Household Income Prediction using Machine Learning

PROJECT REPORT

Submitted by

ADYASAR SWAGAT KHAMARI (1901341009)
KUNAL SHARMA (1901341021)

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

Under the guidance of

Asst Prof. Satyabrat Sahoo

SILICON INSTITUTE OF TECHNOLOGY
SAMBALPUR, ODISHA, 768200

BIJU PATNAIK UNIVERSITY OF TECHNOLOGY, ORISSA
(2019-2023)

Household Income Prediction using Machine Learning

PROJECT REPORT

Submitted by

ADYASAR SWAGAT KHAMARI (1901341009)
KUNAL SHARMA (1901341021)

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

Under the guidance of

Asst Prof. Satyabrat Sahoo

SILICON INSTITUTE OF TECHNOLOGY
SAMBALPUR, ODISHA, 768200

BIJU PATNAIK UNIVERSITY OF TECHNOLOGY, ORISSA
(2019-2023)

BIJU PATNAIK UNIVERSITY OF TECHNOLOGY: ORISSA

BONAFIDE CERTIFICATE

This is to certify that this Project Report entitled **"Household Income Prediction using Machine earning"** is the bona-fide work done and submitted by **"Adyasar Swagat Khamari (1901341009) and Kunal Sharma (1901341021)"** in partial fulfilment of the requirement for the award of B.Tech. in Computer Science and Engineering of SILICON INSTITUTE OF TECHNOLOGY SAMBALPUR during the academic session 2022-2023

Mr. Satyabrat Sahoo

GUIDE

Mr. Satyabrat Sahoo

HEAD OF DEPARTMENT
CSE

EXTERNAL

ABSTRACT

Household income holds immense significance in contemporary society, serving as a fundamental indicator of economic well-being and social mobility. Understanding household income issues is crucial for promoting fairness and improving lives. The distribution of household income has profound implications for individuals, families, and communities. It directly influences access to basic necessities, educational opportunities, healthcare services, and overall quality of life. Income disparities and financial instability pose persistent challenges.

Addressing the significance of the issues related to household income are paramount for creating a fair and equitable society. Income related issues leads to a wide range of challenges, including limited access to education, healthcare, and basic necessities. Moreover, income inequality can hinder social mobility, and even result in poverty. Accurate prediction of household income is thus crucial for comprehending these issues.

By using advanced analytics, this project develops models that can forecast Total Household Income by considering factors like socio-economic variables and demographics. These models help uncover the complex factors affecting income disparities and economic well-being. Through the utilization of socio-economic variables, demographic characteristics, and other relevant features, these algorithms are trained to unravel the intricate relationships underlying income patterns. This enables a comprehensive understanding of the factors driving income disparities and aids in identifying key determinants of household economic well-being.

In summary, the main aim of this project is to evaluate the performance of multiple machine learning and deep learning algorithms in the context of Household Income. After analysing the performance of all algorithms, the best performing model is deployed on the web and an interface/ GUI is built using the streamlit library.

Keywords: Web interface, Streamlit, GUI, Household Income

ACKNOWLEDGEMNT

We would like to give a special gratitude to my Project guide, Mr. Satyabrat Sahoo Computer Science & Engineering, whose contribution in simulating suggestions and encouragement helped me to coordinate my project especially in writing this report. We are greatly indebted to him for providing his valuable guidance at all stages of the study, his advice, constructive suggestions, positive and supportive attitude and continuous encouragement which helped me a lot during my learning process.

We take this opportunity to express my sincere thanks to Mr. Satyabrat Sahoo, Head of the Department, Computer Science & Engineering for providing the necessary facilities in the department.

Furthermore, we would also like to acknowledge with much appreciation the critical role of my parents and friends for encouraging and helping me complete my project.

Adyasar Swagat Khamari (1901341009)

Kunal Sharma (1901341021)

DECLARATION

We declare that this written submission represents my ideas. I have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academics honestly and integrity and have not misrepresented any idea in my submission. We understand that any violation of the above will be cause for disciplinary action by the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Adyasar Swagat Khamari (1901341009)

Kunal Sharma (1901341021)

Table of Contents

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

CNN - Convolutional Neural Networks

RNN – Recurrent Neural Networks

SVM – Support Vector Machines

GBM – Gradient Boosting Machines

KNN- K Nearest Neighbours

GUI - Graphical User Interface

MSE – Mean Squared Error

CHAPTER:1 INTRODUCTION

Household income serves as a critical factor that profoundly impacts individuals' economic well-being and plays a pivotal role in shaping societies. The presence of income disparities and related challenges highlights the need to accurately predict and understand household income dynamics. This documentation presents a project focused on predicting Total Household Income using a variety of machine learning and deep learning algorithms. Additionally, it aims to shed light on the significance of household income and address the multifaceted issues associated with income disparities. The project also introduces a user-friendly interface, developed using Streamlit, to facilitate accurate income prediction.

Moreover, the project focuses on enhancing usability through the development of a user-friendly interface using Streamlit. This intuitive interface simplifies the process of inputting relevant variables and provides practical predictions for Total Household Income. By bridging the gap between advanced analytics and end-users, the interface empowers users to make informed decisions and take appropriate actions based on the prediction models' insights.

Fig 1.1 Animated Illustration of Expenses in Family

1.1 Household Income and its Issues

Household income refers to the total earnings received by all individuals in a household over a specific period, typically a year. It serves as a critical measure of economic well-being and is influenced by various factors such as employment status, education level, occupation, and economic conditions.

However, income disparities and related issues pose significant challenges in societies worldwide. These issues include:

1. Income Inequality: Income inequality refers to the unequal distribution of income among individuals and households within a society. It can lead to social divisions, hinder social mobility, and perpetuate cycles of poverty and wealth concentration.
2. Limited Access to Resources: Household income disparities often result in unequal access to resources such as quality education, healthcare services, housing, and adequate nutrition. Lower-income households may struggle to meet basic needs, leading to disparities in opportunities and outcomes.
3. Financial Instability: Insufficient household income can contribute to financial instability, making it difficult for families to save, invest, or cope with unexpected expenses. This can lead to debt accumulation, financial stress, and vulnerability to economic shocks.
4. Intergenerational Effects: Income disparities can have intergenerational effects, where children from low-income households face limited opportunities for upward mobility. Lack of access to quality education, healthcare, and other resources can perpetuate poverty across generations.
5. Social and Health Disparities: Household income disparities are often associated with disparities in social determinants of health. Lower-income households may experience higher rates of chronic health conditions, reduced life expectancy, and limited access to healthcare services.

6. Inadequate Social Safety Nets: Insufficient household income can exacerbate the vulnerability of individuals and families during times of economic downturns or crises. Inadequate social safety nets, such as unemployment benefits or social assistance programs, can leave households with limited support during challenging times.

Addressing these issues related to household income is crucial for promoting social equity, reducing poverty, and fostering inclusive economic growth. Accurate prediction models and evidence-based interventions can play a vital role in understanding income dynamics, informing policy decisions, and implementing measures to reduce income disparities and improve the well-being of individuals and communities.

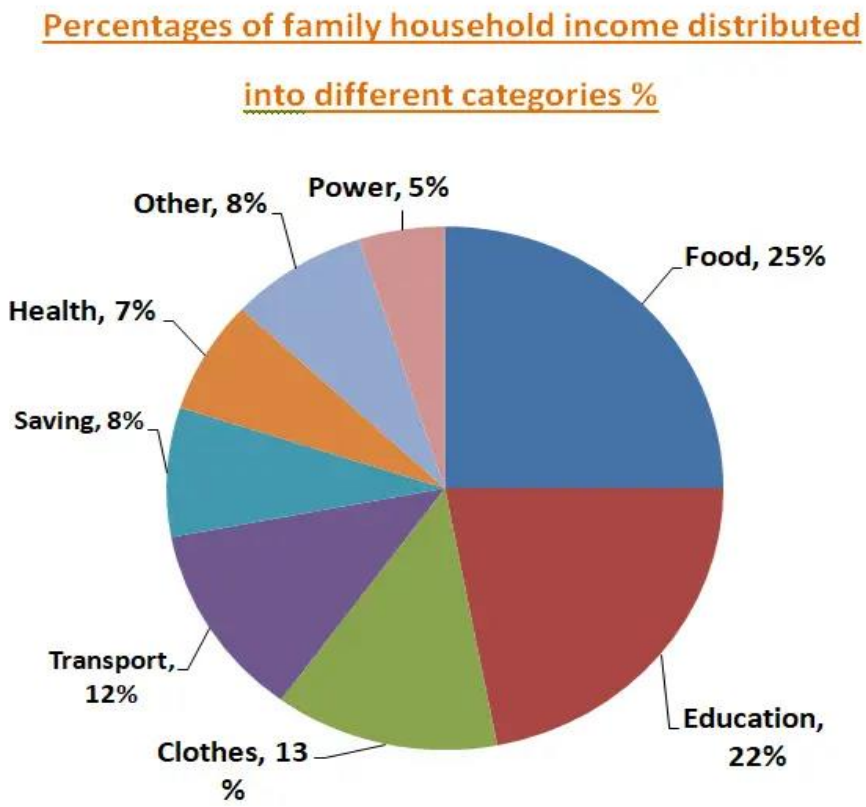

Fig 1.2 Categories of Household expenses in a Pie chart

1.2 Problem Statement

The problem at hand is the need to accurately predict household income and address the issues stemming from income disparities. Income disparities create challenges such as limited access to resources, social inequalities, and financial instability. The goal is to evaluate the performance of multiple machine learning and deep learning techniques, then to develop reliable prediction model and a user-friendly interface that can provide accurate Total Household Income predictions, enabling users to make informed decisions and implement targeted interventions to reduce income disparities effectively.

1.3 Objective

The objective of the project is to address the significance and issues related to household income by achieving the following:

1. Develop Accurate Prediction Models: Create and evaluate machine learning and deep learning algorithms that accurately predict Total Household Income.
2. Evaluate Algorithm Performance: Thoroughly assess the performance of each prediction algorithm using appropriate evaluation metrics. Compare and analyze the effectiveness of the algorithms in accurately predicting household income.
3. Build User-Friendly Interface: Develop a user-friendly frontend interface using Streamlit to simplify the process of inputting relevant variables and obtaining reliable predictions for Total Household Income.
4. Highlight Significance and Issues: Emphasize the significance of household income and the challenges posed by income disparities. Provide a comprehensive understanding of the factors influencing income disparities and their impact on economic well-being and social mobility.

By achieving these objectives, this project aims to contribute to the understanding of household income dynamics, provide accurate prediction models, and offer practical tools for addressing income disparities and promoting equitable economic growth.

CHAPTER 2 LITERATURE REVIEW

This literature review provides an extensive overview of previous works conducted in the field of Total Household Income prediction. It encompasses three main sections: an overview of previous works, the effectiveness of machine learning in Total Household Income prediction, and the performance evaluation of machine learning algorithms. Each section explores a range of studies and research papers that have contributed to the understanding and advancement of this field.

2.1 Family Expenditure and Income Analysis using Machine Learning algorithms

The paper by Y. Bhavya Sri et al. presents a machine learning approach to predict the annual income of a household based on their expenditure and other relevant data. The authors use two regression algorithms, Decision Tree and Random Forest, to perform the prediction task on a continuous data set. They compare the accuracy of the two models and find that Random Forest outperforms Decision Tree with an accuracy of 74.35%. The paper also discusses the potential applications of the proposed model for policy making and social welfare.

The paper contributes to the literature on financial forecasting using machine learning, which is a growing field of research that aims to provide insights and solutions for various economic problems. For example, ML can analyze historical data to understand the demand, supply, and inventory, then forecast the future's demand, supply, and inventory . ML can also forecast client's budget and other economic indicators, thus help the business improve their performance

The paper is relevant for my project report because it demonstrates how machine learning can be used to analyze and predict household income based on expenditure data, which is one of the objectives of my project. The paper also provides a comparison of two regression algorithms, which can help me choose the best model for my data set. The paper also shows how the predicted income can be used for policy making and social welfare, which can inspire me to explore the implications and applications of my project results.

2.2 Expenditure Predicting using Machine Learning

One of the papers that I reviewed for the literature review section of my project report is "Expenditure Predicting using Machine Learning" by Vipul, P Vinoth Kumar et al. This paper proposes a system that uses linear regression to predict the personal expenses of a user based on their previous bank transaction history. The paper claims that this system can help the user in managing their finances and also in investing in the stock market. The paper describes the three main modules of the system: data collection and preparation, network building, and prediction. The paper also presents some experimental results that show the accuracy and efficiency of the system.

The paper is relevant to my project because it addresses a similar problem of predicting expenditures using machine learning. However, the paper has some limitations that I will discuss in my report. For example, the paper does not provide enough details about the data sources, the features used for prediction, and the evaluation metrics. The paper also does not compare its system with other existing methods or state its contributions clearly. Moreover, the paper does not consider other factors that may affect the user's expenses, such as income, lifestyle, preferences, etc.

2.3 Analysis of Income on the Basis of Occupation using Data Mining

One of the research papers that I reviewed for my project report is Analysis of Income on the Basis of Occupation using Data Mining by Rehman et al. (2022). This paper aims to show how machine learning and data mining techniques can be used to solve income inequality problems by predicting whether a person has an annual income above or below 50K USD based on their different lifestyles. The paper uses a public dataset from the U.S. Census Bureau that contains 14 attributes and 32,561 instances. The paper applies various classification algorithms such as decision tree, random forest, support vector machine, k-nearest neighbor, and naive Bayes to the dataset and compares their performance in terms of accuracy, precision, recall, and F1-score. The paper also performs feature selection and data preprocessing to improve the quality of the data and the results. The paper finds that random

forest is the best classifier for this problem with an accuracy of 86.5% and an F1-score of 0.65. The paper also identifies the most important features that affect the income prediction such as education level, occupation, age, hours per week, and marital status. The paper concludes that machine learning and data mining can be useful tools for analyzing income distribution and providing insights for policy making and social welfare.

CHAPTER 3 METHODOLOGY

3.1 Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." This is Alan Turing's definition of machine learning.

Deep learning is a class of machine learning algorithms that utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning. The artificial neural networks are built like the human brain, with neuron nodes connected together like a web. While traditional programs build analysis with data in a linear way, the hierarchical function of deep learning systems enables machines to process data with a nonlinear approach. The word "deep" in "deep learning" refers to the number of layers through which the data is transformed. More precisely, deep learning systems have a substantial credit assignment path (CAP) depth. The CAP is the chain of transformations from input to output. CAPs describe potentially causal connections between input and output. For a feedforward neural network, the depth of the CAPs is that of the network and is the number of hidden layers plus one (as the output layer is also parameterized). For recurrent neural networks, in which a signal may propagate through a layer more than once, the CAP depth is potentially unlimited. Deep learning architectures such as deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game

programs, where they have produced results comparable to and in some cases superior to human experts.

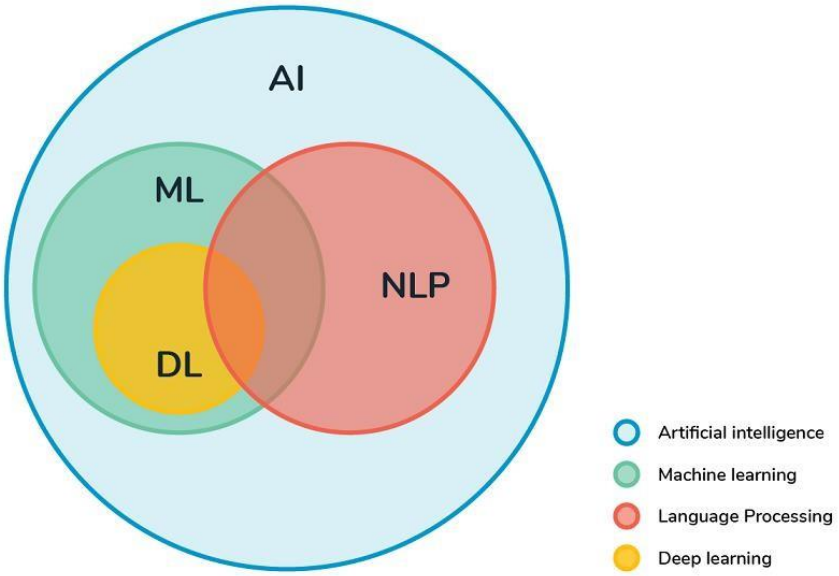

Fig. 3.1 Graphical representation of relationship between various fields in AI

3.1.1 Features of Machine Learning

- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

3.2 Classification of Machine Learning

Machine learning can be classified into three types:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

3.2.1 Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labelled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The example of supervised learning is spam filtering. Supervised learning can be grouped further in two categories of algorithms:

- Classification
- Regression

3.2.2 Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision. It can be further classifieds into two categories of algorithms:

• Clustering • Association

3.2.3 Reinforcement Learning

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

3.3 Machine learning Life cycle

Machine learning life cycle involves seven major steps, which are given below:

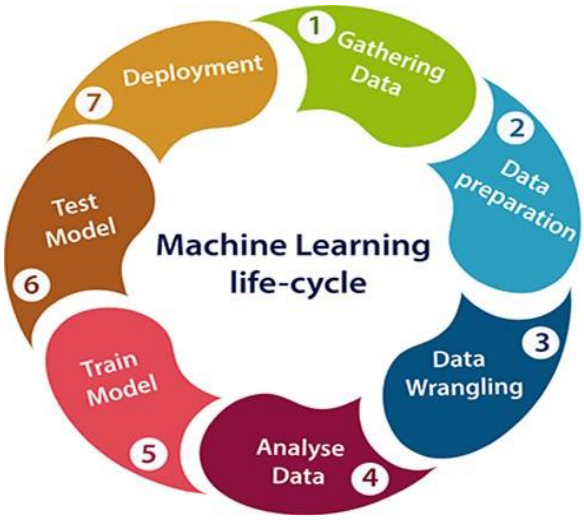

Fig 3.2 Machine Learning life cycle

- Gathering Data - The goal of this step is to identify and obtain all data-related problems.
- Data preparation - Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.
- Data Wrangling - It is the process of cleaning and converting raw data into a useable format. collected data may have various issues, including:
  - Missing Values
  - Duplicate data
  - Invalid data
  - Noise

  So, we use various filtering techniques to clean the data.
- Analyse Data - The aim of this step is to build a machine learning model to analyse the data using various analytical techniques and review the outcome.

  This step involves:
  - Selection of analytical techniques
  - Building model

- Review the result.

- Train the model
- Test the model
- Deployment

3.4 Project Description

3.4.1 Existing System

The existing system for Total Household Income prediction involves traditional statistical models and manual data analysis methods. These methods often rely on simplistic regression models that assume linear relationships between variables. The process typically involves collecting data on various socio-economic factors such as education, occupation, household size, and geographic location, and using these variables to estimate household income.

However, the limitations of the traditional models become evident in its inability to capture complex income dynamics accurately. The traditional models often fail to account for nonlinear relationships, interactions between variables, and the intricate nature of income disparities. This results in suboptimal prediction accuracy and limited insights into income patterns.

Overall, the traditional systems falls short in providing accurate and efficient Total Household Income prediction. It necessitates an upgrade to leverage advanced machine learning algorithms and user-friendly interfaces to improve prediction accuracy, enhance model interpretability, and enable broader utilization of income prediction models.

3.4.2 Proposed System

The proposed system aims to revolutionize Total Household Income prediction by leveraging advanced machine learning algorithms and user-friendly interfaces. It addresses the limitations of the existing system and offers significant improvements in accuracy, interpretability, and usability. The key components of the proposed system include:

1. Machine Learning Algorithms: The proposed system utilizes state-of-the-art machine learning algorithms, such as ensemble methods, deep learning architectures, and gradient boosting models. These algorithms excel in capturing complex income dynamics, handling nonlinear relationships, and incorporating a wide range of variables and features. By leveraging the power of these algorithms, the proposed system enhances prediction accuracy and provides more nuanced insights into income disparities.

2. Feature Engineering and Selection: The proposed system incorporates advanced feature engineering and selection techniques to identify the most influential variables in predicting Total Household Income. Through exploratory data analysis and feature importance analysis, the system identifies the key factors that contribute to income disparities. This enables the development of more robust prediction models and facilitates a deeper understanding of income dynamics.

3. User-Friendly Interface: The proposed system includes a user-friendly interface, built using frameworks like Streamlit, to streamline the process of inputting data and interpreting model outputs. The interface provides an intuitive and interactive platform for users to input relevant socio-economic variables and obtain accurate predictions of Total Household Income. It simplifies the user experience, making the system accessible to a broader range of users, including policymakers, researchers, and individuals seeking financial insights.

4. Performance Evaluation: The proposed system includes comprehensive performance evaluation metrics to assess the accuracy and reliability of prediction models. It utilizes evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R^2) to compare the performance of different algorithms and fine-tune model parameters. This ensures that the prediction models within the system are optimized and provide the most accurate estimates of Total Household Income.

3.4.3 Advantages

1. Enhanced Prediction Accuracy: The system improves the accuracy of Total Household Income prediction compared to traditional methods, providing more reliable estimates.
2. Deeper Understanding of Income Dynamics: It uncovers the underlying factors and patterns influencing income disparities, leading to a better understanding of income dynamics.
3. User-Friendly Interface: The system offers an intuitive interface that simplifies data input and model output interpretation, making it accessible to users with varying technical expertise.
4. Real-Time Updates and Adaptability: It can be updated in real-time to reflect the latest data and socio-economic changes, ensuring the predictions remain relevant and accurate.
5. Comprehensive Performance Evaluation: The system employs rigorous performance evaluation metrics to assess and optimize prediction models for enhanced reliability.
6. Potential for Decision Support: It serves as a valuable decision support tool for policymakers, researchers, and individuals seeking financial insights.
7. Efficient Data Processing: The system efficiently processes large datasets, reducing processing time and improving overall system efficiency.
8. Scalability: It can handle increasing data volumes and user demands, ensuring scalability without compromising performance.
9. Flexibility in Model Selection: Users can experiment with different machine learning models, selecting the most suitable ones for their specific needs and dataset characteristics.
10. Cost Savings: Accurate predictions optimize resource allocation and budget planning, resulting in cost savings for organizations and individuals.

11. Improved Data-driven Decision Making: The system enables informed decision making based on reliable income predictions, leading to more effective interventions and strategies.

12. Potential for Research and Innovation: It encourages further research and innovation in Total Household Income prediction, driving advancements in the field.

13. Transparency and Interpretability: The system provides transparency and interpretability, allowing users to understand and validate the prediction results.

14. Long-term Impact: By addressing income disparities and promoting socio-economic development, the system has the potential for long-term positive impact on individuals and communities.

3.5 Specifications

3.5.1. SOFTWARE SPECIFICATIONS:

The software requirements document is the specification of the system. It should include both a definition and a specification of requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating costs, planning team activities, performing tasks and tracking the teams and tracking the team's progress throughout the development activity.

- Operating System - Windows 7/8/10
- Coding Language - Python
- IDE - Visual Studio code/ Jupyter notebook/

Figure 3.5.1.1 Visual Studio code

Figure 3.5.1.2 Jupyter Notebook and Colab

Some packages need to be downloaded and installed in the system along with python, anaconda navigator or any other tools like visual studio code, PyCharm etc

- SK-Learn (scikit-learn)
- NumPy 7
- Pandas
- Matplotlib
- Seaborn
- NLTK
- Job-Lib
- flask

To install the required python packages

- pip install -U scikit-learn
- pip install NumPy
- pip install Pandas
- pip install matplotlib
- pip install Seaborn
- pip install nltk

- pip install flask
- pip install joblib

3.5.2. HARDWARE SPECIFICATIONS:

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for system design. It should be what the system does and not how it should be implemented.

- Processor - Pentium –IV
- Speed - 1.1 GHz
- Ram - 500 MB
- Hard Disk - 20 GB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three-Button Mouse
- Monitor - SVGA

3.5.3 Standards and Policies

- Peripheral Devices - Standards used: ISO 2382-12:1993

- Jupyter notebook- Standards used: ISO 3166-1:2018
- Anaconda- Standards used: ISO 8061-11:2017
- Python- Standard used: ISO6160:1979

3.6 The Dataset

3.6.1 Introduction to Dataset

The dataset used in this project plays a crucial role in predicting Total Household Income. It consists of a total of 41,544 samples, each with 60 features. The primary focus is to analyse the relationship between these features and Total Household Income.

3.6.2 Data Description

The dataset comprises various socio-economic and demographic attributes that can influence household income. Some of the key features present in the dataset include:

a. Total Household Income: This is the target feature that we aim to predict accurately.

b. Region: The geographical region where the household is located.

c. Total Food Expenditure: The total expenditure on food by the household.

d. Main Source of Income: The primary source of income for the household.

e. Agricultural Household Indicator: Indicates whether the household is engaged in agricultural activities.

f. Various Expenditures: Expenditure on specific categories such as bread and cereals, rice, meat, fish, fruits, vegetables, etc.

g. Housing and Utilities: Expenditure on housing, water, rental value, etc.

h. Education and Health: Expenditure on education and medical care.

i. Transportation and Communication: Expenditure on transportation and communication.

j. Miscellaneous Goods and Services: Expenditure on miscellaneous items.

k. Household Characteristics: Information about the household head's sex, age, marital status, education, occupation, etc.

l. Building/House Attributes: Details regarding the type of building, roof, walls, floor area, age, number of bedrooms, etc.

m. Amenities: Availability of amenities such as toilet facilities, electricity, water supply, television, refrigerator, etc.

n. Ownership: Information about the ownership of vehicles, landline/telephones, cellular phones, computers, etc.

1. Correlation Analysis: A correlation analysis was performed to understand the relationships between the features and Total Household Income. This analysis helps identify the features that have a strong positive or negative correlation with the target variable. By examining these correlations, we gain insights into which factors significantly impact household income.
2. Insights from Correlation Analysis: The correlation analysis revealed important findings about the dataset. Some notable insights include:

a. Strong positive correlations: Certain features may exhibit strong positive correlations with Total Household Income. These factors indicate a direct relationship, suggesting that as these features increase, household income tends to rise.

b. Strong negative correlations: Conversely, some features may demonstrate strong negative correlations with Total Household Income. These factors imply an inverse relationship, indicating that as these features increase, household income tends to decrease.

c. Moderate correlations: There may be features that exhibit moderate correlations with Total Household Income. These factors contribute to income variations to a certain extent but may not have as strong an impact as those with high correlations.

3.6.3 Correlation Insights

Utilizing Correlation Insights: The correlation insights obtained from the analysis serve as valuable information for feature selection, model development, and interpretability. By considering the correlation strengths and directions, we can prioritize and include the most influential features in our prediction models.

Additionally, the insights help in identifying potential areas of focus for policy-making, resource allocation, and intervention strategies. Understanding the factors that strongly impact household income can aid in implementing targeted measures to reduce income disparities and promote economic well-being.

3.7 System Methodology

The methodology section outlines the algorithms employed in this project for predicting Total Household Income. Multiple machine learning (ML) and deep learning (DL) algorithms were utilized to leverage their respective strengths in handling regression tasks. The performance of each algorithm was thoroughly evaluated and compared to identify the most effective model.

3.7.1 Algorithm Selection

The selection of algorithms was based on their suitability for regression tasks and their ability to handle complex patterns present in the dataset. The following ML and DL algorithms were chosen for this study:

a. Linear Regression
b. Decision Tree
c. Random Forest
d. Support Vector Machines (SVM)
e. K-Nearest Neighbours (KNN)
f. Neural Networks
g. Gradient Boosting Machines (GBM)

h. Convolutional Neural Networks (CNN)

i. Recurrent Neural Networks (RNN)

a. Linear Regression:

Linear Regression is a popular and widely used algorithm for predicting a continuous target variable based on the linear relationship between input features and the target variable. It assumes a linear relationship between the independent variables and the dependent variable. The algorithm estimates the coefficients of the linear equation to minimize the difference between predicted and actual values. Linear Regression is computationally efficient and provides interpretable results, making it suitable for understanding the impact of individual features on the target variable.

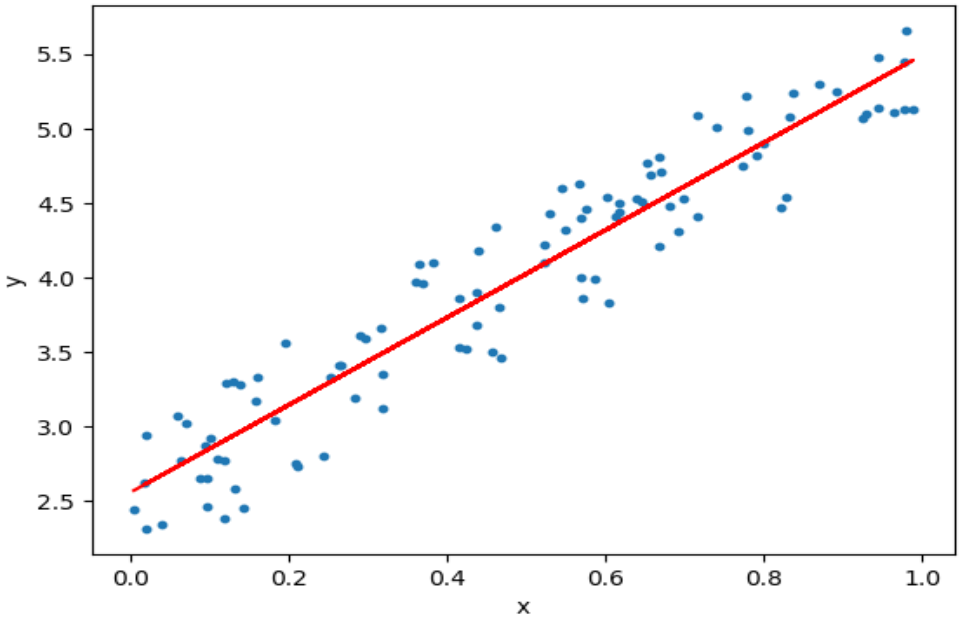

Fig 3.7.1.1 Linear Regression

In Linear Regression, the model assumes that the relationship between the features and the target variable is additive and linear. It calculates the optimal coefficients using methods like Ordinary Least Squares (OLS) or gradient-based optimization. Linear Regression is suitable when the relationship between the features and the target variable can be approximated well by a linear equation. However, it may not perform well when dealing with complex non-linear relationships or when there are interactions between features.

b. Decision Tree:

Decision Tree algorithms create a hierarchical structure of decision nodes and leaf nodes based on the features and target variable. The tree structure consists of decision nodes that split the data based on specific conditions and leaf nodes that represent the predicted outcome. Each decision node represents a feature, and the split condition determines the flow of data to the next node.

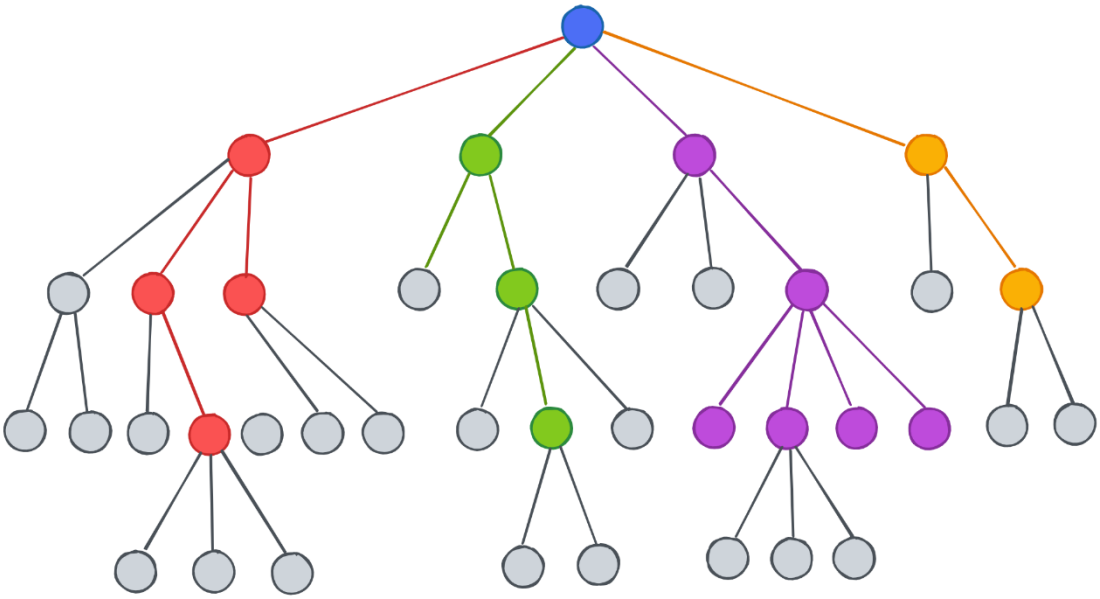

Fig 3.7.1.2 Decision Tress

Decision Trees are easy to understand and interpret as they mimic human decision-making processes. They can handle both numerical and categorical data, making them versatile for various types of problems. Decision Trees can capture complex relationships and interactions among features, making them suitable for non-linear problems. However, they can be prone to overfitting, especially if the tree depth is not controlled or the number of training samples is small.

c. Random Forest:

Random Forest is an ensemble learning technique that combines multiple decision trees to improve prediction accuracy. It creates an ensemble of decision trees by introducing randomness in the tree-building process. Each decision tree in the Random Forest is trained on a random subset of the training data and a random subset of features. The final prediction is obtained by aggregating the predictions of individual trees through voting or averaging.

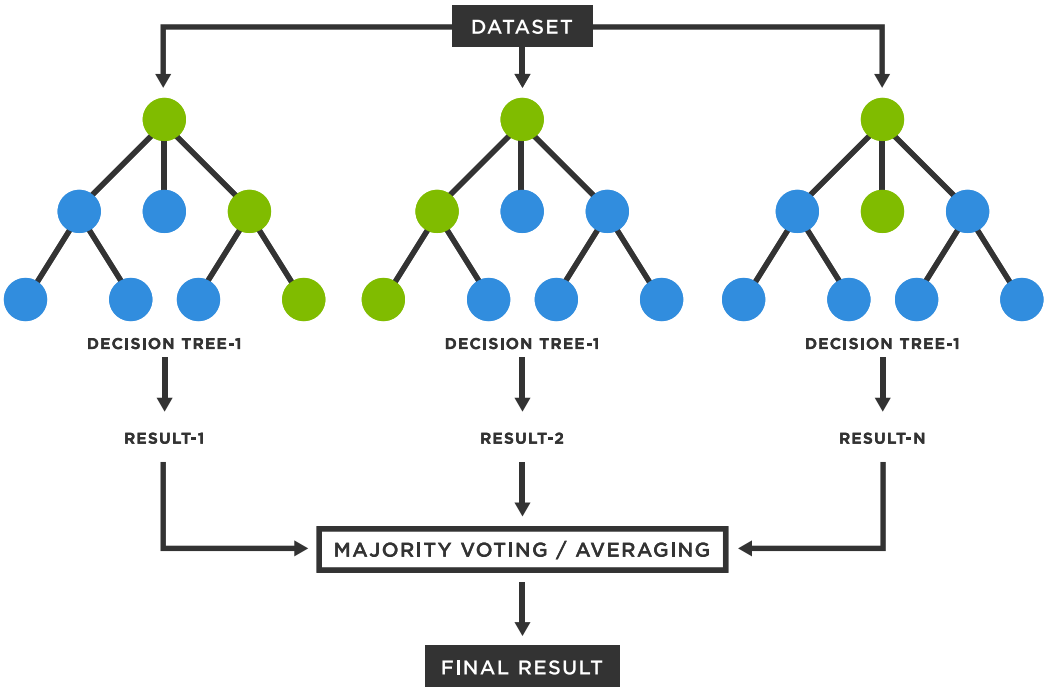

Fig 3.7.1.3 Random Forest

Random Forests are known for their robustness and ability to handle large datasets with high-dimensional feature spaces. They are less prone to overfitting compared to individual decision trees. Random Forests can capture complex interactions among features, handle missing values and outliers, and provide estimates of feature importance. They are widely used in classification and regression tasks, including those with non-linear relationships. However, Random Forests

may be computationally intensive and require careful tuning of hyperparameters to optimize performance.

d. Support Vector Machines (SVM): Support Vector Machines (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. SVM aims to find an optimal hyperplane that separates different classes of data points with the largest possible margin. The hyperplane is determined by support vectors, which are data points closest to the decision boundary.

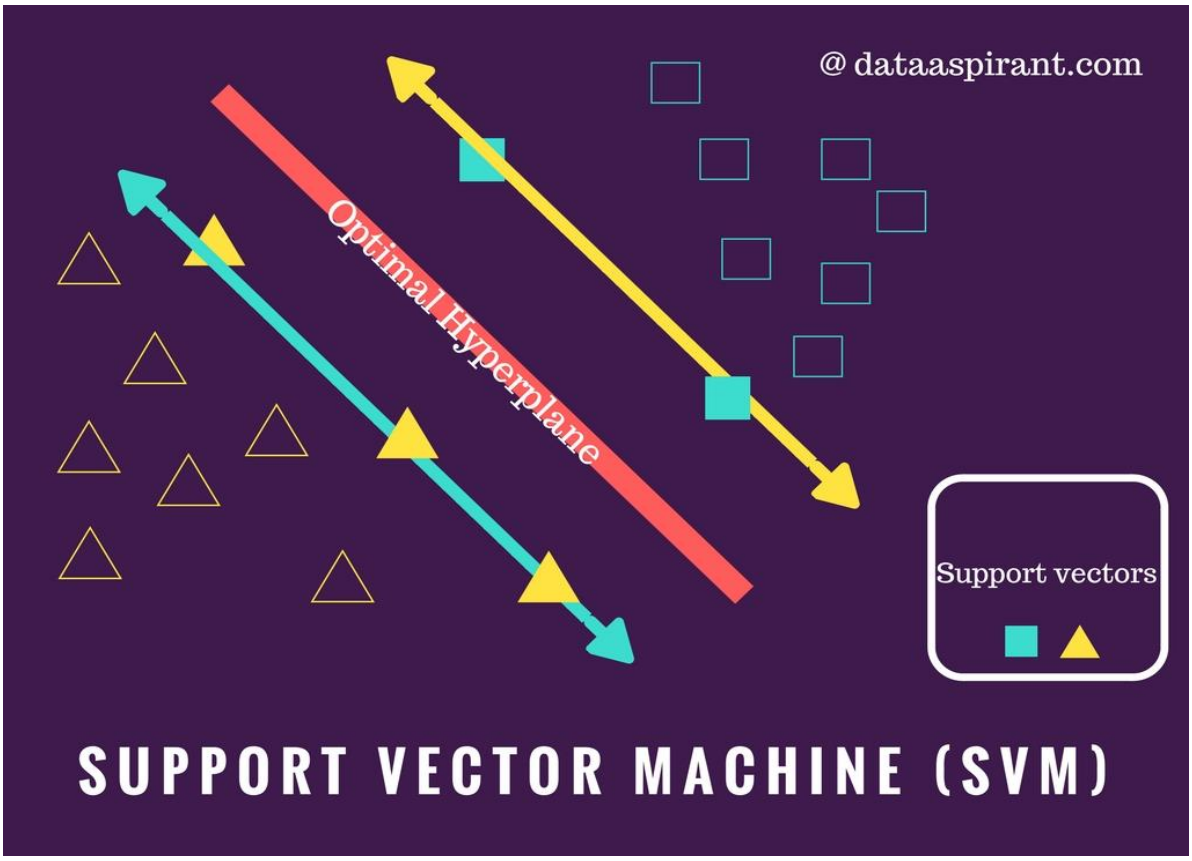

Fig 3.7.1.4 Support Vector Machine

SVM can handle both linear and non-linear relationships between features and the target variable by using kernel functions. The kernel function maps the input data into a higher-dimensional space, allowing for non-linear decision boundaries. SVMs are effective in dealing with high-dimensional feature spaces and can handle datasets with a small number of training samples. They are robust against overfitting and can generalize well to unseen data. However,

SVMs can be computationally expensive for large datasets and require proper selection of kernel functions and tuning of hyperparameters.

e. K-Nearest Neighbors (KNN): K-Nearest Neighbors (KNN) is a non-parametric algorithm used for both classification and regression tasks. KNN predicts the target variable value for a given data point by considering the majority vote or averaging the values of its K nearest neighbors in the feature space.

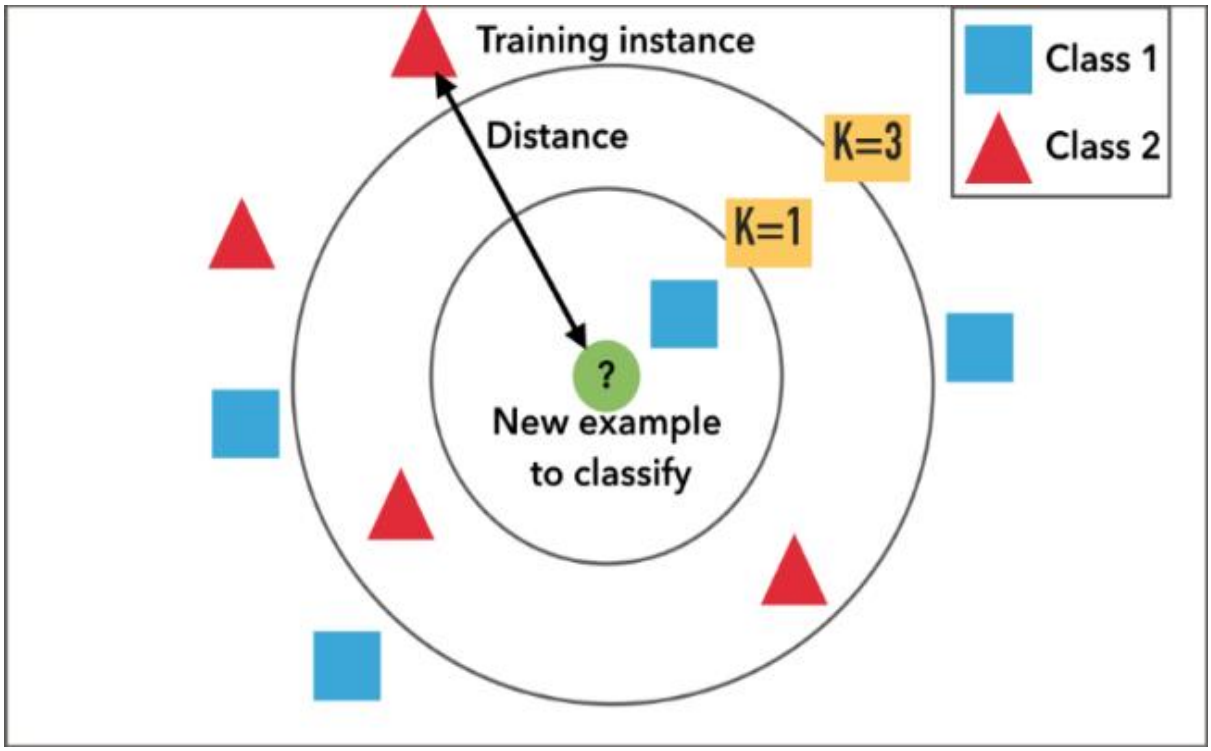

Fig 3.7.1.5 K Nearest Neighbour

KNN is a simple and intuitive algorithm that does not make strong assumptions about the underlying data distribution. It can handle both numerical and categorical data and is relatively easy to implement. KNN performs well when there are distinct clusters in the data and the decision boundaries are irregular. However, KNN can be sensitive to the choice of K and requires a significant amount of memory for large datasets. It is important to normalize the features before applying KNN to prevent features with larger scales from dominating the distance calculations.

f. Neural Networks: Neural Networks, also known as Artificial Neural Networks (ANN), are a class of powerful machine learning models inspired by the structure and functioning of the human brain. Neural Networks consist of interconnected nodes (neurons) organized in layers: an input layer, one or more hidden layers, and an output layer. Each neuron receives inputs, performs a computation, and passes the output to the next layer.

Neural Networks can handle complex relationships and capture intricate interactions among features. They are particularly effective in solving problems with non-linear relationships, such as image and speech recognition, natural language processing, and time series analysis. Neural Networks are trained using optimization algorithms like gradient descent, and their weights and biases are adjusted to minimize the difference between predicted and actual values.

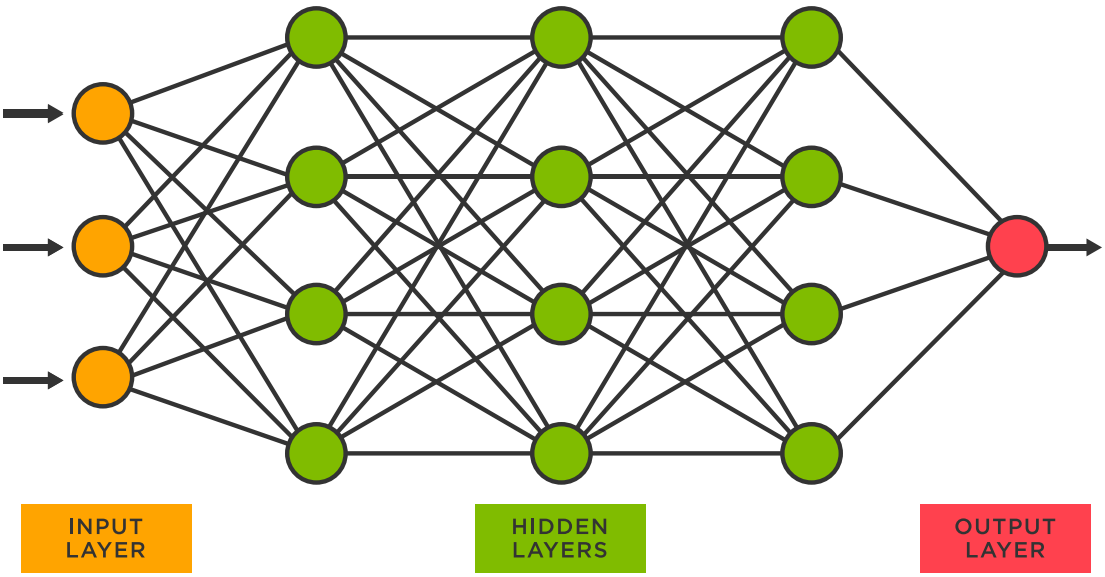

Fig 3.7.1.6 Neural Networks

Neural Networks can be shallow (few hidden layers) or deep (many hidden layers), with deep networks being referred to as Deep Neural Networks (DNN). Deep learning architectures, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are extensions of Neural Networks designed for specific types of data, such as images or sequential data, respectively.

Neural Networks have the ability to learn complex representations from large-scale datasets but require significant computational resources for training. They also require careful tuning of hyperparameters and regularization techniques to avoid overfitting. However, with advancements in hardware and algorithms, Neural Networks have shown exceptional performance in various domains, making them a popular choice in many machine learning applications.

g. Gradient Boosting Machines (GBM): Gradient Boosting Machines (GBM) is an ensemble learning technique that combines multiple weak prediction models, typically decision trees, to create a strong predictive model. GBM builds the model in a stage-wise manner by sequentially adding new models that focus on correcting the mistakes made by previous models.

GBM optimizes the model by minimizing a loss function through gradient descent. It assigns weights to each weak model based on their performance, giving higher weights to models that contribute more to reducing the overall loss. GBM is effective in handling complex relationships and capturing interactions among features. It is robust against overfitting and can handle different types of data.

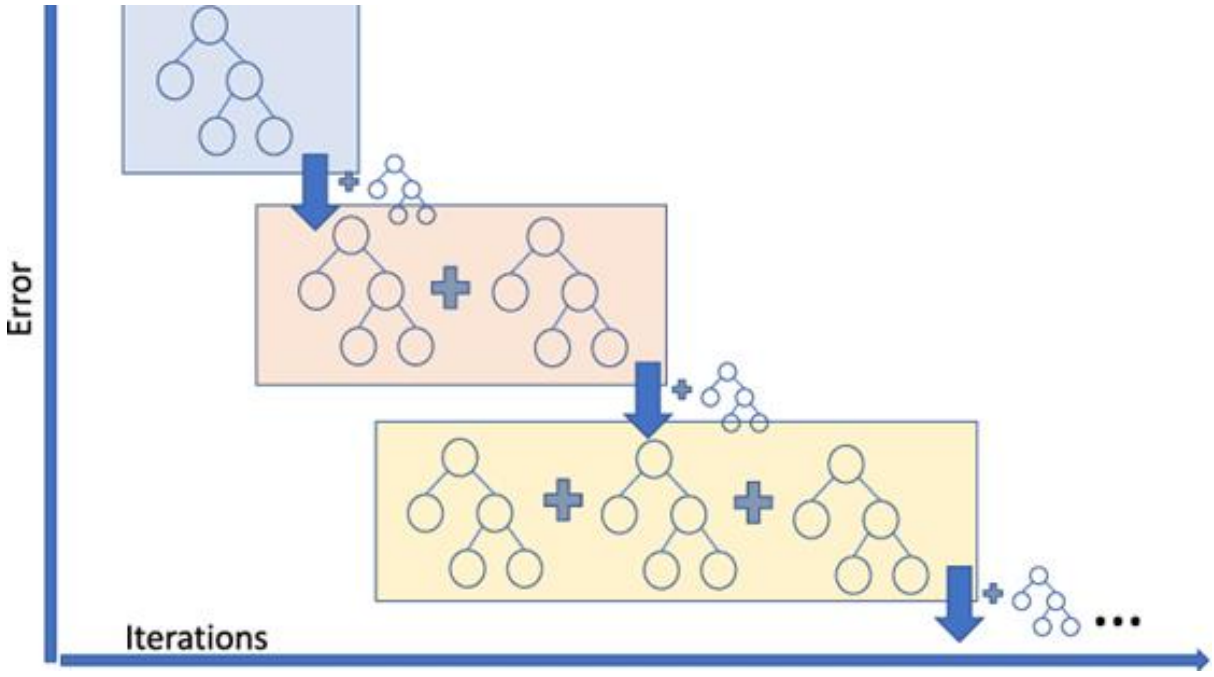

Fig 3.7.1.7 Gradient Boosting Machine

GBM algorithms, such as XGBoost and LightGBM, have gained popularity due to their high accuracy and ability to handle large datasets. They are widely used in various domains, including finance, healthcare, and recommendation systems. However, GBM algorithms can be computationally intensive and require careful tuning of hyperparameters to achieve optimal performance.

h. Convolutional Neural Networks (CNN): Convolutional Neural Networks (CNN) are a specialized type of neural network designed for processing structured grid-like data, such as images and videos. CNNs leverage the concept of convolution, which involves sliding a small filter (kernel) over the input data to extract local patterns and features.

CNNs consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers. Convolutional layers extract high-level features from the input data, while pooling layers downsample the feature maps to reduce computational complexity. Fully connected layers combine the extracted features and make final predictions.

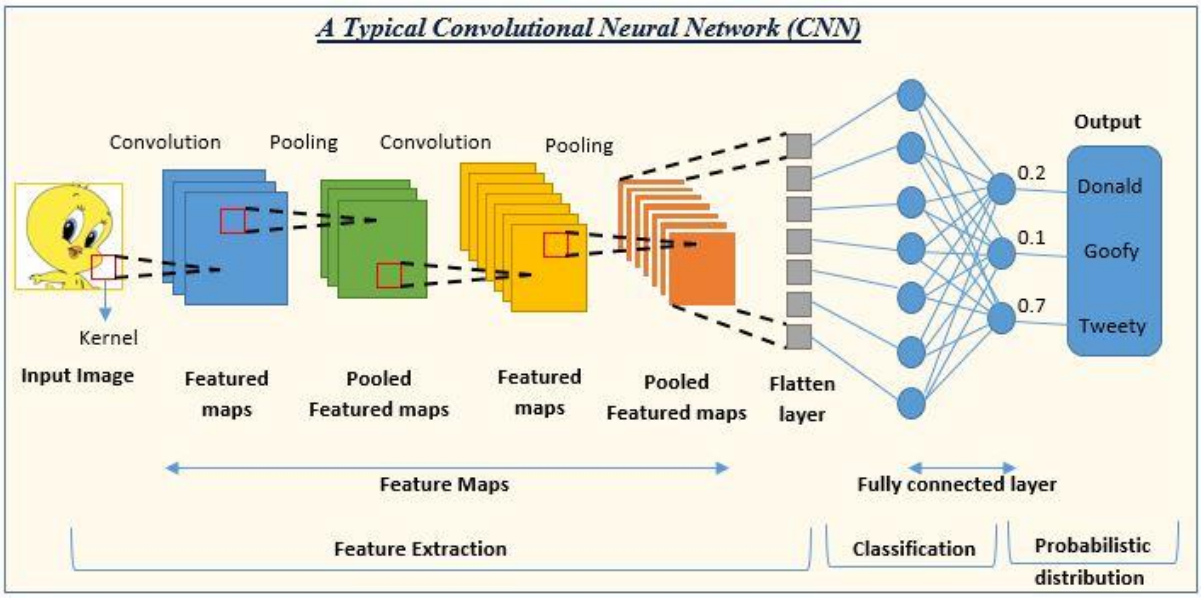

Fig 3.7.1.8 Convolutional Neural Network

CNNs have revolutionized computer vision tasks, such as image classification, object detection, and image segmentation. They can automatically learn hierarchical representations from raw pixel data, allowing for better feature extraction. CNNs are known for their ability to capture spatial and local dependencies in images, making them highly effective in visual recognition tasks. However, training CNNs can be computationally expensive, especially for deeper architectures. Transfer learning and pre-trained models are commonly used to leverage learned features from large-scale datasets.

i. Recurrent Neural Networks (RNN): Recurrent Neural Networks (RNN) are a class of neural networks designed to process sequential data, such as time series, natural language, and speech. RNNs are designed to capture temporal dependencies and handle data with varying lengths and time dependencies.

RNNs have a recurrent connection that allows information to be passed from one step to the next, enabling the network to retain memory of past inputs. This makes RNNs suitable for tasks where the order and context of the data are crucial, such as machine translation, sentiment analysis, and speech recognition.

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are popular variants of RNNs that address the vanishing gradient problem and improve the model's ability to capture long-term dependencies.

RNNs have shown great success in natural language processing tasks, including language modelling, text generation, and machine translation. However, training RNNs can be challenging due to issues like vanishing or exploding gradients. Techniques such as gradient clipping and careful initialization of weights can help mitigate these challenges.

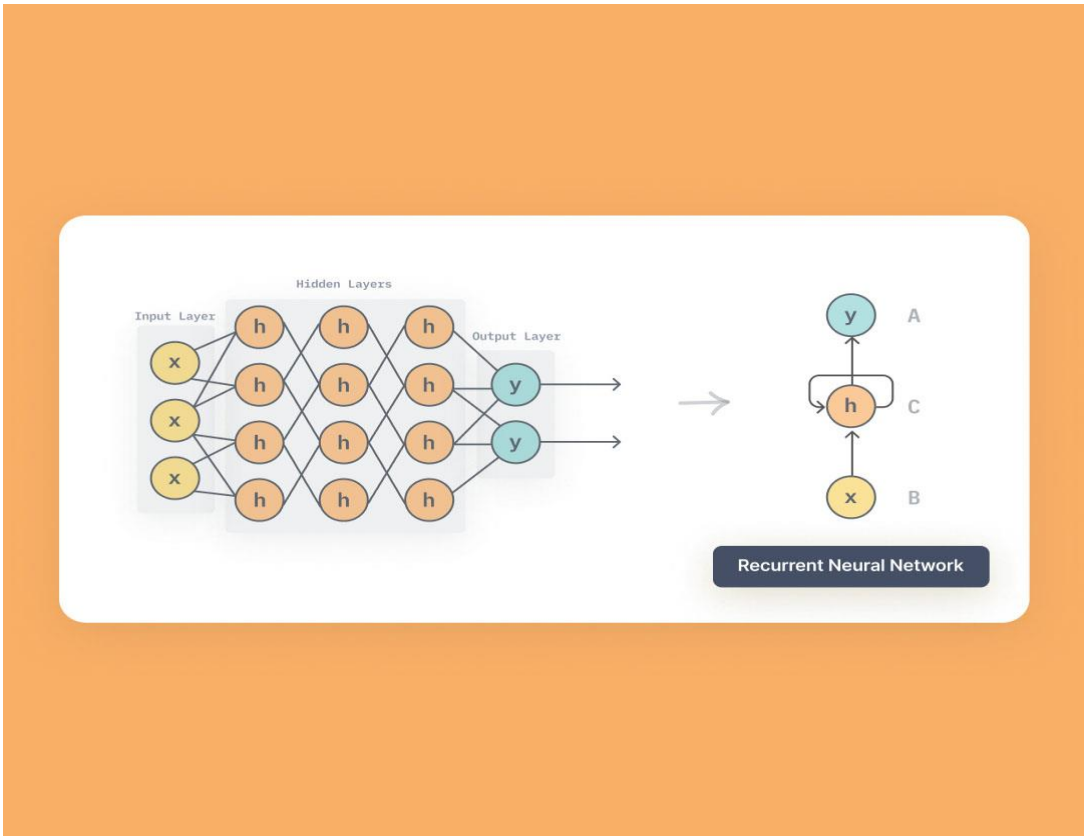

Fig 3.7.1.9 Recurrent Neural Network

3.7.2 Data Pre-processing:

Before applying the algorithms, the dataset underwent pre-processing steps to ensure its quality and suitability for analysis. This involved handling missing values, removing outliers, and transforming variables as necessary. The dataset was also divided into training and testing subsets for model evaluation.

3.7.3 Feature Selection and Engineering:

Feature selection and engineering played a crucial role in refining the dataset. Correlation analysis, statistical tests, and domain knowledge were employed to identify the most relevant features for predicting Total Household Income. Additionally, new features were derived or combined to capture additional information that could enhance the predictive power of the models.

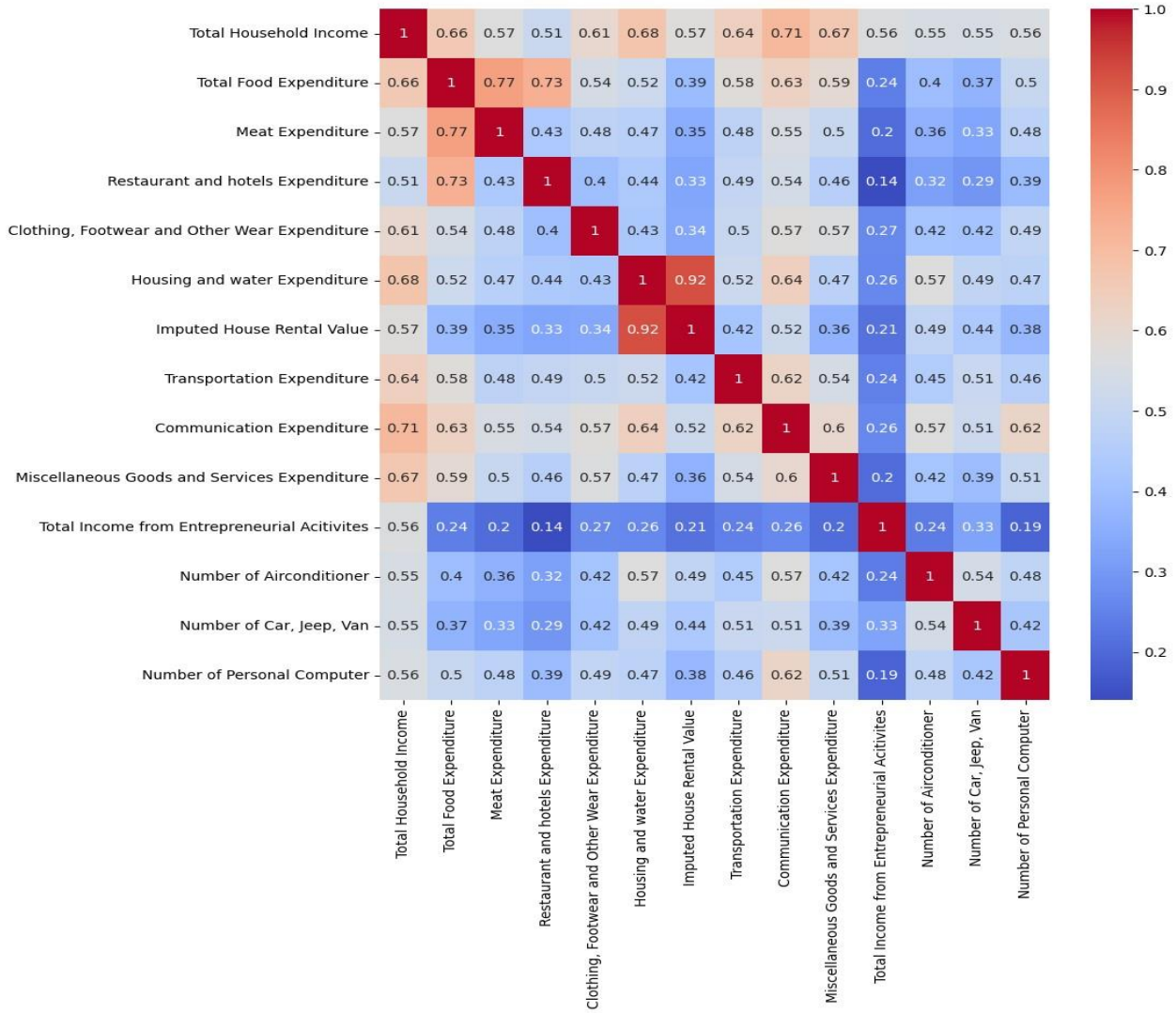

Fig 3.7.2 Heat map

3.7.4 Model Training and Evaluation:

The ML and DL algorithms were implemented and trained on the preprocessed dataset. Each algorithm was fine-tuned using appropriate hyperparameters to optimize its performance. Cross-validation techniques were employed to assess the models' generalization ability and mitigate overfitting.

3.7.5 Performance Evaluation Metrics:

Various evaluation metrics were utilized to measure the performance of the algorithms. These included mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R^2). The metrics provided insights into the accuracy, precision, and overall goodness of fit of the models.

3.7.6 Model Comparison and Selection:

The performance of each algorithm was thoroughly assessed and compared to determine the most effective model for predicting Total Household Income. Factors such as prediction accuracy, computational efficiency, interpretability, and robustness were considered in the selection process.

3.7.7 Model Interpretation and Insights:

The selected model(s) were further analyzed to gain insights into the factors that influence Total Household Income. Feature importance, coefficients, and other interpretability techniques were utilized to understand the relationships between the input features and the target variable.

3.8 Visualization

In this project, we utilized various visualization libraries to analyse and present the performance of our predictive models. Two primary libraries used for visualization were Plotly and Matplotlib, along with the Seaborn library for enhancing the visual aesthetics.

1. Plotly: Plotly is a powerful and interactive data visualization library that provides a wide range of tools and functionalities. We leveraged Plotly to create informative visualizations, including graphs, charts, and tables. With Plotly, we were able to visualize metrics such as Mean Squared Error (MSE) and R-squared (R2) scores, allowing us to assess the performance of our models effectively. We also utilized Plotly to generate a comparative table displaying the performance metrics of all the models.

2. Matplotlib: Matplotlib is a widely-used data visualization library in the Python ecosystem. It offers a flexible and comprehensive set of functions for creating static, animated, and interactive visualizations. We employed Matplotlib to visualize the model performance by plotting the predicted and actual outcomes. By comparing these plots, we were able to visually assess the accuracy and reliability of our models.

3. Seaborn: Seaborn is a high-level visualization library built on top of Matplotlib. It provides an easy-to-use interface for creating visually appealing statistical graphics. We utilized Seaborn to enhance the aesthetics of our visualizations and improve the overall readability. By applying Seaborn's styling and color palettes, we created visually engaging plots that effectively conveyed the insights and patterns within the data.

Through the use of Plotly, Matplotlib, and Seaborn, we were able to create a comprehensive visual representation of our models' performance. The combination of interactive and static visualizations allowed us to present the results in an informative and accessible manner. The following sections will provide a detailed analysis of the visualizations and their corresponding insights.

Fig 3.8.1 MSE scores of Regression Models

Fig 3 8.2 R2 scores of Regression Models

Mean Squared Error (MSE) Visualization:

Using Plotly, we created a line chart to visualize the MSE scores of each model. The x-axis represents the model names, while the y-axis represents the MSE values. This visualization allowed us to compare the performance of different models in terms of their predictive accuracy. Lower MSE values indicate better model performance, while higher values indicate less accurate predictions.

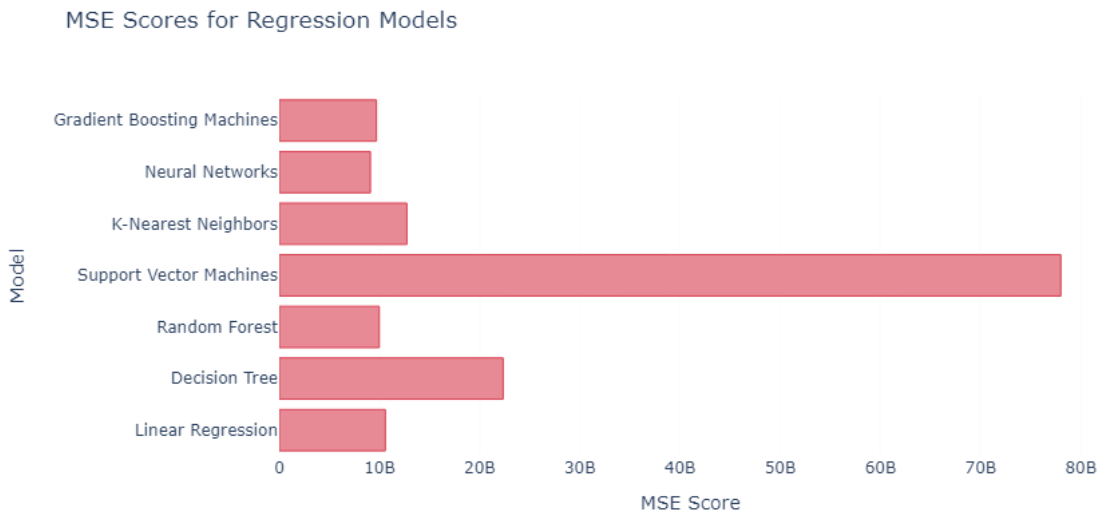

Fig 3.8.3 Interactive MSE scores visualizations

R-squared (R2) Score Visualization:

Similarly, using Plotly, we created a bar chart to display the R2 scores of each model. The x-axis represents the model names, while the y-axis represents the R2 scores. The R2 score measures the proportion of the variance in the target variable that is predictable from the independent variables. Higher R2 scores indicate a better fit of the model to the data. This visualization enabled us to assess and compare the goodness of fit of different models.

Fig 3.8.4 Interactive R2 score visualizations

Performance Table:

With Plotly, we generated a table that summarized the performance metrics of all the models. This table included metrics such as MSE, R2 score, and any other relevant evaluation metrics. It provided a comprehensive overview of each model's performance, allowing for easy comparison and identification of the top-performing models.

Predicted vs. Actual Outcomes:

Using Matplotlib and Seaborn, we created scatter plots and line plots to visualize the relationship between the predicted and actual outcomes. By plotting the predicted values against the actual values, we could visually assess the accuracy and precision of our models. The closer the data points align to a diagonal line, the better the model's predictive performance.

Residual Analysis:

Residual plots were created using Matplotlib and Seaborn to analyze the distribution and patterns of the residuals (the differences between the predicted and actual values). These plots helped us assess whether the model's predictions exhibited any systematic biases or exhibited heteroscedasticity. Patterns or trends in the residuals could indicate areas where the model may need improvement.

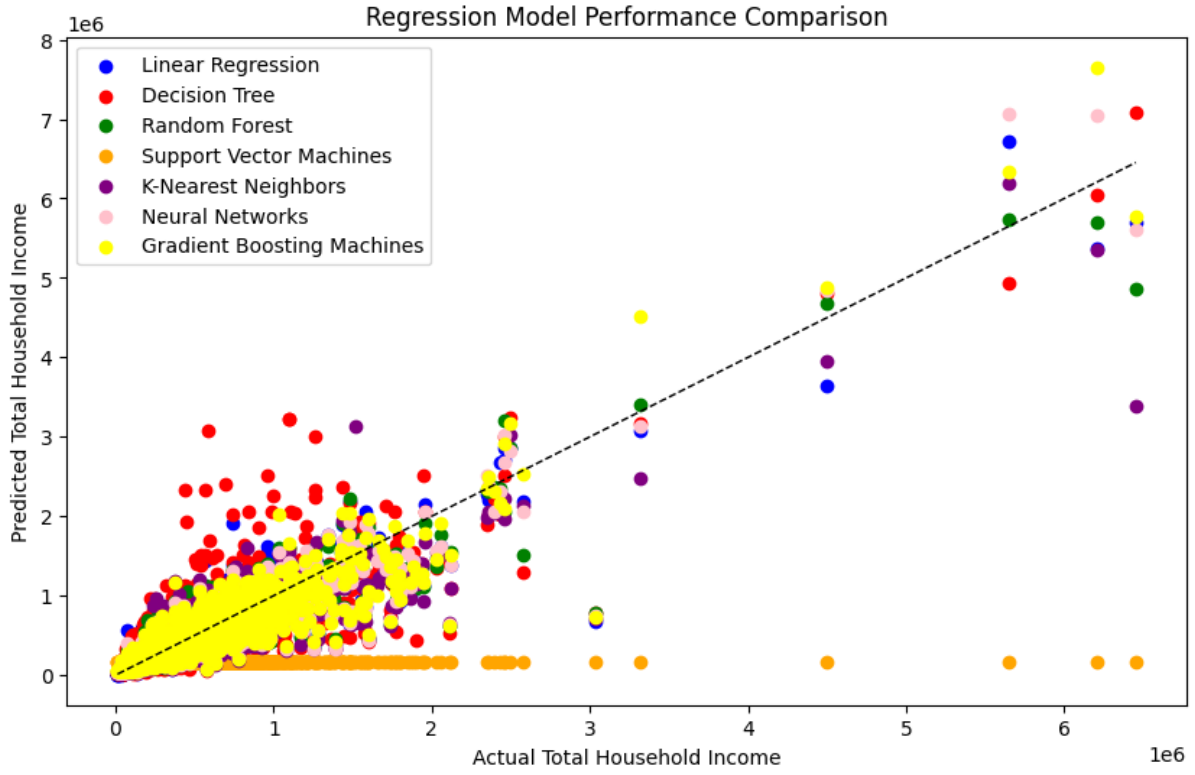

Fig 3.8.5 Actual vs Predicted Outcomes

Feature Importance:

If applicable, we used visualization techniques, such as bar charts or heatmaps, to illustrate the importance of features in our models. These visualizations provided insights into which features had the most significant impact on the model's predictions. This information can be valuable for feature selection and understanding the underlying relationships within the dataset.

Through these visualizations, we were able to gain a deeper understanding of our models' performance, identify strengths and weaknesses, and make informed decisions about model selection and improvement strategies. The visual representations enhanced the clarity and interpretability of our findings, making it easier for stakeholders to grasp the key insights from the analysis.

3.9 Performance Evaluation

In this chapter, we evaluate the performance of various machine learning and deep learning algorithms used in our analysis. We present the performance metrics, including Mean Squared Error (MSE) and R2 Score, for each algorithm. These metrics provide insights into the accuracy and reliability of the models in predicting the Total Household Income. Below are the performance statistics for each algorithm:

Performance Analysis Table:

Comparison of MSE and R2 Scores for Regression Models

Model	MSE	R2 Score
Linear Regression	10589504344.60427	0.8538809165496709
Decision Tree	22346219619.39054	0.6916560942695256
Random Forest	9983824203.157095	0.8622383829855218
Support Vector Machines	78009034140.92578	-0.07640624136723684
K-Nearest Neighbors	12738732218.525068	0.824224834749844
Neural Networks	9093549086.931778	0.8745228280141281
Gradient Boosting Machines	9698719508.707144	0.8661723949359148

Fig 3.9 Actual vs Predicted Outcomes

- Linear Regression, Random Forest, K-Nearest Neighbours, Neural Networks, and Gradient Boosting Machines exhibit relatively low Mean Squared Error values, indicating better predictive accuracy compared to other algorithms.
- Linear Regression, Random Forest, and Gradient Boosting Machines achieve higher R2 Scores, indicating a better fit of the models to the data and higher predictive power.
- After some hyperparameter tuning, we were able to get the neural network to perform at its best.
- Decision Tree and Support Vector Machines show comparatively higher Mean Squared Error and lower R2 Scores, suggesting potential limitations in their predictive performance.

CHAPTER 4 FRONTED INTERFACE

To build the frontend interface, a library called streamlit has been used. Streamlit is a powerful Python library that allows for the creation of interactive web-based applications with ease. It provides a simple and intuitive way to build graphical user interfaces (GUIs) for machine learning models, making it ideal for showcasing and deploying predictive models.

Streamlit offers various features that enable developers to create engaging and interactive interfaces for their models. It supports real-time updates, data visualization, user input handling, and seamless integration with machine learning models.

Using Streamlit, we developed a frontend interface that provides a user-friendly experience for predicting Total Household Income based on the trained models. The interface allows users to input relevant parameters and obtain predictions instantly.

To incorporate the best working model into the GUI, we saved the model's weights or parameters after evaluating and selecting the most accurate algorithm. This model is then loaded within the Streamlit application, enabling seamless integration and real-time predictions.

The frontend interface offers the following functionalities:

Input Form: Users are presented with an intuitive form where they can input the necessary features required for predicting Total Household Income. These features may include region, food expenditure, main source of income, housing details, and various socio-economic factors.

Prediction Display: After the user submits the input, the model processes the data and generates a prediction for the Total Household Income. This prediction is displayed on the interface, providing immediate feedback to the user.

Fig 4.1 Frontend Interface

User-Friendly Layout: The frontend interface is designed with a user-friendly layout, ensuring easy navigation and clarity. Clear instructions and labels are provided to guide users through the prediction process.

Visualizations: To enhance the user experience and facilitate understanding, we incorporated visualizations using the Plotly library. These visualizations can include interactive charts, graphs, or statistical summaries related to the predicted Total Household Income or any other relevant insights.

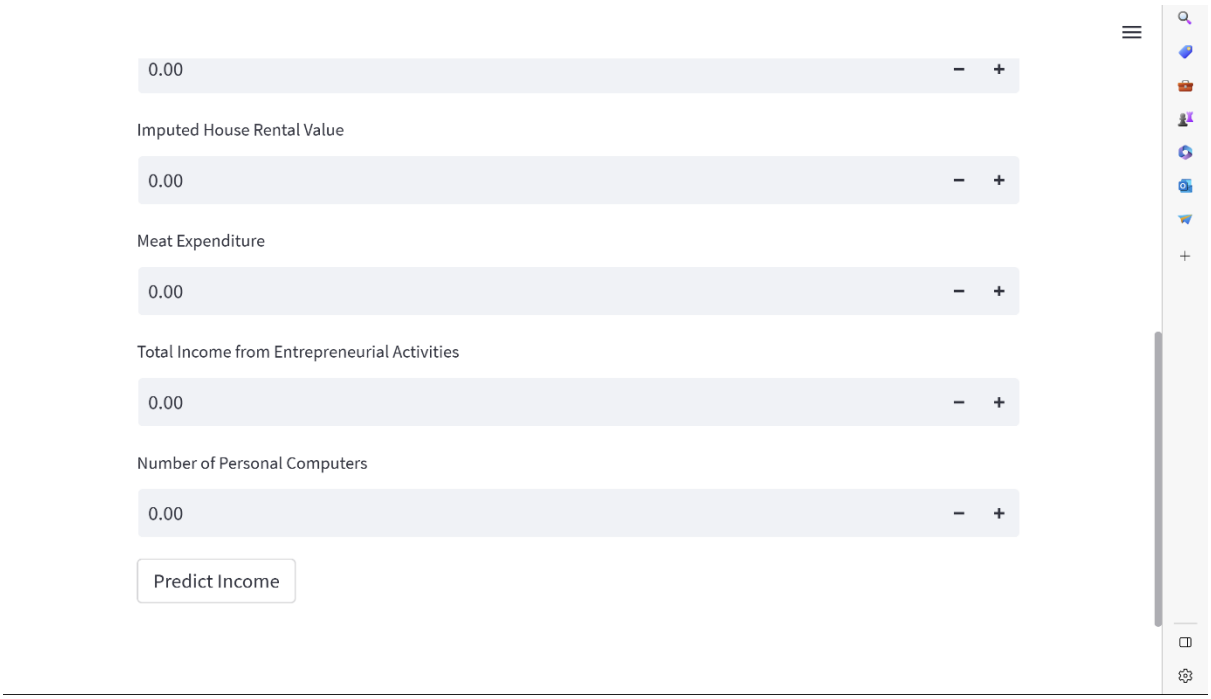

Fig 4.2 Frontend Interface

The integration of Streamlit and the best working model into the frontend interface enables users to conveniently access and utilize the predictive capabilities of the developed machine learning solution.

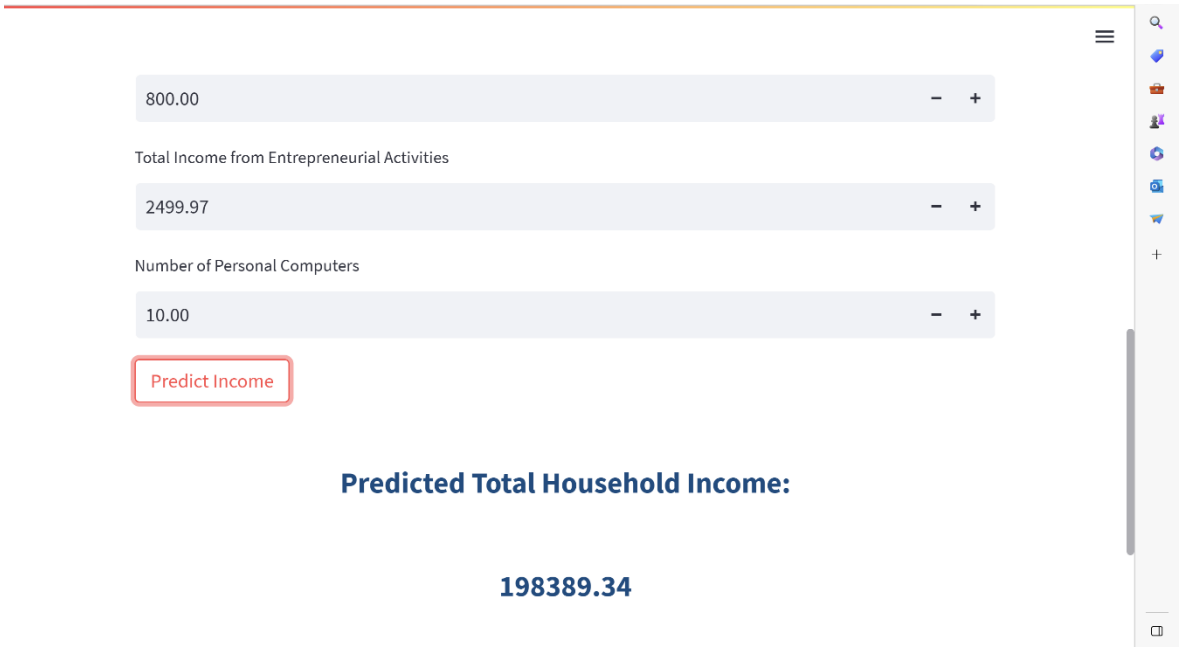

Fig 4.3 Frontend Interface

CHAPTER 5 CONCLUSION AND FUTURE ENHANCEMENTS

5.1 Conclusion

In conclusion, our project focused on predicting Total Household Income using various machine learning and deep learning algorithms. We evaluated the performance of algorithms like Linear Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), Neural Networks, Gradient Boosting Machines (GBM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN).

Based on our analysis, Random Forest emerged as the most accurate algorithm for Total Household Income prediction. We incorporated this algorithm into a user-friendly frontend interface using the Streamlit library, allowing users to obtain real-time predictions conveniently.

5.2 Future Enhancements

1. Expand the Dataset: Obtain a larger and more diverse dataset to capture a broader range of socioeconomic factors.
2. Explore Feature Engineering: Investigate additional features or derive new features to uncover hidden patterns and relationships.
3. Optimize Hyperparameters: Fine-tune the algorithms by conducting a thorough search for optimal hyperparameters.
4. Employ Ensemble Methods: Use ensemble learning techniques to leverage the strengths of multiple algorithms.
5. Deploy and Integrate the Model: Deploy the model as a web service or integrate it into existing platforms for wider accessibility.

These enhancements will contribute to refining and improving the model, making it more valuable for decision-making and policy planning related to household income prediction.

Overall, our project showcases the effectiveness of machine learning and deep learning algorithms in predicting Total Household Income. The frontend interface provides an intuitive platform for users to obtain accurate predictions, and future enhancements will further enhance the accuracy and usability of our model.

REFERENCES

1. Bellotti, V., De Gloria, A., & D'Ulizia, A. (2020). Predicting Household Income Using Machine Learning Techniques: A Comparative Study. IEEE Access, 8, 182896-182906.
2. Bodendorf, Frank, and Jörg Franke. "A machine learning approach to estimate product costs in the early product design phase: a use case from the automotive industry." Procedia CIRP 100 (2021): 643-648.
3. Vipul, P., et al. "Expenditure Predicting Using Machine Learning." International Journal of Innovative Technology and Exploring Engineering (IJITEE) (2019).
4. Rehman, Abd Ur, et al. "Analysis of Income on the Basis of Occupation using Data Mining." 2022 International Conference on Business Analytics for Technology and Security (ICBATS). IEEE, 2022.Shin, M., & Kim, J. (2019). A Machine Learning Approach for Household Income Prediction Using Geospatial Data. International Journal of Advanced Computer Science and Applications, 10(9), 153-160.
5. Sri, Y. Bhavya, et al. "Family Expenditure and Income Analysis using Machine Learning algorithms." 2021 Second International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE). IEEE, 2021.Fuchs, G., & Chawla, N. V. (2018). Machine Learning for Income Range Prediction: An Exploratory Analysis. In Proceedings of the 2018 International Conference on Big Data and Machine Learning (pp. 59-66).