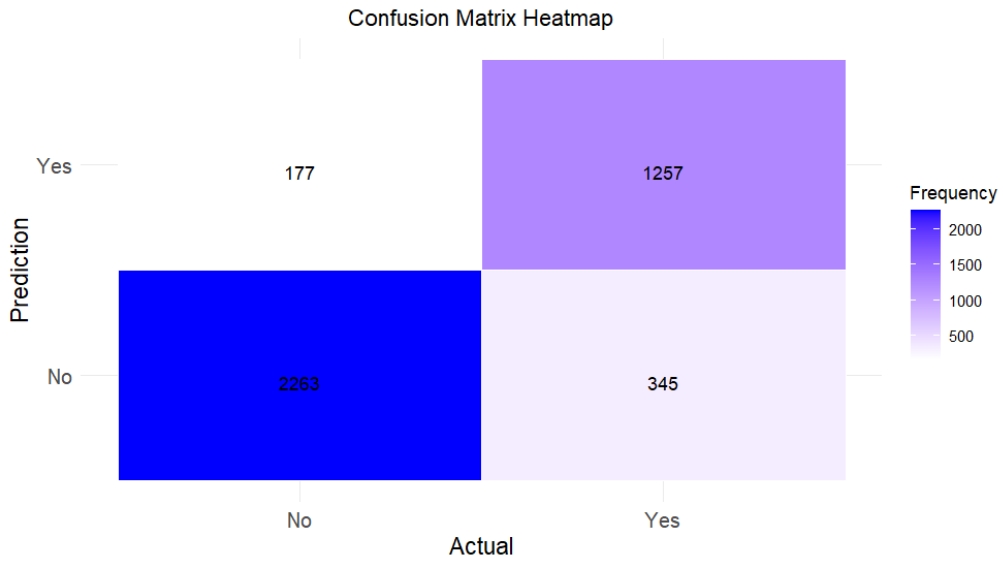

Figure 5.4: Heatmap for the Decision Tree Confusion Matrix

According to Figure 5.4, the model correctly predicts the social security coverage status for approximately 87.09% of the observations. This model got a high specificity of 92.75%, indicating the model's effectiveness in identifying individuals without social security coverage. Moreover, the tree got a sensitivity of 78.46%, indicating the model is reasonably good at identifying individuals with social security coverage.

ROC Curve for Decision Tree

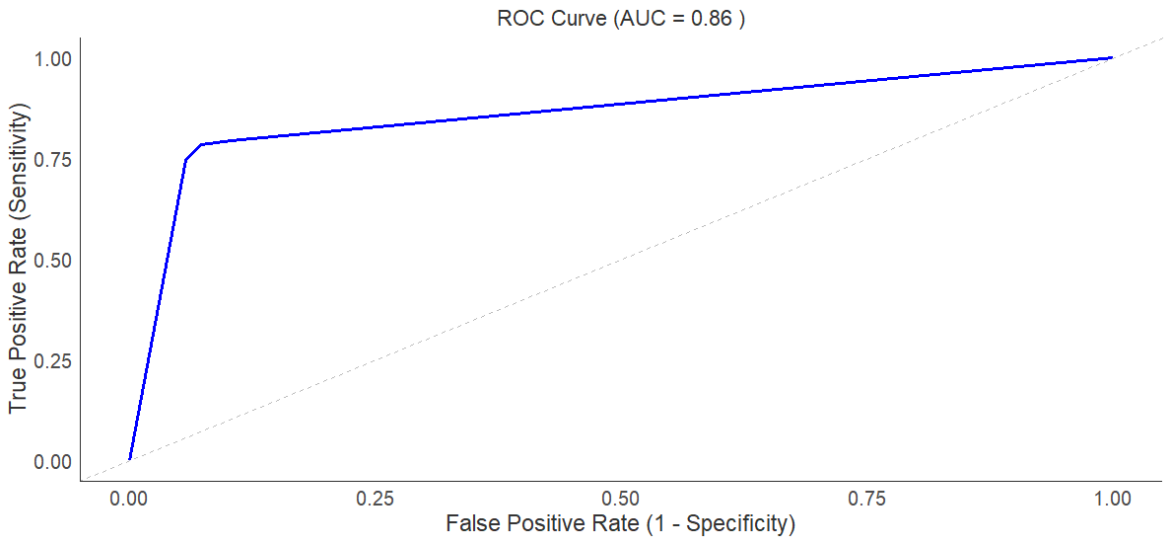

Figure 5. 5: Decision Tree ROC Curve

The ROC curve, in Figure 5.5, rises towards the top left corner of the plot, suggesting that the decision tree model performs well in distinguishing between individuals with and without social security coverage. AUC values range from 0.5 (no discrimination, equivalent to random guessing) to 1 (perfect discrimination). An AUC of 0.86 means there is an 86% chance that the model will correctly distinguish between a randomly chosen positive instance and a randomly chosen negative instance, indicating that the model has a good level of overall performance.

Feature Importance

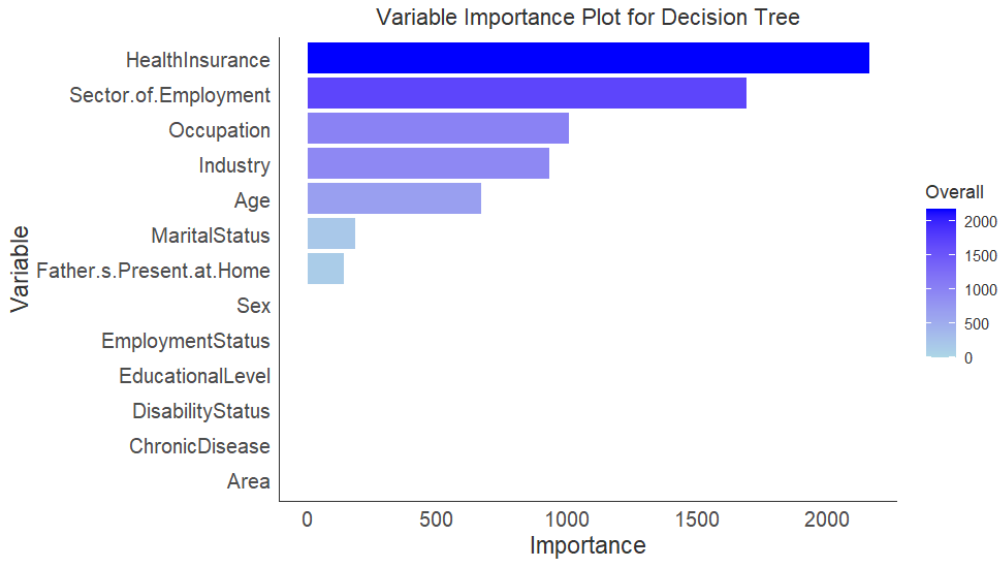

Figure 5. 6: Important Features for Decision Tree

The decision tree model's feature importance analysis shown in Figure 5.6, highlights key predictors of social security coverage among the Egyptian labor force. Health insurance emerged as the most critical factor, followed by sector of employment, occupation, and industry. Age and father's presence at home also significantly influenced social security coverage, while marital status showed moderate importance. The rest of the variables did not significantly contribute to the model's predictions in this context.

These findings suggest that policy interventions should prioritize enhancing health insurance access and targeting specific sectors and occupations to improve social security coverage. Additionally, demographic factors like age and family structure should be considered when designing social security programs.

5.2 Random Forest Algorithm

In this section, a random forest algorithm is introduced, how it works, how to implement it, and the advantages and drawbacks of this algorithm. Random Forest is implemented, investigating its predictive power. Finally, the important variables are detected according to the random forest algorithm.

5.2.1 Introduction to Random Forest

As *Mbaabu (2020)* mentioned, Random Forests is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. This method, developed by *Leo Breiman and Adele Cutler*, enhances the performance and robustness of individual decision trees by combining their predictions.

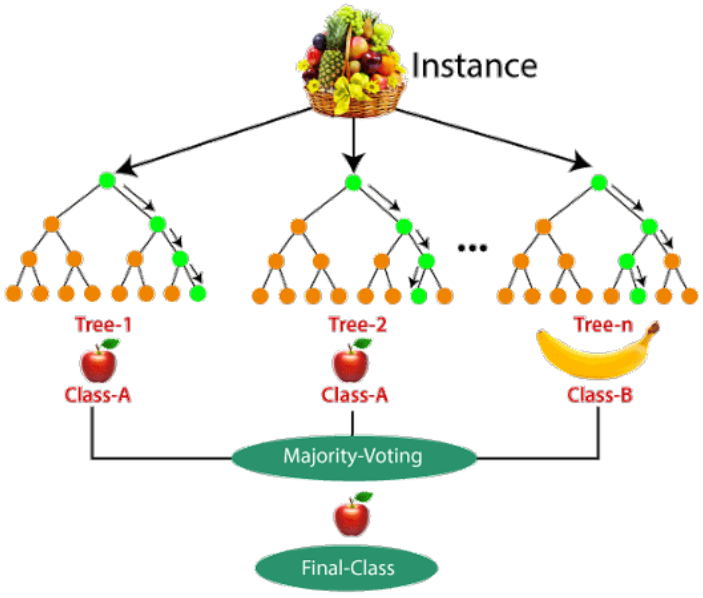

Figure 5. 7: Random Forest Structure

Source: ***Mbaabu (2020)***

5.2.2 How Random Forest Work

From Figure 5.7, Random Forests start by creating multiple subsets of the original training data through a process called **bootstrap sampling**. Each subset is generated by randomly selecting samples from the training set with replacements. For each bootstrap sample, a **decision tree is constructed**. However, unlike standard decision trees, Random Forests introduce an additional layer of randomness in the tree-building process, where, at each node, rather than considering all available features to determine the best split, a random subset of features is selected. This ensures that the trees are more diverse. Each decision tree is grown to its maximum extent without pruning, using the Gini index or another criterion to determine the **best splits** based on the randomly selected subset of features. For classification tasks, each decision tree in the forest provides a class prediction, and the final output is determined by the **majority vote** (the class with the most votes across all trees). For regression tasks, each tree provides a numerical prediction, and the final output is the average of these predictions.

5.2.3 How to Implement the Random Forest

1. Parameter Tuning:
    - Number of Trees: The number of trees in the forest. A higher number generally improves performance but increases computational cost.
    - Number of Features: The number of features to consider when looking for the best split. Common choices include the square root of the total number of features or the logarithm of the total number of features.
    - Maximum Depth: The maximum depth of the trees. Limiting depth can prevent overfitting.
2. Training and Validation:
    - Train the Random Forest model on the training data and validate its performance on a separate validation set.
3. Feature Importance:
    - Random Forests provide an estimate of feature importance, indicating how much each feature contributes to the prediction, which is useful for understanding the model and making decisions based on the most important features.

5.2.4 The Implemented Random Forest

In this section, the random forest was implemented, where the optimal parameters that resulted in the highest accuracy were the maximum number of trees was 200, the number of features to consider when looking for the best split was 4, and the maximum depth of the trees was 40. Finally, there is a measure of feature importance, indicating how much each feature contributes to the prediction accuracy.

Confusion Matrix

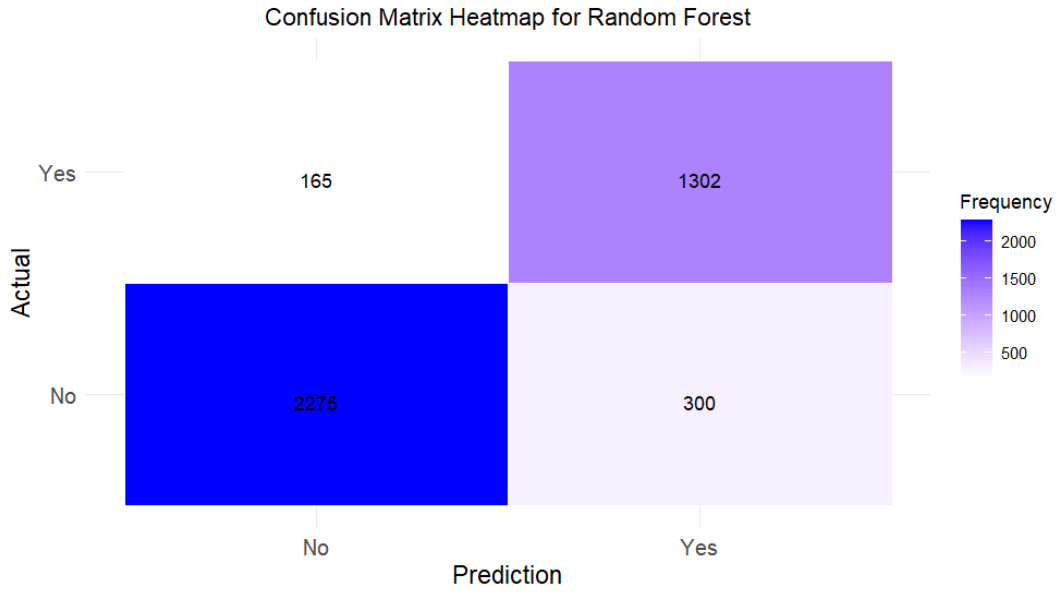

Figure 5.8: Heatmap for the Random Forest's Confusion Matrix

According to Figure 5.8, the high specificity of 93.24% indicates that the model is very effective at identifying individuals without social security coverage, minimizing the number of false negatives. The random forest is reasonably good at identifying individuals with social security coverage with a sensitivity of 81.27%. As a result, the model achieves an accuracy of 88.5%, demonstrating its high overall performance in predicting social security coverage.

ROC Curve for Random Forest

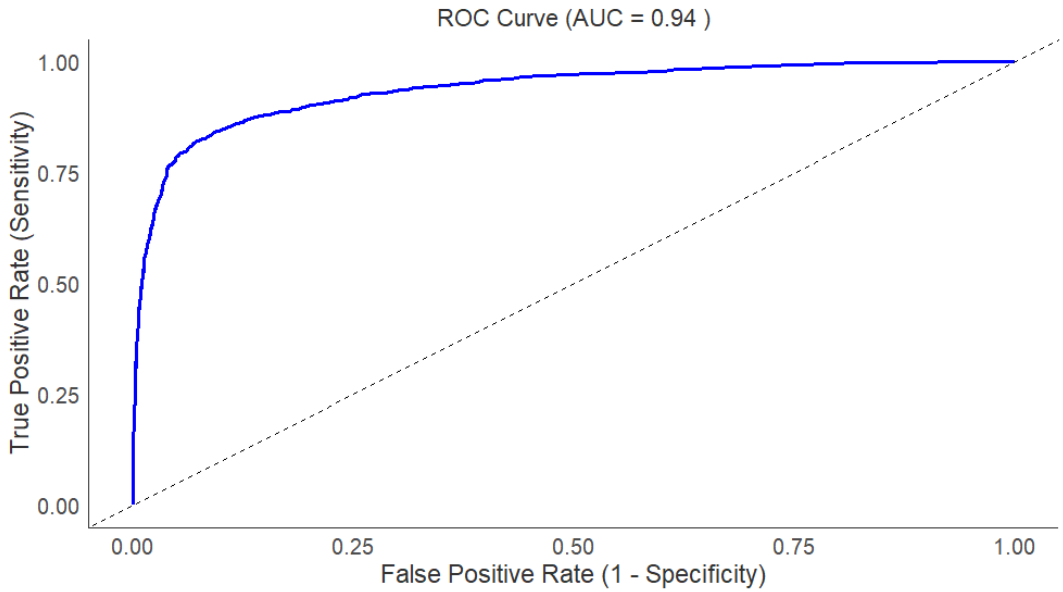

Figure 5. 9: Receiving Operating Characteristic (ROC) Cure for the Random Forest

The ROC curve in Figure 5.9 rises steeply towards the top left corner of the plot, showing a strong discriminatory power in predicting social security coverage among the Egyptian labor force. An AUC of 0.94 means there is a 94% chance that the model will correctly distinguish between a randomly chosen positive instance and a randomly chosen negative instance, indicating that the model has excellent overall performance.

Feature Importance

Random Forests provide a measure of feature importance, indicating how much each feature contributes to the prediction accuracy. This is calculated based on the mean decrease accuracy, which is how much the model's accuracy would decrease if a particular variable were excluded.

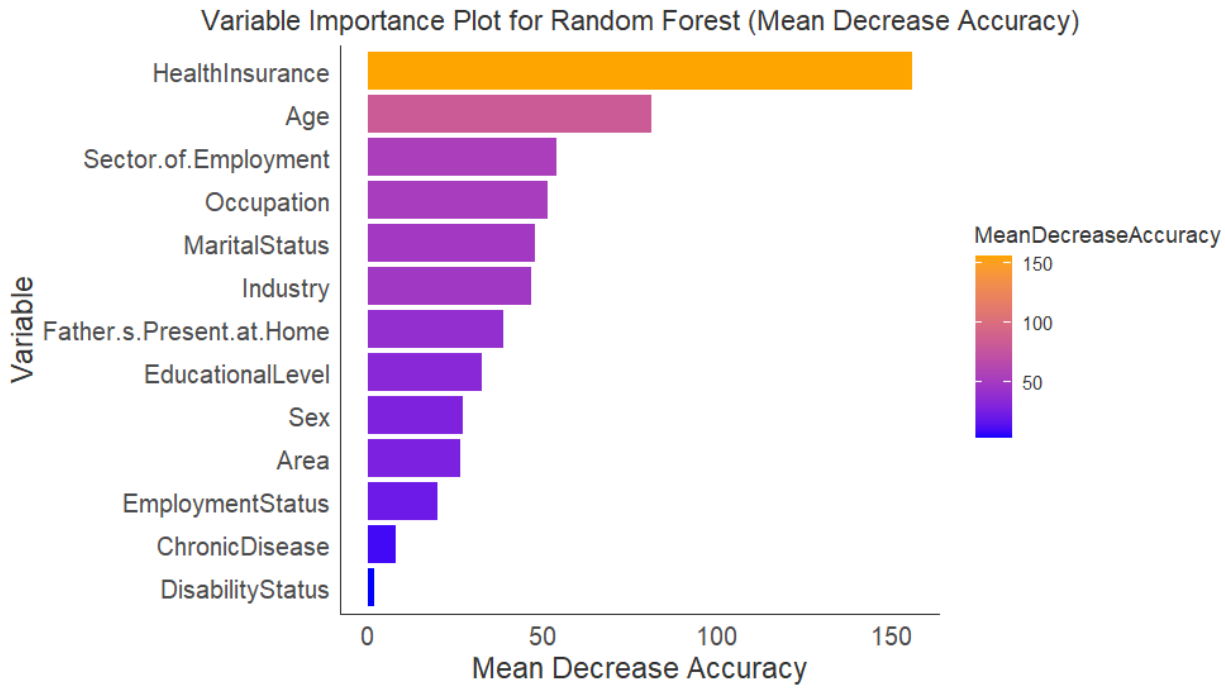

Figure 5.10: Variable Importance Plot for Random Forest

The variable importance plot in Figure 5.10 reveals that Health Insurance is the most influential factor in predicting social security coverage among the Egyptian labor force. Age, sector of employment, and occupation also significantly impact the model's accuracy. Marital status, industry, and father's presence at home are notable predictors. Educational level, Gender, and Area of residence also contribute to the model's predictions, but to a lesser extent. Employment status, chronic disease, and disability status have relatively minor impacts.

Chapter 6:

Deep Learning: Convolution Neural Networks (CNN) Model

In this chapter, we explore the application of Deep Learning, specifically Convolutional Neural Networks (CNNs), to classify the social security status among the Egyptian labor force. The objective is to benefit from the powerful feature extraction and classification capabilities of CNNs to analyze the tabular data related to social security. We will briefly explain the fundamental concepts of Deep Learning and Neural Networks, followed by a detailed examination of CNN architecture. Furthermore, we will describe the process of building, training, and evaluating a CNN model tailored to our classification task, providing insights into its performance and effectiveness in predicting social security coverage. Finally, we will compare all the applied models together with the pros and cons of each one of them.

6.1 What is Deep Learning?

Deep Learning is a subset of machine learning that involves neural networks with many layers. These models are designed to simulate the human brain's structure and function, enabling them to learn and make decisions from complex, high-dimensional data. Deep learning models are particularly effective for tasks involving large datasets and complex patterns, such as image and speech recognition, natural language processing, and autonomous driving.

6.2 What are Neural Networks?

Neural Networks are computational models inspired by the human brain. They consist of interconnected layers of nodes (neurons), each performing a simple computation. The main components of a neural network, as shown in Figure 6.1, include:

1. Input Layer: Receives the input data.
2. Hidden Layers: Intermediate layers that process inputs received from the input layer. Each hidden layer consists of multiple neurons.
3. Output Layer: Produces the final output or prediction.

Each connection between neurons has an associated weight, which is adjusted during training to minimize the error between the predicted and actual outputs.

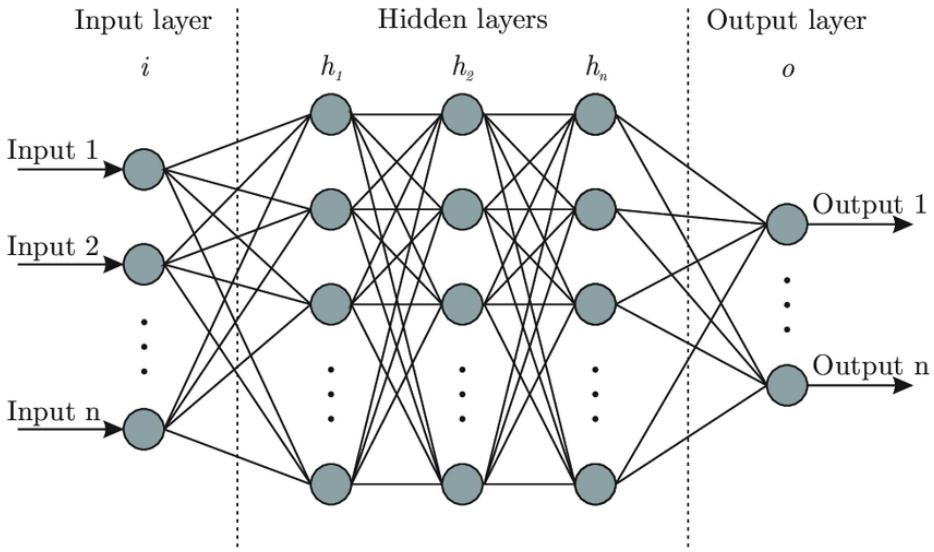

Figure 6.1: Architecture of Neural Network

Source: ***Ketkar (2021)***

6.3 Convolution Neural Network (CNN) Model

Convolutional Neural Networks (CNNs) are a specialized type of neural network designed to process data with a grid-like topology, such as images. CNNs are particularly effective in image classification, object detection, and other computer vision tasks. Additionally, it can be applied to tabulated data as we will do in our case.

6.3.1 CNN Architecture

The architecture of a CNN, as shown in Figure 6.2, typically consists of the following layers:

1. Convolutional Layers: These layers apply a set of small learnable convolutional filters (kernels) to the input data, creating feature maps that highlight different aspects of the input.
2. Pooling Layers: These layers perform down-sampling operations to reduce the spatial dimensions of the feature maps, retaining the most important information. Common pooling operations include max pooling.

3. Fully Connected (Dense) Layers: After several convolutional and pooling layers, the output is flattened and passed through fully connected layers that perform the final classification based on the extracted features.
4. Activation Functions: Activation functions like ReLU (Rectified Linear Unit) introduce non-linearity to the network, allowing it to learn complex patterns.
5. Output Layer: The final layer of the network outputs the probabilities for each class.

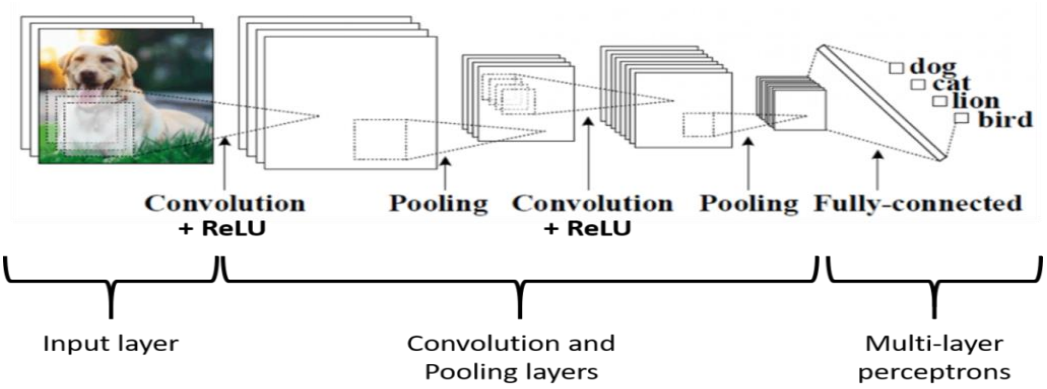

Figure 6.2: Architecture of CNN

Source: *Alzubaidi* et al. (2021)

6.3.2 CNN for Classification Task

Here, according to *Alzubaidi et al. (2021)*, we detail each component of the CNN architecture and explain the process involved at each step, from input to output.

1. Conv1D Layer

This layer applies convolutional filters to extract features. First, the Conv1D layer performs a convolution operation by applying a set of filters (also known as kernels) to the input data. Each filter slides over the input data, computing a dot product between the filter weights and the input values. Secondly, it makes a feature extraction, this operation captures local patterns and features in the data, such as edges or textures in image data. In the context of tabular data, it helps to learn local dependencies between features.

For our case:

- Filters: 64 filters were used, each responsible for detecting different features.

- Kernel Size: A kernel size of 3 indicates that each filter looks at three consecutive input elements at a time.
- Output: The output of the Conv1D layer is a set of feature maps, each representing a different learned feature.

2. **MaxPooling1D Layer**

This layer reduces spatial dimensions and computational load. First, the pooling operation, the MaxPooling1D layer performs down-sampling by dividing the input into non-overlapping regions and taking the maximum value from each region. Then it makes a dimensionality reduction.

For our case:

- Pool Size: A pool size of 2 is used, which means that the max operation is applied to every two adjacent values, reducing the dimension of the feature map by half.

3. **Dropout Layer**

This layer prevents overfitting by randomly dropping units. Here, a regularization technique is done, where the dropout layer randomly sets a fraction of the input units to zero at each update during training time. This prevents the network from becoming too dependent on specific neurons and helps to generalize better on unseen data.

For our case here:

- Dropout Rate: A dropout rate of 0.5 indicates that 50% of the units are dropped out at each update during training.

4. **Flatten Layer**

This layer flattens the multi-dimensional input to a single vector. The flattening layer converts the multi-dimensional output of the convolutional and pooling layers into a one-dimensional vector which is necessary to connect the convolutional layers with the fully connected dense layers.

5. **Dense Layer**

This layer performs classification with fully connected layers. Here, the dense layers (also known as fully connected layers) perform high-level reasoning and classification. Each neuron in

a dense layer is connected to every neuron in the previous layer. Then, the activation function, in which the first dense layer uses ReLU activation to introduce non-linearity, while the final dense layer uses a sigmoid activation function to output probabilities for binary classification.

For our case:

- First Dense Layer: Consists of 64 units with ReLU activation, which allows the network to learn complex representations.
- Output Dense Layer: Consists of 1 unit with sigmoid activation, producing a probability value between 0 and 1 for binary classification.

6.4 The Implemented CNN Model

After training and validating the Convolutional Neural Network (CNN) model, it is essential to evaluate its performance on the test set to understand how well it generalizes to new, unseen data. The following section interprets the results of the model's evaluation, including the confusion matrix, accuracy, sensitivity, specificity, and the ROC AUC score.

Confusion Matrix

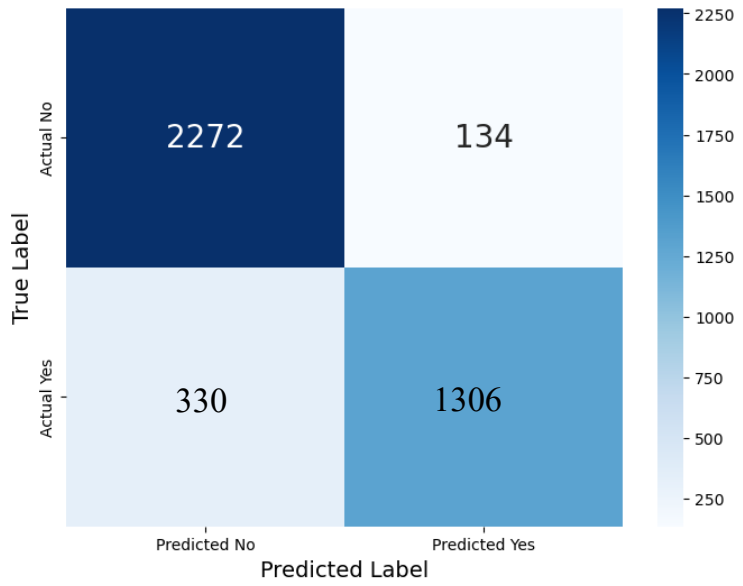

Figure 6.3: Heatmap for CNN Confusion Matrix

From Figure 6.3, We can conclude that the model is accurate in 88.52% of the cases, indicating good overall performance. With a sensitivity of 79.83%, the model correctly identified 79.83% of the instances where social security coverage was present. With a specificity of 94.43%, the model correctly identified 94.43% of the instances where social security coverage was not present.

ROC Curve for CNN Model

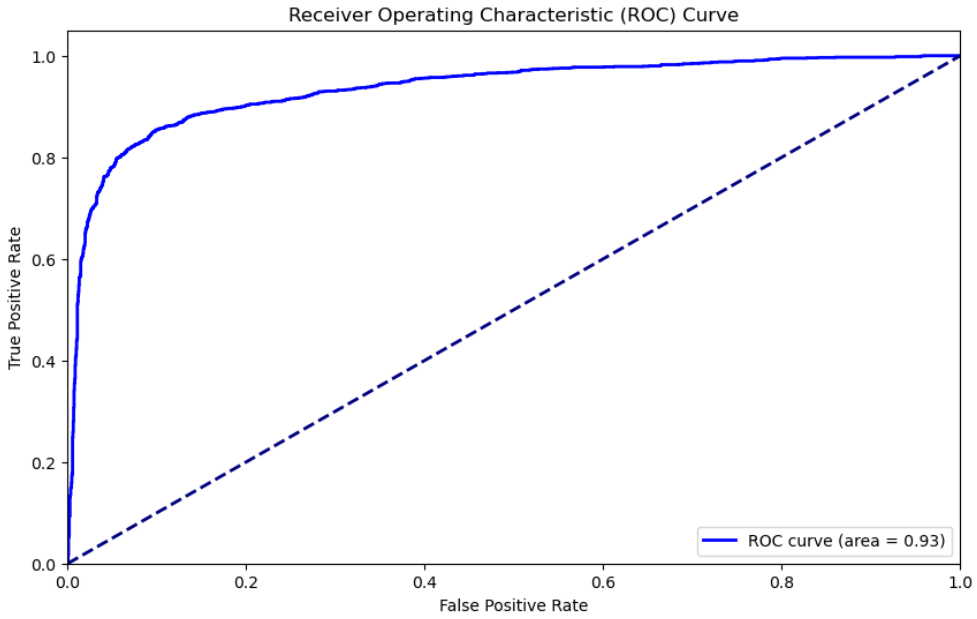

Figure 6.4: Receiving Operating Characteristic (ROC) Curve for CNN

The ROC curve for the CNN model in Figure 6.4 shows a significant rise towards the top-left corner of the plot. The curve's position above the diagonal line indicates that the model performs better than a random classifier. The AUC score is 0.93, which is a high value and indicates excellent discriminatory power. This means that there is a 93% chance that the model will correctly distinguish between a randomly chosen positive instance and a randomly chosen negative instance.

6.5 Comparing the 4 Applied Classification Models

In this section, we compare the performance of four different classification models: Binary Logistic Regression, Decision Tree, Random Forest, and Convolutional Neural Network (CNN). The comparison includes key metrics such as accuracy, precision, recall, F1-score, and ROC AUC score.

Performance Metrics

Table 6.1: Performance Matrix for All Applied Models

	Logistic Regression	**Decision Tree**	**Random Forest**	**CNN**
Overall Accuracy	88.25%	87.09%	88.5%	88.52%
Sensitivity	82.02%	78.46%	81.27%	79.83%
Specificity	92.33%	92.75%	93.24%	94.43%
AUC	0.94	0.86	0.94	0.93

From Table 6.1, the CNN model achieves the highest accuracy at 88.52%, closely followed by the Random Forest at 88.5%. Logistic Regression has the highest sensitivity at 82.02%, indicating it is the best at identifying positive instances (individuals with social security coverage), making it suitable for scenarios where detecting all instances of social security coverage is crucial. The CNN model has the highest specificity at 94.43%, meaning it is most effective at identifying negative instances (individuals without social security coverage). Logistic Regression and the Random Forest models both have the highest AUC at 0.94, indicating excellent discriminatory power. Decision Tree, while interpretable and straightforward, shows lower performance metrics compared to other techniques.

6.6 Advantages and Disadvantages of Each Model

Table 6.2: The Advantages and Drawbacks of the Applied Models

	Binary Logistic Regression	Decision Tree	Random Forest	Convolution Neural Network (CNN)
Advantages	1. Provides a measure of how significant the predictor is and the direction of the association either positive or negative. 2. Easy to implement and interpret. 3. Provides good accuracy for many simple data.	1. Simple to interpret and can be visualized. 2. No data preparation is required. 3. No statistical assumptions are required. 4. Make feature selection automatically.	1. Provides higher accuracy. 2. Reduce overfitting problem. 3. Less sensitive to outliers and noise in the data.	1. Can handle large datasets easily. 2. Capture the non-linear relationship between features and response variable. 3. Capture complex patterns and interactions, leading to high predictive power.
Disadvantages	1. The presence of multicollinearity can distort the estimates of the coefficients. 2. Assumes a linear relationship between the predictors and the log odds of the response. 3. Requires careful feature selection.	1. Individual trees are prone to overfitting. 2. Small changes in the data can result in a completely different tree structure.	1. More complex and hard to interpret. 2. Requires more computational power and memory, especially with large datasets.	1. Difficult to understand and interpret. 2. Requires large datasets to prevent overfit. 3. Requires longer time to train the model.

Chapter 7:

Conclusion and Recommendations

Social security is a very crucial service that each individual in our country needs to have, and the government's responsibility is to research and investigate ways to apply social security in a way that can reach the individuals who need it yet have social, demographic or economic barriers that decrease their chance of accessing these benefits. This study aimed to investigate the determinants of social security coverage among the Egyptian labor force using exploratory analysis and classification models. The analysis utilized data from the Household Income, Expenditure, and Consumption Survey (HIECS) 2019/2020, focusing on demographic, labor, health, and social characteristics. Four models were applied namely: Binary Logistic Regression, Decision Tree, Random Forest, and Convolutional Neural Network (CNN).

7.1 Main Findings

For the demographic variables, older individuals have higher odds of social security coverage, in addition to individuals in rural and urban areas are less likely to have coverage compared to those in frontier governorates. For the social factors, we found that workers with a father not in the household are more likely to have coverage, and higher-educated individuals are associated with increased odds of coverage, moreover, the never-married and widowed individuals show significant differences in coverage compared to divorced or separated individuals. For the health characteristics, it was found that disabled people are more likely to be covered, and having health insurance significantly increases the likelihood of coverage. Finally, regarding the labor factors, it was identified that employers have higher odds of coverage compared to unpaid family workers. In addition, employment in private or other sectors reduces the odds of coverage compared to government employment. Workers in craft, elementary, and agricultural occupations have lower odds of coverage compared to clerical workers. Workers in electricity and utilities, manufacturing, and transportation sectors have higher odds of coverage compared to those in agriculture and fishing.

About the applied models, the machine learning algorithms show that health insurance is the most influential factor. Age, sector of employment, and occupation significantly impact model accuracy. Furthermore, the CNN model achieves the highest accuracy and highest specificity, whereas random forest comes close behind it. The logistic regression showed the highest

sensitivity, making it highly effective at identifying individuals with social security coverage. While, the decision tree, although interpretable and straightforward, it has lower performance metrics compared to other techniques.

7.2 Recommendations

Based on the analysis and findings, the following recommendations are proposed to improve social security coverage among the Egyptian labor force:

1. Enhance access to health insurance as it emerged as the most critical factor influencing social security coverage. Policies aimed at increasing access to health insurance can significantly improve coverage rates.
2. Developing the social security system in the private sectors and other non-government sectors to strengthen the social security policies to ensure better coverage in these sectors. Provide incentives for employers in these sectors to facilitate social security enrollment for their employees.
3. Implementing policies to improve coverage among workers in craft, elementary, and agricultural occupations.
4. Developing industry-specific social security schemes to enhance coverage in the agriculture and fishing industry.
5. Making targeted awareness campaigns focusing on younger workers to raise awareness about the importance and benefits of social security coverage, encouraging early enrollment.
6. Conduct targeted campaigns in rural areas to raise awareness and facilitate access to social security benefits.
7. Incorporate marital status and family structure factors, such as the presence of the father at home, into policy designs to address the diverse needs of different household compositions.
8. Provide support mechanisms for educational advancement, which can indirectly increase social security enrollment through better job opportunities.
9. Utilize advanced models like sentiment analysis to understand workers' attitudes towards social security and identify barriers to enrollment.

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., and Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(53). https://doi.org/10.1186/s40537-021-00444-8.

Assaad, R., and Wahby, S. (2023). Why is social insurance coverage declining in Egypt? A decomposition analysis (Working Paper No. 1658). *Economic Research Forum*. https://www.erf.org.eg.

Barsoum, G., and Selwaness, I. N. (2022). Egypt's reformed social insurance system: How might design change incentivize enrolment? *International Social Security Review*, 75(2), 47-74. https://onlinelibrary.wiley.com/doi/10.1002/jid.3434.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387.

Kassem, N. (2021). Roles, rules, and controls: An analytical review of the governance of social protection in Egypt [Master's thesis, The American University in Cairo]. *AUC Knowledge Fountain*. https://fount.aucegypt.edu/etds/1583.

Ketkar, N., and Moolayil, J. (2021). *Convolutional neural networks. In Deep Learning with Python* (pp. 133-150). Apress. https://doi.org/10.1007/978-1-4842-5364-9_6.

Loewe, M. (2024). Social security in Egypt: An analysis and agenda for policy reform. *German Institute of Development and Sustainability (IDOS)*. Retrieved from https://www.idos-research.de/uploads/media/2024.pdf.

Merouani, W., El Moudden, C., and Hammouda, N. E. (2021). Social security enrollment as an indicator of state fragility and legitimacy: A field experiment in Maghreb countries. *Social Sciences*, 10(7), 266. https://doi.org/10.3390/socsci10070266.

Merouani, W., and Lassassi, M. (2021). The willingness to pay for social insurance: A field experiment in the Algiers Governorate. *Revue internationale des études du développement*, 247, 199-229. https://doi.org/10.3917/ried.247.0199.

OAMDI. (2023). Harmonized Household Income and Expenditure Surveys (HHIES), http://www.erfdataportal.com/index.php/catalog. Version 3.0 of Licensed Data Files; HIECS 2019/2020 - Central Agency for Public Mobilization and Statistics (CAPMAS). Egypt: *Economic Research Forum (ERF)*.

Roushdy, R., and Selwaness, I. (2017). Who is covered and who underreports: An empirical analysis of access to social insurance on the Egyptian labor market (GLO Discussion Paper No. 29). *Global Labor Organization (GLO)*. https://www.econstor.eu/bitstream/10419/155755/1/GLO_DP_0029.pdf.

Rupp, K., and Stapleton, D. (1995). Determinants of the growth in the Social Security Administration's disability programs: An overview. *Social Security Bulletin*, 58(4), 43-70. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6e8f51692b4c82fc32b9c47ea9cf93549cce43dc

Selwaness, I., and Barsoum, G. (2024). Social insurance in Egypt: Between costly formality and legal informality. *Economic Research Forum*. Retrieved from https://theforum.erf.org.eg/2024/01/14/social-insurance-in-egypt-between-costly-formality-and-legal-informality/

Sieverding, M., & Selwaness, I. (2012). Social protection in Egypt: A policy overview (Gender and Work in the MENA Region Working Paper No. 23). *Cairo: Population Council*. https://knowledgecommons.popcouncil.org/departments_sbsr-pgy/120/.