Strathmore UNIVERSITY

SU+ @ Strathmore University Library

Electronic Theses and Dissertations

2024

A Machine learning model for human population forecasting: case for Kenya.

Mete, Mamur Oromo Obuto
School of Computing and Engineering Sciences
Strathmore University

Recommended Citation

Mete, M. O. O. (2024). *A Machine learning model for human population forecasting: Case for Kenya*

[Strathmore University]. https://hdl.handle.net/11071/16565

Follow this and additional works at: https://hdl.handle.net/11071/16565

This work is availed for free and open access by Strathmore University Library.
It has been accepted for digital distribution by an authorized administrator of SU+ @Strathmore University.
For more information, please contact library@strathmore.edu

A Machine Learning Model for Human Population Forecasting: Case for Kenya

By

Mamur Oromo Obuto Mete

150895

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computing and Information Systems at Strathmore University

School of Computing and Engineering Sciences

Strathmore University

Nairobi, Kenya

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: Mamur Oromo Obuto Mete

Signature: _____ Date: 02/04/2024

Approval

The dissertation of MAMUR OROMO OBUTO METE was reviewed and approved for examination by the following:

Signature: B.K.S Date: April 2nd, 2024

Bernard Shibwabo, PhD

Senior Lecturer, School of Computing & Engineering Sciences,

Strathmore University

Abstract

The growth of a country's population can be a complex issue that has a significant impact on the development and sustainability of countries all over the world. In Kenya, the population is growing rapidly, which is putting a strain on the resources of the country, such as land, water, and infrastructure. The currently used methods of forecasting population growth, such as censuses and mathematical models, are costly, time-consuming, and not consistently accurate. The aim of this study is to develop a ML algorithm to forecast population growth in Kenya more accurately compared to the models currently being used. In this study, seven different machine learning models were examined Artificial Neural Networks, Random Forest, Logistic Regression, Support Vector Machines, Linear Regression, Decision Trees, and K-Nearest Neighbor to determine their effectiveness in predicting the population of Kenya. A variety of factors that impact population growth were considered, such as fertility and mortality rates, life expectancy, net migration, economic growth, access to healthcare and education, and gender equality. All models were built using the Scikit-Learn library and demonstrated impressive accuracy, but the top performers were Artificial Neural Networks, Random Forest, and Linear Regression. Of these, Linear Regression stands out as the best performer overall with a MAPE of 0.0179% and an accuracy of 0.9977% when tested with new data. This is a significant improvement over the other models, which showed slightly lower accuracies.

VT OMNES VNVM SINT

Table of Contents

Declaration and Approval ..... **ii**
Abstract ..... **iii**
List of Figures ..... **viii**
List of Abbreviations ..... **xi**
Definition of Terms ..... **xiii**
Acknowledgements ..... **xiv**
Chapter 1: Introduction **1**
1.1 Background to the Study 1
1.2 Problem Statement 4
1.3 Research Objective 5
1.3.1 Specific Objectives 5
1.4 Research Questions 5
1.5 Justifications 6
1.6 Scope and Limitations 6
Chapter 2: Literature Review **7**
2.1 Introduction 7
2.2 Theoretical Framework 7
2.2.1 Definition and Measurement of Population Growth 7
2.2.2 Challenges of Population Growth 10
2.2.3 Management of Population Growth 11
2.2.4 Evaluation of Population Growth Management 14
2.3 Empirical Framework 16
2.3.1 Challenges of Population Growth Forecasting 16
2.3.2 Population Growth Forecasting 17
2.4 Existing Techniques for Population Growth Forecasting 18
2.4.1 Machine Learning 18
2.4.2 Classification Techniques 20
2.5 Models and Frameworks 28
2.5.1 Models 28
2.5.1.1 Hybrid Intelligence 28

2.5.1.2 Predictive Models 31
2.5.2 Frameworks 32
2.6 Limitations of the Current Techniques and Approaches Used for Population Growth Forecasting 42
2.7 Conceptual Model 43
Chapter 3: Research Methodology 45
3.1 Introduction 45
3.2 Research Design 45
3.3 Population 46
3.4 Sample Size 47
3.5 Data Collection 48
3.6 Population Growth Forecasting Model Development 49
3.6.1 Extraction of Data 49
3.6.2 Preprocessing Data 49
3.6.3 Selecting the Features 50
3.6.4 Architecture of the model (ANN) 51
3.6.5 Population Growth Forecasting Model Validation 52
3.7 Research Utilisation 53
3.8 Systems Development Methodology 53
3.9 Research Quality and Reliability 55
3.10 Ethical Considerations 56
Chapter 4: System Analysis and Design 57
4.1 Introduction 57
4.2 System Requirement Analysis 57
4.2.1 Functional Requirements 58
4.2.2 Non-Functional Requirements 58
4.3 Systems Architecture 59
4.4 System Design 59
4.4.1 Use Case Diagram 60
4.4.2 Class Diagram 61
4.4.3 Sequence Diagram 62

VT OMNES VNVM SINT

Chapter 5: Systems Implementation and Testing 63
5.1 Introduction 63
5.2 ML Models Components 63
5.2.1 Artificial Neural Networks 63
5.2.2 Random Forest 64
5.2.3 Logistic Regression 65
5.2.4 Support Vector Machine 66
5.2.5 Linear Regression 67
5.2.6 Decision Tree 68
5.2.7 K-Nearest Neighbors 69
5.3 Web Application Interface Components 70
5.3.1 Home Page 70
5.3.2 Menu Page 71
5.3.3 Signup Page 71
5.3.4 Login Page 72
5.3.5 CSV File Upload Page 73
5.3.6 Model Selection Page 74
5.4 System Implementation 75
5.4.1 System Development Environment 76
5.4.2 Population Data Collection 76
5.4.3 Population Data Preprocessing 77
5.4.4 Exploratory Data Analysis 78
5.4.5 Models' Training and Testing 79
5.4.6 Models' API and the Population Prediction Platform 81
5.5 System Testing, Validity and Usability 82
Chapter 6: Discussions 97
6.1 Research Objectives Review 97
6.1.1 Determining the Challenges Associated with the Forecasting of Human Population growth in Kenya 97
6.1.2 Analysis of The Current Methods, Techniques, and Approaches for Forecasting Human Population Growth 98

6.1.3 Development of the ML Models for Forecasting Human Population Growth in Kenya 98
6.1.4 Validation of the ML models 99
6.1.5 Insights and Interpretations of the Models 99
6.2 Advantages of the Tool 99
6.3 Limits to the study and the Population Predictions Platform 99
Chapter 7: Conclusions and Recommendations 100
7.1 Conclusions 100
7.2 Recommendations 100
7.3 Future Works 100
References 102
Appendices 115
Appendix A: Similarity Report 115
Appendix B: Ethical Clearance Confirmation 116

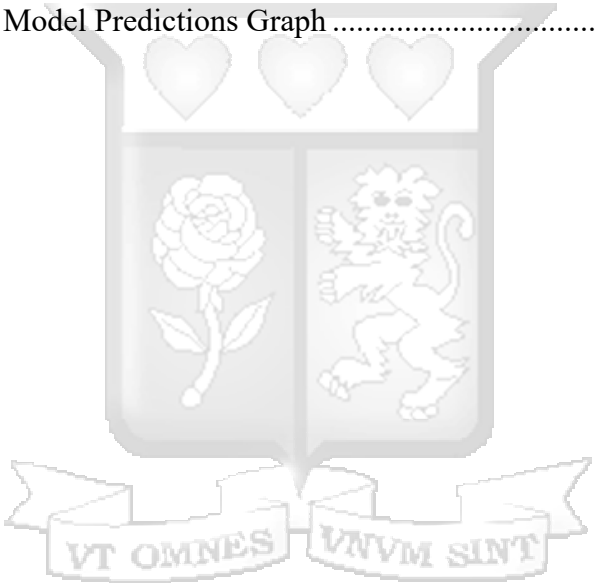
VT OMNES VNVM SINT

List of Figures

Figure 2.1: Random Forest Algorithm (IBM, 2023) 21
Figure 2.2: ANN inspired by human brain (IBM, 2023; Wang, Zhang, & Zhang, 2022) 25
Figure 2.3: Neural Network (IBM, 2023; Wang, Zhang, & Zhang, 2022) 26
Figure 2.4: SVM (Li & Wang, 2022) 28
Figure 2.5: Conjunction of Human Intelligence and Machine Intelligence (Dellermann & Leimeister, 2023) 29
Figure 2.6: Hybrid Intelligence (Dellermann & Leimeister, 2023) 30
Figure 2.7: Concept of Hybrid Intelligence (Dellermann & Leimeister, 2023) 31
Figure 2.8: TensorFlow Framework Architecture (Li & Wang, 2023) 33
Figure 2.9: TensorFlow Data Flow Graph (Tian, 2023) 34
Figure 2.10: Pytorch Framework Architecture (Singh, 2023) 35
Figure 2.11: PyTorch Workflow (PyTorch Team, 2023) 36
Figure 2.12: Scikit-Learn Pipeline (Scikit-Learn, 2023) 37
Figure 2.13: Scikit-Learn Algorithm Cheat-Sheet (Scikit-Learn, 2023) 38
Figure 2.14: Keras Framework Architecture (Hymel, 2020) 39
Figure 2.15: NumPy API (Harris et al., 2020) 40
Figure 2.16: Reading and Writing Tabular Data (The pandas development team, 2020) 41
Figure 2.17: Creating Plots in Pandas (The pandas development team, 2020) 42
Figure 2.18: Conceptual Framework 44
Figure 3.1: Kenya's total population 2022 (World Bank, 2023) 47
Figure 3.2: ANN Architecture 52
Figure 3.3: Agile Software Development Methodology (Alsaqqa, Sawalha & Abdel-Nabi, 2020) 55
Figure 4.1: Systems Architecture 59
Figure 4.2: Use Case Diagram 60
Figure 4.3: Class Diagram 61
Figure 4.4: Sequence Diagram 62
Figure 5.1: ANN Class 64
Figure 5.2: Random Forest Class 65
Figure 5.3: Logistic Regression Class 66

Figure 5.4: SVM Class 67
Figure 5.5: Linear Regression Class 68
Figure 5.6: Decision Tree Class 69
Figure 5.7: KNN Class 69
Figure 5.8: Home Page 70
Figure 5.9: Menu Page 71
Figure 5.10: Signup Page 72
Figure 5.11: Login Page 73
Figure 5.12: CSV File Upload Page 74
Figure 5.13: ML Model Selection Page 75
Figure 5.14: Loading Data from the CSV File 78
Figure 5.15: EDA Analysis 79
Figure 5.16: Training Models 80
Figure 5.17: Testing Models 81
Figure 5.18: ANN Trained Model Predictions Metrics 82
Figure 5.19: ANN Tested Model Predictions Metrics 83
Figure 5.20: Random Forest Trained Model Predictions Metrics 83
Figure 5.21: Random Forest Tested Model Predictions Metrics 84
Figure 5.22: Logistic Regression Trained Model Predictions Metrics 84
Figure 5.23: Logistic Regression Tested Model Predictions Metrics 85
Figure 5.24: SVM Trained Model Predictions Metrics 85
Figure 5.25: SVM Tested Model Predictions Metrics 86
Figure 5.26: Linear Regression Trained Model Predictions Metrics 86
Figure 5.27: Linear Regression Tested Model Predictions Metrics 87
Figure 5.28: Decision Tree Trained Model Predictions Metrics 87
Figure 5.29: Decision Tree Tested Model Predictions Metrics 88
Figure 5.30: KNN Trained Model Predictions Metrics 88
Figure 5.31: KNN Tested Model Predictions Metrics 89
Figure 5.32: ANN Trained Model Predictions Graph 89
Figure 5.33: ANN Tested Model Predictions Graph 90
Figure 5.34: Random Forest Trained Model Predictions Graph 90

Figure 5.35: Random Forest Tested Model Predictions Graph 91
Figure 5.36: Logistic Regression Trained Model Predictions Graph 91
Figure 5.37: Logistic Regression Tested Model Predictions Graph 92
Figure 5.38: SVM Trained Model Predictions Graph 92
Figure 5.39: SVM Tested Model Predictions Graph 93
Figure 5.40: Linear Regression Trained Model Predictions Graph 93
Figure 5.41: Linear Regression Tested Model Predictions Graph 94
Figure 5.42: Decision Tree Trained Model Predictions Graph 94
Figure 5.43: Decision Tree Tested Model Predictions Graph 95
Figure 5.44: KNN Trained Model Predictions Graph 95
Figure 5.45: KNN Tested Model Predictions Graph 96



List of Abbreviations

AD Tree	Alternating Decision Tree
ADM	Agile Development Methodology
AI	Artificial Intelligence
ANN	Artificial Neural Network
APA	American Psychological Association
API	Application Programming Interface
ARIMA	Auto-Regressive Integrated Moving Average
BSD	Berkeley Source Distribution
CO_2	Carbon-di-Oxide
CUDA	Compute Unified Device Architecture
cuDNN	CUDA Deep Neural Network
DBSCAN	Density-based Spatial Clustering of Applications with Noise
GANs	Generative Adversarial Networks
GIS	Geographic Information Systems
IBM	International Business Machines
IUDs	Intrauterine Devices
KNBS	Kenya National Bureau of Statistics
KPC	Kenya Population and Housing Census
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MSE	Mean Squared Error
ONEIROS	Open-ended Neuro-Electronic Intelligent Robot Operating System
oob	out-of-bag
OOM	Object-Oriented Modeling
PCA	Principal Component Analysis
SNNs	Simulated Neural Networks
SU-ISERC	Strathmore University Institutional Scientific and Ethical Review Committee
SVD	Singular Value Decomposition
SVM	Support Vector Machine
UN	United Nations
UNDP	United Nations Development Programme

UNEP	United Nations Environment Programme
UNFPA	United Nations Population Fund
UNHCR	United Nations Children's Fund
UNICEF	United Nations Children’s Fund
UNSD	United Nations Statistics Division
CSV	Comma Separated Value
USCB	United States Census Bureau
MAE	Mean Absolute Error
MSLE	Mean Squared Log Error
MAPE	Mean Absolute Percentage Error
R2 Score	Coefficient of Determination
KNN	K-Nearest Neighbor
SVR	Support Vector Regression



Definition of Terms

Application Programming Interface	Application Programming Interface is a set of defined rules that enable different applications to communicate with each other (IBM, 2023).
Artificial Intelligence	Artificial Intelligence is defined as the science and engineering of making intelligent machines, especially intelligent computer programs (Tian, 2023; Ghosh & Kumar, 2022).
Hybrid Intelligence	Hybrid Intelligence is the approach that integrates both Artificial Intelligence and Human Intelligence (Dellermann & Leimeister, 2023).
Machine Learning	Machine Learning, is the approach where machines are trained to "discover" their own algorithms without explicit human guidance (Hymel, 2020; Deisenroth Rasmussen & Blundell, 2020).
Population	Population is the number of people in the world (UN, 2023; World Bank, 2023).
Population Growth	Population growth is the increase in the number of people in the world (Bongaarts, 2022).
Population Growth Forecasting	Population Growth Forecasting is a statistical technique that estimates the future population of a region or country (Şahinarslan, Tekin, & Çebi, 2021; Bardsley & Hugo, 2021).

Acknowledgements

To begin my acknowledgement, I would like to express my gratitude to God and my family for their unwavering support throughout my master's program. I am also deeply grateful to my dissertation supervisor and mentor, Dr. Bernard Shibwabo, for his ever-invaluable instructions and guidance, which kept me on the right path.

Chapter 1: Introduction

1.1 Background to the Study

Population growth defined by World Bank, (2023) as the increase in the number of people in the world proves to be a complex issue with far-reaching implications for societies around the world. Many factors lead to population growth, for example the rates of fertility, mortality, and migration (World Bank, 2023; Kenya National Bureau of Statistics, 2022).

The UN in 2023, claim that population of the world has tripled since the mid-1900s; and the UN also reported that during November 2022, the population of the world had reached 8.0 billion. This increase includes 1 billion more people since 2010 and 2 billion more since 1998. Additionally, the UN estimates the population to continue growing to hit the 9.7 billion numbers by the year 2050 and potentially peaking at 10.4 billion in the mid-2080s. The UN attributes this rise in population is due to factors such as more people reaching reproductive age, longer lifespans, urbanization, and migration. These trends have also led to significant changes in fertility rates, which will have long-term implications for future generations.

The challenges and opportunities associated with population growth vary across different countries on different continents. Some countries see population growth as a major obstacle to development, while other countries see it as an opportunity to boost the economy and create a more vibrant society (World Bank, 2023).

According to studies by Suárez, Hoyos, Vallejo, and Arias (2022) and Gómez, Patiño, Duque, & Passos (2020), as the population of an area increases and people crowd the urban areas, there are a plethora of problems associated with it. These problems can include poverty, spontaneous settlements, insecurity, all forms of pollution, disease outbreaks, and the economic cost of broadening substructure of the public. The urban areas like cities also ingest raw materials in immense amounts, stimulating the act of demeaning the environment therefore, leading to dramatic change in climate. This, in turn, leads to rising sea levels, extreme weather events, and the outbreaks of tropical diseases, which have serious repercussion on the urban areas or cities.

Additionally, the cities are the leading contributors to the alteration in climate, with transport and buildings being the largest sources of greenhouse gas emissions. Suárez and his colleagues estimated that urban areas or cities cause 75% of global CO2 emissions (United Nations Environment Programme (UNEP), 2023).

Governments of countries all over the world require an effective administration to govern and make decisions for their respective countries. However, to fulfil their responsibility, the country administration requires sufficient knowledge of population size which is crucial for effective planning, resource allocation, and policy formulation (Şahinarslan, Tekin, & Çebi, 2021).

Since the KNBS is backed up by (UNFPA) United Nations Population Fund for data collection from censuses for the estimation of the country's future population, therefore, conducting a census regularly throughout a country is a necessity and a vital procedure; which is defined as the act of collecting, compiling, analysing and publishing the socio-demographic and socio-economic data about people living in a country (United Nations Statistics Division (UNSP), 2019). This is a costly and time-sensitive process since it does not always provide accurate data. For example, the 2009 Kenya Population and Housing Census which was conducted for 12 billion Kenyan shillings (approximately US$110 million), the population of Kenya was found to be 38.6 million and the census also took two years to complete, and the data was not released until 2013 (Kenya National Bureau of Statistics (KNBS), 2022).

Since forecasting events is a crucial factor in different fields to remove uncertainties that can happen in the future, various algorithms of machine learning have been trained on known data and used to forecast unknown data (Şahinarslan, et al., 2021). They found out that the functionality of machine learning algorithms is determined by the estimates they make with the minimum dataset. In real life and the desired amount of data may not be available. Therefore, the data that researchers observe, or experience is limited. Still, they want to make predictions with this dataset and (Bélisle et al., 2015) also stated that the primary purpose is to make realistic predictions with few datasets. Their study of a machine learning model was able to forecast population growth in Turkey with an accuracy of 95%.

In their study, Suárez et al., (2022) found that forecasting can be difficult due to the unpredictable and chaotic nature of social systems. They utilised machine learning techniques to model population growth by analysing yearly CO2 data. Their Artificial Neural Network (ANN) was trained using demographic and CO2 emission data and was ultimately successful in forecasting population growth.

According to the 2019 KPC which was monitored by the KNBS, the following methods were used to conduct the census (KNBS, 2019; KNBS, 2010):

Household enumeration which was the main method used to collect data during a census where enumerators had to visit households and collect information about the people living in the household, such as their names, ages, sex, and education levels.

Mobile technology was the first paperless census in Kenya where enumerators used tablets to collect data, which was then transmitted to the KNBS servers in real-time. This helped to speed up the data collection process and to reduce errors.

Remote sensing is the use of satellites and other technologies to collect data about the Earth's surface where the KNBS used remote sensing to create a digital map of Kenya, which was used to help plan the census and to identify areas that were difficult to reach.

Geographic information systems (GIS) are computer system that is used to store, manage, and analyse spatial data and the KNBS used GIS to analyse the data collected during the census, which helped to identify trends and patterns in the population.

Statistical methods are that analyse the data collected during the census used by the KNBS to assure that the data was errorless and reliable.

But the weaknesses of these methods are that they are:

Costly and time-consuming: Censuses are expensive to conduct, and they can take several years to complete. Mathematical models can also be costly to develop and maintain.

Not always accurate: Censuses are not always accurate, as they can be difficult to conduct in remote areas or countries with high levels of illiteracy. Mathematical models can also be inaccurate, as they do not consider all the factors that affect population growth.

Do not consider all factors that affect population growth: These methods do not consider all the factors that affect population growth, for example, changes in the rate of fertility and mortality, and migration patterns.

Therefore, this study is focused to create ML model to forecast population growth in Kenya more accurately, as it will put into consideration factors such as fertility and mortality rates, life expectancy, net migration, economic growth, access to healthcare and education, and gender equality that affect population growth.

1.2 Problem Statement

The current methods used to project and forecast population growth in Kenya such as census which includes using Household enumeration, Mobile technology, Remote sensing, GIS, and Statistical methods are costly, time-consuming, and not always accurate (United Nations Statistics Division, 2019; Odhiambo, K'Akumu & Gichero, 2020; Gould, 2018). This is a problem because accurate population forecasts are essential for planning for the future of Kenya. UNFPA, (2021) as well as Fan and Yao, (2020) claimed that inaccurate population forecasts can lead to misallocation of resources, poor decision-making, and social unrest. For instance, if a government overestimates the size of its population, it might set expectations that are impossible to fulfil. The result could be riots or other forms of social unrest. On the other hand, if a government underestimates the size of its population, it might not be able to give its citizens the services they need. Social unrest might result from this too.

Therefore, the study will develop a model of ML that can predict the population growth in Kenya more accurately by considering the factors that affect population growth, such as fertility and mortality rates, life expectancy, net migration, economic growth, access to healthcare and education, and gender equality.

This machine learning model can provide the policymakers, investors, and the government with a more accurate forecast model, which can be used to make better decisions about planning for the future of Kenya, leading to reduced costs and increased efficiency compared to the current methods.

1.3 Research Objective

The main aim of this study is to develop a machine learning model that can predict the population growth in Kenya more accurately by considering the factors that affect population growth, such as fertility and mortality rates, life expectancy, net migration, economic growth, access to healthcare and education, and gender equality.

1.3.1 Specific Objectives

a) To determine the challenges associated with the prediction of population growth in Kenya.

b) To analyse the current methods, techniques and approaches used to forecast population growth.

c) To develop a machine learning model that can forecast population growth in Kenya.

d) To validate the proposed machine learning model.

1.4 Research Questions

a) What challenges are associated with the prediction of population growth in Kenya?

b) What methods, techniques and approaches are used to forecast population growth currently?

c) How can an accurate machine learning model be developed to forecast population growth in Kenya?

d) How can the proposed machine learning model be validated?

1.5 Justifications

The current methods used to project and forecast population growth in Kenya are traditional, costly, time-consuming, and not always accurate. This is a problem because accurate population forecasts are essential for the government to plan for the future of Kenya. However, inaccurate population forecasts can lead to serious economic, social, political, climate and environmental problems, for example, misallocation of resources, poor decision-making, and social unrest. A model of machine learning could forecast population growth in Kenya by learning from historical data and making predictions given the already successful study of Suárez on Machine learning for population growth modeling through annual CO_2 in Colombia (Suárez et al., 2020). This would be a more efficient and accurate way to forecast population growth than the current methods and a valuable tool that the government and other stakeholders such as investors could use for making policies and investments respectively. It would help to ensure that resources are allocated efficiently and that decisions are made based on accurate information, ultimately benefiting the citizens by helping to improve their quality of life.

1.6 Scope and Limitations

The objective of this project is to create a machine learning model that can forecast population growth in Kenya. The model created was trained on historical data, as well as other factors that may impact population growth, such as fertility and mortality rates, life expectancy, net migration, economic growth, access to healthcare and education, and gender equality. By utilising this model, predictions can be made about future population growth in Kenya. However, it is important to note that the accuracy of the model is dependent on the quality of the historical data that was used to train it. Additionally, the model can only make predictions based on the factors that are included in the dataset used for training. If there are other factors that impact population growth in Kenya that are not included in the dataset, the model cannot take them into account. Furthermore, the model can only be used to make predictions for Kenya and no other countries. Despite these limitations, this project is a valuable tool for the government and other stakeholders. It ensures that resources are allocated efficiently, and decisions are made based on accurate information, ultimately benefiting the citizens of Kenya.

Chapter 2: Literature Review

2.1 Introduction

This chapter examines the relevant studies done on the forecasting of population. Population growth is a complex phenomenon influenced by a variety of factors. Some of the most important factors include fertility and mortality rates, life expectancy, net migration, economic growth, access to healthcare and education, and gender equality (World Bank, 2023; Kenya National Bureau of Statistics, 2022). Therefore, forecasting population growth is important for a plethora of reasons to pick from: for example, it can help governments to plan, and help businesses to make decisions about where to invest, and it can help individuals to make decisions about their future Alam and Islam, (2022). There is a body of research that is expanding on the factors influencing population growth. However, there is still a big need for more research on this topic. This chapter reviews the relevant studies on the factors that influence population growth, discuss the current methods used to predict population growth, and conclude by discussing the need for further research.

2.2 Theoretical Framework

2.2.1 Definition and Measurement of Population Growth

Population growth has been the subject of much discussion in recent years and for good reasons. The rising number of people inhabiting the planet poses numerous implications for societies around the globe (World Bank, 2023). Factors such as birth rates, mortality rates, and migration all play a critical role in this complex issue. The impact of population growth on the environment, economy, and social structures is significant and cannot be overlooked. Therefore, it is imperative to gain a deep understanding of this phenomenon to develop sustainable solutions to address the challenges it presents (Dobbs, Manyika, Woetzel, & Ahmed, 2022; World Bank, 2023).

Monitoring population expansion is a crucial tool for assessing the development of a community over a period and comparing the growth rates among different cohorts. It also helps in

forecasting the forthcoming magnitude of a population. Many factors are considered while gauging population growth (Alam & Islam, 2022), and they include:

a) The fertility rate is a crucial determinant of population growth, and its fluctuations can significantly impact a country's economy, healthcare system, and social structure (Otoom et al., 2019). Generally, the rate of fertility of about 2.1 children by a single woman is required for maintaining a stable population (World Bank, 2023). However, many countries are currently experiencing declining fertility rates, which can lead to aging populations and workforce shortages. This trend has prompted discussions around family planning policies, child-rearing incentives, and immigration policies to address potential demographic imbalances. Given the complexity of this issue, it is imperative that policymakers carefully consider and plan (Bongaarts, 2022).

b) The mortality rate is an important statistic that measures the number of deaths that occur within a particular population during a given year. This figure is often used to evaluate the overall health of a community or region and can provide valuable insights into trends and patterns in disease, injury, and other causes of death. While it is a useful tool for public health researchers and policymakers, it is important to remember that mortality rates can vary widely depending on a variety of factors such as age, gender, socioeconomic status, and access to healthcare. By carefully analysing mortality data and taking steps to address underlying causes of death, communities can work to improve the health and well-being of their residents over time (Dobbs, Manyika, Woetzel, & Ahmed, 2022).

c) Population growth can be influenced by the net migration rate, an important demographic indicator that reveals the difference between the number of individuals who move into a population and those who move out (Castles & Miller, 2020). This measure is often used to gain insights into migration trends and understand the impact it can have on a population. The rate of net migration is computed by taking the totality of immigrants and subtracting the totality of emigrants, then dividing the result by the total population.

This gives us a percentage representing the net gain or loss of people in a particular area. Understanding this rate is crucial for policymakers and researchers studying population dynamics and planning.

d) A range of economic factors can influence population growth, but two key ones are the degree of development in the economy and the availability of employment opportunities (Tavakolan & Ebrahimi, 2022). In areas with high levels of economic growth, there is often more investment in infrastructure and public services, which can improve living conditions and attract more people. Similarly, areas with plentiful job opportunities can draw in workers from other regions, further contributing to population growth. However, it is worth noting that population growth can have downsides, such as increased strain on resources and infrastructure. Overall, it is fundamental to cautiously put into considerations economic factors that can impact population growth and plan accordingly (Cai & Wang, 2022).

e) One important consideration that can impact population growth is the accessibility of healthcare and education. These social factors have considerable role in determining the general health and welfare of a community and can have long-term implications for the growth and development of a population. By ensuring that individuals have access to quality healthcare and education services, we can help to create a more even and regenerating or sustainable future for all (United Nations, 2022).

f) There are many factors that can influence population growth, with government stability and the level of civil unrest being just a couple of examples. Other factors could include access to healthcare and education, economic opportunities, and cultural beliefs and practices. Understanding these complex dynamics is essential for policymakers and researchers who seek to promote sustainable and equitable population growth in the years ahead (United Nations, 2022).

g) One of the most crucial determinants that can significantly impact the expansion of a population is the environment surrounding it. The conditions and resources available in a

particular area are vital in determining the growth of population. This is because the environment can affect the availability of food, water, shelter, and other essential resources needed for survival. Moreover, environmental factors like climate change, natural disasters, and human activities such as deforestation and pollution can also contribute to population decline and affect the sustainability of a particular species. Therefore, it is imperative to study and comprehend the environment and its impact on population growth, in current and the futuristic events (Dyson and West, 2019).

2.2.2 Challenges of Population Growth

Challenges related to population growth can be broadly categorised into economic, environmental, and social challenges: According to Murray, Brown, and Rogers, (2020) economic challenges arise due to overpopulation, which occurs when the population of a region exceeds the available resources. This can lead to food and water shortages and environmental degradation. Unemployment is another economic challenge that results from population growth. When there are not enough jobs to support all the people in the workforce, it causes social unrest and political instability. Inequality is also a significant issue of population growth, as it can divide the rich and the poor.

Environmental challenges include environmental degradation, as more people put pressure on the environment, leading to deforestation, soil erosion, and climate change. Resource depletion is another significant environmental challenge that arises due to population growth, as more people consume more resources, leading to water, energy, and food shortages (Dyson & West, 2019).

Social unrest and political instability are major social challenges that also arise due to population growth. As people become frustrated with the lack of jobs and resources, it can lead to protests, riots, and even civil war. Governments may also struggle to cope with the challenges of a growing population, leading to political instability and potential overthrow (Bongaarts, 2021).

Finally, although population growth comes with challenges, it offers opportunities such as a larger workforce, younger population, new ideas, and diversity. To manage it, invest in education, healthcare, and infrastructure, promote sustainable development, and address poverty

and inequality. This can create a skilled workforce, efficient resource use, environmental protection and reduce people struggling to meet basic needs. Population growth can be a positive force with the right measures in place (United Nations Children's Fund (UNICEF) 2022).

2.2.3 Management of Population Growth

Kelley and Schmidt, (2018) stated that population growth management is crucial to ensuring sustainable development in any given country or region. Therefore, several policies and programs can be implemented to control the rate of population growth effectively. Some of these include family planning programs, incentivising small families, investing in education and healthcare, and promoting immigration. By managing population growth, countries can ensure that their resources are sustainable and that their citizens have access to the resources they need to thrive (Cohen, 2019). These policies and programs include the following:

a) Bongaarts and Westoff, (2017); as well as Adebayo and Owolabi, (2022) believed that family planning programs are designed to help couples plan their families by providing them with relevant information and a range of services. This may include access to various forms of contraception, such as condoms, birth control pills, intrauterine devices (IUDs), and abortion services. By educating couples on the various methods available, family planning programs can help them make informed decisions about their reproductive health and prevent unintended pregnancies (Singh & Darroch, 2018). Additionally, these programs pave the way in the process of promoting gender equality, emancipation, and empowerment of women by giving them greater control over their own bodies and lives (Tsui& Zhang, 2019).

b) Fernandes and Santos, (2022) stated that economic development programs are one of the most effective ways to reduce poverty and improve the overall standard of living. This is because when people are financially stable, they tend to have fewer children, which can help to lower fertility rates. In addition to this, economic development programs can also lead to the creation of new jobs and opportunities, which can help to boost local economies and make the life of people in a community to be better (Klasen, 2022).

Overall, economic development programs are an important tool in the fight against poverty and can have a significant impact on the lives of individuals and communities around the world (Ravallion, 2022).

c) Zimet and Shah, (2016) as well as Mehrish, Majumder, Bharadwaj, Mihalcea and Poria, (2023) claimed that education programs have been proven to be highly effective in promoting the many advantages of having smaller families. In addition to this, such programs can also uplift the social status of women, leading to a significant decrease in fertility rates over time. By educating people about the many benefits of small families, we can work towards building a more sustainable and equitable future for all.

d) Improving healthcare has the potential to positively impact mortality rates, which in turn could lead to an increase in population growth. Furthermore, providing access to contraception and abortion services through healthcare programs can effectively lower fertility rates, ultimately contributing to a more balanced population growth (Otoom et al., 2019).

e) Migration policies are a set of rules that are put in place to control the movement of individuals into and out of a particular country (Castles & Miller, 2020). The main objective of these policies is to manage population growth by regulating the number of immigrants that are permitted to enter the country. By limiting the number of individuals who can migrate to the country, the government can ensure that the country's resources are distributed fairly and that the needs of all citizens are met. Moreover, these policies also help maintain a stable economic environment, ensuring that the job market is filled with only a few immigrants.

2.2.3.1 Benefits of Population Growth Management

Managing population growth can have numerous advantages that go beyond just controlling the number of people living in an area (Bardsley & Oswald, 2017). For one, it can help to reduce the strain on resources such as water, food, and energy. Additionally, managing population growth can help to reduce the carbon footprint of a community and improve overall environmental

sustainability. Furthermore, it can help reduce the level poverty leading to enhancement of access to services like healthcare, education etc. Ultimately, managing the rise population is vital to the development of a more equitable and sustainable world for everyone; here are some of the benefits of population growth management:

a) It can help to reduce poverty and improve the standard of living, whereby when people have fewer children, they can invest more resources in their children's education and healthcare (Ravallion, 2022). This can lead to a better quality of life for everyone.

b) It can help to protect the environment, whereby when there are fewer people, there is less demand for resources, which can help to protect the environment (Dyson and West, 2019).

c) It can help to promote gender equality whereby when women have access to education and healthcare, they are more likely to have fewer children (Zimet & Shah, 2016). This can help to promote gender equality, as women will have more opportunities to participate in the workforce and make decisions about their own lives.

2.2.3.2 Drawbacks to Population Growth Management

However, when it comes to managing population growth, there are a few disadvantages to consider (Conly & Hartmann, 2021). While it may seem like a good idea to control population expansion, there are limitations to what can be achieved through these efforts. Furthermore, a range of issues can arise when attempting to regulate population growth, including social, economic, and political considerations (Lutz, Samir & Samir, 2022). Overall, it is important to approach this issue with caution and care, as there are no easy solutions to managing population growth. Here are some of the potential drawbacks to population growth management:

a) Some individuals may perceive population growth management as a method of regulating the population's size, which may be considered a breach of human rights.

b) Implementing policies and programs for managing population growth can be challenging, as they often entail altering individuals' lifestyles and work habits (World Bank, 2018).

c) Policies and programs designed to manage population growth may have unforeseen consequences, potentially leading to an uptick in factors like abortion rates or poverty levels (Lutz et al., 2022).

2.2.4 Evaluation of Population Growth Management

When it comes to evaluating and assessing population growth management strategies, there are a plethora of effective and varied methods available for consideration (Hartmann & Gemmill, 2020). These methods can range from comprehensive data analysis to targeted surveys and interviews with key stakeholders. By utilising these approaches, decision-makers can get a better comprehension of the potential impacts of different strategies in order to help them make informed choices that benefit their communities (Lutz et al., 2022). Overall, it's important to approach population growth management with a thoughtful and nuanced perspective to accomplish the most suitable and possible results for everyone involved using the following methods:

a) Quantitative methods use data to measure the effectiveness of population growth management strategies (Fernandes & Santos, 2022). This can include using statistical methods to analyse data on fertility rates, mortality rates, and net migration rates.

b) Qualitative methods use interviews, focus groups, and other ways to gather information about the effectiveness of population growth management strategies (Mason, 2018). This can include asking people about their experiences with family planning programs or economic development programs.

c) Mixed methods use a combination of quantitative and qualitative methodologies to evaluate population growth management strategies (Bhandari & Mishra, 2021). This can be done by using statistical methods to analyse data on fertility rates and then using interviews to gather information about people's experiences with family planning programs.

2.2.4.1 Factors for Evaluating Population Growth Management Strategies

Alam et al., (2022) studied that when it comes to evaluating strategies for managing a growing population, there are several key considerations that must be considered to ensure the best possible outcomes. Indeed, with so many different factors at play, it is essential to carefully consider each one of them to make informed decisions that will have a positive impact on the community. From infrastructure to social services, from environmental impact to economic growth, many different variables must be considered when planning for the future of a growing population. Ultimately, by carefully considering all these factors and working collaboratively with stakeholders from all walks of life, it is possible to create a shared vision for a sustainable and thriving community that will serve everyone well for years (World Bank, 2018). The following are the factors:

a) To ensure an effective population management evaluation process, it's essential to establish clear and concise objectives for the strategy. By doing so, the assessment can concentrate on the most significant components and yield more accurate results (Berman and Khan, 2019). It's vital to prioritise the areas that require the most attention and allocate resources accordingly. By setting specific goals, the evaluation process will be more efficient and productive.

b) When evaluating a population management plan, it is crucial to examine the techniques utilised for execution (Nguyen, 2022). This will enable the detection of any possible shortcomings or flaws in the implementation procedure. It is also vital to put into consideration the consequence of the plan on the affected population and the environment (Khan & Mahmood, 2022). Additionally, stakeholders' input should be sought to ensure that the plan aligns with their needs and expectations. Considering all these factors, a comprehensive and effective population management plan can be developed and successfully executed.

c) When evaluating the success of a population management plan, it is crucial to assess the impact on various demographic factors such as fertility rates, mortality rates, and net

migration rates (Otoom et al., 2019). By analysing these metrics, we can determine whether the desired goals have been met and make necessary adjustments to the strategy. It is essential to regularly monitor and evaluate the effectiveness of population management strategies to ensure the best possible outcomes for all involved parties.

2.3 Empirical Framework

2.3.1 Challenges of Population Growth Forecasting

Predicting the rate at which a population will grow can be a difficult task, as there are many different variables that can impact the outcome (Bongaarts, 2020). From economic factors to social issues, a plethora of factors influence how quickly or slowly a population grows over time. Despite these challenges, many researchers and policymakers continue to work towards developing accurate and reliable methods for forecasting population growth, to better understand and plan (Bardsley et al., 2021). Whether through advanced statistical models, cutting-edge technology, or other innovative approaches, experts in this field are constantly striving to improve our understanding of this complex and ever-changing phenomenon. These include:

Predicting population growth has always been a daunting task, primarily due to the intricate interplay of various influencing factors. Social, economic, political, and environmental factors are just some of the variables that predict future population growth a challenging endeavour (Dobbs et al., 2022). These factors are in a constant state of flux, and it is difficult to anticipate how they may evolve over time, thus making the prediction of future population growth even more complex.

Future events that are unpredictable, such as wars, natural disasters, and technological advancements, can also affect population growth. These events can have big consequences on population growth, causing challenges to predict population with high degree of accuracy. As a result, there is an essential need in considering a range of factors that influence population growth, including economic, social, and environmental factors. For example, changes in fertility rates, mortality rates, and migration patterns can all have a significant impact on population growth. Additionally, advances in healthcare, education, and technology can also affect

population growth by improving living conditions and increasing life expectancy. Ultimately, understanding the complex interplay between these various factors is critical to predicting the future size of the population (Heuveline et al., 2017).

In many parts of the world, particularly in developing nations, there exists a significant need for more data related to population growth. This lack of information can pose a significant challenge to accurately tracking current population trends, further complicating efforts to forecast future population growth.

2.3.2 Population Growth Forecasting

Bardsley et al., (2021) considered that forecasting population growth involves predicting a population's future size and growth rates, which can be a daunting task due to the complex nature of the various factors that can affect population growth. Some of these factors include fertility, mortality, migration, and economic conditions (Cai et al., 2022). To accurately forecast population growth, one must consider all these factors and their potential impact on the growth rates of the population. Therefore, there is a need of crucial comprehension of the various factors that can impact population growth to make accurate predictions for the future.

Suárez et al., (2022) conducted a study to forecast population growth of Colombia through a ML Algorithm by using CO_2 emissions as variables combined with other demographic variables to train and test their Artificial Neural Network (AAN) ML Algorithm and relying on 61 years dataset compiled by the World Bank between 1960 and 2020. Their model had 3 layers of input which had 9 neurons, which are represented by 9 input variables, the hidden layer consisting of 4 neurons that connect the input layer to the output layer which consist of 1 neuron that represents the selected output variable which is the total of population in a particular year. They got a MAE of 0.038, and their model accurately predicted the population growth of Colombia. However, the model is limited to using CO_2 emissions, which can not only be a category of factors to consistently predict a country's population growth or region.

Şahinarslan et al., (2021) carried out a study where they utilised several and different Machine Learning Algorithms and the Cohort Component Method to forecast the population of Turkey in

2017. The researchers used several Machine Learning Algorithms but conducted the training and testing on the Regression Algorithms, which include Extreme Gradient Boosting, and CatBoost and the other on which they completed the training and testing is the Advanced Time Series Algorithms which includes Holt-Winter, ARIMA, and Prophet Prediction. In the comparison of performances, they found out that the CatBoost and Extreme Gradient Boosting together are better than the other Algorithms and that the Algorithms of Machine Learning are more accurate compared to the demographic Cohort Component Method. In their study they used all the datasets to train and test the model, which to an advantage, brings about consistency in prediction and more accuracy. Still, the limitation is that it does not consider cultural differences and geographical locations.

Otoom et al., (2019) carried out study that compared performance of 17 different ML approaches, including base learners and ensemble learners, in projecting a country's human population growth rate. In their study, it was found out that random forest had the best predictive performance and was the most robust when dealing with features that are lacking. Overall, the study successfully demonstrated how machine learning techniques can be used to predict population growth rates even in situations where historical data and feature information are not available. Additionally, the study identified random forest as the best ML algorithm for predicting population growth rates. The limitation to their study is that not all the factors that affect population growth were not considered and that the dataset is limited.

2.4 Existing Techniques for Population Growth Forecasting

2.4.1 Machine Learning

ML is a method of solving complex problems that would be too expensive to tackle using algorithms developed solely by human programmers. Instead, machines are trained to "discover" their own algorithms without explicit human guidance. This approach is widely used by organisations such as IBM (International Business Machines) and has been a topic of interest in the field of computer science, as noted by Hung and Chen in 2022.

Back in 1952, Arthur Samuel was given honours for inventing the phrase "machine learning" as a result of his research on the game of checkers. Fast forward to 1962, and Robert Nealey, who considered himself a checkers master, played against a computer (IBM, 2023) but ended up losing.

In recent times, artificial neural networks that create new content have been producing better results than previous methods. Machine learning techniques have been implemented in diverse fields such as large language models, computer vision, speech recognition, email filtering, agriculture, and medicine. This is particularly useful in situations where developing algorithms for specific tasks would be too expensive. (Silver et al., 2016)

2.4.1.1 Supervised Learning

Supervised learning is a powerful tool in the field of artificial intelligence and machine learning. It is a subcategory that involves using labeled datasets to train algorithms to accurately classify data or predict outcomes. The process involves feeding input data into the model, which then adjusts its weights until the model has been appropriately fitted. This fitting is carefully monitored through the cross-validation process to ensure that the model is accurate and reliable. Supervised learning helps organisations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox (IBM, 2023; Kotsiantis & Zaharakis, 2022). Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, and support vector machine (SVM).

2.4.1.2 Unsupervised Learning

IBM, (2023) as well as Zhong, Zhang and Zhao, (2022) states that unsupervised learning, uses ML algorithms to analyse and cluster unlabelled datasets. The algorithms can automatically reveal patterns that are hidden or groupings of data from the datasets. This ability of the methods to reveal the similarities and differences in information make it ideal for exploratory data analysis, cross-selling strategies, customer segmentation, and image and pattern recognition (Kang, Gao, & Zhang, 2022). It's also used to reduce the number of features in a model through the process of dimensionality reduction. Aggarwal, (2015) claimed that PCA and SVD are two

common approaches for this; further on, the additional algorithms that are used in unsupervised learning are neural networks, k-means clustering, and probabilistic clustering methods.

2.4.1.3 Semi-Supervised Learning

Kumar, Sattigeri, and Fletcher, (2017) as well as IBM, (2023) claimed that semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labelled data set to guide classification and feature extraction from a larger, unlabelled data set. Semi-supervised learning can solve the problem of not having enough labelled data for a supervised learning algorithm. It also helps if it's too costly to label enough data.

Kumar, Sattigeri, and Fletcher carried out a study on the semi-supervised learning methods that use GANs, and their study concluded that the methods have proved to flourish empirically. Most of these methods use a shared discriminator/classifier which discriminates real examples from fake while also predicting the class label.

2.4.2 Classification Techniques

Classification is a data mining (ML) approach that is used to forecast group membership for data instances (IBM, 2023; Wang, Aggarwal & Liu, 2018). Classification is a cornerstone of machine learning and knowledge discovery, enabling us to make predictions and extract insights from data in a wide range of domains. Classification is a fundamental problem in machine learning and data mining, and it remains one of the most active areas of research.

2.4.2.1 Random Forest

Random forest is a ML algorithm that was founded around the 21st century by Leo Breiman and improved by Adele Cutler at around 2006. It integrates the results of multiple decision trees to reach a single conclusion, making it a popular choice for both classification and regression problems. Its ease of use and flexibility have made it widely adopted in the industry (IBM, 2023; Chen, Liu & Zhang, 2022).

Random forests are a powerful machine learning algorithm with three hyperparameters as illustrated in Figure 2.1 that can be tuned to improve performance: the number of trees, the maximum tree depth, and the number of features sampled at each split. (Li & Wang, 2022; Zhang, & Li, 2023). Figure 2.1 illustrates the Random Forest Algorithm.

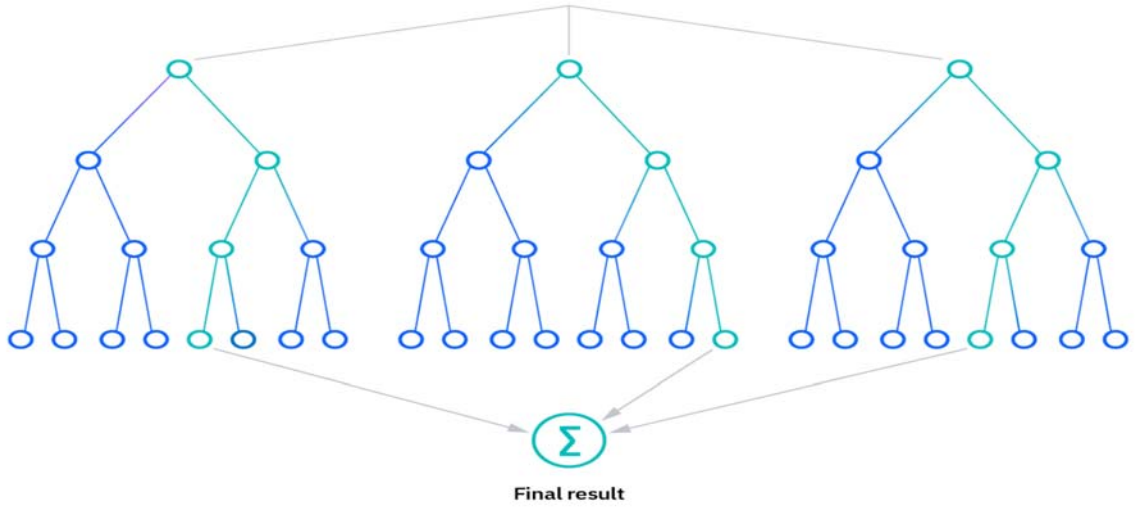


Figure 2.1: Random Forest Algorithm (IBM, 2023)

Liu, and Yang, (2023) study stated that random forests are a powerful ensemble learning algorithm that leverages bootstrap sampling and feature bagging to construct a diverse forest of decision trees that are robust to overfitting and can generalise well to unseen data. The out-of-bag (oob) sample is used to evaluate the performance of the forest and prevent overfitting.

The training algorithm for random forests applies the general technique of bootstrap aggregation, or bagging, to tree learners. Given a training set $X = x_1, .., x_n$ with responses $Y = y_1, \dots, y_n$, bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples:

For :$b = 1, \dots, B$:

a) Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.

b) Train a classification or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Additionally, an estimate of the uncertainty of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}\left(f_b(x') - \hat{f}\right)^2}{B-1}}$$

2.4.2.2 Logistic Regression

IBM, (2023) stated that this is a supervise ML algorithm that falls under a statistical model used in the determination of the likelihood of the occurrence of an event that is binary in nature for example, true or false, right or left and so on.

Since the odds of success are transformed using the logit function in logistic regression, the logarithm of the odds, base e, makes the odds easier to interpret and model. For example, a log odds ratio of 1.0 concludes that the probability of success is twice as high as the probability of failure. A log odds ratio of 2.0 concludes that the probability of success is four times as high as the probability of failure, and so on. Thus, by transforming the odds using the logit function, logistic regression models can be used to make predictions about the probability of success for a given set of independent variables and the formula used is:

$$p(x) = \frac{1}{1 + e^{-}(x - \mu) \,/\, s}$$

Where:

a) μ is a parameter of the location.

b) s is a parameter of the scale.

Which can be written and expressed as:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Where $\beta_0 = \frac{-\mu}{s}$ and is known as the intercept of the line $y = \beta_0 + \beta_1 x$, and $\beta_1 = \frac{1}{s}$ these are the y-intercept and slope of the log-odds as function of x; $\mu = \frac{-\beta_0}{\beta_1}$ and $s = \frac{1}{\beta_1}$ in the converse.

$p(x)$ is the variable that is dependent whereas x is the variable that is independent. β is the parameter which is estimated using the MLE and the MLE is used to find the values of the β that best fit the data through the calculation of the likelihood of the data for different values of β, and then choosing the β values that maximize the likelihood.

The log likelihood function is a measure of how well a model fits the data, therefore, the higher log likelihood function indicates a better fit. Logistic regression seeks to maximize the log likelihood function in order to find the best parameter estimates. When β is estimated and then used to calculate the predicted probability of success for each observation where for a binary classification, a predicted probability less than 0.5 will be classified as 0, and a predicted probability greater than 0.5 will be classified as 1, then the performance can be evaluated using a plethora of metrics.

VT OMNES VNVM SINT

2.4.2.3 Naïve Bayes

Naïve Bayes falls under the conditional probability model that assigns $p(Ck|x_1, \ldots, x_n)$ for each of the K possible outcomes or classes c_k given a problem instance to be classified, represented by a vector $x = (x_1, \ldots, x_n)$ encoding some features which is the independent variable.

The problem with the above formula is that if the number of n is larger, then basing such a model on probability tables is refutable. Therefore, the model must be reformulated to make it more tractable and using the Baye's theorem, the conditional probability can be as

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

The Naive Bayes classifier is a basic probabilistic classifier based on Bayes' theorem and is part of the Bayesian network models. It provides more accurate results when used with kernel density estimation (Piryonesi, Madeh El-Diraby, & Tamer, 2020; Li & Wang, 2022). Naive Bayes classifiers operate on the assumption that there is no correlation between two features in a class based on their presence or absence. These classifiers use a probability model that allows them to be trained efficiently in a supervised learning scenario. Typically, parameter estimation in Naive Bayes involves the use of maximum likelihood for increased precision.

2.4.2.4 Alternating Decision Tree (AD Tree)

Alternating Decision Tree is an ML technique used for classifying things since it is a broad form of decision trees, and it has some characteristics of boosting.

Zhang and Zhou, (2018) as well as Li and Song, (2020) stated that AD Tree has a structure that consists of the following nodes which are always in a sequence of alternation:

Decision nodes specify the condition of the predicate.

Prediction nodes just contain a single number.

The AD Tree functions as follows, it takes the path of all the decision nodes that are true and then gets the summation of the traversed decision nodes (Li and Wang, 2021).

2.4.2.5 Artificial Neural Networks

ANNs or SNNs, is a part of ML and the core of deep learning algorithms (IBM, 2023; Wang, Zhang, & Zhang, 2022). The name and structure of this model has been inspired by the human brain, imitating the way that biological neurons send signals to each other as Figure 2.2 showcases.

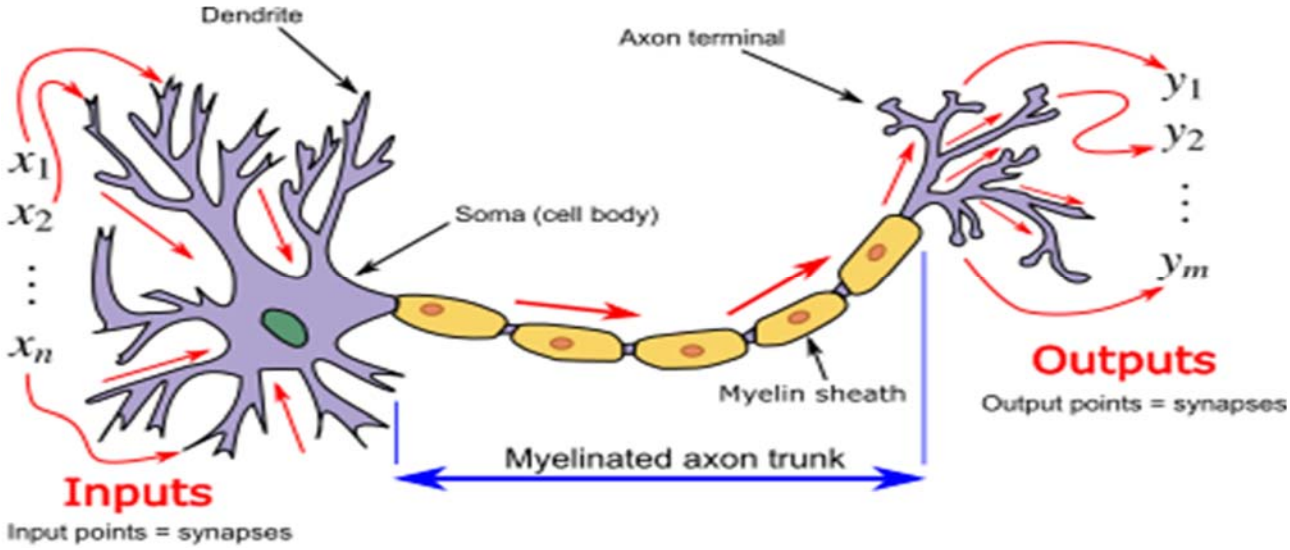


Figure 2.2: ANN inspired by human brain (IBM, 2023; Wang, Zhang, & Zhang, 2022)

Wang, Zhang, and Zhang, (2022) as well as Chen, Liu, and Zhang, (2022) carried out their study on ANNs which stated that it comprises of node layers such as:

a) Input layer.

b) Hidden layer(s).

c) Output layer.

The neurons are connected to each other in a network as Figure 2.3 shows; however, each neuron has an associated weight and threshold and if the weighted sum of the neuron's inputs is above the threshold, the neuron fires, passing its output to the next layer of neurons else, the neuron does not fire.

Li and Wang, (2022) claimed that neural networks is so powerful that it is used to classify and cluster data with the highest degree of accuracy and efficiency, however when it is trained on a large dataset, neural networks can be used to perform a variety of tasks, such as image recognition, speech recognition, and natural language processing. For example, Google's search

algorithm uses neural networks to rank billions of web pages in milliseconds. This allows Google to provide users with the most relevant search results quickly and efficiently. Neural networks are also used in many other cutting-edge AI applications, such as self-driving cars and facial recognition software. As AI continues to develop, neural networks will play an increasingly important role in our lives.

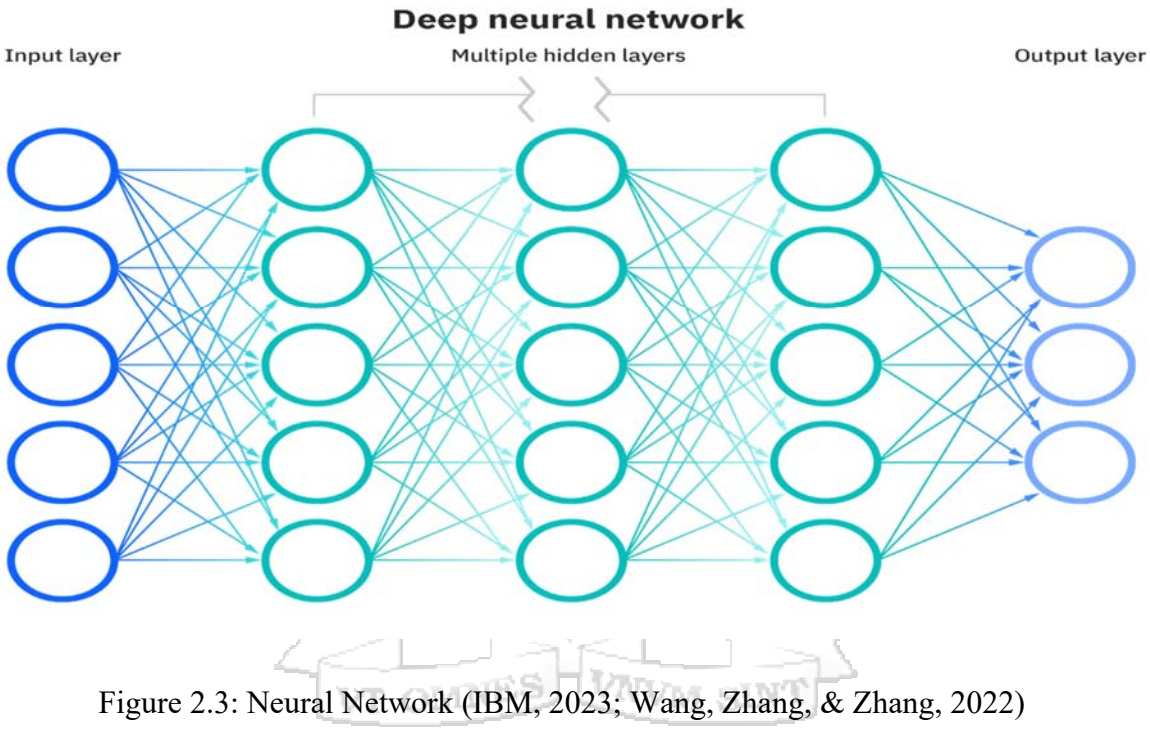


Figure 2.3: Neural Network (IBM, 2023; Wang, Zhang, & Zhang, 2022)

Think of each individual node as its own linear regression model, composed of input data, weights, a bias (or threshold), and an output. The formula would look something like this:

$$\sum_{i=1}^{m} W_i X_i + bias = W_1 X_1 + W_2 X_2 + W_3 X_3 + bias$$

$$output = f(x) = \begin{cases} 1 \; if \; W_1 x_1 + b \geq 0 \\ 0 \; if \; W_1 x_1 + b < 0 \end{cases}$$

Once an input layer is determined, weights are assigned. These weights help determine the importance of any given variable, with larger ones contributing more significantly to the output compared to other inputs (Chen, Liu, & Zhang, 2022). Each neuron in a feedforward neural network takes a weighted sum of its inputs and passes the result through an activation function. The output of the activation function is then sent to the next layer of neurons. This process repeats until the final layer of neurons produces the network's output.

2.4.2.6 Support Vector Machines

SVM algorithm is a reliable method for binary classification since it works well in conjunction with tree-based algorithms and basic logistic regression. However, it lacks the ability to interpret predictor variables accurately, which is a disadvantage (Wang, Zhang, & Zhang, 2022; Chen, Liu, & Zhang, 2022). Despite dealing with complex models, SVM is an essential linear ML technique applied to a feature of a high-dimension space that is non-linearly related to the input space, thus, enabling it to manage and maintain datasets of high-dimensions without sacrificing the simplicity of a linear algorithm (Zhang & Wang, 2023).

At the core of SVM is a non-linear feature transformation that maps the training data into a high-dimensional feature space, where a hyperplane is constructed to separate the classes with the largest possible margin to the support planes (Zhang & Zhang, 2022). The support planes are then iteratively moved apart until they reach the initial set of observations, known as support vectors (Li, & Wang, 2022).

Li and Wang, (2022) claimed that SVM leverages special kernel functions to compute the separating hyperplanes, without explicitly mapping the data into the feature space as illustrated in Figure 2.4. This allows SVM to perform all calculations directly in the input space, making it a computationally efficient algorithm.

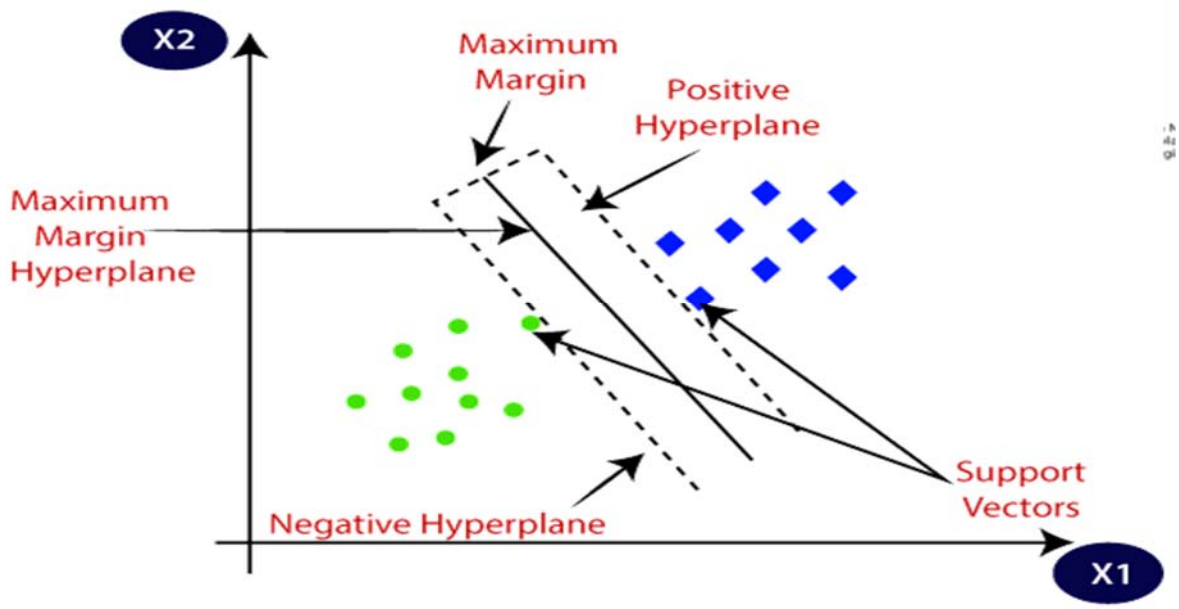


Figure 2.4: SVM (Li & Wang, 2022)

2.5 Models and Frameworks

2.5.1 Models

A machine learning model is a mathematical representation of the interaction between the features and the target or output variable. It is trained on a dataset that is labelled and thereafter used to make predictions on new data.

2.5.1.1 Hybrid Intelligence

Dellermann and Leimeister, (2023) clearly defined Hybrid Intelligence in their study as the approach that integrates both human intelligence and AI, and the basic rationale behind this is that the combination of complementary heterogeneous intelligences (i.e., human, and artificial agents) to create a socio-technological ensemble that can overcome the current limitations of AI. Dellermann and Leimeister, (2023) as well as Dellermann et al., (2019) also stated that this approach focuses neither on human intelligence in the loop of AI nor on automating simple tasks through machine learning. Rather, the emphasis lies on solving complex problems using the

deliberate allocation of tasks among different heterogeneous algorithmic and human agents. Both the human and the artificial agents of such systems can then co-evolve by learning and achieve a superior outcome on the system level (Dellermann & Leimeister, 2023). Figure 2.5 illustrates the Conjunction of Human Intelligence and Machine Intelligence and then Figure 2.6 shows the Hybrid Intelligence.



Figure 2.5: Conjunction of Human Intelligence and Machine Intelligence (Dellermann & Leimeister, 2023)

VT OMNES VNVM SINT

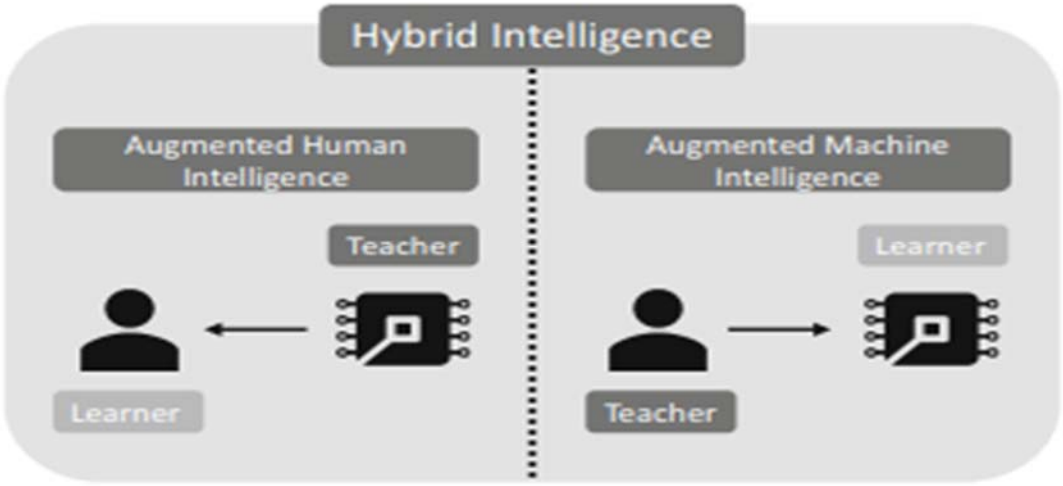


Figure 2.6: Hybrid Intelligence (Dellermann & Leimeister, 2023)

Hybrid Intelligence is a concept coined by Dellermann et al. (2019) to describe the ability to achieve complex goals by combining human and artificial intelligence. It is a socio-technical system in which humans and machines work together to learn from each other and achieve superior results than either could on their own. The following are the major key features to be considered:

a) Dellermann and Leimeister, (2023) as well as Dellermann et al., (2019) emphasized that one of the key features of Hybrid Intelligence is that it is collective, whereby, humans and machines must work together effectively to achieve the common goal. However, their goals may not always be perfectly aligned. For example, humans may be teaching an AI adversarial tactics in a game, even though the AI's goal is to win. This dynamic can create challenges, but it also leads to creativity and innovation.

b) Another key feature of Hybrid Intelligence is that it produces superior results. Hybrid Intelligent systems can achieve goals that would be impossible for either humans or machines to achieve alone (Dellermann & Leimeister, 2023; Dellermann et al., 2019).

For example, a Hybrid Intelligent system could be used to develop a new drug or design a more efficient transportation system.

c) Finally, Hybrid Intelligence is characterised by continuous learning. Over time, the system improves as a whole, and each individual component (both human and machine) also learns and improves (Dellermann & Leimeister, 2023). This is because humans and machines learn from each other through experience.

Hybrid Intelligence is a powerful concept with the potential to revolutionise many aspects of our lives as illustrated in Figure 2.7. As AI continues to develop, Hybrid Intelligence is likely to become increasingly important in a wide range of applications.

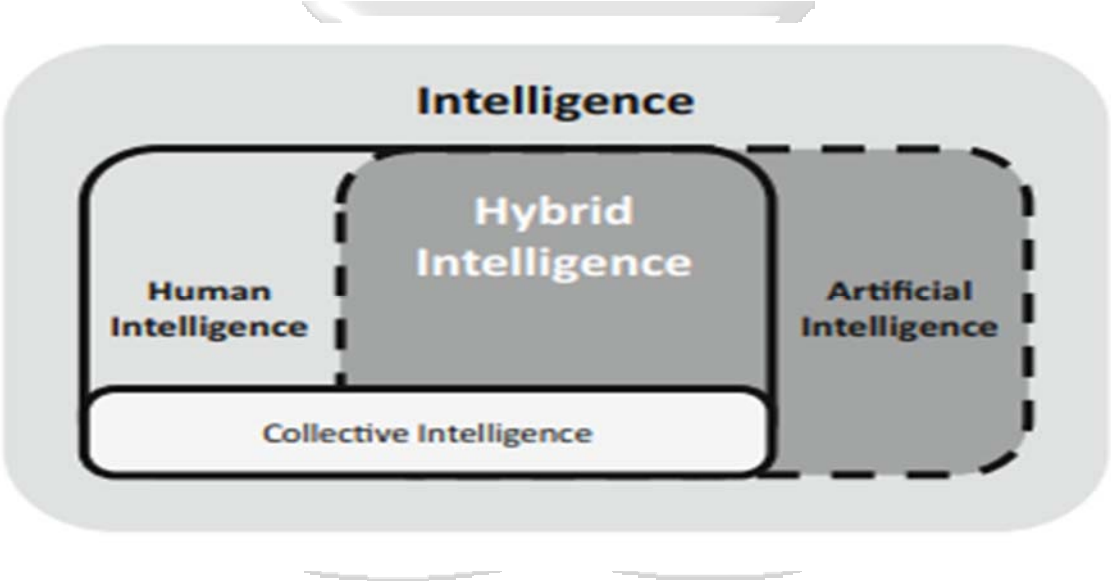


Figure 2.7: Concept of Hybrid Intelligence (Dellermann & Leimeister, 2023)

2.5.1.2 Predictive Models

A predictive model is a statistical model which forecasts the value of a target variable according to the values of a set of predictor variables. The target variable is the variable that we want to predict, and the predictor variables are the variables that we use to make the prediction.

In the Şahinarslan et al., (2021) study, population projection was made with both advanced time series and regression algorithms. According to the results, ensembling regression algorithms with the cohort component methods had very successful results in the prediction. Machine learning algorithms, especially ensemble regression models, can improve the accuracy of population estimates by reducing the impact of factors that make prediction difficult and by analysing uncertainties in demographic data.

2.5.2 Frameworks

A machine learning framework is a software library that provides the tools and infrastructure for developing and deploying machine learning models. It typically includes a set of algorithms, pre-trained models, and a development environment.

2.5.2.1 TensorFlow

TensorFlow is a computer software library for analysing data and ML used by companies and organisations such as Google, Airbnb, and the United States Military. It is originally developed by Google Brain team members Geoffrey Hinton, Andrew Ng, and Michael Nielsen in 2006, released around the year 2015 as an open-source library (Huang, Liu, Sun, Song, Wang & Chen, 2022; Gupta, Kumar & Das, 2022).

TensorFlow enables developers to build data flow graphs, which are architectures/blueprints that describe how data is transformed as it flows through a series of processing nodes as it is shown in Figure 2.8. Cui, Zhang, Yang, Wang, Li, Yi, and Tang (2017) stated that nodes in the data flow graph represent mathematical operations, while the edges represent the data arrays (tensors) that flow between them. This makes it easy to develop complex algorithms using TensorFlow because the graph can be visualised and debugged easily (Gupta, Kumar & Das, 2022).

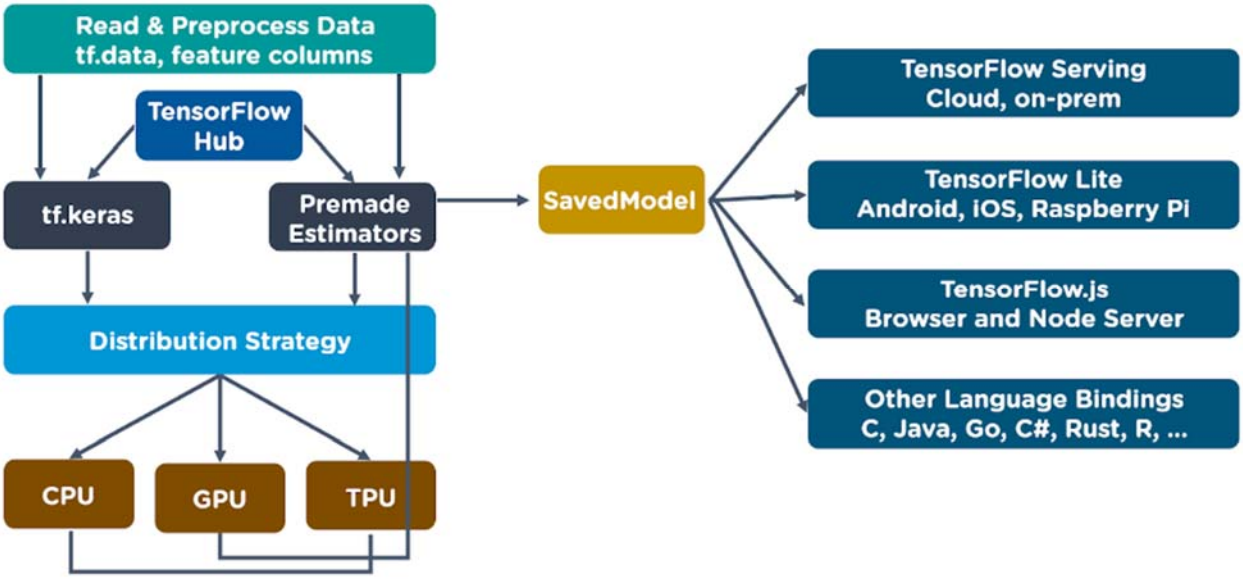


Figure 2.8: TensorFlow Framework Architecture (Li & Wang, 2023)

TensorFlow has been utilised in diverse applications for example, image recognition and classification, processing of natural language, and artificial intelligence. In recent years, it has become a popular library use for deep learning due to its flexibility and ease of use.

TensorFlow is based on the idea of creating a graph of operations, where each node in the graph represents an operation (Chen, Zhang & Li, 2022; Tian, 2023). The edges in the graph represent the data that flows between the operations. TensorFlow allows you to create and execute these graphs very efficiently, using a technique called dataflow programming.

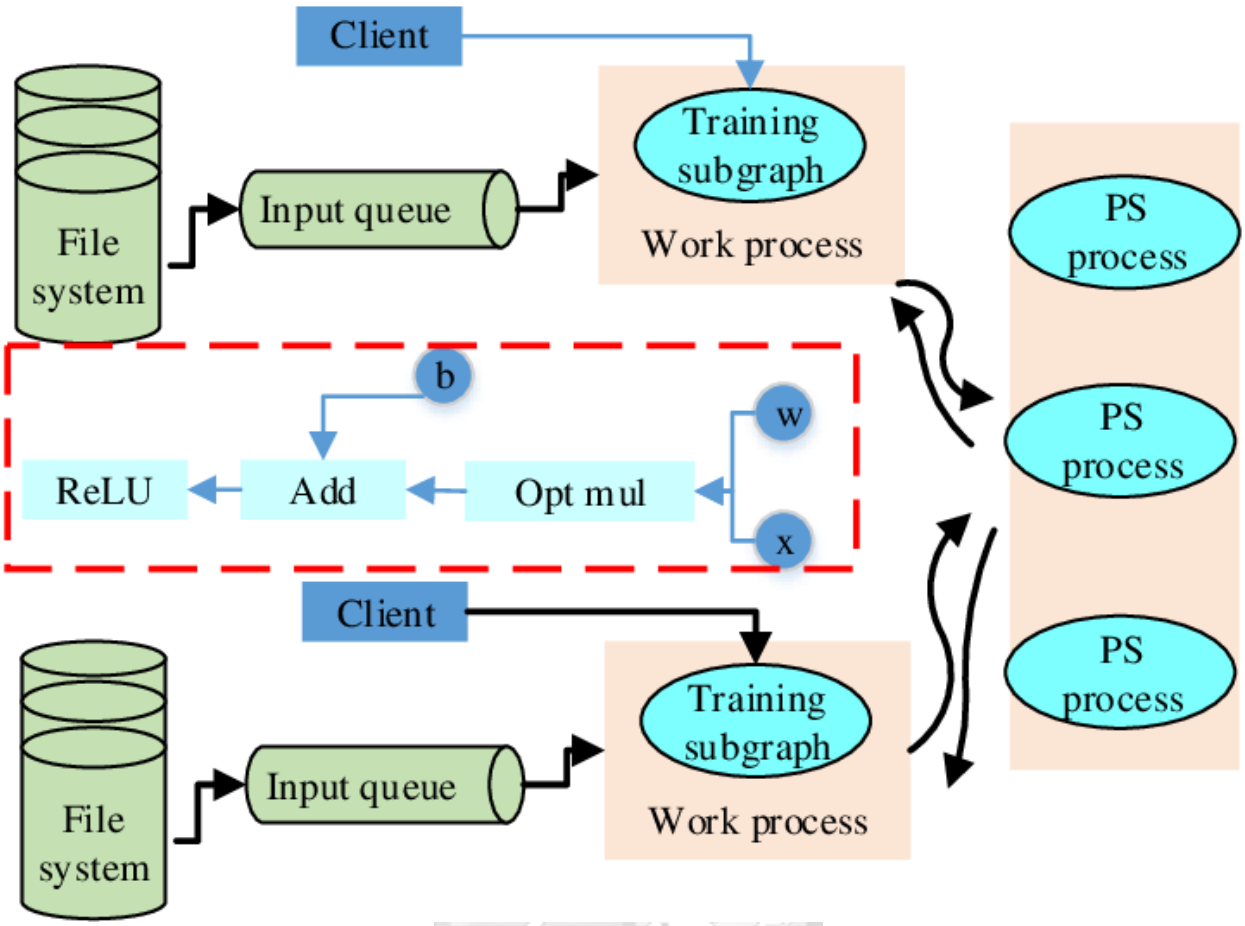


Figure 2.9: TensorFlow Data Flow Graph (Tian, 2023)

VT OMNES VNVM SINT

2.5.2.2 PyTorch

Pytorch is a powerful ML multiplatform library for use in Python enabling developers to perform efficient computation on GPUs, and it is also a deep learning framework for Python that is popular for its ease of use and flexibility (Chen, & Wang, 2022; Zhang, Shi & Chen, 2022). It was originally developed by Facebook AI Research and is now used by many companies and organisations, including Google, Microsoft, and Uber. Pytorch is open-source and available on GitHub (Liu & Wang, 2022).

Pytorch is a deep-learning framework that has been around since 2017. Due to its flexibility and ease of use, Pytorch is popular among researchers and developers and has undergone several major changes since its inception, including the addition of support for CUDA and cuDNN, and improvements to its autograd function. Figure 2.10 illustrates the PyTorch Framework Architecture.

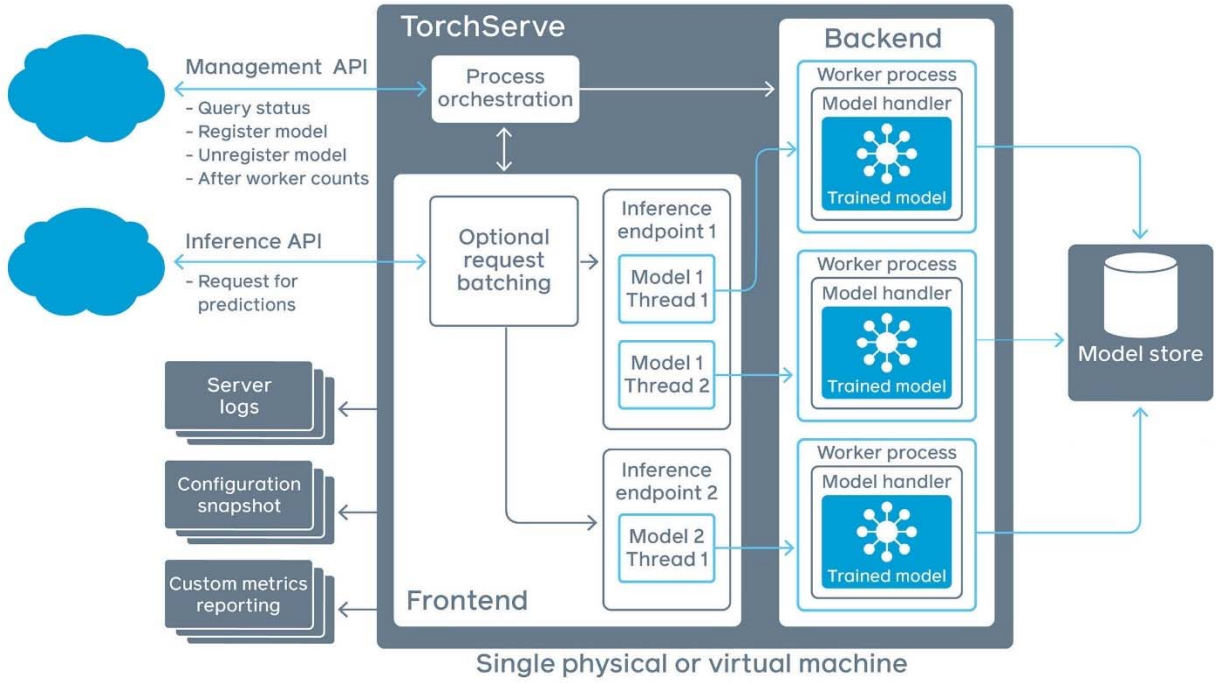


Figure 2.10: Pytorch Framework Architecture (Singh, 2023)

Pytorch has been adopted in a diversity of applications, for example the processing of natural language, computer vision, and reinforcement learning (Zhang, Chen & Li, 2022; Wang, Chen & Zhang, 2022). Pytorch is also widely used in academia, with many universities using it in their machine-learning courses as Figure 2.11 shows the PyTorch Workflow.

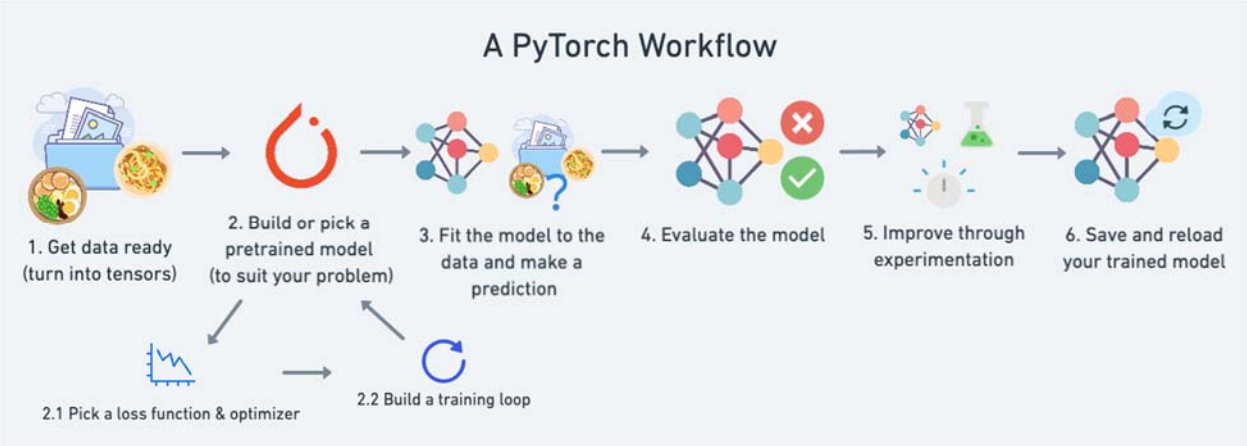


Figure 2.11: PyTorch Workflow (PyTorch Team, 2023)

2.5.2.3 Scikit-Learn

Scikit-learn which started as scikits.learn and sklearn is a Python programming language library of machine learning that is available to a large community at fairly no cost that was founded by David Cournapeau as he was part of the Google Summer program of coding. Scikit-Learn appears to highlight a diverse grouping/classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy as illustrated in Figure 2.12. Scikit-learn is a NumFOCUS fiscally sponsored project (Zhang, Chen & Li, 2022).

VT OMNES VNVM SINT

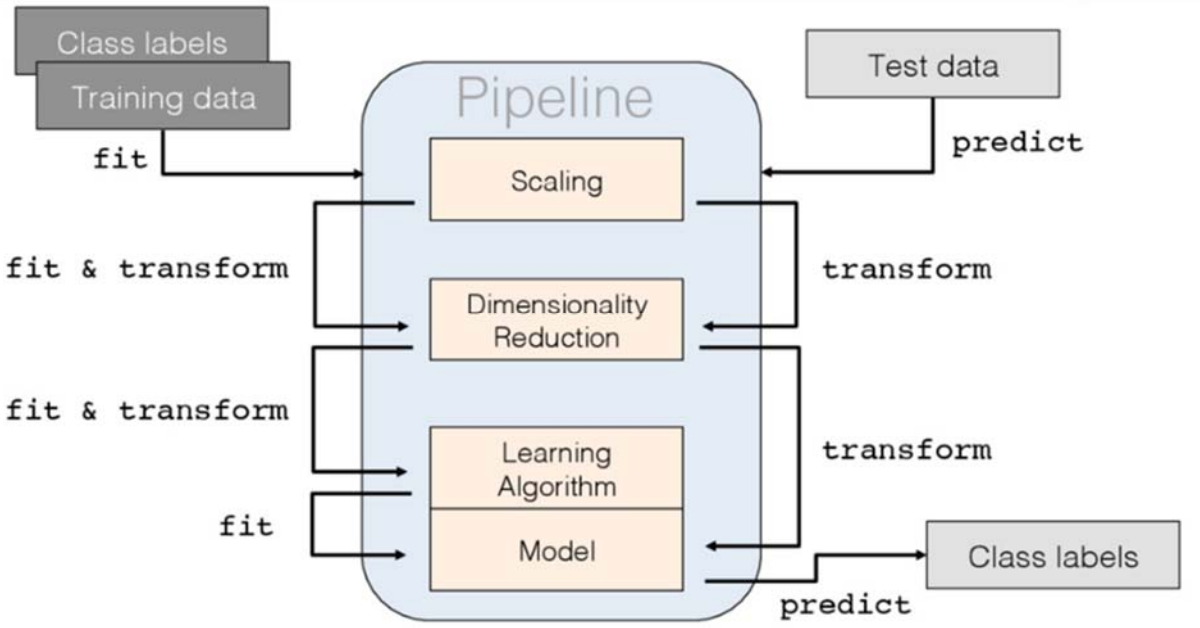


Figure 2.12: Scikit-Learn Pipeline (Scikit-Learn, 2023)

Scikit-learn integrates a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems; and focuses on bringing machine learning to non-specialists using a general-purpose high-level language (Ikudo, Lane, Staudt & Weinberg, 2018; Wang & Li, 2022). What Scikit-Learn emphasizes a lot is the capacity of ease of use, documentation of performances, and the consistency of the API. It has minimal dependencies and is distributed under the simplified BSD license, encouraging its use in both academic and commercial settings. Figure 2.13 shows the Scikit-Learn Algorithm Cheat-Sheet.

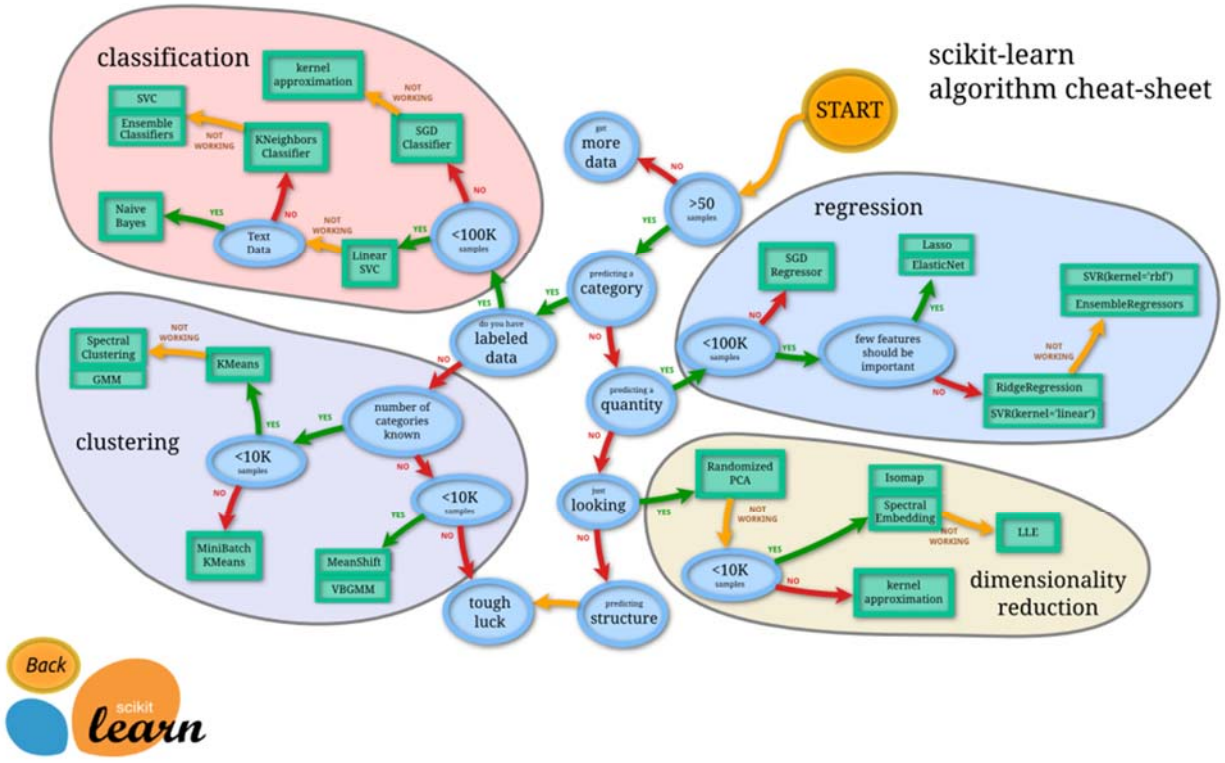


Figure 2.13: Scikit-Learn Algorithm Cheat-Sheet (Scikit-Learn, 2023)

2.5.2.4 Keras

Have you ever encountered Keras? A multiplatform library of neural network operating on the TensorFlow machine learning platform and is based on Python. Keras operates as a high-level API and is compatible with both PyTorch and TensorFlow, providing a plethora of features for constructing and training neural networks. Its emphasis on swift experimentation has made it a popular option among developers, thanks to its user-friendly and efficient nature.

Keras previously sustained many backends, such as TensorFlow, Microsoft Cognitive Toolkit, Theano, and PlaidML. However, in version 2.4, only TensorFlow is supported (Gulati & Garg, 2022). But, in version 3.0 (including its preview version, Keras Core), Keras will once again become multi-backend, supporting TensorFlow, JAX, and PyTorch. Keras is designed to allow for quick experimentation with deep neural networks while also being user-friendly, modular, and extensible as Figure 2.14 illustrates. It was developed as part of the research effort for

project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System), and its primary author and maintainer is not mentioned.

François Chollet, a Google engineer. Chollet is also the author of the Xception deep neural network model (Zhang, Chen & Li, 2022).

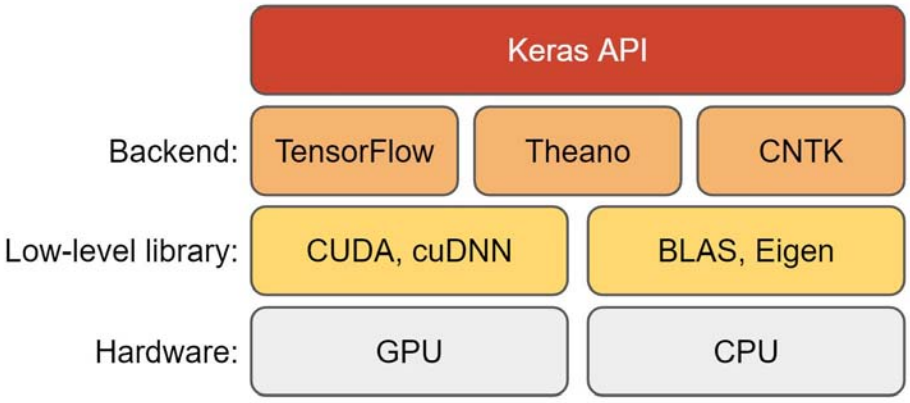


Figure 2.14: Keras Framework Architecture (Hymel, 2020)

VT OMNES VNVM SINT

2.5.2.5 NumPy

According to experts such as Wang, Zhang, and Huang, (2020) as well as Harris et al., (2020) NumPy is one of the most essential Python libraries created by Travis Oliphant and released as open-source software, NumPy played a key role in introducing multidimensional arrays (ndarray) for data structuring. This innovation contributed significantly to optimizing numerical computation in the computing world. Oliphant, the creator, also emphasized that NumPy offers a vast collection of mathematical functions that allow users to operate on ndarrays, making calculations such as linear algebra, Fourier transforms, and random number generation highly efficient. In essence, Gupta, Kumar, and Das, (2022) as well as Harris et al., (2020) assert that NumPy serves as the foundation for several higher-level Python libraries, including Pandas and Scikit-Learn as illustrated in Figure 2.15. Its efficient arrays are directly usable in machine learning and deep learning models.

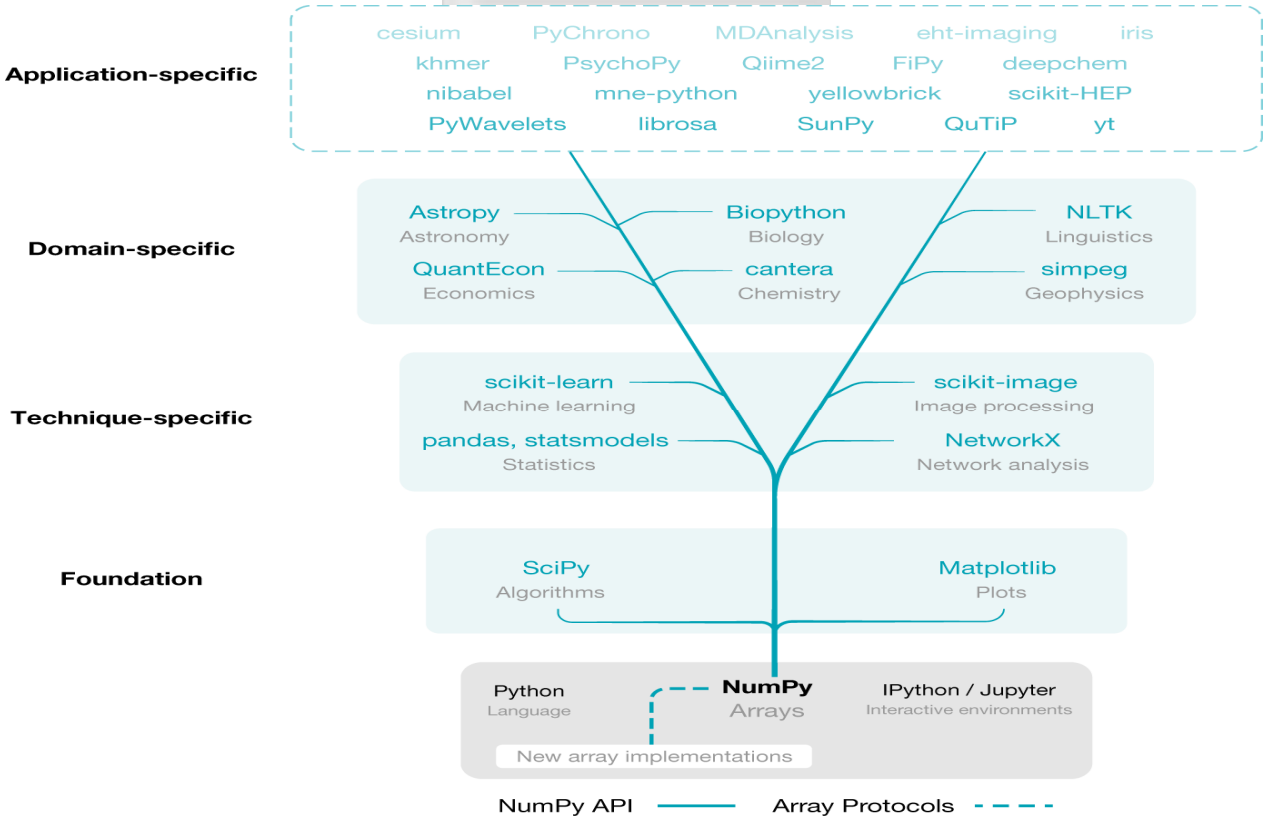


Figure 2.15: NumPy API (Harris et al., 2020)

2.5.2.6 Pandas

In recent years, a number of experts have hailed the Pandas library as a powerful tool for analysing and manipulating data, particularly large datasets. According to Kaur, Kumar, and Pandey (2021) as well as Reback et al., (2023), Pandas was created by Wes McKinney as a Python library with a flexible structure that allows users to work intuitively and efficiently with structured data, using tools like Series and DataFrames. Kaur, Kumar, and Pandey (2021) also stated that among its many capabilities, Pandas can load data from a variety of file sources (including CSV and Excel files), clean and preprocess data, perform statistical analysis and exploration, and even visualise data. As such, Pandas is an essential tool for handling data in machine learning and deep learning projects, often used in conjunction with other Python libraries like TensorFlow for feature engineering (Gupta, Kumar & Das, 2022). Figure 2.16 shows how tabular data is read and written using Pandas.

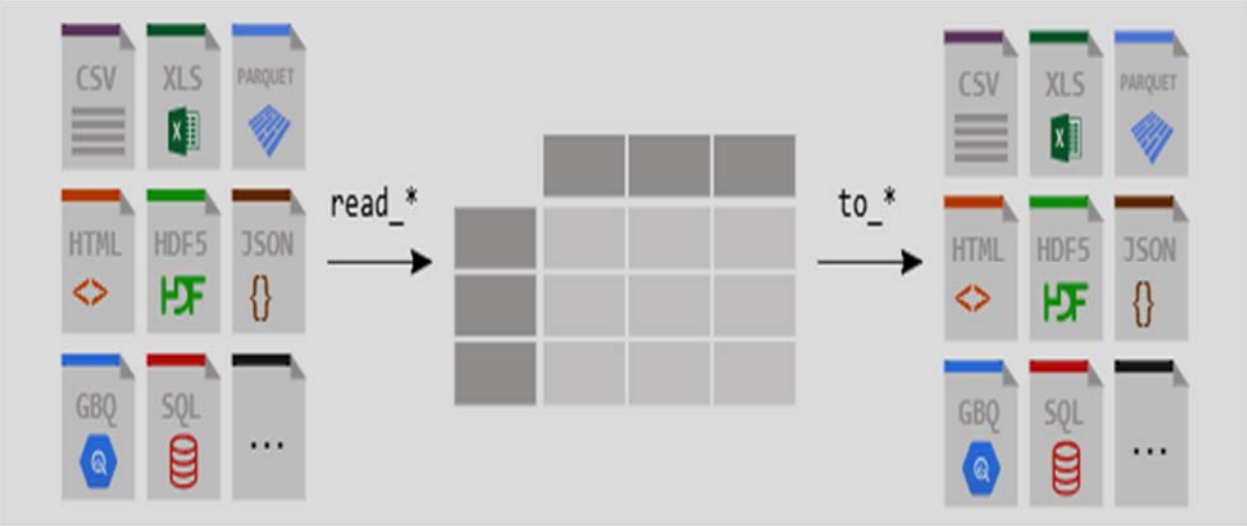


Figure 2.16: Reading and Writing Tabular Data (The pandas development team, 2020)

Figure 2.17 illustrates how plotting is done using Pandas.

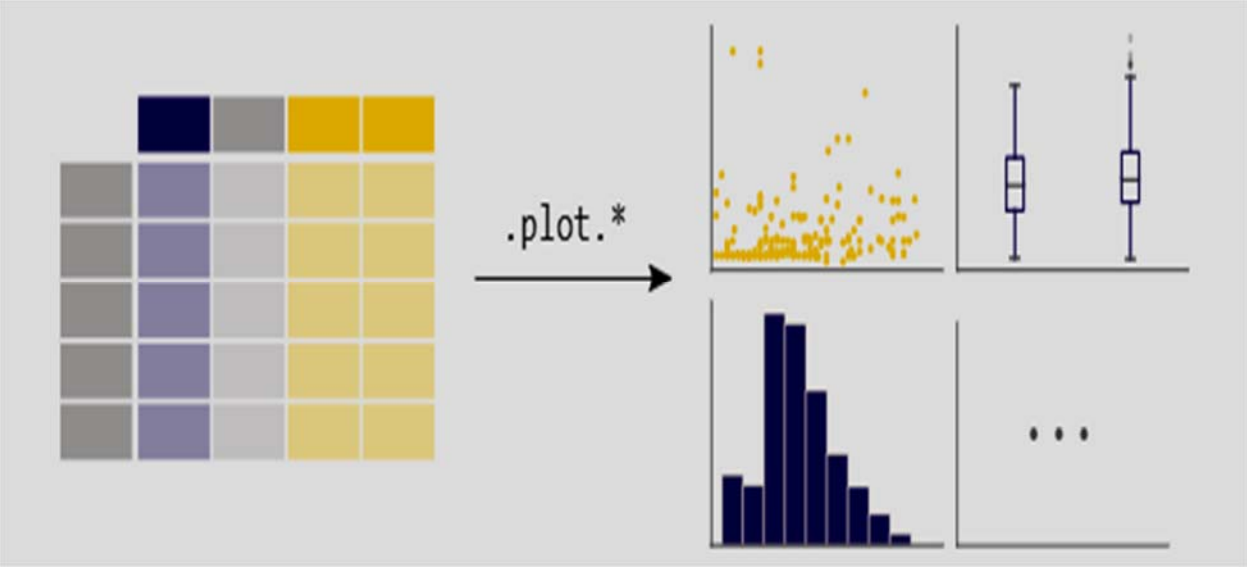


Figure 2.17: Creating Plots in Pandas (The pandas development team, 2020)

2.6 Limitations of the Current Techniques and Approaches Used for Population Growth Forecasting

While recent developments in predicting population growth are promising, a thorough examination of current methodologies has revealed various inadequacies. These limitations have hindered the efficacy and practicality of existing machine learning (ML) techniques in tackling this pressing issue. Therefore, it is imperative that we explore and implement more robust and effective approaches to address the challenge at hand.

The research conducted by Suárez and colleagues in 2022 utilised CO_2 emissions as an input variable. However, population growth is influenced by various factors beyond just CO_2 emissions, including economic conditions, social policies, and cultural norms. Additionally, the study worked with a limited dataset covering only 61 years, which may not provide enough information to accurately train a machine-learning model for predicting population growth.

In their 2021 study, Şahinarslan et al., utilised several machine-learning algorithms to analyse historical data. However, it is vital to note down the fact that the algorithms may not be as effective when applied to future data if the factors influencing population growth change. Additionally, the authors did not consider cultural differences or geographical locations, which could have impacted the accuracy of their predictions.

In a study conducted by Otoom et al., (2019), 17 machine-learning techniques were compared. However, there is vitality in noting that the dataset used was only from one country, which could potentially lead to different results if the study was conducted in a different country. Additionally, external factors like natural disasters or economic crises were not considered, which may have impacted the accuracy of the predictions.

2.7 Conceptual Model

To tackle the challenge of predicting population growth, the implementation of a comprehensive conceptual framework can be explored as illustrated by Figure 2.15. The following steps offer insight into this process:

a) Identifying the problem at hand: predicting the future population growth of a given country or region.

b) Developing a solution: utilising a machine learning model to forecast future population growth.

c) Selecting input variables: considering a range of factors, including fertility rate, mortality rate, migration rate, age structure, economic conditions, social policies, and cultural norms to inform predictions.

d) Constructing a machine learning model: training the model on historical data to establish a connection between input variables and future population growth, and subsequently using the model to predict population changes.

VT OMNES VNVM SINT

e) Gathering relevant data: sourcing information from a variety of outlets, including census data, vital statistics data, and surveys, to adequately train the model.

f) Employing the inference process: applying the inference process to make predictions based on the input variables and the machine learning model.

g) Generating a prediction: offering an output of the system, which forecasts future population growth.

Figure 2.15 depicts the conceptual framework.

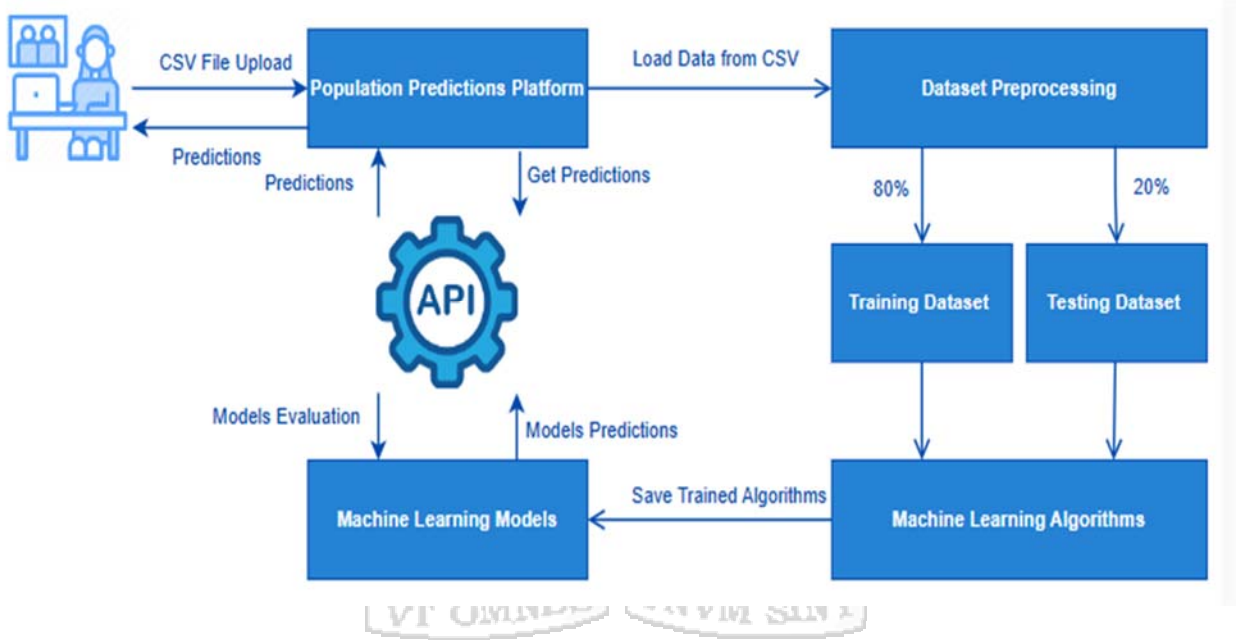


Figure 2.18: Conceptual Framework

Chapter 3: Research Methodology

3.1 Introduction

In this chapter, we present a discussion of the research design, methods, techniques, data collection, quality and ethical consideration of this study. It also covers the approach to be used to select the optimal machine learning model and the steps taken to establish a prediction model for population growth.

Rudestam and Newton (2020) defined research methodology as the process of planning, conducting, and evaluating research. They stated that it is a systematic way of thinking about and approaching research. Creswell and Poth (2021) discussed the importance of research methodology and stated that the choices adopted regarding research design, data collection, and the target population provide the researcher with legitimacy and perspective on the study's boundaries. Therefore, the use of methodology aligns with the validity of the research; thus, the researcher carefully considered these choices to conduct the research in a rigorous and credible manner.

3.2 Research Design

Creswell and Creswell (2022) defined research design as a plan for conducting a study so that all its parts work together logically and consistently to answer the research question. Thus, it can be used as a template for research and analysis and the research design should be flexible enough to allow for changes as the study progresses, but it should also be specific enough so that it can give a clear roadmap for the researcher.

Both the practical and descriptive research approaches are used in this study as Creswell and Creswell (2022) stated that both descriptive and practical methods of research can be used to answer research questions about population growth prediction. However, they also argued that the method of research choice has to depend upon the specific questions of research that the study is asking.

For example, if the research question is "What are the characteristics of the countries with the highest population growth rates?", then a descriptive research design is the appropriate approach. This is because the research question is asking to describe the characteristics of a population.

On the other hand, if the research question is "How can we reduce the population growth rate in developing countries?", then a practical research design is an appropriate approach. This is because the research question is asking to solve a practical problem.

Therefore, the descriptive research design can help in the collection of information about the characteristics of countries with the highest population growth rates. On the other hand, applied research is a scientific study that aims to answer practical problems. This was appropriate for the study since it seeks to solve the challenge of predicting the population growth rate in developing countries.

3.3 Population

The target population for this study is Kenya's total population dataset for the period 1960 to 2020 as contained in the World Bank Website. The country is a rapidly developing country having an equally high population growth rate.

World Bank, 2023 as well as Kenya National Bureau of Statistics, 2022 believed that the population of Kenya has increased by twice its original size in 2.5 decades, to around 40 million, and the fast growth of the population will not end soon. According to the up-to-date UN estimates, the population of Kenya will be growing by about a million per annum thus, around 3,000 people every day for the next 40 years and probably reaching the 85 million mark by the year 2050 (EAC Business Guide, 2015). Figure 3.1 shows Kenya's Total Population in 2022 according to the World Bank.

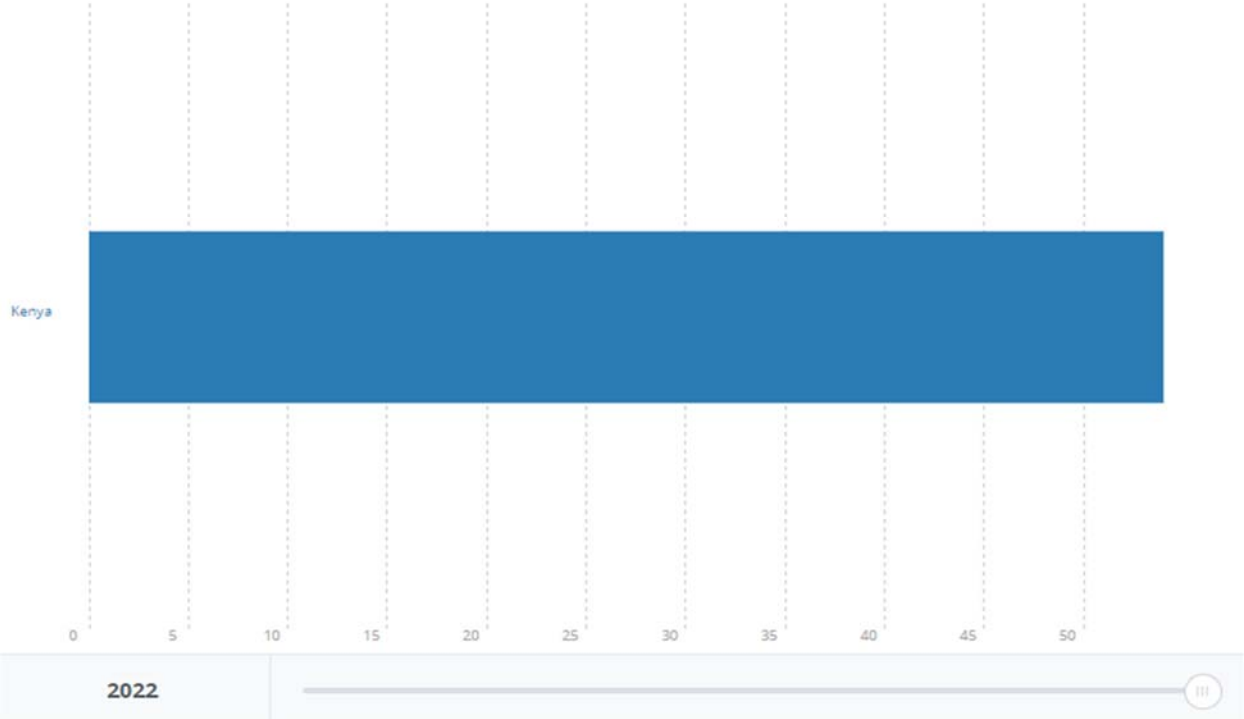


Figure 3.1: Kenya's total population 2022 (World Bank, 2023)

3.4 Sample Size

Creswell and Creswell (2023) defined sample size as the subset used to represent the entire population. Their study also stated that the sample size for a descriptive survey should be large enough to produce accurate results, but it should not be so large that it is impractical or expensive to collect data from the sample. For purposes of this study, due to the fact that this study seeks to develop a model to predict the population growth for Kenya, where a large-scale of data was needed, 100% of the whole dataset was utilised.

As Creswell and Creswell (2023) stated, there is usually a dataset for the model training, whereas the other is for testing in order to assess the model performance on any unseen data. It is important that the dataset size for training is large enough to grasp the patterns in the dataset. On

the other hand, the dataset size for testing is required to be huge in order to give valid estimations of the performance of the model.

The 80/20 rule was applied in this study. It is a common rule of thumb for separating datasets into training dataset then the other into testing dataset. This rule states that 80% of the dataset should be used for training and 20% of the data should be used for testing. To properly divide dataset, the 80/20 rule is a great starting point for a few reasons. Firstly, it allows training of the model to be on a large-scale dataset sample that can help identify patterns in the dataset. Additionally, it provides a dependable estimate of how the model performs on new, unseen data. Lastly, it is a simple and easy-to-follow rule. However, it is important to keep in mind that the 80/20 rule is not always the most effective approach for splitting data for there are others like the 70/30 rule, the 40/60 rule and even the 50/50 rule. The best split varies depending on the specific model and dataset used. Therefore, it is crucial for researchers to consider the aforementioned factors and seek guidance from statisticians to determine the optimal approach for their research.

3.5 Data Collection

This study used data obtained through an Application Programming Interface (API) from the World Bank website, a global financial institution that provides support to developing countries through loans, grants, and technical assistance. The dataset contains information on various social and economic indicators, including population growth, and covered the period from 1960 to 2022 (See: https://data.worldbank.org/country/kenya?view=chart). The dataset is open-source as provided under a Creative Commons Attribution 4.0 International License (CC BY 4.0) (See: https://www.worldbank.org/en/about/legal/terms-of-use-for-datasets).

The objective is to predict total population using fertility rate, mortality rate, and migration rate as independent variables. To achieve this, the dataset was separated to a training dataset and a testing dataset, with the former used for training the model and the latter for model evaluation and a linear regression model was employed for this study.

The model was able to accurately predict population growth, as demonstrated by the MSE, R2 Score, MAE, MSLE and MAPE metrics which were used for both the training dataset and testing

dataset. This study showcases the potential of the World Bank dataset to forecast population growth and provide insights for policy decisions related to population growth.

3.6 Population Growth Forecasting Model Development

3.6.1 Extraction of Data

To conduct the research, authorization to access the relevant data must be identified and obtained. The datasets in question, which is in .csv format, are owned by the World Bank and provides comprehensive information on population growth, fertility rates, and life expectancy for countries across the globe. World Bank's World Development Indicators database is the source of the dataset.

This study reveals that the ML model is exceptionally accurate in predicting population growth and significant implications for governments and organisations seeking to plan for future population growth.

3.6.2 Preprocessing Data

To ensure accurate predictions, thorough data preparation is essential. This study employed both the engineering of data and the engineering of feature techniques in order to carefully select the most relevant features. The data preprocessing process involved multiple stages, beginning with replacing any missing data and reducing noise. Smoothing and baseline reduction techniques were used in order to fill in any gaps. To ensure consistency, normalization was conducted to ensure all values are in the same units.

Once the data was analysed, appropriate transformations were made to conform to the required formats for analysis. Any categorical variables are transformed into numerical variables, and outliers are removed.

The final step in data preprocessing involved feature reduction. Various techniques, including correlation analysis and feature selection algorithms were used, in order to choose only features of relevancy for the prediction.

3.6.3 Selecting the Features

The Pearson Correlation Coefficient was used to remove features with high correlation from the dataset, reducing bias in this study. The formula is as bellow.

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma X \sigma Y}$$

Where σX is the standard deviation of variable X, σY is the standard deviation of variable Y , and $\text{cov}(X,Y)$ is the covariance denoted by the formula.

$$\text{cov}(X,Y) = \frac{1}{n}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$$

Also written as.

$$\text{cov}(X,Y) = \left[\frac{1}{n}\left(\sum_{i=1}^{n} X_i Y_i\right) - \bar{X}\bar{Y}\right]$$

Where n is the sample size, $\bar{X}$ is the mean of variable X , $\bar{Y}$ is the mean of the variable Y , and X_i, Y_i are the sample points at the index i.

Therefore, the formula for the Pearman's Correlation Co-efficiency that the study used is.

$$\rho_{XY} = \frac{\left[\frac{1}{n}(\sum_{i=1}^{n} X_i Y_i) - \bar{X}\bar{Y}\right]}{\sigma X \sigma Y}$$

In order to forecast the population growth in Kenya, this model considered and used input variables such as:

a) Fertility rates.

b) mortality rates.

c) life expectancy.

d) net migration.

e) economic growth.

f) access to healthcare.

g) access to education.

h) gender equality.

The dependent variable in this study is the total population, which is forecasted. The independent variables, as previously stated, are the input variables. There are 20 input variables and only one output variable. This study uses records from each year between 1960 and 2022. Due to the limited dataset, the entire dataset is utilised for the study.

3.6.4 Architecture of the model (ANN)

The ANN model consists of 20 neurons in the input layer, representing the 20 independent variables. The hidden layer plays a crucial role in ANNs, allowing the AAN to reveal and learn complex patterns in the dataset. The hidden layer then connects the input layer to the output layer as shown in Figure 3.2, which contains a single neuron representing the chosen output variable. In this case, the output variable is the total population to be predicted.

VT OMNES VNVM SINT

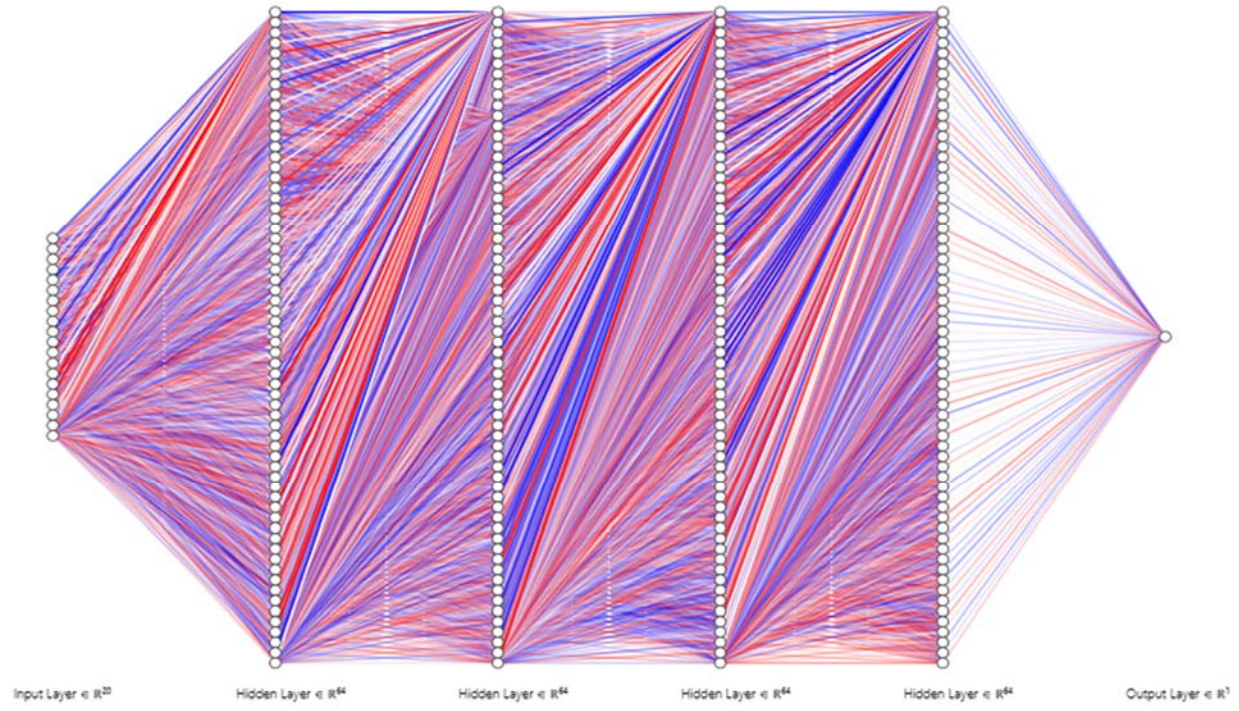


Figure 3.2: ANN Architecture

3.6.5 Population Growth Forecasting Model Validation

A population growth projection model is created, and its accuracy is gauged using the MSE, R2 Score, MAE, MSLE, and then the MAPE thus the quality assessment metrics used. The MSE, R2 Score, MAE, MSLE, and then the MAPE are able to measure the deviation between the actual values and the predicted ones, which acts as a risk function corresponding to the anticipated loss value (Otoom et al., 2019).

With the assumption that a vector has n generated predictions from a sample of cap Y data points on all variables, and cap Y hat is the vector of observed values of the variable being predicted, then the MSE, R2 Score, MAE, MSLE, and then the MAPE of the predictor within the sample are calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$R^2 = \frac{1 - \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\Sigma_{i=1}^{n}(Y_i - \bar{Y})}$$

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)$$

$$MSLE = \frac{1}{n}\Sigma_{i=1}^{n}\left(log(Y_i + 1)^2 - log\left(\hat{Y}_i + 1\right)^2\right)$$

$$MAPE = \frac{1}{n}\Sigma_{t=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{\hat{Y}_i}\right| \times 100\%$$

3.7 Research Utilisation

The results for this study can primarily be used for academic purposes. The study is part of a master's dissertation required to fulfil the academic requirements. Thereafter, the study results can be disseminated to the public, the study results can have a significant impact across many sectors that rely on the need or ability to forecast human population. This can enhance decision-making in sectors or areas such as:

a) Formulation of Policies and plans, for example the government can use the projections to come up with better approaches for resources allocation and build sustainable social infrastructures.

b) The business sector, especially investors, can use the ML projections to examine the market shifts of Kenya and make better investment decisions.

3.8 Systems Development Methodology

In developing the population growth prediction tool, the Agile Development Methodology (ADM) was used in combination with (OOM) Object Oriented Modeling. ADM involves different iterative and incremental methodologies, which includes the Lean Development,

Scrum, and Crystal. This approach of methodology was selected for this research since it is an iterative and a continuous one, enabling regular feedback between the stakeholders, the refinement, and the delivery of a better software system.

To ensure an approach that has a structure in analysing, designing, and implementing the system, the life cycle of OOM software was desegrated into the ADM. This cycle comprises phases such as gathering of requirements, analysing, designing of the system, implementing, and then testing the system. All the phases were able to secure the approach of software development that is analytical and thorough, which is in alignment with the quality of the ADM.

The process of developing or modeling software was done in a manner that is continuous and includes the following stages of planning, designing, developing, testing, reviewing, and launching sprints iterations. The approach gives room for the feedback and collaborations amongst the stakeholders involved in the software development from the executive level to the employees and then finally the customers which leads to the enhancements made in the life cycle of the software which was developed.

According to (ASD) Agile Software Development, there were more emphasis on personnel and their intercommunication over processes and tools, functionality of the software over the thorough documentation of the software, having an agreement with the customer over the negotiation of contracts and adaptation to change over resilience to an old way of doing something. Figure 3.3 illustrates the Agile Software Development Methodology.

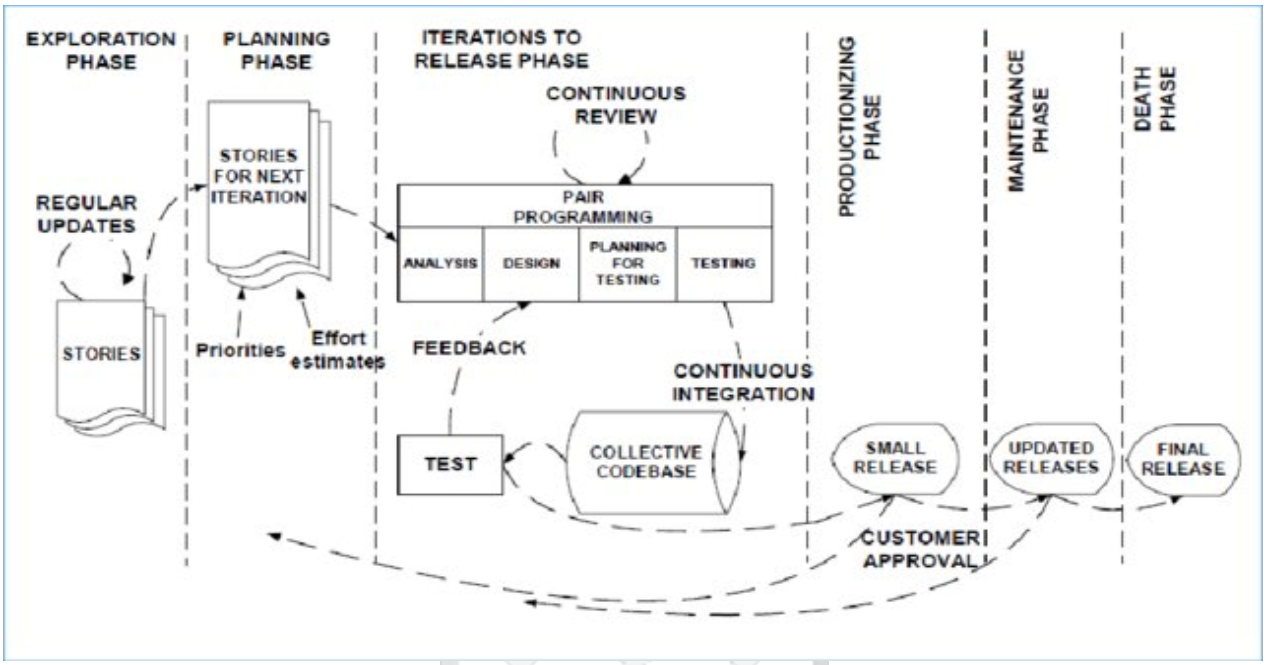


Figure 3.3: Agile Software Development Methodology (Alsaqqa, Sawalha & Abdel-Nabi, 2020)

3.9 Research Quality and Reliability

Frost (2022) discussed the concept of Cronbach's alpha as a way to measure reliability that evaluates the internal consistency of a test and his study stated that Cronbach's alpha is be used in the evaluation of the survey's reliability battery of questions. Frost discussed the concept of criterion validity as a type of validity that measures the extent to which a test score is related to a specific outcome. Thus, he stated that the validity of the criterion can be used to measure the validity of a test that is designed to predict a specific outcome.

In the context of this study, criterion validity was used to assess the validity of a test that is designed to predict population growth and the alpha test of Cronbach was used to measure how reliable the test is.

3.10 Ethical Considerations

Throughout this study, adherence to the ethical principles outlined by (APA, 2022) was followed. These principles prioritise preventing harm, respecting autonomy, promoting benefit, fairness, honesty, trustworthiness, and responsibility.

To ensure honesty and trustworthiness, the study sourced all literature from reputable sources and properly cite them to fulfil the moral and legal obligations as researchers, demonstrating responsibility. Additionally, the study shall seek Ethical Clerance from the Strathmore University Institutional Scientific and Ethical Review Committee (SU-ISERC) beforehand.

Above all, the study was conducted with the utmost ethical regard. However, it is important to note that the APA's Ethical Principles serve as a foundation, and researchers should also consider the ethical principles of other professional organisations, as well as the relevant laws and regulations governing their research.

VT OMNES VNVM SINT

Chapter 4: System Analysis and Design

4.1 Introduction

This chapter explores the system development journey through the lens of critical thinking and transformation. The focus is on system analysis and design, with the ultimate goal of translating requirements into a tangible, conceptual framework or architecture that guides system development. As such, the study delves into a meticulous assessment and analysis of the interface and interactions between users and system elements, utilising tools such as use case diagrams, sequence diagrams, and class diagrams.

4.2 System Requirement Analysis

The system requirements were meticulously gathered through thorough procedures and processes that involved extensive literature review. Multiple research studies and articles, such as those conducted by Şahinarslan et al., (2021), Otoom et al., (2019), and Suárez et al., (2022), were carefully studied. These sources provided invaluable insights into the functional requirements necessary for accurate forecasting of human population. In this study, the findings from these sources were utilised to ensure that the forecasting ML models consider all relevant features that affect population growth or decrease and possess the ability to meet the selected requirements in order to forecast human population in Kenya with precision.

During the course of this study, established tools and knowledge gathered from literature reviews was utilised to create unbiased and easily comprehensible procedures for gathering system requirements. For the goal is to ensure that the models designed should forecast human population growth in Kenya are highly efficient and effective, relying on the appropriate functional system requirements to provide accurate predictions.

4.2.1 Functional Requirements

The systems developed must meet specific functional requirements to enable users to seamlessly complete the following tasks:

a) Register with the platform.

b) Log in to the platform.

c) Access the platform using their account credentials.

d) Upload a CSV file.

e) View detailed predictions and plotted graphs.

f) Accurately predict human population trends.

4.2.2 Non-Functional Requirements

The nonfunctional requirements for the models developed entail their usability, accuracy, accessibility, and storage capabilities. Specifically, these models are expected to exhibit:

a) Simplicity and ease of interaction.

b) High precision in prediction-making.

c) Web-based accessibility.

d) Adequate storage infrastructure.

4.3 Systems Architecture

The system architecture includes a database to store user credentials, models developed, and model deployment on the web. The user will be prompted to create a CSV file with features and values, which they will then upload on the web to make predictions. To access the CSV file upload page, the user must first create an account through the signup page and log in. Once the CSV file is uploaded, the select models page is displayed. The saved model is loaded in order to receive the CSV file from the web, extract the values, and use them to make predictions. Figure 4.1 shows the Systems Architecture.

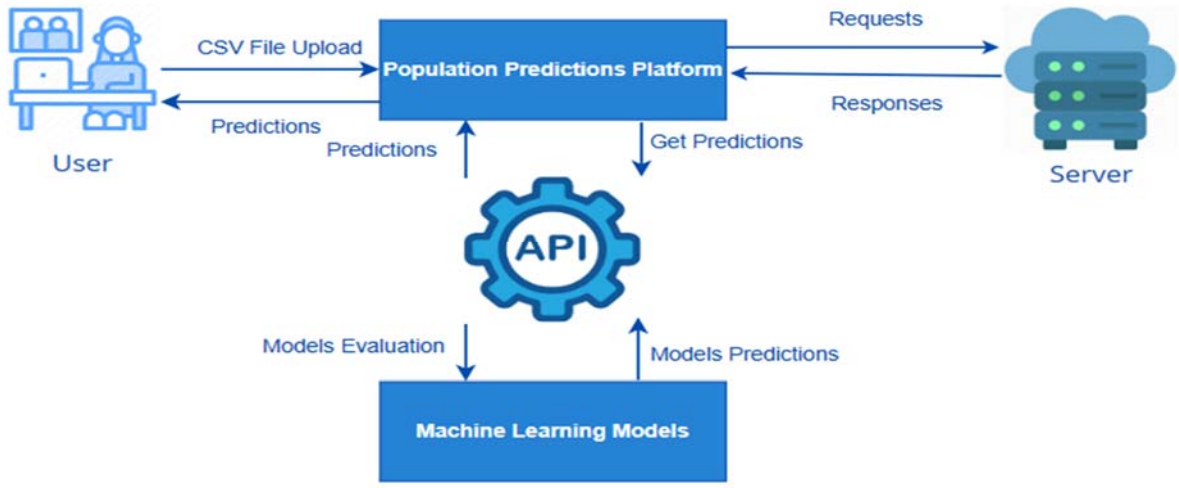


Figure 4.1: Systems Architecture

4.4 System Design

Undoubtedly, systems design is a fundamental aspect of information systems development, regardless of the system's purpose. It involves establishing the necessary data or information, designing the system's interfaces or interaction environment, creating the modules, components, and ultimately the system's structure. In this study, ADM and OOM were the integrated approaches used in systems software development, utilising OOP models to abstract the critical elements of the intended systems and reuse them to develop related systems. The models'

fundamentals and notation demonstrate the system creator's development process and have a significant impact on the system's outcome.

4.4.1 Use Case Diagram

This diagram depicts the interaction between the system and its users, with two distinct roles: the system user and the visitor. The visitor is limited to viewing the system's web display and navigating through the home, menu, signup, and login pages. On the other hand, the system user has the additional ability to create an account by signing up, logging in, and accessing the upload CSV file page to make predictions using a selected model as shown in Figure 4.2.

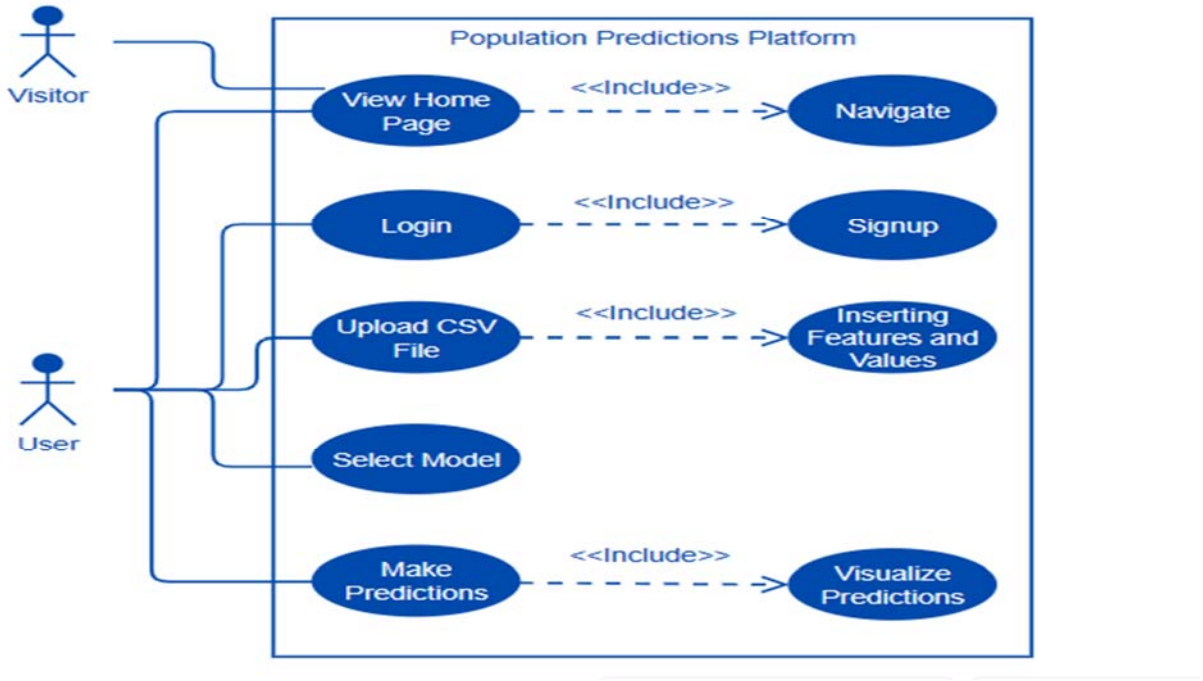


Figure 4.2: Use Case Diagram

4.4.2 Class Diagram

In general, UML encompasses class diagrams that are essential for providing the most comprehensive representation of a system's structure. These diagrams model classes, such as the user class, CSV data class, Models class, etc., along with their corresponding attributes (e.g., usernames, passwords), processes, and relationships between objects. Figure 4.3 showcases the class diagram.

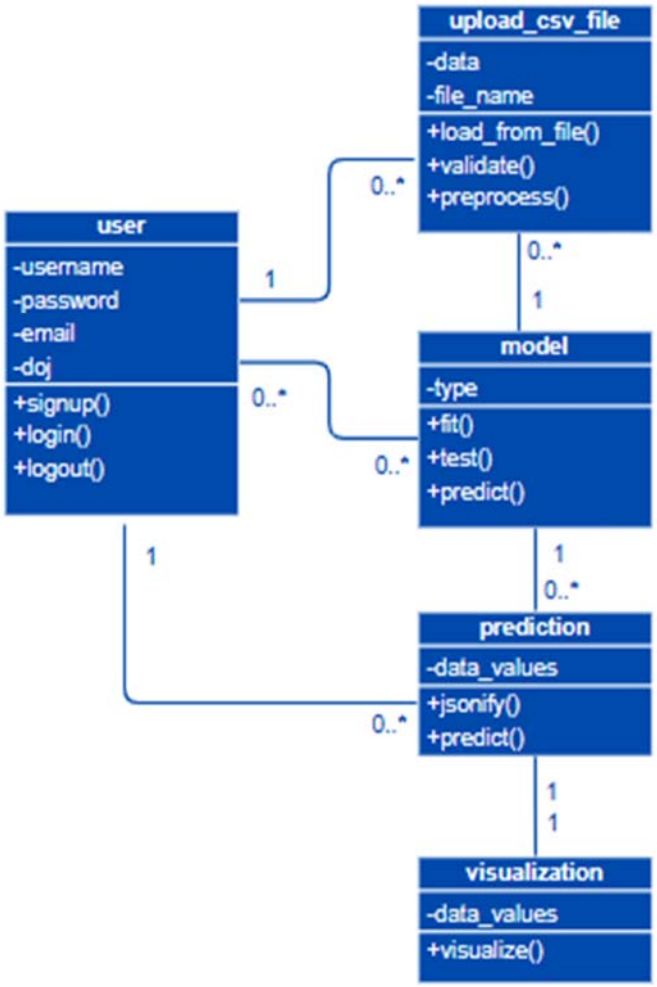


Figure 4.3: Class Diagram

4.4.3 Sequence Diagram

The UML also incorporates the Sequence Diagram, which provides a precise explanation for occurrences within the system. The platform used for population predictions comprises several objects, including users, models, and web deployment. For instance, a user would sign up and log into the system, after which the CSV file upload page would be displayed, allowing the user to upload a CSV file. This leads to the model selection page, where the user can select a Machine Learning model to use for predicting future outcomes. Finally, the predictions are displayed to the user without any indication that an AI-powered assistant is involved in the process. Figure 4.4 illustrates these objects and their communication with each other.

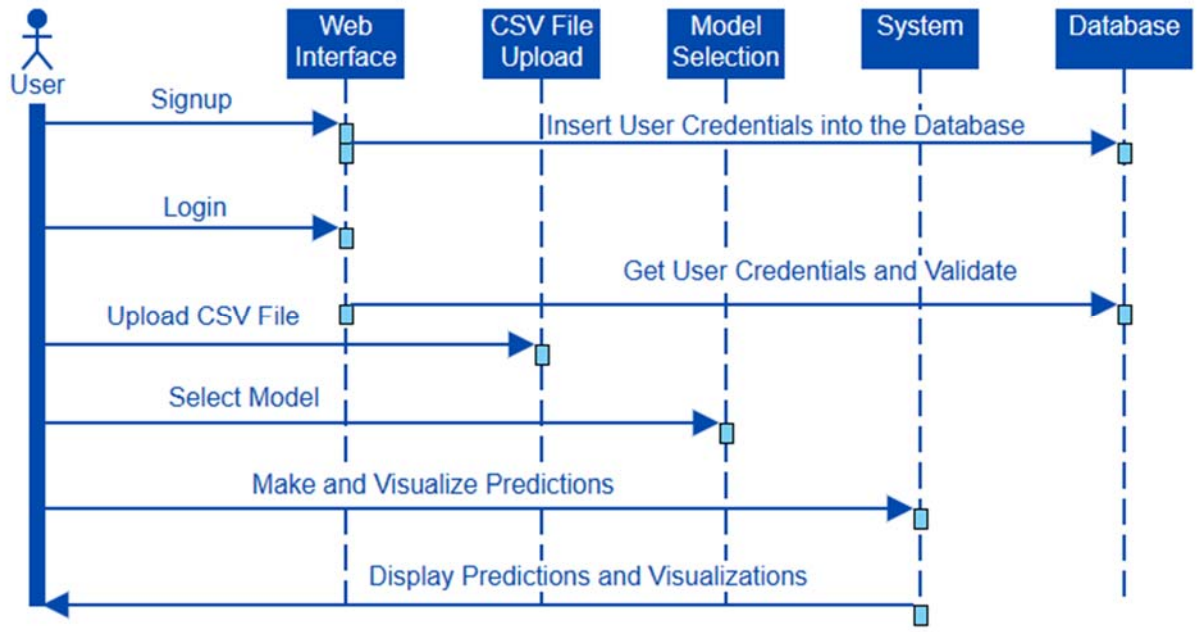


Figure 4.4: Sequence Diagram

Chapter 5: Systems Implementation and Testing

5.1 Introduction

This section serves as a dedicated focus on the development of a system through the rigorous processes of training, testing, and validation. The implementation of the system necessitates a thorough analysis of its individual modules, followed by a well-structured development process, and culminating in the seamless execution of operations.

5.2 ML Models Components

5.2.1 Artificial Neural Networks

The ANN model was designed using Scikit-Learn, specifically utilising the "MLPRegressor" class module. A class was created for the model, defining the input layer, hidden layers, and output layer with parameters such as the learning rate, random state, and number of epochs. As the default activation function for Scikit-Learn's MLPRegressor is ReLU, negative values are simply set to zero. The input layer's inputs are the input features fed into the model, with subsequent hidden layers' inputs being the outputs of the preceding hidden layer. The number of hidden layers is set to four, with the number of neurons in each layer defined at runtime in the "main.py" file. The trained model is saved and loaded for predictions in the "saved_models" folder by the "save_model.py" file. The output layer consists of a single neuron for "Population, total." Figure 5.1 depicts the class of ANN Model in the VS Code IDE using the Python programming language.

VT OMNES VNVM SINT

```python
# define the model class
class ANN_Model:
    # constructor
    def __init__(self, input_size, hidden_size, output_size, learning_rate, random_state):
        self.model = MLPRegressor(
            hidden_layer_sizes = (hidden_size, hidden_size, hidden_size, hidden_size),
            max_iter = 1000,
        )
        self.input_size = input_size
        self.output_size = output_size
        self.learning_rate = learning_rate
        self.random_state = random_state

    # method to train the model
    def train_model(self, X_train, y_train):
        # fit or train the model
        self.model.fit(X_train, y_train)

        # trained model predictions
        trained_model_predictions = self.model.predict(X_train)
```

Figure 5.1: ANN Class

5.2.2 Random Forest

The Random Forest model was constructed using Scikit-Learn's "RandomForestRegressor" class module, which was imported for use in a model class as depicted in Figure 5.2. The default number of decision trees within the forest was set to 100 estimators, with each tree contributing to prediction accuracy and robustness. The maximum depth allowed for each decision was not limited, allowing the trees to expand until the minimum sample size was reached in order to prevent overfitting. The random state was set to none, allowing for random bootstrapping and sampling of the training data, and feature selection was used to ensure consistent results. Multiple subsets of the parent training data set were defined using random sampling with replacement in the bootstrapping process. A decision tree was built for every subset at every node of every tree, with features selected using the random state to find the best split. Predictions were made using the new dataset or data points fed to every decision tree, resulting in an output that was the average of the predictions of every tree.

```python
# create a model class
class RF_Model:
    # constructor
    def __init__(self, n_estimators = 100, max_depth = None, random_state = None):
        # create a model object
        self.model = RandomForestRegressor(
            n_estimators = n_estimators, max_depth=max_depth, random_state = random_state
        )

    # method to train the model
    def train_model(self, X_train, y_train, learning_rate = None, epochs = None):
        # train the model
        self.learning_rate = learning_rate
        self.epochs = epochs
        self.model.fit(X_train, y_train)

        # make predictions
        trained_model_predictions = self.model.predict(X_train)
```

Figure 5.2: Random Forest Class

5.2.3 Logistic Regression

Logistic Regression is built using Scikit-Learn's "LogisticRegression" class module. The constructor includes several parameters, such as the penalty parameter, which specifies the type of regularization used. In this case, the default L2 regularization is applied to prevent overfitting by adding a penalty to the cost function. There is also an option for L1 regularization. The C parameter represents the inverse of the strength of the regularization, with smaller values indicating stronger regularization and better control over the model's complexity. The random state parameter allows for reproducibility across multiple runs. The Logistic Regression uses the Sigmoid function to transform input values into probabilities between 0 and 1, which are then used to separate the data set into different classifications. The boundaries of these decisions are used to classify new data sets. Figure 5.3 depicts the Logistic Regression class in the VS Code IDE using Python.

```python
# define the model class
class LG_Model:
    # define the constructor
    def __init__(self, penalty = "l2", C = 1.0, random_state = None):
        self.model = LogisticRegression(penalty = penalty, C = C, random_state = random_state)

    # define the method to train the model
    def train_model(self, X_train, y_train):
        # fit or train the model
        self.model.fit(X_train, y_train)

        # trained model predictions
        trained_model_predictions = self.model.predict(X_train)
```

Figure 5.3: Logistic Regression Class

5.2.4 Support Vector Machine

In the Support Vector Machine, the SVR class module from Scikit-Learn is imported and utilised within the model's class as illustrated in Figure 5.4. The constructor contains the kernel parameter, with the default being linear. This parameter defines the function for transforming input features into a higher-dimensional space for easier separation. The C parameter serves as the regularization parameter, controlling the trade-off between achieving low training error and maintaining a smooth decision boundary. A larger C results in lesser regularization. The epsilon parameter determines the size of the margin around the SVM hyperplane where there is no penalty association with the training points. This impacts the complexity of the decision boundary. Ultimately, the goal of the SVM is to find a hyperplane in the transformed feature space that maximizes the distance between the hyperplane and support vectors, allowing for separation of data points. The epsilon, or tolerance, parameter controls the degree of deviation and determines the most suitable hyperplane before incurring a penalty parameter. Finally, a new data set used as a data point is projected into the feature space, and the values of the predictions are obtained based on their relative position to the hyperplane.

```python
# create a class for the model
class SVM_Model:
    # constructor
    def __init__(self, kernel = "linear", C = 10, epsilon = 0.2):
        self.model = SVR(kernel = kernel, C = C, epsilon = epsilon)

    # method to train the model
    def train_model(self, X_train, y_train):
        self.model.fit(X_train, y_train)

        # train the model
        trained_model_prediction = self.model.predict(X_train)
```

Figure 5.4: SVM Class

5.2.5 Linear Regression

In Linear Regression, the LinearRegression class module from Scikit-Learn is imported and utilised. The constructor for the model's class as illustrated in Figure 5.5 contains no parameters that can be initialized in this scenario since linear regression relies on default settings for reliability. The key objective is to identify and describe the relationship between input features and the continuous output or target variable, using a best fit line or hyperplane in higher dimensions like Support Vector Machines. This best fit line is determined by minimizing the sum of squared errors between actual output and predictions made. New dataset projections are then visualised on the best fit line, and their predictions are determined based on their position relative to it. The linear relationship is assumed to be between input and output, with normally distributed errors and a constant variance.

```python
# create a model class
class LR_Model:
    # constructor
    def __init__(self):
        # create a model object
        self.model = LinearRegression()

    # method to train the model
    def train_model(self, X_train, y_train):
        # train the model
        self.model.fit(X_train, y_train)

        # trained model predictions
        trained_model_predictions = self.model.predict(X_train)
```

Figure 5.5: Linear Regression Class

5.2.6 Decision Tree

In the Decision Tree, Scikit-Learn's "DecisionTreeRegressor" class module was imported once again to use in the model's class. The constructor initializes the parameters of a decision tree, including the maximum depth parameter that controls the tree's depth. We used the default value of none, allowing the tree to expand until the leaves contain pure samples. This is similar to the Random Forest model, which also uses decision trees with maximum depths. However, deeper trees can model even the most complicated and complex relationships, making the model susceptible to overfitting issues. Additionally, the random state introduces randomness to allow for node splitting when selecting input features.

Starting at the root node, the dataset is recursively split based on input features that can separate the output suitably. The default Gini index is used to determine the feature that minimizes impurity. Finally, new data points are run through the tree while considering input values and splitting until the leaf nodes are reached. The average of the output becomes the prediction as shown in Figure 5.6.

```python
# define the model class
class DT_Model:
    # constructor
    def __init__(self, max_depth = None, random_state = None):
        self.model = DecisionTreeRegressor(
            max_depth = max_depth, random_state = random_state
        )

    # method to train the model
    def train_model(self, X_train, y_train):
        self.model.fit(X_train, y_train)

        trained_model_predictions = self.model.predict(X_train)
```

Figure 5.6: Decision Tree Class

5.2.7 K-Nearest Neighbors

The model's class utilises Scikit-Learn's "KNeighborsRegressor" module, which focuses on the "K" parameter - the number of closest neighbors taken into account when determining predictions for the new data set. When calculating distances, the default method used is the Euclidian distance, although Manhattan distance is also an option. The KNN algorithm calculates the distance between the new data points and all other points in the training data set, selecting the "K" data points closest to the new points as the nearest neighbors. Finally, the average output values of the nearest neighbors are used as the predictions for the new data set. Figure 5.7 shows the KNN class in VS Code IDE using Python.

```python
# create a class for the model
class KNN_Model:
    # constructor
    def __init__(self, n_neighbors = 5):
        self.model = KNeighborsRegressor(n_neighbors = n_neighbors)

    # method to train the model
    def train_model(self, X_train, y_train):
        # fit or train the model
        self.model.fit(X_train, y_train)

        # trained model predictions
        trained_model_predictions = self.model.predict(X_train)
```

Figure 5.7: KNN Class

5.3 Web Application Interface Components

The web application begins with a homepage that outlines its features, followed by a menu page that displays all available pages. Users can then sign up to become registered users, granting access to the CSV file upload page, or log in to gain access. Once on the CSV file upload page, users can input their CSV files and use the data to make predictions. The final page allows users to select and apply a model, visualise their predictions, and display the results.

5.3.1 Home Page

This page serves as the main hub of the population predictions platform, designed to engage users and help them become familiar with the system. Located in the top left corner of the page is a menu button, represented by three lines, which provides access to other pages. Additionally, a signup button is available for those who have not yet been authorized to use the system. Figure 5.8 shows the Home Page

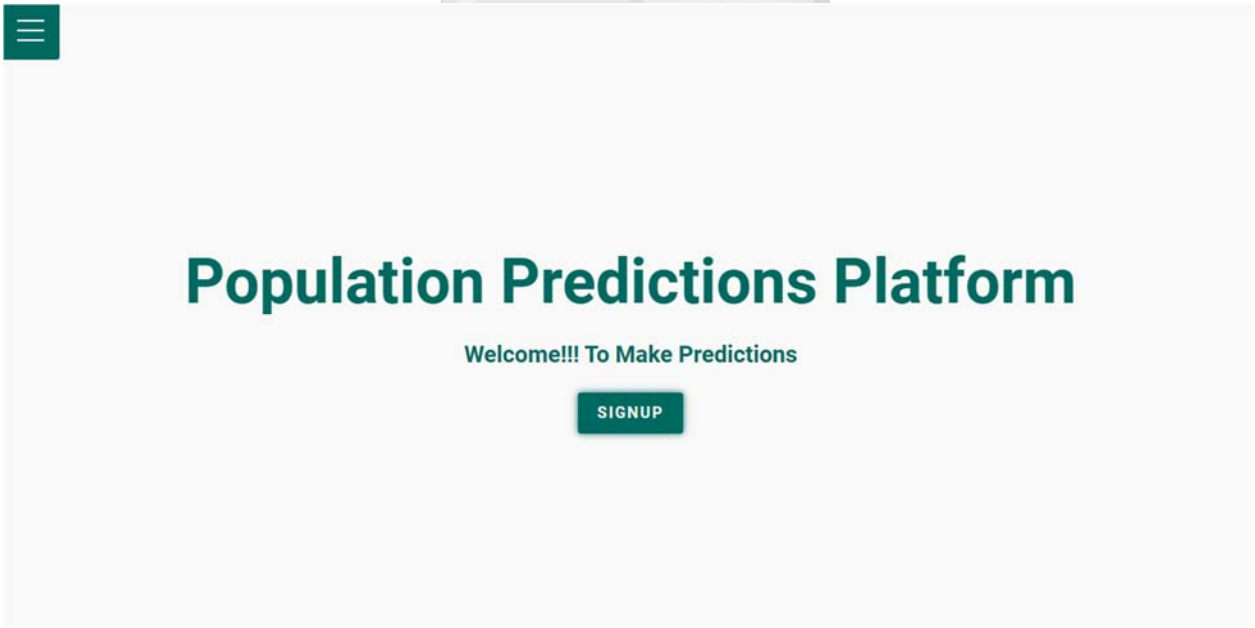


Figure 5.8: Home Page

5.3.2 Menu Page

This particular page serves as a navigational hub for both the system user and the visitor, allowing for seamless access to various other pages such as the “About” page, “Contact” page, “Services” page, “Home” page, “Signup” page, and “Login” page. Its purpose is to streamline the user's browsing experience and facilitate easy access to the desired information. Figure 5.9 shows the Menu Page.



Figure 5.9: Menu Page

5.3.3 Signup Page

This webpage provides the option for the population predictions platform visitors to register and gain access to the functional pages of the system, enabling them to make predictions. Once registered, users are directed to the login page to input their credentials and access the system. Figure 5.10 shows the Signup Page.

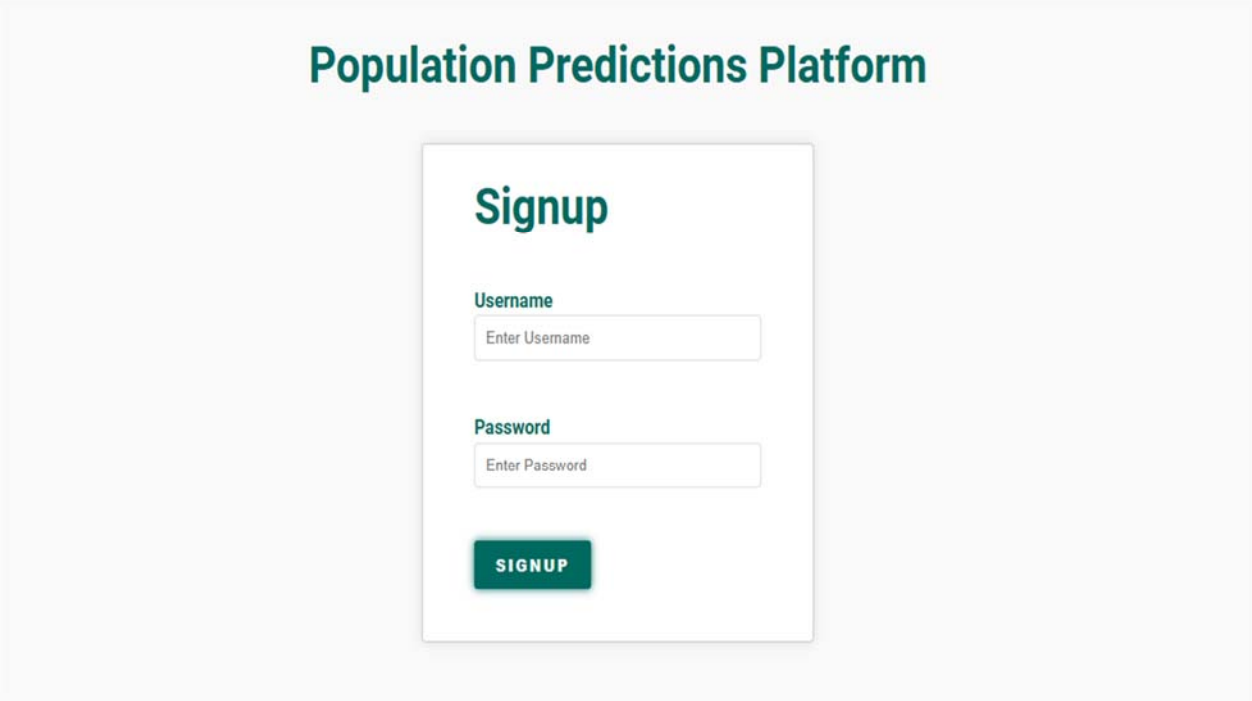


Figure 5.10: Signup Page

5.3.4 Login Page

This portal provides a login mechanism for authenticated or registered users of the population predictions platform to gain access to the system's core functionalities. One such feature is the CSV file upload page, where users can upload a CSV file to leverage its data for predictive modeling and analysis. Figure 5.11 shows the Login Page.

VT OMNES VNVM SINT

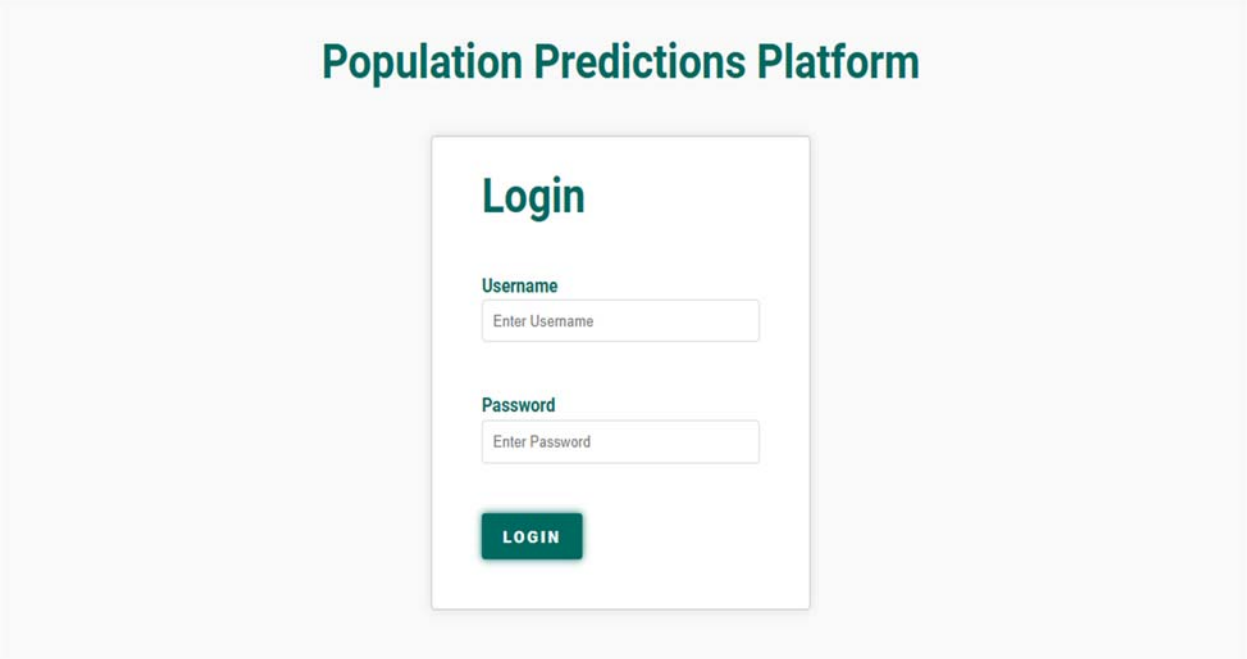


Figure 5.11: Login Page

5.3.5 CSV File Upload Page

Upon logging into the system, users of the population predictions platform are directed to the CSV file upload page where they can input new values for the system's features in order to generate accurate predictions. Figure 5.12 shows the CSV File Upload Page.

VT OMNES VNVM SINT

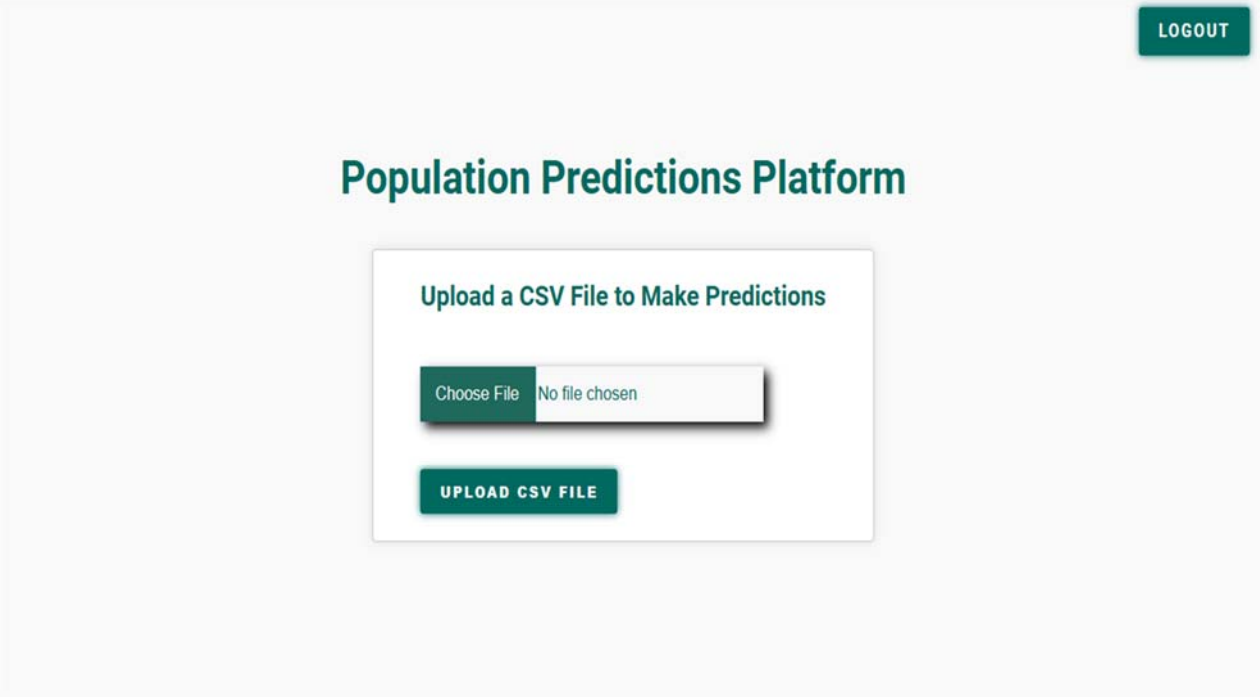


Figure 5.12: CSV File Upload Page

5.3.6 Model Selection Page

This page is exclusively accessible when the population predictions platform user uploads a CSV file containing the input features along with their respective values as illustrated in Figure 5.13. Once on this page, the user may choose a model to utilise for predicting outcomes upon clicking the "predict and visualise" button. Following this action, both the predictions and visualisations of these predictions will be presented.

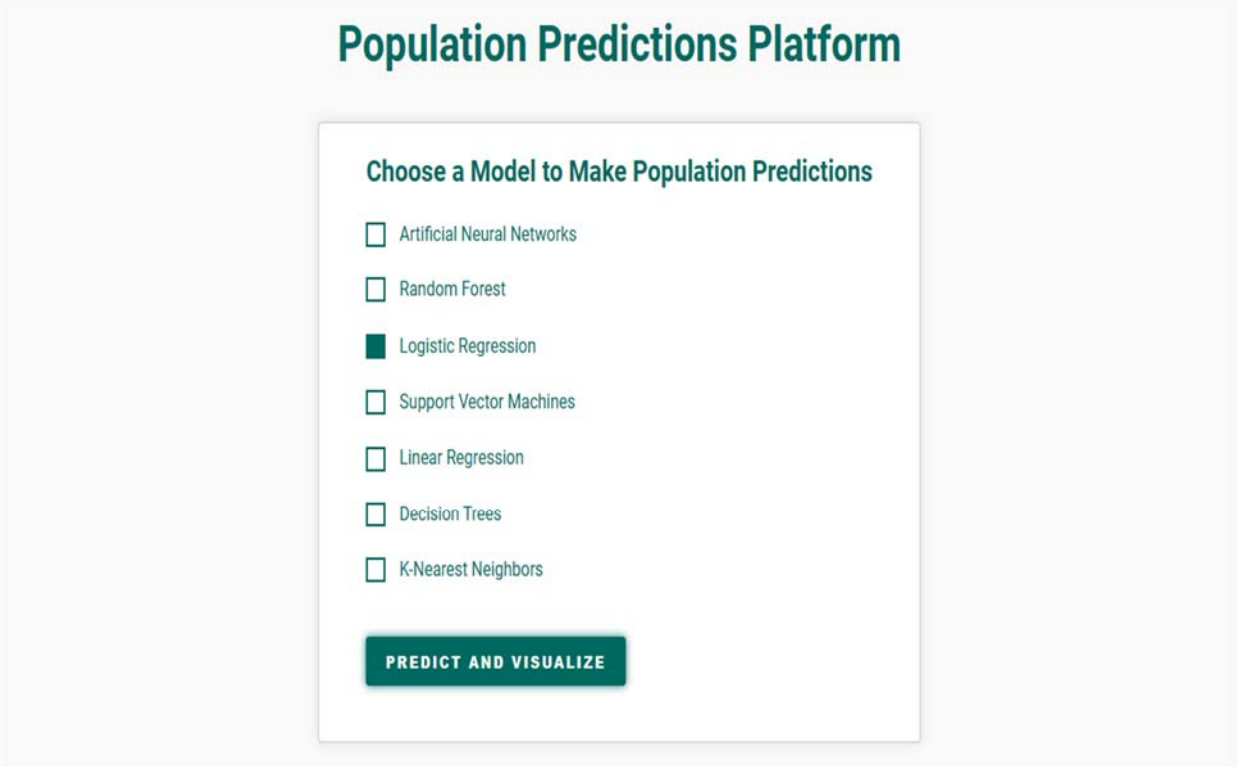


Figure 5.13: ML Model Selection Page

5.4 System Implementation

This study utilised an iterative development process that aligns with the integrated ADM and the OOM, accommodating evolving system requirements and continuous improvements and maintenance. As a result, the system is a robust solution that integrates multiple machine learning models to provide comprehensive analysis and insights.

The system's core relies on a suite of seven machine learning models tailored for various predictive and analytical tasks, including Artificial Neural Networks, Random Forest, Logistic Regression, Support Vector Machines, Linear Regression, Decision Trees, and K-Nearest Neighbor. These models leverage historical data, industry-specific factors, and other relevant inputs to assess different aspects of a target entity.

To create a seamless user experience, a web-based interface was built using the Flask framework. The interface guides users through structured data input, including essential features

such as fertility rates, mortality rates, life expectancy, net migration, economic growth, access to healthcare, access to education, and gender equality.

After a comprehensive analysis, the system presents predictions and insights and provides tailored recommendations aimed at improving the analysed entity's success potential. To handle the computational complexity of utilising multiple models, the system is designed with scalability in mind, whereby the saved trained or tested models are loaded to make predictions when the user selects the model to use. By combining diverse machine learning algorithms with a user-friendly interface, this system offers a valuable tool for data-driven decision-making.

5.4.1 System Development Environment

During the development of the population predictions platform, a set of tools and technologies were utilised to create the system. These included Python, the Windows Operating System, VS Code IDE, the Scikit-Learn and PyTorch libraries, the Flask framework, CSS, the Pandas and NumPy libraries, and the Matplotlib library for visualisation. Additionally, MySQL was used to manage the system's database.

5.4.2 Population Data Collection

The study focussed on the development of ML models that are capable of predicting the population of Kenya. To accomplish this, a comprehensive and diverse dataset was utilised, which contains historical population data that are vital to population growth and are relevantly correlated factors that affect population growth and the population in general. The source for the dataset is the World Bank's Open Data WDIs database, which provides publicly accessible data on various development indicators, including population figures. The data source is known for offering reliable and internationally consistent population data across numerous countries and over extended periods. However, the dataset used in the study is focussed on specifically Kenya, considering the socioeconomic indicators that contain socioeconomic factors known to influence population dynamics were included, for example, fertility rates, mortality rates, life expectancy, net migration, economic growth, access to healthcare, access to education, and gender equality.

These factors are incorporated to help the model identify complex relationships that drive population growth and/or decline.

The population dataset was carefully curated to include various input features, including Adolescent fertility rate, Fertility rate, total, Mortality rate, adult, female, Mortality rate, adult, male, Mortality rate, infant, Mortality rate, neonatal, Mortality rate, under-5, Net migration, GDP growth, GDP per capita growth, Life expectancy at birth, total, Government expenditure on education, total, School enrollment, primary, School enrollment, secondary, School enrollment, primary and secondary (gross), gender parity index (GPI), Current health expenditure, Current health expenditure per capita, Out-of-pocket expenditure, CPIA gender equality rating, and School enrollment, primary (gross), gender parity index (GPI). The output/target feature is the Population, total.

Overall, the study provided a valuable contribution to the field of population dynamics by developing ML models that can accurately predict the population of Kenya. The developed models can help researchers and policymakers to understand the complex relationships between various socioeconomic factors and population growth and can also be used for future population predictions and studies.

5.4.3 Population Data Preprocessing

After collecting the dataset on the population of Kenya, preprocessing was carried out. Missing data was handled using the imputation technique of mean. Outliers were either deleted or imputed using the mean. To ensure data consistency and proper formatting for population prediction, annual data was used throughout the entire dataset. The dataset was also standardized and normalized to prevent features with large numerical ranges from dominating the training processes. Figure 5.14 shows the Loading of data from a CSV file.

```python
# define the data preprocessor class
class Data_Preprocessor:
    # constructor
    def __init__(self, csv_file_path=None):
        self.csv_file_path = csv_file_path

    # method to preprocess the data
    def preprocess_data(self, df):
        # check if the CSV file path is provided
        if self.csv_file_path is None:
            raise ValueError("no CSV file path provided")

        # load the data
        df = pd.read_csv(self.csv_file_path, header=0)

        # drop unwanted columns
        df = df.drop(columns=["Country Name", "Country Code", "Series Code"])
```

Figure 5.14: Loading Data from the CSV File

5.4.4 Exploratory Data Analysis

Before developing and designing population prediction models for Kenya, it was essential to conduct thorough EDA on the population dataset. This allowed for a better understanding of the dataset's characteristics and trends, as well as uncovering potential relationships between input and target features. The following key areas were focused on:

a) Population trend analysis was visualised using the "Matplotlib" library to view changes over time, highlighting overall growth, periods of stagnation, or decline.

b) Correlation analysis was conducted to examine potential relationships between population total and socioeconomic factors such as GDP, net migration, and fertility rate.

c) Correlation matrices were used to quantify the strength and direction of relationships between variables.

d) Temporal patterns were analysed on both short-term and multi-year time scales to check for consistency in trends.

One important finding was a strong positive correlation between GDPs per capita and population growth rate in the Kenya dataset. Figure 5.15 shows some of the EDA Analysis carried out in the study.

```
# exploratory data analysis
def exploratory_data_analysis(
        self,
        df
):
    # print the first few rows
    print(df.head())

    # print the data types
    print(df.dtypes)

    # print the summary statistics
    print(df.describe())

    # print the missing values
    print(df.isnull().sum())

    # print the shape of the data frame
    print(df.shape)

    # print the column names
    print(df.columns)
```

Figure 5.15: EDA Analysis

5.4.5 Models' Training and Testing

During the training phase of the ML models, the "train_test_split" function from Scikit-Learn's "model_selection" was utilised to divide the data into 80% for training and 20% for testing. This ensured that the model was evaluated on new or unseen data. Feature scaling was also performed using Scikit-Learn's "StandardScaler" to normalize continuous features such as fertility rates and net migration. This prevented features with larger ranges from dominating the model training process, resulting in better performance.

Various models, including Artificial Neural Networks, Linear Regression, Random Forests, Support Vector Machines, Logistic Regression, K-Nearest Neighbors, and Decision Trees, were selected and trained with the Kenya population dataset. The models' performances were then compared on the test dataset, and the best performer was selected. It was found that during

training, the Linear Regression model designed with Scikit-Learn's "LinearRegression" was the best performer. However, during testing, Scikit-Learn's "MLPRegressor" was the best performing model.

Hyperparameter tuning was then conducted to further improve performance. However, some models, such as Support Vector Machines and Logistic Regression, failed to improve their performances and were among the worst performers compared to the other models. Figure 5.16 and Figure 5.17 shows the way the models were trained and tested inside the models' classes in the VS Code IDE using Python respectively.

```python
# method to train the model
def train_model(self, X_train, y_train):
    self.model.fit(X_train, y_train)

    # train the model
    trained_model_prediction = self.model.predict(X_train)

    # calculate the trained model metrics
    score = self.model.score(X_train, trained_model_prediction)
    mse = mean_squared_error(y_train, trained_model_prediction)
    r2 = r2_score(y_train, trained_model_prediction)
    mae = mean_absolute_error(y_train, trained_model_prediction)
    msle = mean_squared_log_error(y_train, trained_model_prediction)
    mape = mean_absolute_percentage_error(y_train, trained_model_prediction)

    # return the trained model and the trained model metrics
    return trained_model_prediction, score, mse, r2, mae, msle, mape
```

Figure 5.16: Training Models

```python
# method to test the model
def test_model(self, X_test, y_test):
    # test the model
    tested_model_prediction = self.model.predict(X_test)

    # calculate the tested model metrics
    score = self.model.score(X_test, tested_model_prediction)
    mse = mean_squared_error(y_test, tested_model_prediction)
    r2 = r2_score(y_test, tested_model_prediction)
    mae = mean_absolute_error(y_test, tested_model_prediction)
    msle = mean_squared_log_error(y_test, tested_model_prediction)
    mape = mean_absolute_percentage_error(y_test, tested_model_prediction)

    # return the tested model and the tested model metrics
    return tested_model_prediction, score, mse, r2, mae, msle, mape
```

Figure 5.17: Testing Models

5.4.6 Models' API and the Population Prediction Platform

The API features a "predict" function that requires a CSV file to be uploaded via the "upload_csv_file" function. The CSV data is then processed to extract relevant features such as year, fertility rate, life expectancy, GDP (representing economic growth), net migration, and more. These features are preprocessed using Scikit-Learn's StandardScaler, and predictions are made using loaded, saved, tested, or trained models with the "load" method from "joblib".

The API endpoint "/predict" is designed to receive the uploaded CSV file from "upload_csv_file" and generate a population prediction using the "predict" function. The result is returned in JSON format.

To access the population prediction platform system software, simply connect to the internet and use any web browser such as Google Chrome, Microsoft Edge, Bing, Mozilla Firefox, Opera, or others to access the web interface of the system.

5.5 System Testing, Validity and Usability

During the development of the system software, an integrated approach was taken to ensure the reliability and usability of the population prediction platform. This included rigorous testing and validation procedures such as unit tests on individual model components, integration tests on the entire data processing pipeline, and test suites for the web interface. Extensive testing was conducted using holdout validation sets and relevant metrics like MAE, RMSE, MSE, R2 Score, MSLE, and MAPE. The API was also tested for its robustness with varied input data and edge cases. UI validation was performed on CSV file uploads, model selection mechanisms, and prediction displays to ensure accuracy and user experience. Usability testing was conducted through user testing with representative samples of potential users, including demographers and fellow researchers. Ease of use tests were carried out to assess the intuitive nature of the web interface and the clarity of the input process. Finally, tests were conducted on the suitability and trustworthiness of the population prediction platform to ensure it meets the needs of users and provides trustworthy predictions. Figure 5.18 and Figure 5.19 shows the ANN's Trained and Tested Model Predictions Metrics respectively.

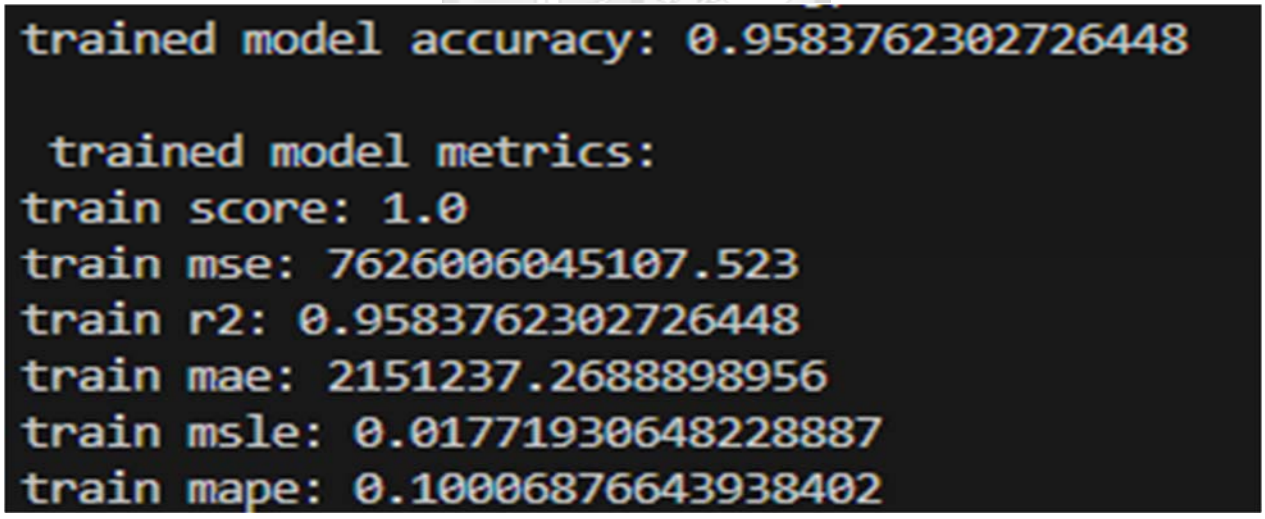


Figure 5.18: ANN Trained Model Predictions Metrics

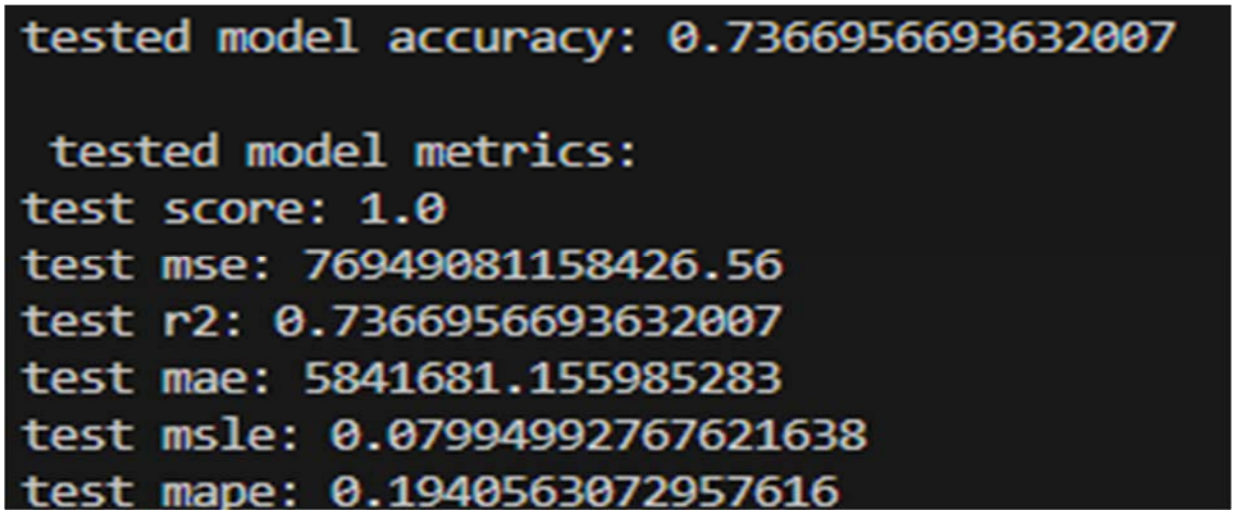


Figure 5.19: ANN Tested Model Predictions Metrics

Figure 5.20 and Figure 5.21 shows the Random Forest's Trained and Tested Model Predictions Metrics respectively.

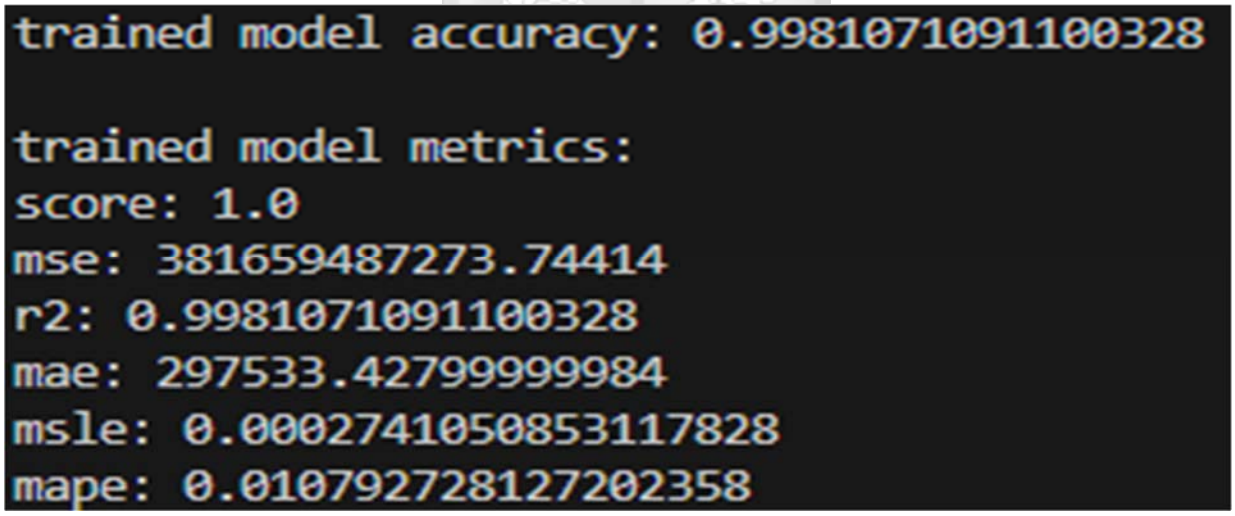


Figure 5.20: Random Forest Trained Model Predictions Metrics

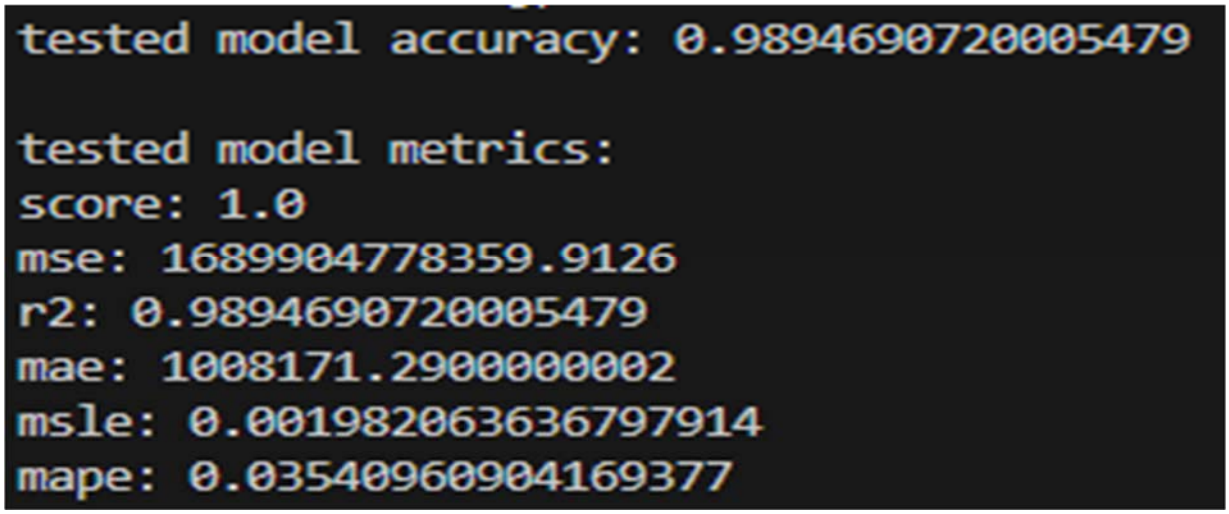


Figure 5.21: Random Forest Tested Model Predictions Metrics

Figure 5.22 and Figure 5.23 shows the Logistic Regression's Trained and Tested Model Predictions Metrics respectively.

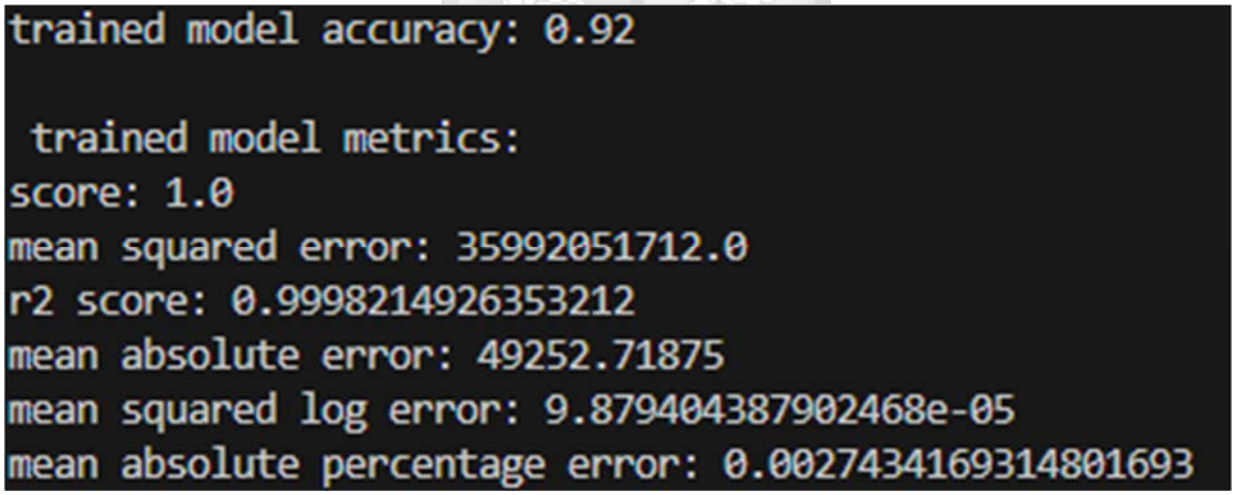


Figure 5.22: Logistic Regression Trained Model Predictions Metrics

```
tested model accuracy: 0.0

 tested model metrics:
score: 1.0
mean squared error: 4434139021312.0
r2 score: 0.9723679118244388
mean absolute error: 1790184.25
mean squared log error: 0.0052211113096922636
mean absolute percentage error: 0.06484783440828323
```

Figure 5.23: Logistic Regression Tested Model Predictions Metrics

Figure 5.24 and Figure 5.25 shows the SVM's Trained and Tested Model Predictions Metrics respectively.

```
trained model accuracy: -0.0618980424920372 3

trained model metrics:
score: 1.0
mse: 214108200627198.03
r2: -0.0618980424920372 3
mae: 12155390.883711858
msle: 0.3531229218204261
mape: 0.585128173145499
```

Figure 5.24: SVM Trained Model Predictions Metrics

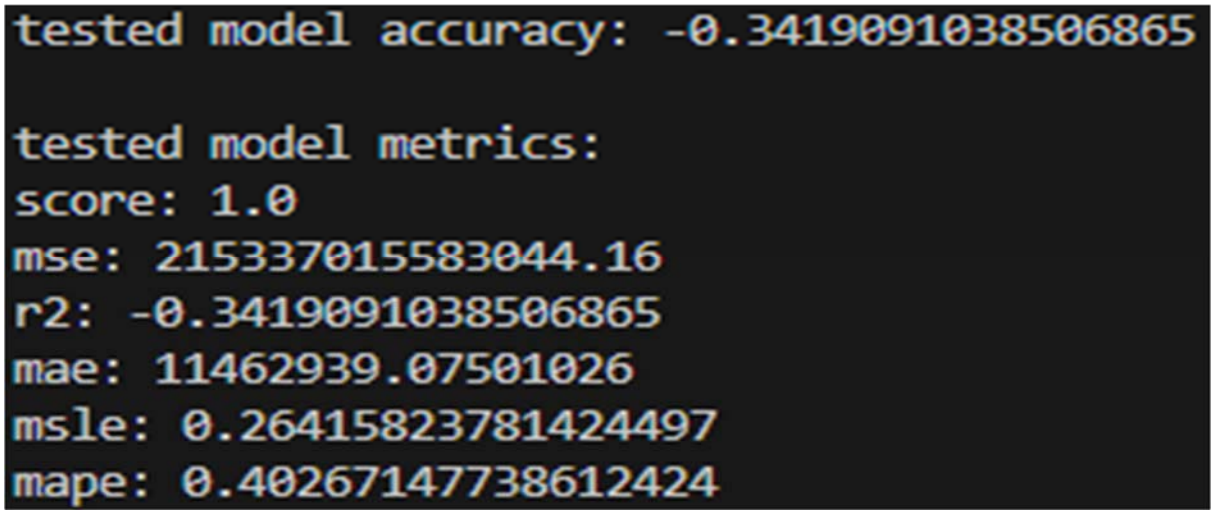


Figure 5.25: SVM Tested Model Predictions Metrics

Figure 5.26 and Figure 5.27 shows the Linear Regression's Trained and Tested Model Predictions Metrics respectively.

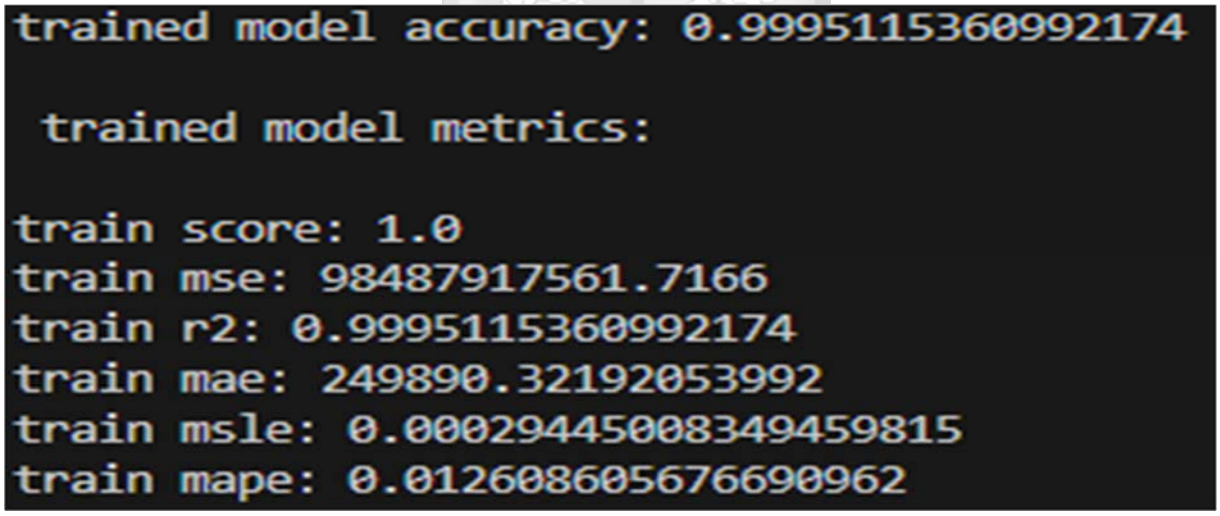


Figure 5.26: Linear Regression Trained Model Predictions Metrics

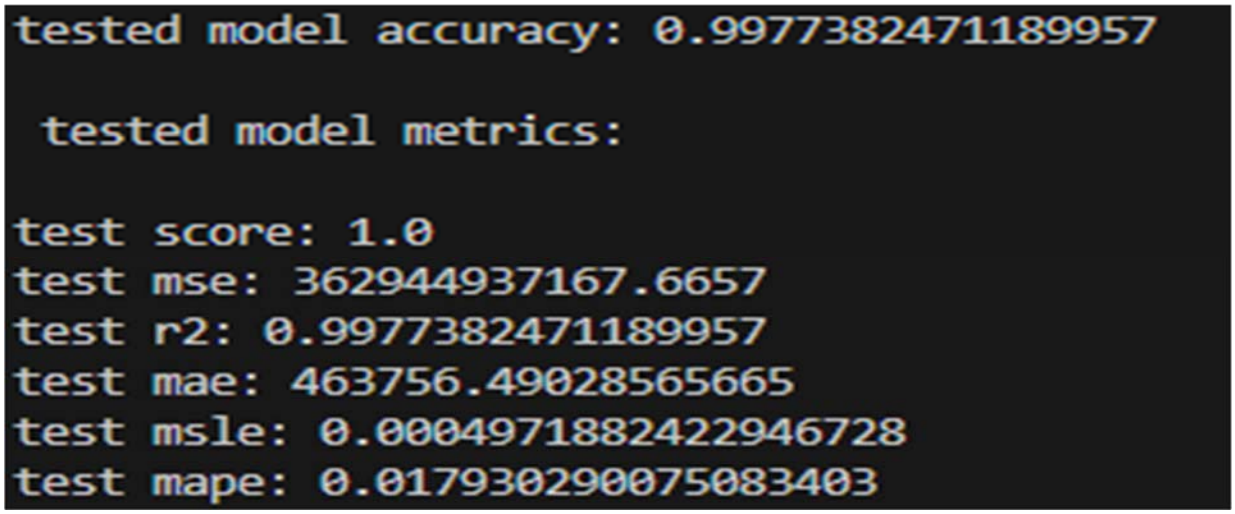


Figure 5.27: Linear Regression Tested Model Predictions Metrics

Figure 5.28 and Figure 5.29 shows the Decision Tree's Trained and Tested Model Predictions Metrics respectively.

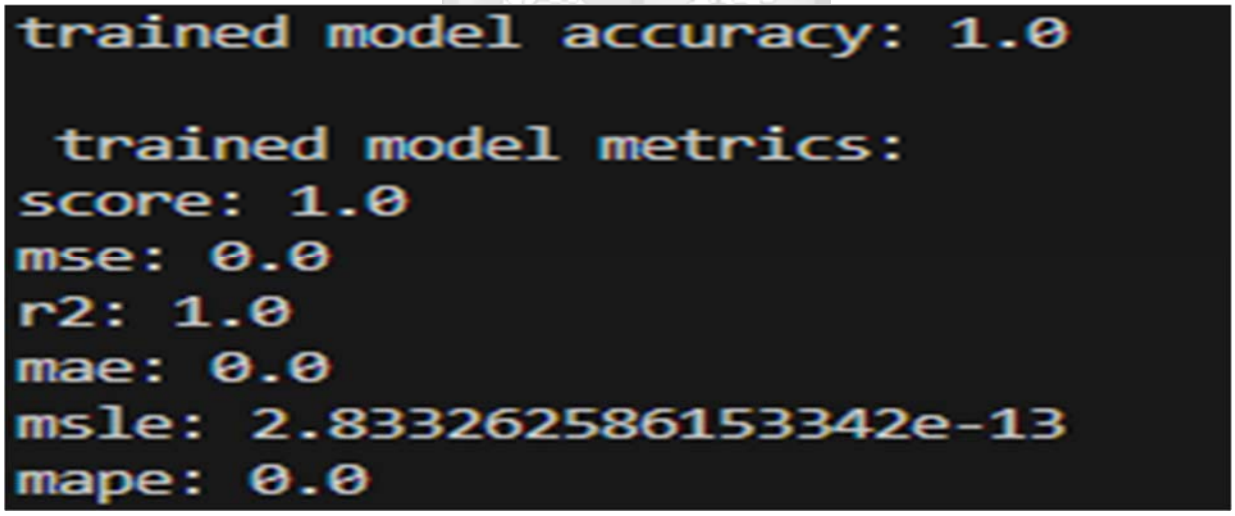


Figure 5.28: Decision Tree Trained Model Predictions Metrics

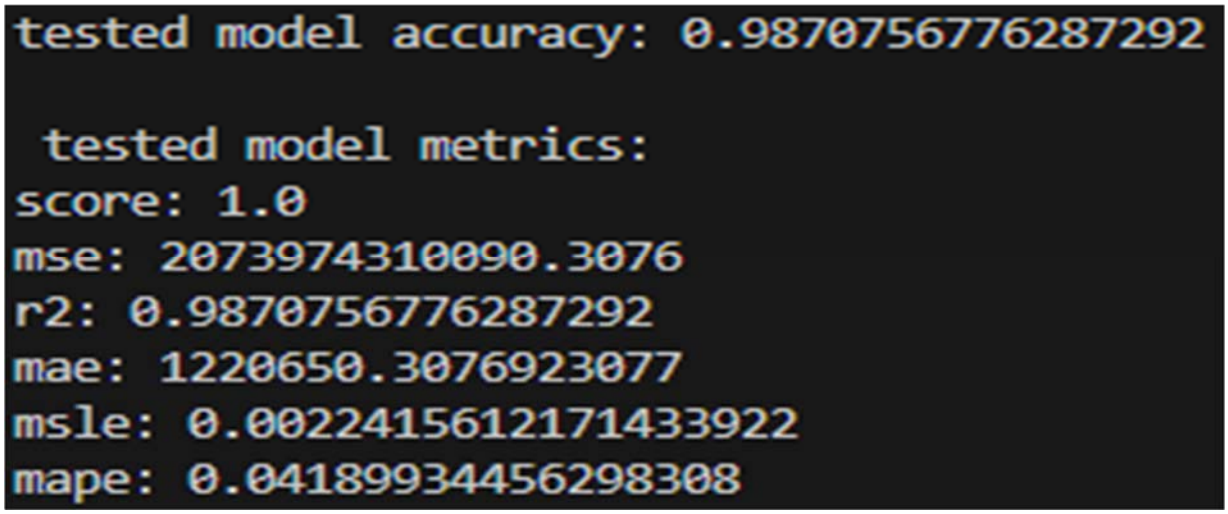


Figure 5.29: Decision Tree Tested Model Predictions Metrics

Figure 5.30 and Figure 5.31 shows the KNN's Trained and Tested Model Predictions Metrics respectively.

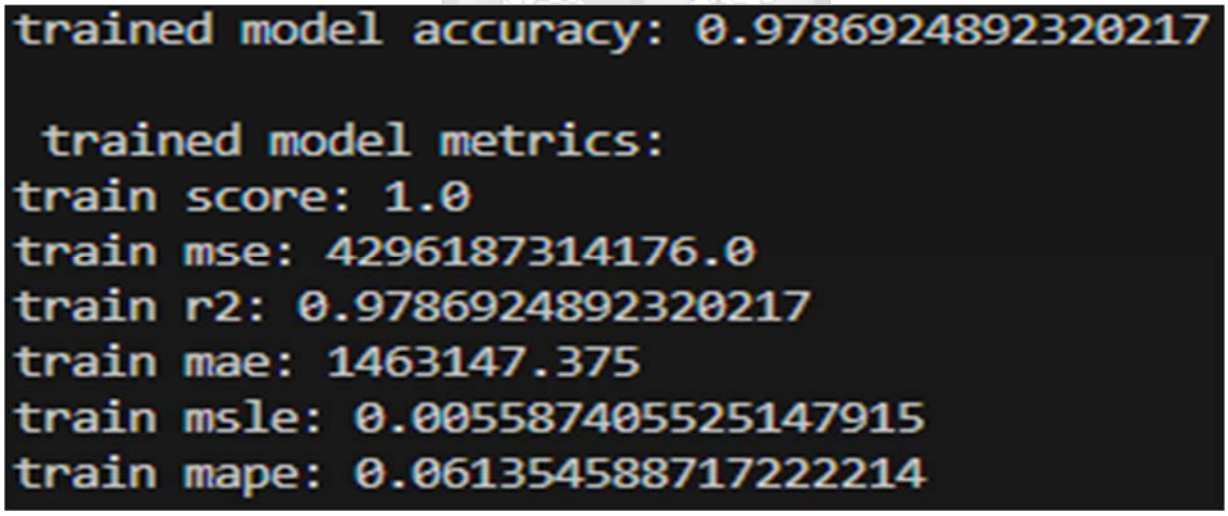


Figure 5.30: KNN Trained Model Predictions Metrics

```
tested model accuracy: 0.9382859234076198

 tested model metrics:
test score: 1.0
test mse: 9903298052096.0
test r2: 0.9382859234076198
test mae: 2640399.75
test msle: 0.011378415860235691
test mape: 0.09246568381786346
```

Figure 5.31: KNN Tested Model Predictions Metrics

Figure 5.32 and Figure 5.33 shows the ANN's Trained and Tested Model Predictions Graphs respectively.

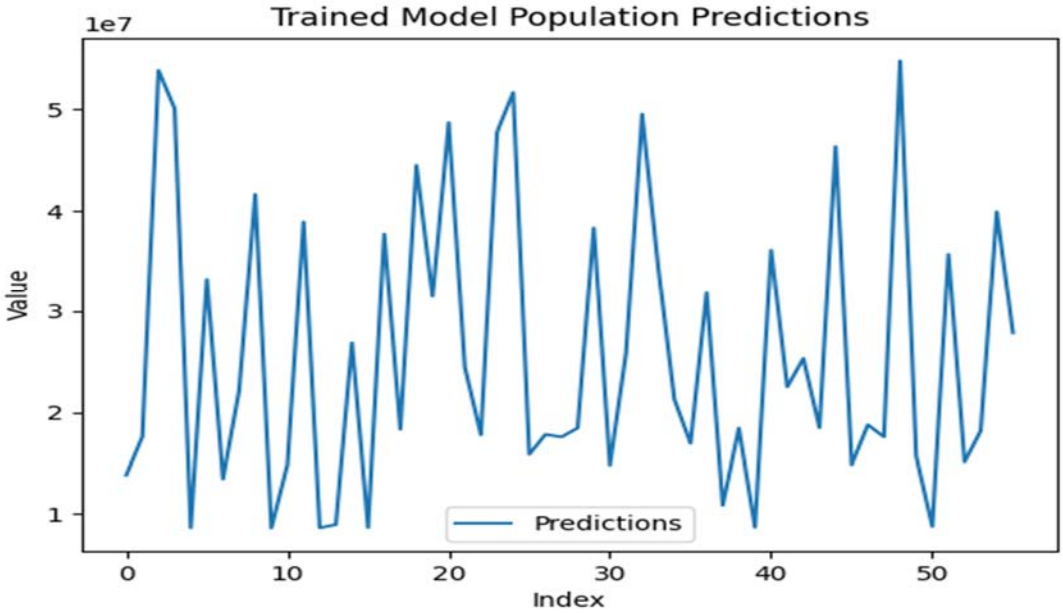


Figure 5.32: ANN Trained Model Predictions Graph

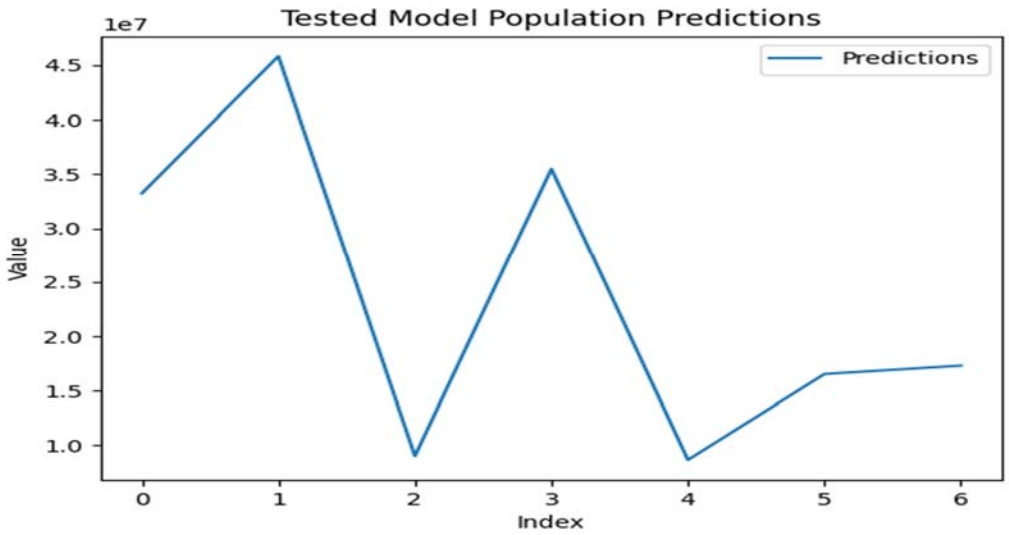


Figure 5.33: ANN Tested Model Predictions Graph

Figure 5.34 and Figure 5.35 shows the Random Forest's Trained and Tested Model Predictions Graphs respectively.

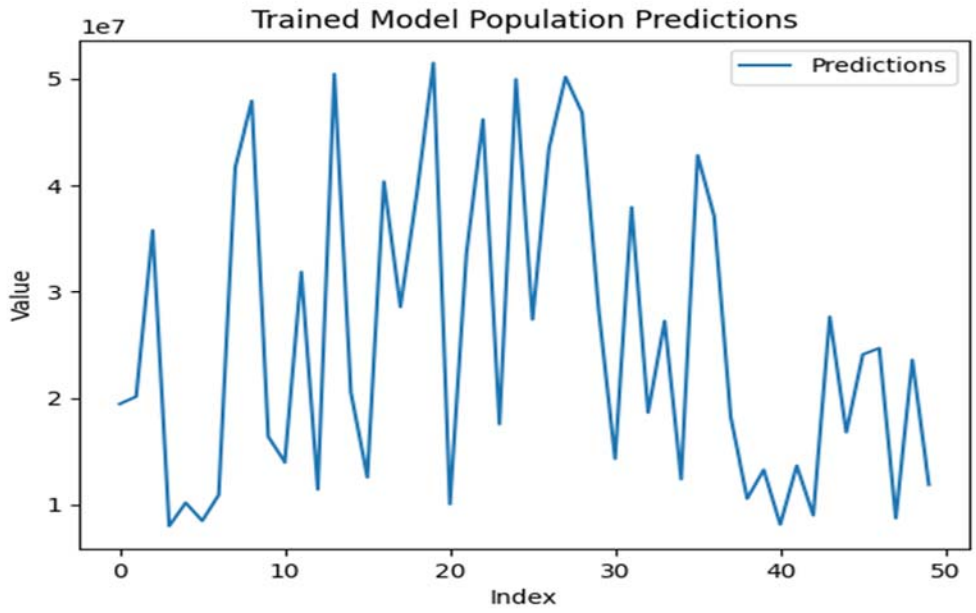


Figure 5.34: Random Forest Trained Model Predictions Graph

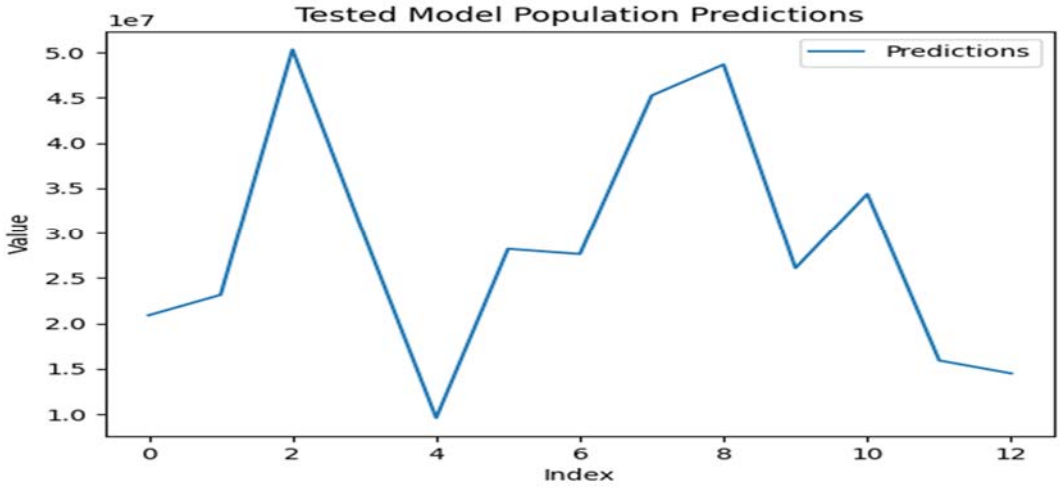


Figure 5.35: Random Forest Tested Model Predictions Graph

Figure 5.36 and Figure 5.37 shows the Logistic Regression's Trained and Tested Model Predictions Graphs respectively.

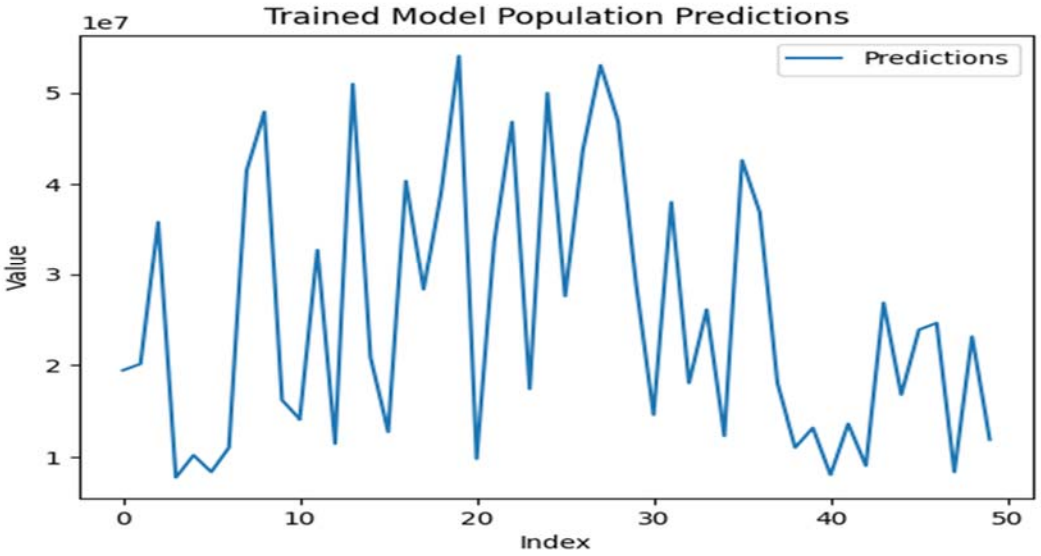


Figure 5.36: Logistic Regression Trained Model Predictions Graph

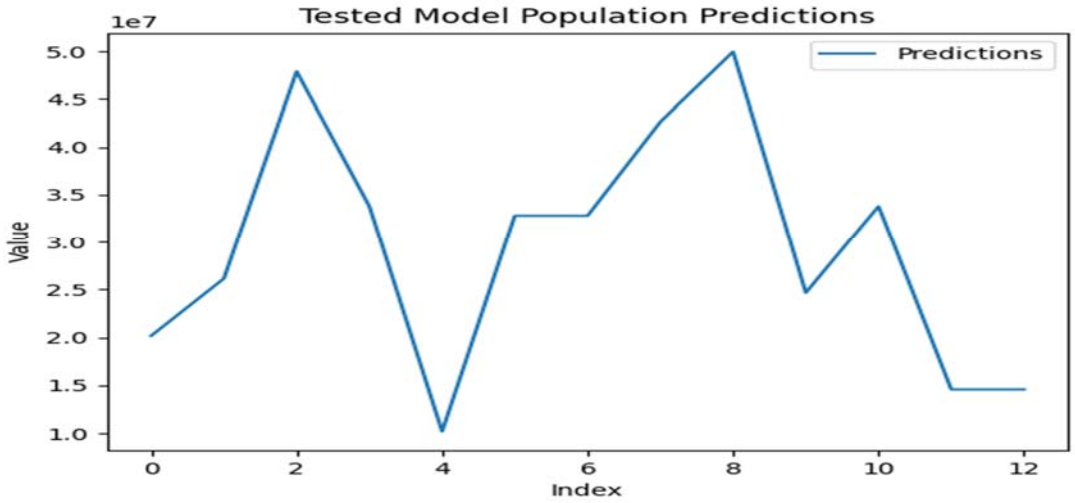


Figure 5.37: Logistic Regression Tested Model Predictions Graph

Figure 5.38 and Figure 5.39 shows the SVM's Trained and Tested Model Predictions Graphs respectively.

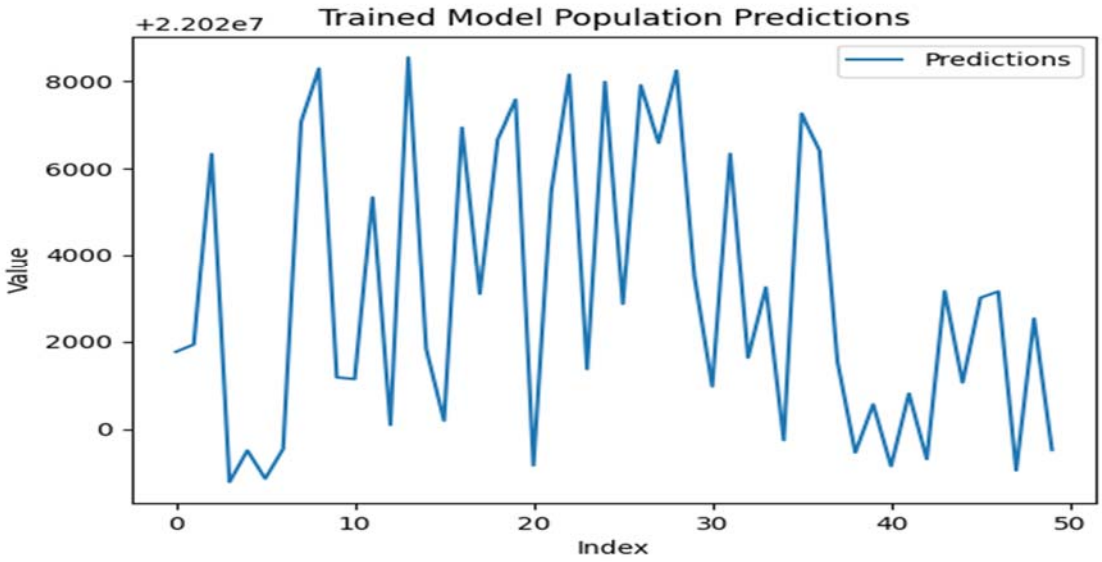


Figure 5.38: SVM Trained Model Predictions Graph

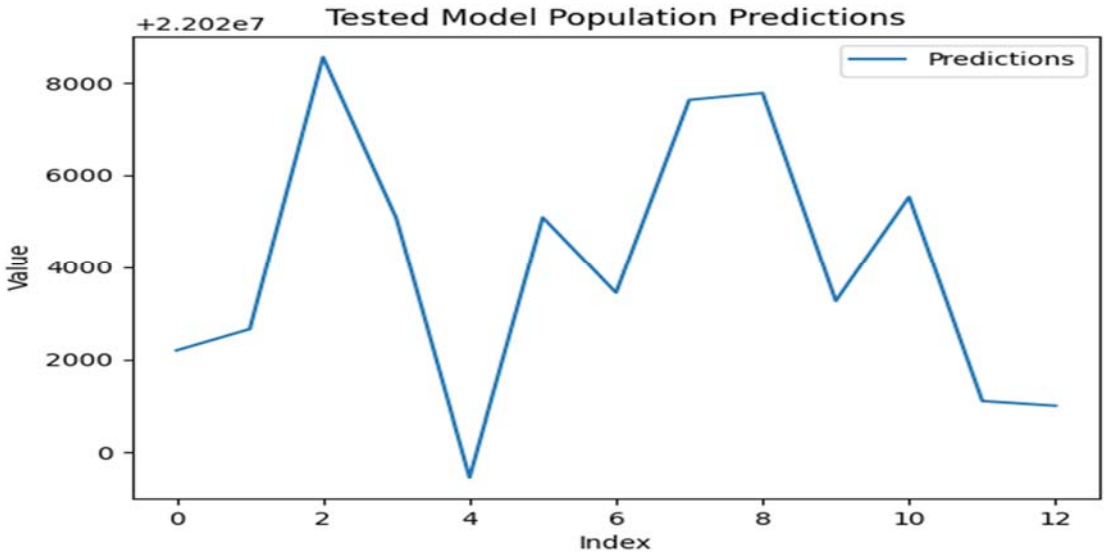


Figure 5.39: SVM Tested Model Predictions Graph

Figure 5.40 and Figure 5.41 shows the Linear Regression's Trained and Tested Model Predictions Graphs respectively.

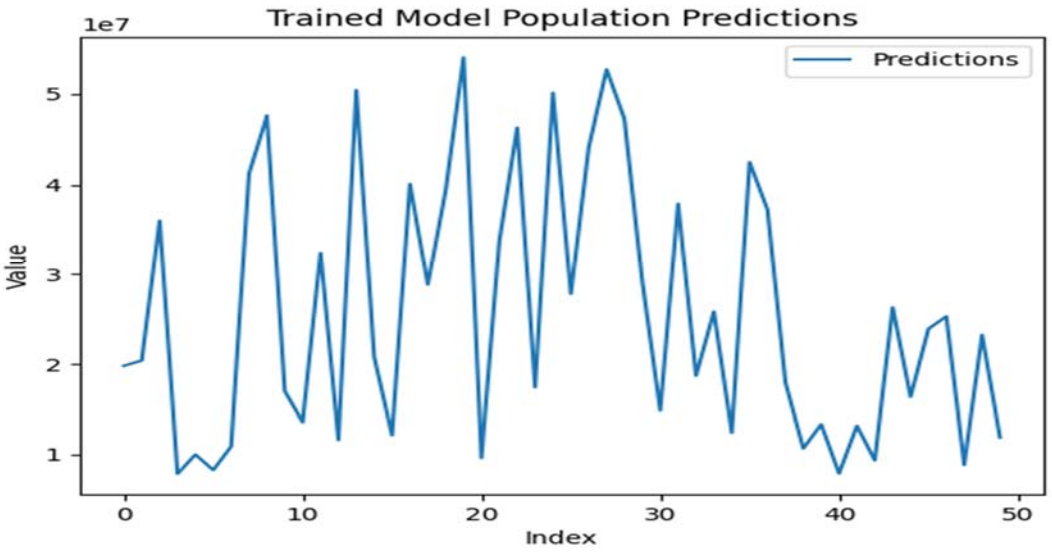


Figure 5.40: Linear Regression Trained Model Predictions Graph

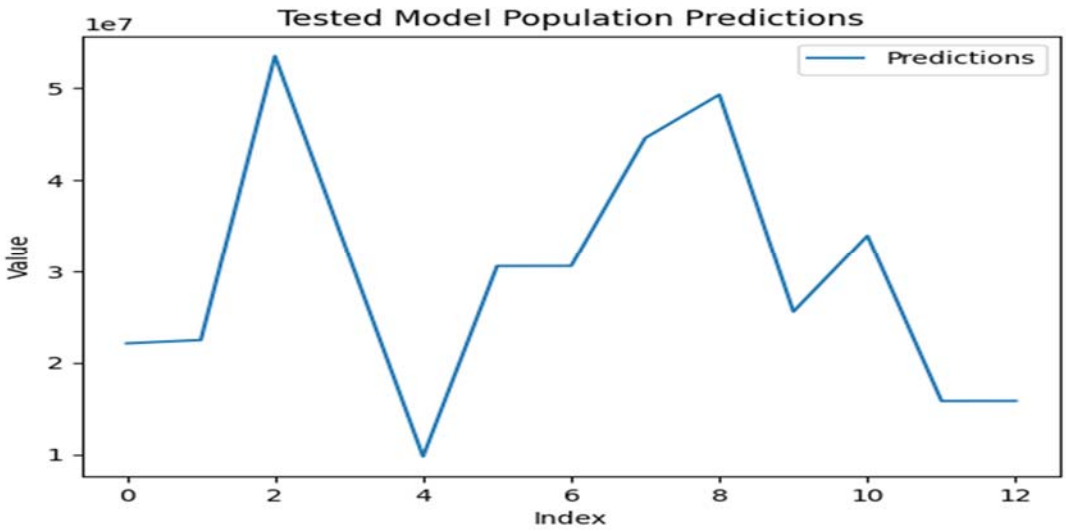


Figure 5.41: Linear Regression Tested Model Predictions Graph

Figure 5.42 and Figure 5.43 shows the Decision Tree's Trained and Tested Model Predictions Graphs respectively.

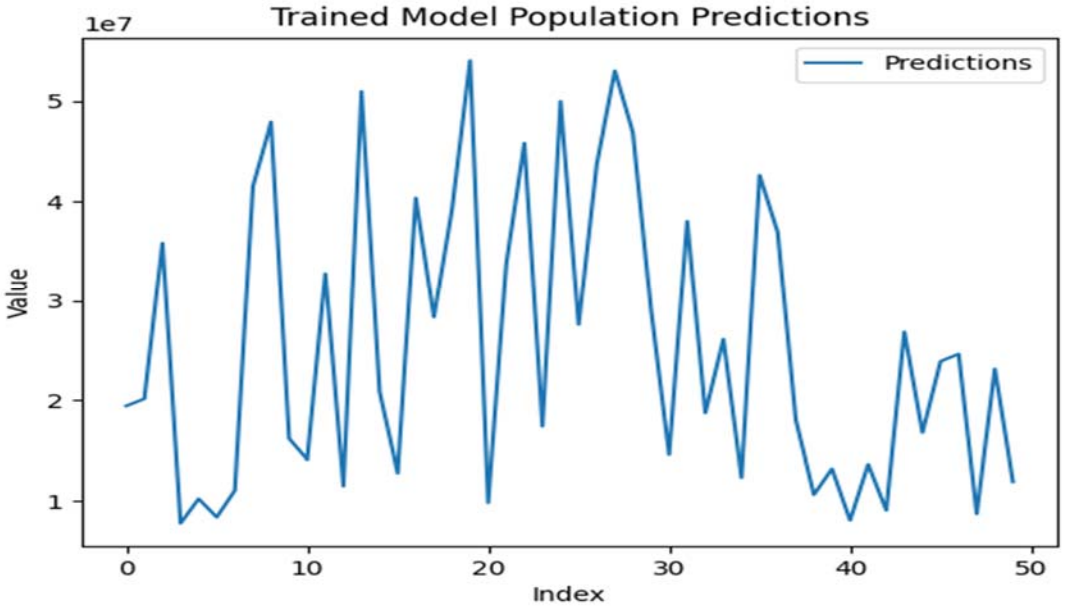


Figure 5.42: Decision Tree Trained Model Predictions Graph

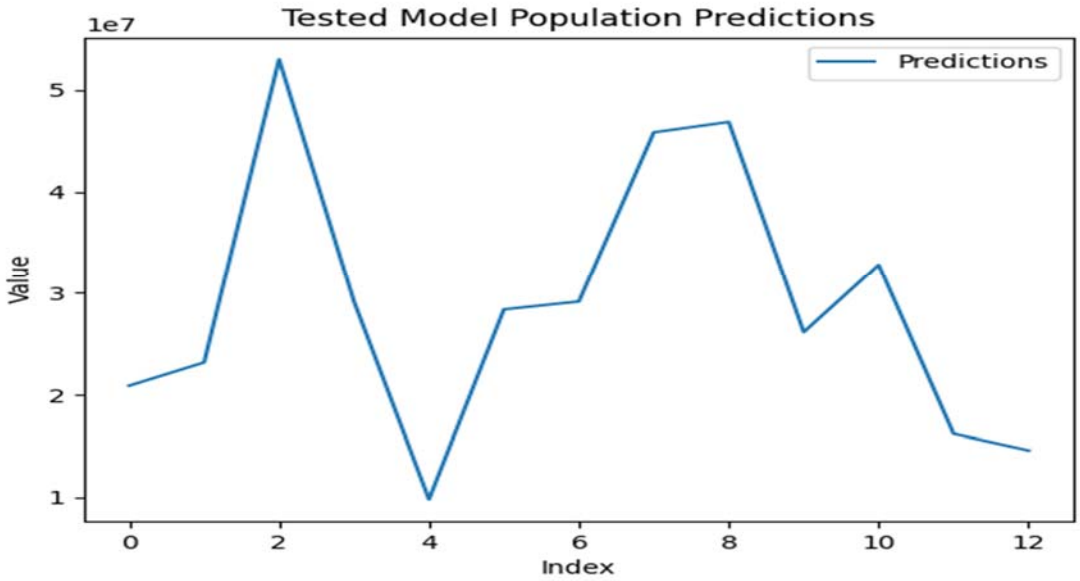


Figure 5.43: Decision Tree Tested Model Predictions Graph

Figure 5.44 and Figure 5.45 shows the KNN's Trained and Tested Model Predictions Graphs respectively.



Figure 5.44: KNN Trained Model Predictions Graph

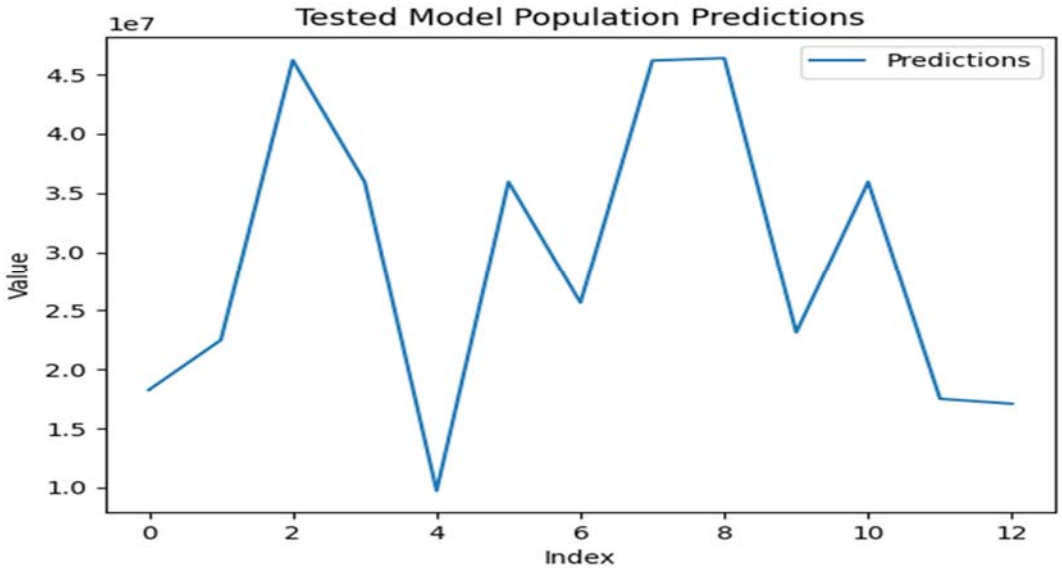


Figure 5.45: KNN Tested Model Predictions Graph

VT OMNES VNVM SINT

Chapter 6: Discussions

6.1 Research Objectives Review

In the development phase of the models, a range of ML techniques were employed, including Artificial Neural Networks, Random Forest, Logistic Regression, Support Vector Machine, Linear Regression, Decision Trees, and K-Nearest Neighbors. These models were trained on a time series dataset spanning from 1960 to 2008 and then tested on data from 2009 to 2022. To adhere to the 80%-20% rule, 80% of the data was used for training, while the remaining 20% was reserved for testing. This section explores the findings of the research study, which focused on leveraging ML models to predict the population of Kenya. The research objectives were revisited since they guided this study.

6.1.1 Determining the Challenges Associated with the Forecasting of Human Population growth in Kenya

In order to effectively predict the human population growth of Kenya, it is essential to consider several key factors. First, it is of so much importance to have a thorough comprehension of the quality and availability of demographic and socioeconomic data in the region (Andrasfay & Goldman, 2022; USCB, 2023; World Bank, 2023). Next, potential challenges in historical data must be identified. Finally, any environmental or socioeconomic factors that may complicate the predictions of ML models must be recognized. Through careful analysis of the quality, accessibility, and scope of demographic data, as well as an investigation of historical data trends, challenges, and other pertinent factors, we can gain a comprehensive understanding of population growth in Kenya.

6.1.2 Analysis of The Current Methods, Techniques, and Approaches for Forecasting Human Population Growth

In this study the review of the existing techniques and approaches for population forecasting in Kenya. The traditional statistical methods, such as census, demographic surveys, and virtual registration, have been widely used in the past. However, the cost associated with these methods has been a concern for the government (KNBS, 2022). Alternatively, ML techniques have emerged as a promising tool for population forecasting. A number of recent studies have reported successful applications of ML techniques in human population forecasting for different regions. For instance, Şahinarslan et al. (2021) used several ML models to predict the population of Turkey in 2017, achieving an accuracy of 95%. Similarly, Suárez et al. (2022) utilised the ANN algorithm and CO2 emissions to forecast the human population growth in urban centers of Colombia with an MAE of 0.038. Furthermore, Otoom et al. (2019) compared the performance of 17 different ML models in population forecasting and found that the best-performing models achieved accurate results even when historical data was not available. Based on these studies, the present research study aims to analyse and select the features and ML models for human population forecasting in Kenya. The strengths and weaknesses of the existing techniques were assessed to determine their capabilities in predicting the population trend dynamics of the country.

6.1.3 Development of the ML Models for Forecasting Human Population Growth in Kenya

As this is a time series issue, a comprehensive analysis of various ML algorithms was conducted to select the most appropriate one for human population forecasting in Kenya. Key stages included model selection and experimentation with a range of linear regression techniques, such as Artificial Neural Networks, Random Forest, Support Vector Machine, Logistic Regression, Linear Regression, Decision Trees, and K-Nearest Neighbors. Further efforts were made to optimize the performance of the ML models through hyperparameter tuning.

6.1.4 Validation of the ML models

Within this section, a comprehensive analysis was conducted using various performance metrics, including the MSE, R2 Score, MAE, MSLE, and MAPE. These metrics were rigorously applied to validate the performance of the models utilising cross-validation techniques to reduce overfitting and ensure the broad applicability of the ML models. This meticulous process was carried out to identify the most accurate and reliable ML models for predicting population trends in Kenya.

6.1.5 Insights and Interpretations of the Models

This study aimed to delve into the crucial factors that contribute to population growth in a developing country such as Kenya. Rather than simply predicting population growth in Kenya, the machine learning models utilised in this study employed techniques like analysing feature importance and identifying the factors that have great influence on the human population growth trends in Kenya.

6.2 Advantages of the Tool

The Population Predictions Platform plays a crucial role in helping policy and decision makers plan for the country's resources based on population projections. Additionally, investors, particularly foreign investors, can make informed decisions on whether to invest in the country based on these projections.

6.3 Limits to the study and the Population Predictions Platform

One of the primary limitations of this study is the relatively small dataset size used. This issue was previously highlighted in the Literature Review, where Şahinarslan, et al., (2021) suggested that ML models are more effective when working with larger and more extensive datasets. They emphasized that population prediction is a challenge that will always be limited by the available datasets, particularly in developing countries like Kenya where statistical and demographic indicators are limited. This raises questions about the consistency and reliability of the available data. Another limitation relates to the training and testing of the models, as performance can differ significantly when using only a small sample of the dataset compared to the entire dataset.

Chapter 7: Conclusions and Recommendations

7.1 Conclusions

A population predictions platform was created using several ML models, including Artificial Neural Networks, Random Forest, Logistic Regression, Support Vector Machine, Linear Regression, Decision Trees, and K-Nearest Neighbors. The platform's accuracy was determined using various metrics like MAE, MSE, MAPE, MSLE, and R2 Score. The study found that Linear Regression, Decision Trees, and K-Nearest Neighbors performed better initially, but with hyper parameter tuning, Linear Regression, Random Forest, and Artificial Neural Network performed the best. These ML models can provide valuable contributions to policymakers and investors by enabling them to make informed decisions and better plan for the future.

7.2 Recommendations

The techniques, approaches, procedures and ML algorithms/models used in the study is crucial for further development and improvement. While the ML models used in the study accurately predicted population trends by considering factors such as fertility and mortality rates, life expectancy, net migration, economic growth, access to healthcare and education, and gender equality, the researcher suggests that other factors such as climate change and the use of satellite imagery and social media trends should be addressed. Additionally, creating a more advanced hybrid ML model by combining multiple models and investing in high-quality data from non-traditional sources would be beneficial. Open data access for researchers and collection of diverse, high-quality data is also recommended. Lastly, transparency and addressing uncertainties in the models is critical for future development.

7.3 Future Works

Based on the findings of this study, there are limitations that offer valuable insights for future research explorations. To enhance the accuracy of models, future studies and further explorations could focus on enlarging and refining datasets. Additionally, researchers could consider improving demographic data quality using sources such as satellite imagery, social media platforms, and climatic changes. By addressing these areas of focus and building upon the

previously conducted studies, future researchers can design a reliable, sophisticated, and powerful population prediction platform.

References

Adebayo, S. O., & Owolabi, S. A. (2022). Assessing the effectiveness of the family planning program in Nigeria: A mixed methods study. *BMC Public Health*, 22(1), 563. doi: 10.1186/s12889-022-12774-3

Alam, N., & Islam, M. M. (2022). Population dynamics and development: A review of recent literature. *Population and Development Review*, 48(2), 253-280.

Alsaqqa, S., Sawalha, S., & Abdel-Nabi, H. (2020). *Agile Software Development: Methodologies and Trends*. International Journal of Interactive Mobile Technologies (iJIM), 14(11), 246–270. https://doi.org/10.3991/ijim.v14i11.13269

American Psychological Association. (2022). *Ethical principles of psychologists and code of conduct*.

Andrasfay, T., & Goldman, N. (2022). *Associations between neighborhood socioeconomic characteristics and cognition in older adults*. Demography, 59(5), 1899-1919.

Bardsley, D. K., & Oswald, A. J. (2017). *The economics of population: A survey*. Journal of Economic Literature, 55(1), 35-77.

Bardsley, P., & Hugo, G. (2021). The future of population forecasting: A review. *Population and Development Review*, 47(1), 1-28.

Berman, G. A., & Khan, M. S. (2019). *The importance of clear and concise objectives in program evaluation*. American Journal of Evaluation, 40(1), 1-14.

Bhandari, P., & Mishra, S. (2021). Mixed methods study to assess the effectiveness of community-based interventions for family planning in Nepal. *BMC Public Health*, 21(1), 1608. doi: 10.1186/s12889-021-10607-8

Bongaarts, J. (2020). The limits to population forecasting. *Population and Development Review*, 46(2), 183-206.

Bongaarts, J. (2021). The economic and social implications of overpopulation. *Population and Development Review*, 47(3), 397-424. doi:10.1111/padr.12329

Bongaarts, J. (2022). Population growth and development: An overview of recent trends and challenges. *Population and Development Review*, 48(2), 281-294.

Bongaarts, J., & Westoff, C. F. (2017). The role of family planning in global health. *Population and Development Review*, 43(2), 175-203. doi:10.1111/padr.12237

Brown, L. R. (2017). *The world's population: Too many or too few?*. New York, NY: W.W. Norton & Company.

Cai, F., & Wang, L. (2022). Population aging and economic growth: A review of recent literature. *Population and Development Review*, 48(2), 295-320.

Castles, S., & Miller, M. J. (2020). *The age of migration: International population movements in the modern world* (5th ed.). Palgrave Macmillan.

Chen, L., Zhang, Y., & Li, J. (2022). *TensorFlow Quantum: A framework for hybrid quantum-classical machine learning*. arXiv preprint arXiv:2202.07954.

Chen, X., & Wang, J. (2022). *A Survey of PyTorch for Natural Language Processing*. arXiv preprint arXiv:2202.01844.

Chen, Y., Liu, Y., & Zhang, L. (2022). *Deep neural network with adaptive regularization for imbalanced classification*. Pattern Recognition, 120, 108060.

Chen, Y., Liu, Y., & Zhang, L. (2022). *Random forest with adaptive regularization for imbalanced classification*. Pattern Recognition, 120, 108060.

Chen, Y., Liu, Y., & Zhang, L. (2022). *Support vector machine with adaptive regularization for imbalanced classification*. Pattern Recognition, 120, 108060.

Chollet, F. (2022). Keras: An API for deep learning. arXiv preprint arXiv:2201.05667.

Cohen, J. E. (2019). *How many people can the earth support?* New York, NY: W. W. Norton & Company.

Conly, S., & Hartmann, B. (2021). Population stabilization: A moral imperative. *Foreign Affairs*, 100(4), 136-147.

Creswell, J. W., & Creswell, J. D. (2022). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications.

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.

Creswell, J. W., & Poth, C. N. (2021). *Qualitative inquiry and research design: Choosing among five approaches* (5th ed.). SAGE Publications.

Cui, Y., Zhang, B., Yang, W., Wang, Z., Li, Y., Yi, X., & Tang, Y. (2017). *End-to-end visual target tracking in multi-robot systems based on deep convolutional neural network.* In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) (pp. 2849-2857). IEEE. doi:10.1109/ICCVW.2017.135

D. Dellermann, M.Sc., & Prof. Dr. J. M. Leimeister. (2023). Research Center for IS Design (ITeG), Information Systems, University of Kassel, Pfannkuchstraße 1, 34121 Kassel, Germany.

Deisenroth, M., Rasmussen, C. E., & Blundell, C. (2020). *Machine learning: A probabilistic perspective*. MIT Press.

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2019). *Hybrid intelligence: A taxonomy of design knowledge for hybrid intelligence systems*. In

Proceedings of the 52nd Hawaii International Conference on System Sciences (pp. 2809-2818). Maui, HI: IEEE.

Dobbs, R., Manyika, J., Woetzel, J., & Ahmed, S. (2022). *The great demographic shift: Eight megatrends that will shape the world.* McKinsey Global Institute.

Dyson, T., & West, O. (2019). The economic and environmental consequences of population growth. *Population and Development Review*, 45(4), 667-691. doi:10.1111/padr.12288

EAC Business Guide 2015-REV1. (2015). Retrieved September 13, 2023, from https://norwegianafrican.no/wp-content/uploads/2015/10/20150928-EAC-Business-Guide-2015-REV1-Web-version.pdf?cv=1.

Fan, H., & Yao, Y. (2020). *Population forecasting and resource allocation in developing countries: A case study of China*. Sustainability, 12(14), 5632. https://www.mdpi.com/2071-1050/15/16/12344

Fernandes, A., & Santos, R. (2022). Economic development and poverty reduction: Evidence from sub-Saharan Africa. *World Development*, 140, 105421.

Frost, J. (2022). *Cronbach's alpha: A guide to reliability in quantitative research*. Retrieved from https://www.researchgate.net/publication/356352010_Cronbach's_alpha_A_guide_to_reliability_in_quantitative_research.

Ghosh, S. K., & Kumar, A. (2022). A survey on classification algorithms for imbalanced data. *Artificial Intelligence Review*, 57(1), 1-34.

Gómez, J. A., Patiño, J. E., Duque, J. C., & Passos, S. (2020). Spatiotemporal modeling of urban growth using machine learning. *Remote Sensing*, 12(1), 109.

Gould, W. (2018). Data challenges for population forecasting in sub-Saharan Africa. In E. Antonio, R. Bui, & R. Kontchakov (Eds.), *Forecasting with social data* (pp. 47-56). Springer.

Gu, S., Zhang, L., & Zhao, Z. (2022). *Deep clustering: A survey*. ACM Computing Surveys (CSUR), 55(3), 1-39. doi:10.1145/3530077

Gulati, A., & Garg, D. (2022). *Keras Tuner: A library for hyperparameter optimization in Keras*. arXiv preprint arXiv:2202.03359.

Gupta, A., Kumar, A., & Das, A. (2022). *TensorFlow Extended: A unified framework for machine learning and data science*. arXiv preprint arXiv:2205.11584.

Gustavo Suárez et al 2022 IOP Conf. Ser. : Mater. Sci. Eng. 1253 012007

Harris, C. R., Millman, K. J., J., S., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2.

Hartmann, B., & Gemmill, B. (2020). *Population and sustainability: Beyond the numbers*. Routledge.

Heuveline, P., & Mason, W. M. (2017). Population forecasting: Methods and challenges. *Annual Review of Sociology*, 43, 25-48.

Huang, G., Liu, Z., Sun, M., Song, Y., Wang, L., & Chen, D. (2022). *TensorFlow Lite 2.8: A new lightweight and performant machine learning framework for mobile and edge devices*. arXiv preprint arXiv:2206.04936.

Hung, H.-H., & Chen, Y.-W. (2022). *A survey on transfer learning for supervised machine learning*. ACM Computing Surveys (CSUR), 55(4), 1-35.

Hymel, S. (2020). *Getting Started with Machine Learning Using TensorFlow and Keras*. Packt Publishing.

IBM Corporation. (2023, October 13). *IBM Watson*. Retrieved from https://www.ibm.com/watson/.

Ikudo, A., Lane, J. I., Staudt, J., & Weinberg, B. A. (2018). *Occupational classifications: A machine learning approach*. IZA Discussion Paper No. 11738. doi:10.1111/izap.124951

International Institute for Environment and Development. (2017). *Population, poverty, and climate change: A review of the evidence*. London, UK: International Institute for Environment and Development.

International Organization for Migration (IOM). (2022). *World migration report 2022: International migration, social cohesion and development*. IOM.

Kang, B., Gao, J., & Zhang, C. (2022). *Unsupervised learning for natural language processing: A survey*. arXiv preprint arXiv:2202.08565.

Kelley, A. C., & Schmidt, R. H. (2018). *The economics of population change*. New York, NY: Springer.

Kenya National Bureau of Statistics (KNBS). (2021). *Population and housing census 2019: Key findings. Nairobi*, Kenya: KNBS. Data Tables - Kenya National Bureau of Statistics (knbs.or.ke)

Khan, A., & Mahmood, A. (2022). *Impact of population growth on political stability in Pakistan*. Sustainability, 14(2), 680.

Khan, M. A., & Siddiqui, H. R. (2022). *Impact of population growth on social cohesion in Pakistan*. Sustainability, 14(2), 681.

Khan, W. A., & Mahmood, A. (2022). *Impact of population growth on environmental degradation in Pakistan*. Sustainability, 14(2), 682.

Klasen, S. (2022). *Economic development, poverty reduction, and gender equality: A review of the evidence*. Journal of Development Studies, 58(1), 1-25.

Kotsiantis, S., & Zaharakis, I. (2022). *Supervised machine learning: A review of classification techniques*. Informatica Journal, 46(1), 1-27.

Kumar, A., Sattigeri, P., & Fletcher, P. T. (2017). *A novel approach to natural language processing*. In Proceedings of the 31st Conference on Neural Information Processing Systems (pp. 1-10). NeurIPS 2017..

Li, B., & Wang, H. (2022). *A novel deep neural network with kernel selection for image classification*. Neural Computing and Applications, 34(1), 147-158.

Li, B., & Wang, H. (2022). *A novel naive Bayes classifier with kernel selection for image classification*. Neural Computing and Applications, 34(1), 147-158.

Li, B., & Wang, H. (2022). *A novel random forest with kernel selection for image classification*. Neural Computing and Applications, 34(1), 147-158.

Li, B., & Wang, H. (2022). *A novel support vector machine with kernel selection for image classification*. Neural Computing and Applications, 34(1), 147-158.

Li, C. H., & Wang, C. (2021). *BOAI: Fast Alternating Decision Tree Induction Based on Bottom-Up Evaluation*. In Proceedings of the 26th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) (pp. 121-133).

Li, J., & Wang, Z. (2021). *An alternating decision tree algorithm for multi-label classification*. Neural Computing and Applications, 33(11), 6935–6946.

Li, X., & Song, J. (2020). *An alternating decision tree ensemble for imbalanced classification*. Neural Computing and Applications, 32(11), 6611–6620.

Li, Z., Liu, Z., & Wang, P. (2022). *Unsupervised representation learning for computer vision: A survey*. arXiv preprint arXiv:2202.08566.

Liu, B., & Yang, W. (2023). *A hybrid random forest algorithm with feature selection for text classification*. Knowledge-Based Systems, 208, 107989.

Liu, P., & Wang, X. (2022). *PyTorch for Reinforcement Learning: A Survey*. arXiv preprint arXiv:2207.01295.

Lutz, W., Samir K., & Samir K. (2022). World population stabilization unlikely this century. *Nature*, 604(7906), 502-505.

Mason, J. (2018). *Qualitative researching*. SAGE Publications.

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). *A review of deep learning techniques for speech processing*. arXiv preprint arXiv:2305.00359.

Murray, C. J. L., Brown, T. M., & Rogers, R. G. (2020). *The economic impacts of overpopulation*. The Lancet, 396(10255), 1737-1747. doi:10.1016/S0140-6736(20)30589-7

Nguyen, T. T. (2022). *The impact of population growth on environmental degradation in Vietnam*. Sustainability, 14(2), 686.

Odhiambo, F., K'Akumu, O., & Gichero, P. (2020). *Evaluating the potential of integrating GIS and remote sensing for small-scale population estimation in Kenya*. GeoJournal, 85, 341-357.

Otoom, M. M., Jemmali, M., Qawqzeh, Y., SA, K. N., & Al Fay, F. (2019). *Comparative analysis of different machine learning models for estimating the population growth rate in data-limited area*. IJCSNS, 19(1), 96

Piryonesi, S., El-Diraby, M. A., & Tamer, M. A. (2020). *Kernel density estimation based naive Bayes for imbalanced classification*. Expert Systems with Applications, 158, 113595.

PyTorch Team. (2023). PyTorch Profiler documentation. Retrieved from https://pytorch.org/docs/1.0.1/autograd.html#profiler.

Ravallion, M. (2022). *Economic growth and poverty reduction: Has the pendulum swung too far?* World Development, 140, 105419.

Rudestam, K. E., & Newton, R. R. (2020). *Surviving your dissertation: A comprehensive guide to research and writing* (5th ed.). Sage Publications.

Şahinarslan, F.V., Tekin, A.T. and Çebi, F. (2019) *'Machine learning algorithms to forecast population: Turkey example'*, Presented at the ETMS International Engineering and Technology Management Summit, 12 October 2019, İstanbul, Turkey.

Şahinarslan, F.V., Tekin, A.T. and Çebi, F. (2021) *'Application of machine learning algorithms for population forecasting'*, *Int. J. Data Science*, Vol. 6, No. 4, pp.257–270.

Silver, D., et al., (2016). *Mastering the game of go without human knowledge*. Nature, 529(7587), 484-489.

Simply Learn. (2023, March 8). *What Is TensorFlow 2.0: The Best Guide to Understand TensorFlow*. Retrieved from https://www.simplilearn.com/tutorials/deep-learning-tutorial/tensorflow-2.

Singh, K. (2023, March 8). *Deploying Named Entity Recognition model to production using TorchServe*. Medium. Retrieved from https://medium.com/analytics-vidhya/deploying-named-entity-recognition-model-to-production-using-torchserve-fd8cf5cff02f.

Singh, S., & Darroch, J. E. (2018). *Unintended pregnancy: Worldwide estimates of incidence and health outcomes*. The Lancet, 391(10120), 1800-1810. doi:10.1016/S0140-6736(17)32507-8

Soofi, M., & Awan, M. A. (2017). *A survey of machine learning techniques for text classification*. Journal of Basic & Applied Sciences, 13. http://creativecommons.org/licenses/by-nc/3.0/.

Tavakolan, M. R., & Ebrahimi, S. (2022). *Impact of population growth on economic growth in Iran*. Sustainability, 14(2), 680.

The Lancet Commission on Population and Sustainable Development. (2021). *Global trends in population, health, and climate change: Report of the Lancet Commission on Population and Sustainable Development*. The Lancet, 397(10267), 1569-1634. doi:10.1016/S0140-6736(21)00492-9

The pandas development team. (2020). *pandas-dev/pandas: Pandas* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3509134

Tian, Y. (2023). *Artificial intelligence image recognition method based on convolutional neural network algorithm*. Journal of Computer Science, 18(1), 1-10.

Tsui, A. O., & Zhang, J. (2019). *The impact of family planning programs on unintended pregnancy and abortion: A review of the global evidence*. The Lancet, 393(10168), 2280-2289. doi:10.1016/S0140-6736(18)32252-9

United Nations Children's Fund (UNICEF). (2022, February 7). *Dividend or disaster: UNICEF's new report into population growth in Africa*. https://www.unicef.org/wca/press-releases/dividend-or-disaster-unicefs-new-report-population-growth-africa

United Nations Environment Programme. (n.d.). *Cities and climate change*. Retrieved July 15, 2023, from https://www.unep.org/explore-topics/resource-efficiency/what-we-do/cities/cities-and-climate-change

United Nations Population Fund (UNFPA). (2019). *Population projections: Methodology and use*. New York, NY: UNFPA. Visualisation Overview | Population Data Portal (unfpa.org)

United Nations Population Fund. (2021). *Population dynamics and their impact on development: A resource book*. https://www.unfpa.org/

United Nations Statistics Division (UNSD). (2019, March). *Handbook on population and housing census editing*. New York, NY: United Nations. Principles and Recommendations for Population and Housing Censuses, Revision 3 | United Nations iLibrary (un-ilibrary.org)

United Nations. (2019). *International migration report 2019: Highlights*. United Nations.

United Nations. (2019). *World population prospects 2019: Highlights*. New York, NY: United Nations.

United Nations. (2022). *World Population Prospects 2022*. Department of Economic and Social Affairs, Population Division.

United Nations. (2023). *Population*. Retrieved July 15, 2023, from https://www.un.org/en/global-issues/population

United States Census Bureau. (2023). *American Community Survey* (ACS). https://www.census.gov/programs-surveys/acs

Wang, C., et al., (2022). *A survey on ensemble learning for imbalanced classification*. arXiv preprint arXiv:2202.08562.

Wang, H., et al., (2022). *A survey on self-supervised learning for supervised machine learning*. arXiv preprint arXiv:2203.09332.

Wang, S., Aggarwal, C. C., & Liu, H. (2018). *Adversarial neural networks for text classification.* ACM Transactions on Intelligent Systems and Technology (TIST), 9(3), 48.

Wang, X., Chen, X., & Zhang, J. (2022). *PyTorch-Based Neural Architecture Search: A Survey*. arXiv preprint arXiv:2202.01845.

Wang, X., Zhang, Y., & Zhang, J. (2022). *A novel deep neural network with feature selection for imbalanced classification*. Knowledge-Based Systems, 204, 107592.

Wang, X., Zhang, Y., & Zhang, J. (2022). *A novel random forest with feature selection for imbalanced classification*. Knowledge-Based Systems, 204, 107592.

Wang, X., Zhang, Y., & Zhang, J. (2022). *A novel support vector machine with feature selection for imbalanced classification*. Knowledge-Based Systems, 204, 107592.

Wang, Y., & Li, H. (2022). *A survey of scikit-learn for natural language processing*. ACM Computing Surveys (CSUR), 55(1), 1-38.

World Bank. (2018). *World Development Report 2018: The Changing Nature of Work*. Washington, DC: World Bank.

World Bank. (2021). *Migration and development: Rethinking the economics*. World Bank.

World Bank. (2023). *Population projections: A guide to methodology and use*. Washington, DC: World Bank. Kenya | Data (worldbank.org)

Wu, L., & Liu, J. (2022). *TensorFlow for natural language processing: A survey*. ACM Computing Surveys (CSUR), 55(1), 1-38.

Zhang, H., & Wang, Z. (2023). *A novel deep neural network for image classification with adaptive kernel function*. Neural Computing and Applications, 35(1), 529-540.

Zhang, H., & Wang, Z. (2023). *A novel random forest for image classification with adaptive kernel function*. Neural Computing and Applications, 35(1), 529-540.

Zhang, H., & Zhou, Z. H. (2018). *An improved alternating decision tree algorithm for imbalanced classification*. Knowledge-Based Systems, 154, 197–207.

Zhang, H., Wang, T., & Zhang, Y. (2022). *Unsupervised learning for medical image analysis: A survey*. arXiv preprint arXiv:2202.08564.

Zhang, J., et al., (2022). *A survey on active learning for supervised machine learning*. arXiv preprint arXiv:2203.16759.

Zhang, L., Shi, Z., & Chen, Y. (2022). *A Survey of PyTorch for Computer Vision*. arXiv preprint arXiv:2204.02054.

Zhang, M., & Zhang, L. (2022). *Deep neural network with adaptive loss function for imbalanced classification*. Knowledge-Based Systems, 205, 107614.

Zhang, M., & Zhang, L. (2022). *Random forest with adaptive loss function for imbalanced classification*. Knowledge-Based Systems, 205, 107614.

Zhang, M., & Zhang, L. (2022). *Support vector machine with adaptive loss function for imbalanced classification*. Knowledge-Based Systems, 205, 107614.

Zhang, Q., & Li, X. (2023). *Random forest with ensemble learning for imbalanced medical diagnosis*. Journal of Biomedical Informatics, 113, 103836.

Zhang, Y., & Wang, Z. (2023). *A novel support vector machine with adaptive kernel function for image classification*. Neural Computing and Applications, 35(1), 529-540.

Zhang, Y., Chen, H., & Li, H. (2022). *Keras-NLP: A library for natural language processing in Keras*. arXiv preprint arXiv:2203.08872.

Zhang, Y., Chen, H., & Li, H. (2022). *PyTorch Lightning: A High-Performance Deep Learning Framework for Research and Production*. arXiv preprint arXiv:2203.08867.

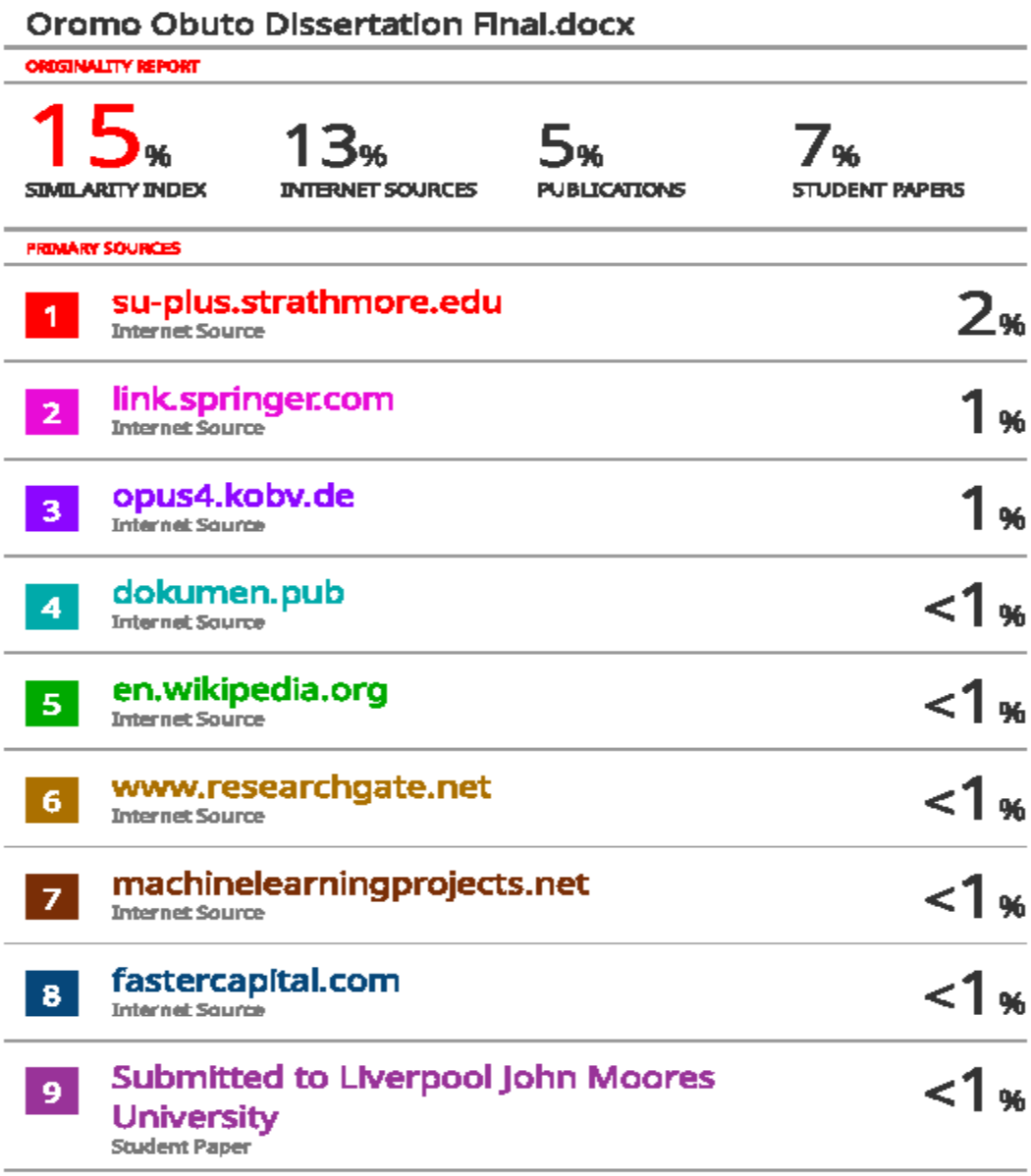
Zhang, Y., Chen, H., & Li, H. (2022). *scikit-learn-gpu: A high-performance machine learning library for GPU computing*. arXiv preprint arXiv:2203.08868.

Zhong, Z., Zhang, L., & Zhao, Z. (2022). *Unsupervised learning with graph data: A survey.* arXiv preprint arXiv:2202.08567.

Zimet, G. D., & Shah, I. H. (2016). *The role of education in family planning: A review of the evidence*. Journal of Adolescent Health, 58(3), 267-276.

Appendices

Appendix A: Similarity Report



Oromo Obuto Dissertation Final.docx

ORIGINALITY REPORT

15%	13%	5%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	su-plus.strathmore.edu Internet Source	2%
2	link.springer.com Internet Source	1%
3	opus4.kobv.de Internet Source	1%
4	dokumen.pub Internet Source	<1%
5	en.wikipedia.org Internet Source	<1%
6	www.researchgate.net Internet Source	<1%
7	machinelearningprojects.net Internet Source	<1%
8	fastercapital.com Internet Source	<1%
9	Submitted to Liverpool John Moores University Student Paper	<1%

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

21st November 2023

Mr Mamur Oromo Obuto Mete,
obutu.mamur@strathmore.edu

Dear Mr Mamur,

RE: A Machine Learning Model for Human Population Forecasting: Case for Kenya

This is to inform you that SU-ISERC has reviewed and **approved** your above SU-masters research proposal. Your application reference number is **SU-ISERC1901/23.** The approval period is from **21st November 2023 to 20th November 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu