ADULT INCOME EXPLORATORY DATA ANALYSIS AND PREDICTIVE ANALYSIS

Submitted by
Pavani Suresh

TABLE OF CONTENTS

1. INTRODUCTION

The 1994 Adult Census Income dataset serves as an essential tool for socio-economic research, reflecting the complexities of financial realities in mid-1990s America. It provides a snapshot of the population's income distribution alongside various socio-demographic attributes. The dataset was specifically created to facilitate machine learning and statistical modeling by providing a clean, pre-processed collection of data points that represent individual adults in the United States. Its creation was part of an effort to support academic research and development in predictive analytics, offering a real-world dataset that poses a challenging yet accessible problem for statistical classifiers and predictive models.

1.1 Background

The 1994 Adult Census Income dataset originates from the U.S. Census Bureau and was meticulously extracted from the 1994 U.S. Census database by Ronny Kohavi and Barry Becker. This dataset has become a staple in the data science community, particularly for those interested in exploring how demographic and employment-related factors correlate with income levels. Its widespread use is primarily aimed at binary classification problems, where the goal is to predict whether an individual earns more than $50,000 per year.

1.2 Objective

The main objective of delving into this dataset is to uncover patterns and insights that can predict income levels based on a variety of factors including age, education, race, gender, work class, and more. Such analyses are crucial for several applications: economists can glean valuable information for policy-making, educational institutions and workforce development programs can tailor their initiatives to better prepare individuals for high-earning roles, and social scientists can study the impact of various socio-demographic factors on economic success.

1.3 Hypothesis

Several hypotheses can be formulated and tested with this dataset. For instance, it is hypothesized that there is a strong correlation between educational attainment and income, where individuals with higher education levels are more likely to earn above $50,000. Similarly, the type of work class might influence income levels, with potentially different outcomes for those in private versus government sectors. Age might also play a significant role, with earnings increasing as individuals gain more work experience. Moreover, the analysis could explore the gender wage gap, hypothesizing that males might earn more than females, and assess income disparities across different races.

1.4 The 1994 Adult Census Income Dataset in Terms of V5 Model

The 1994 Adult Census Income dataset, while valuable for research and analysis, may not typically meet the strictest definitions of "Big Data" when considered in the classical sense of the term. Big Data is often characterized by its size (Volume), the speed at which it is created and needs to be processed (Velocity), the range of data types and sources it encompasses (Variety), the reliability and accuracy of the data (Veracity), and the issues related to security and privacy (Vulnerability). Let's analyze each of these characteristics in the context of the 1994 Adult Census Income dataset:

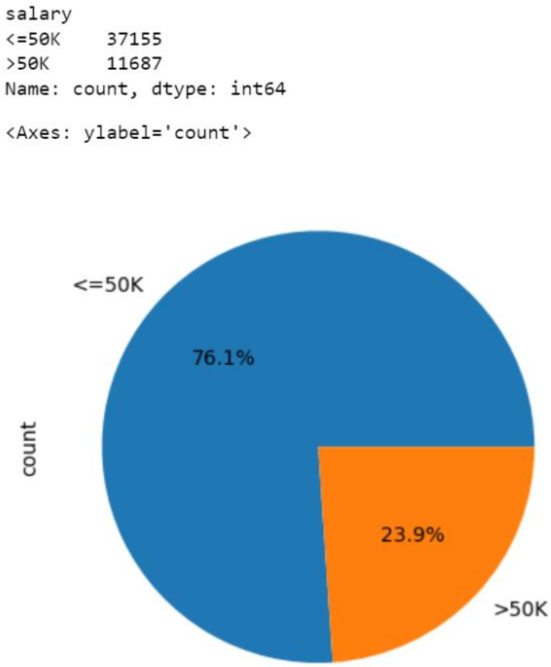
i. **Volume:** The dataset contains over 48,000 entries, each with multiple attributes such as age, work class, education, marital status, occupation, relationship, race, gender, capital gain, capital loss, hours per week, and income. While sizable, this does not necessarily qualify as "Big Data" compared to datasets that are comprised of millions to billions of records commonly seen in contemporary Big Data scenarios.

ii. **Velocity:** The dataset is a static snapshot from the 1994 census and does not continually generate data. There is no ongoing, real-time data input or update, which contrasts with the high-velocity nature of Big Data, where data streams continuously and at high speed (e.g., data from internet of things devices, online transactions).

iii. **Variety:** This dataset has a structured format with predefined types, which is less complex compared to Big Data environments that often include a wide variety of data types and structures — from structured numeric data to unstructured text, videos, and images.

iv. **Veracity:** The 1994 Adult Census Income dataset is likely to have a high degree of veracity as it is derived from the U.S. census, which typically involves rigorous data collection and processing methods to ensure accuracy. However, like any dataset, it can contain errors or biases depending on how the data was collected and processed.

v. **Vulnerability:** Any dataset that includes personal information can be vulnerable to privacy issues and security breaches. The Adult Census Income dataset includes potentially sensitive information (e.g., income levels), which needs to be handled with care to ensure privacy and compliance with data protection laws. However, the public availability of the dataset indicates that identifiers likely have been removed to mitigate privacy concerns, which is a common practice in data release for research purposes.



```
salary
<=50K    37155
>50K     11687
Name: count, dtype: int64

<Axes: ylabel='count'>
```

2. DISCUSSION OF METHODS

2.1 Data Collection and Preprocessing

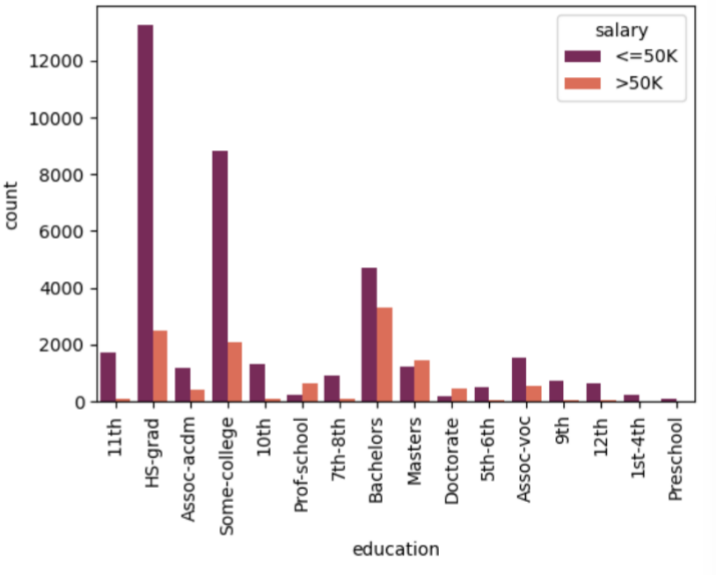
The initial stage involves data collection where the dataset typically contains various categorical and numerical features. However, one of the main challenges faced is the presence of missing values denoted as '?' which require appropriate preprocessing to ensure the data is suitable for analysis.

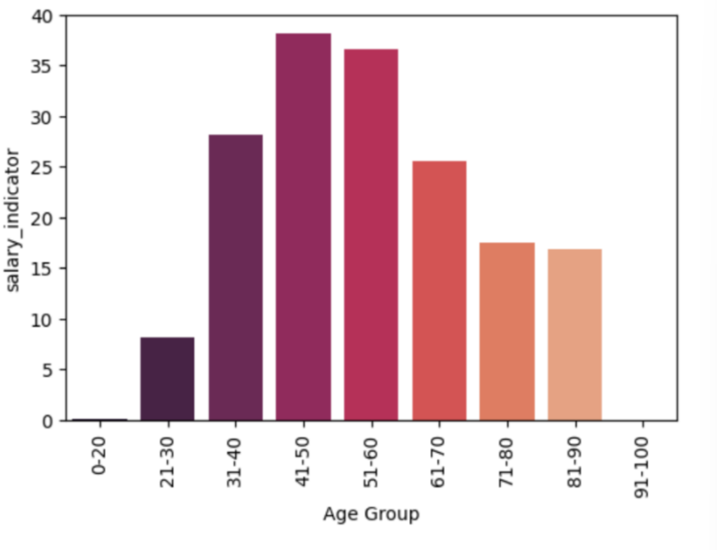
In the preprocessing phase, the approach involves converting all '?' entries to NaN for a consistent handling of missing values. For categorical data, any missing values are imputed using the mode of the column whereas numerical columns are first converted to numeric types, coercing errors to NaN, and then missing values are filled using the column's mean. This strategy ensures that the data remains robust for further analysis.

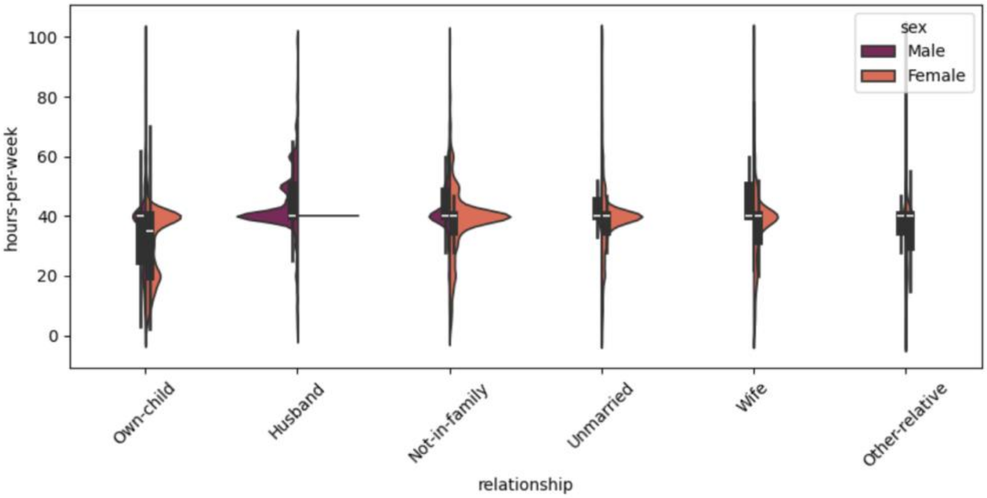
Moreover, the dataset undergoes normalization and encoding processes. Numerical features like age and hours-per-week are standardized (mean-centered and scaled by standard deviation), while categorical features such as workclass and education are transformed using one-hot encoding. This method converts categorical labels into a format that can be provided to ML models to do a better job in prediction.

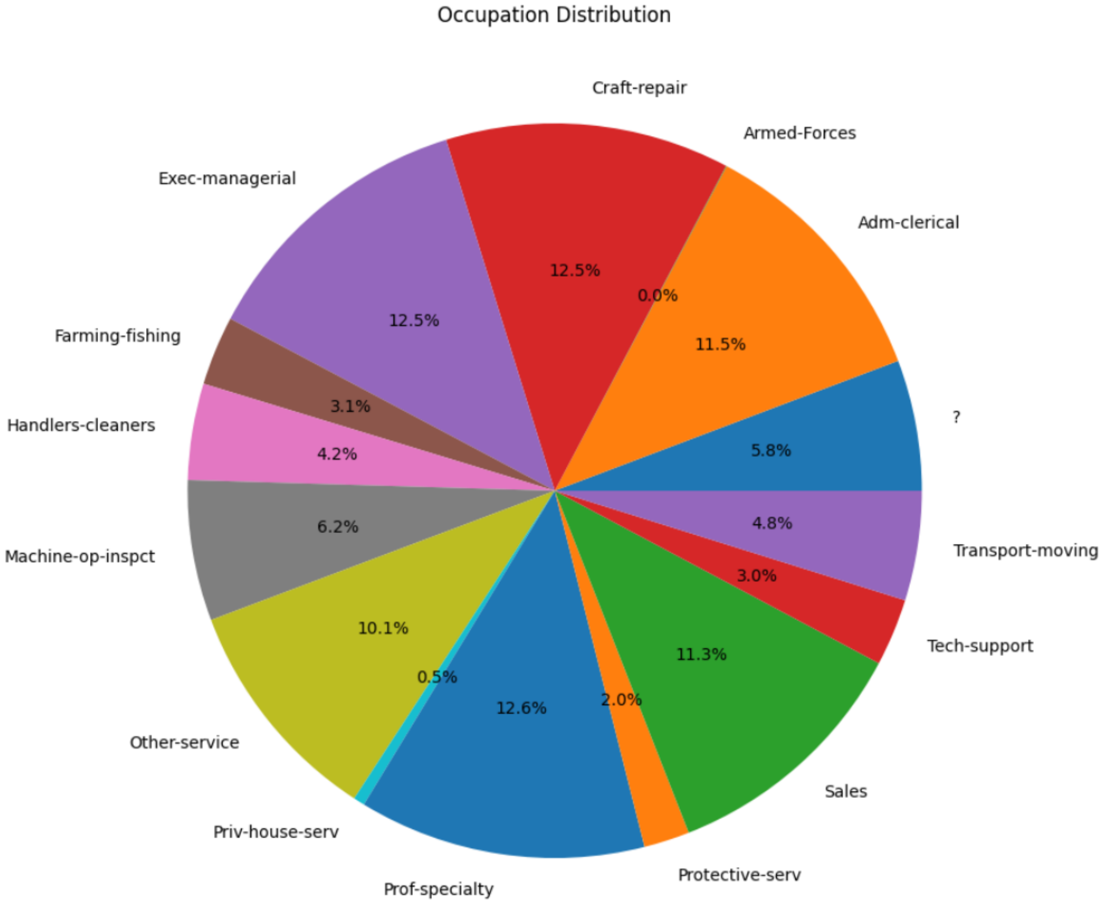
2.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) begins with the investigation of the distribution of key features such as salary, which is the target variable. Visualization tools like pie charts and count plots provide insights into the balance or imbalance of classes (income thresholds). Subsequent analysis includes using scatter plots and pair plots to visually assess the relationships between variables such as age, education, and hours worked per week.

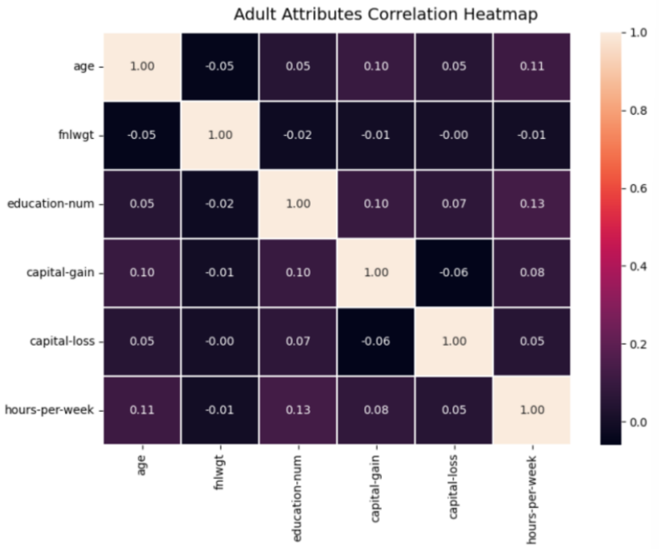


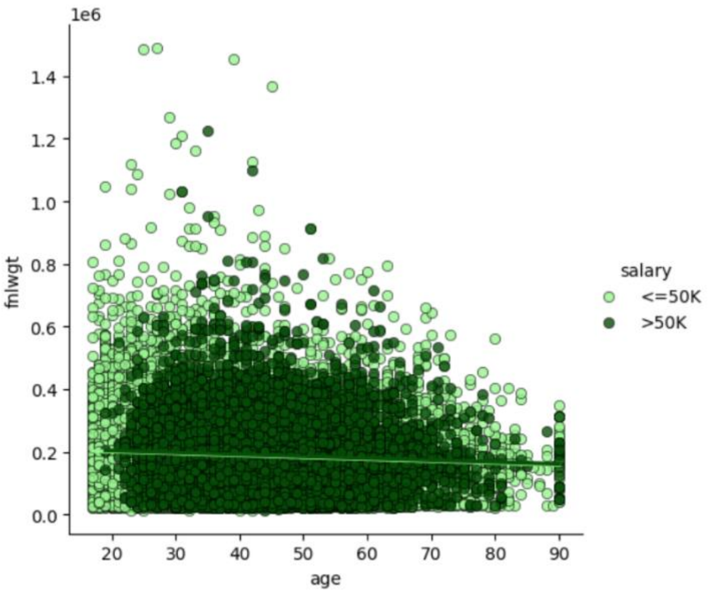






Bar plots and violin plots elucidate how features like education, sex, and relationship status vary with income, providing a nuanced view of potential biases or trends in the data. Correlation matrices further reveal how numerical features correlate with each other, which aids in understanding the interdependencies that could affect model predictions. Spearman and Kendall correlation methods are used due to their ability to handle non-linear relationships better compared to Pearson's method.





2.3 Predictive Modeling

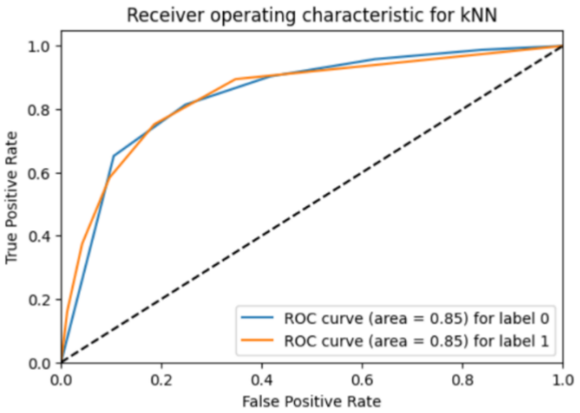
2.3.1 Data Preparation

Data is initially preprocessed and normalized, which is a crucial step for many machine learning algorithms. This normalization typically involves scaling numeric values to ensure that no variable dominates others due to scale differences, making the training process more stable and faster. Splitting Data: The data is split into training and test sets. Here, 35% of the data is reserved for testing to evaluate the model's performance on unseen data, a common practice to avoid overfitting.

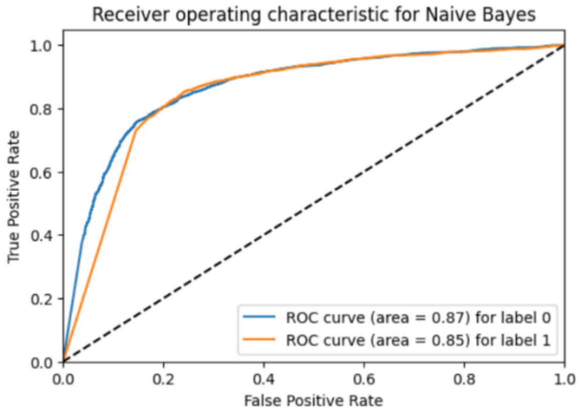
2.3.2 Model Training

Multiple classifiers are employed to handle a binary classification problem, wrapped in MultiOutputClassifier when necessary. This wrapping is particularly useful when dealing with multi-label classification:

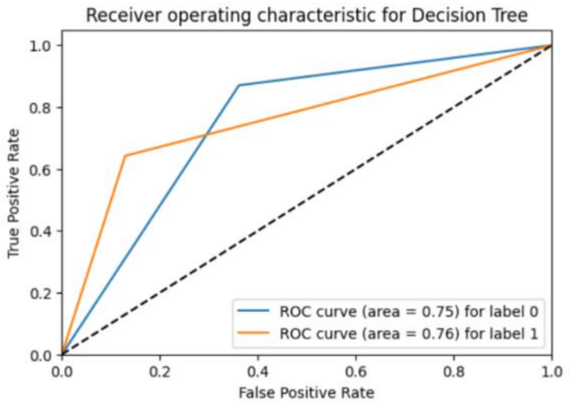
K-Nearest Neighbors (kNN): A non-parametric method used for classification. The output is a class membership determined by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.



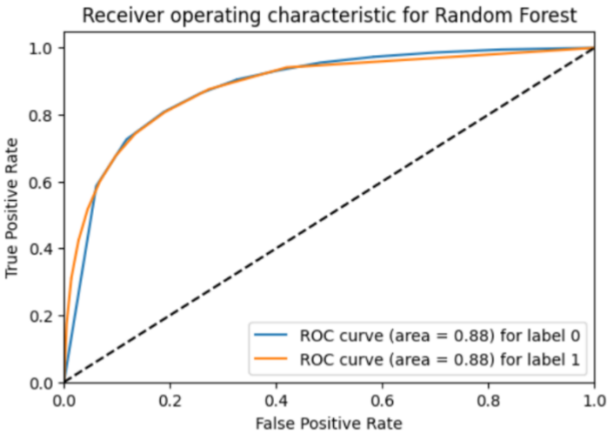
Naive Bayes: A set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.



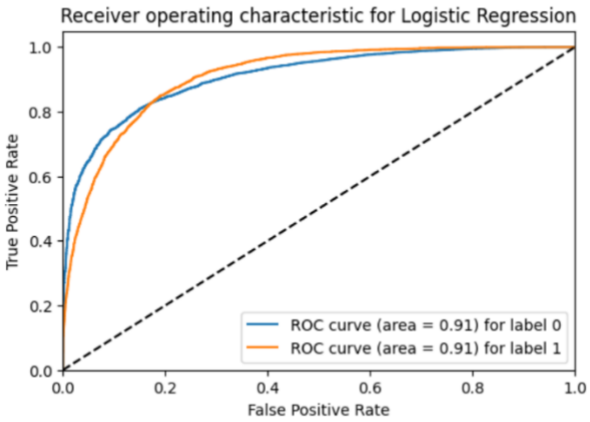
Decision Tree: A decision support tool that uses a tree-like graph of decisions and their possible consequences. It's simple to understand and interpret and can handle both numerical and categorical data.



Random Forest: An ensemble method of decision trees generated on a randomly split dataset. This model improves classification accuracy by averaging multiple decision trees that individually suffer from high variance.



Logistic Regression: Although typically used for binary classification problems, logistic regression can be extended to multiclass problems, either by applying a softmax function for multinomial logistic regression or by using the one-vs-rest (OvR) scheme.



2.3.3 Model Evaluation

Each model's performance is assessed using the test set:

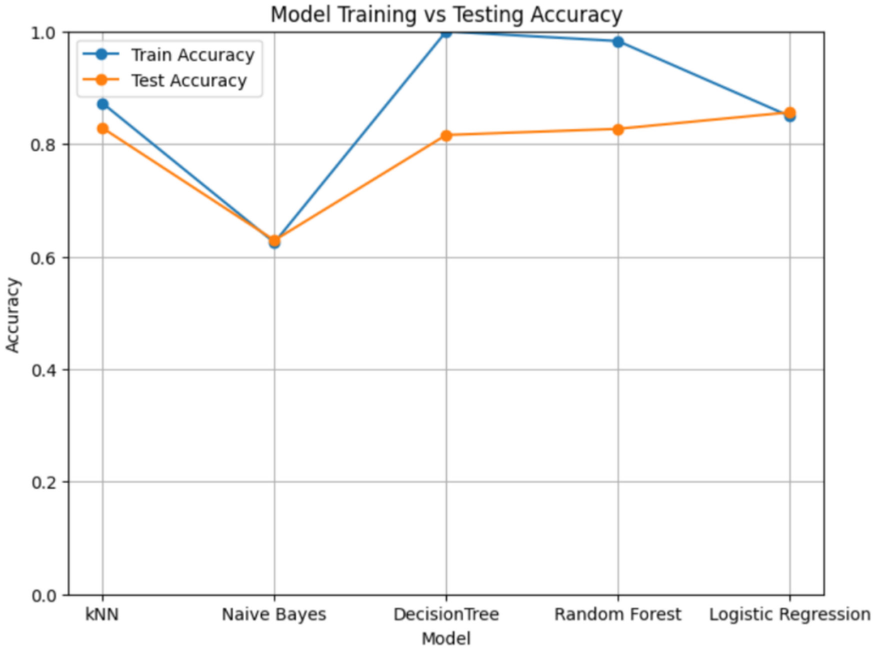
Accuracy: The proportion of true results (both true positives and true negatives) among the total number of cases examined.

ROC Curve and AUC: Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The area under the curve (AUC) represents a measure of the ability of the classifier to distinguish between the classes and is used as a summary of the ROC curve.

3. RESULTS, DATA VISUALIZATION AND ANALYSIS

3.1 Model Training and Evaluation

The code snippet begins with training various machine learning models on a training dataset and subsequently evaluates their performance on a testing dataset using accuracy as the metric. The models used include k-Nearest Neighbors, Naive Bayes, Decision Trees, Random Forest, and Logistic Regression. Accuracy is calculated for each model, and results are stored and printed.
Analysis:
The output will provide a clear indication of which model performs best on the testing data in terms of accuracy. Higher accuracy indicates a model's better generalization ability on unseen data. However, accuracy alone might not always be the best metric, especially if the dataset is imbalanced. Therefore, looking at other performance metrics like ROC AUC can provide deeper insights.

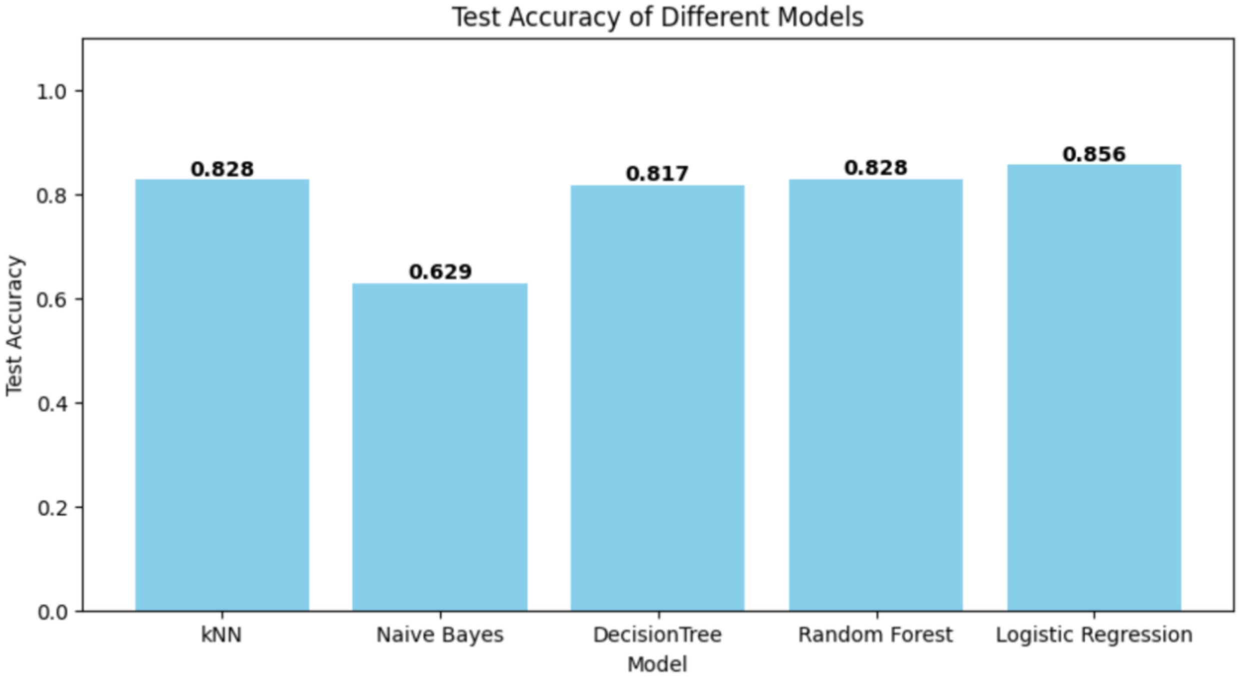


3.2 Plotting Accuracies

A bar plot visualizes the accuracy of each model. This visual representation helps in quickly identifying which models perform relatively better or worse.

Analysis:

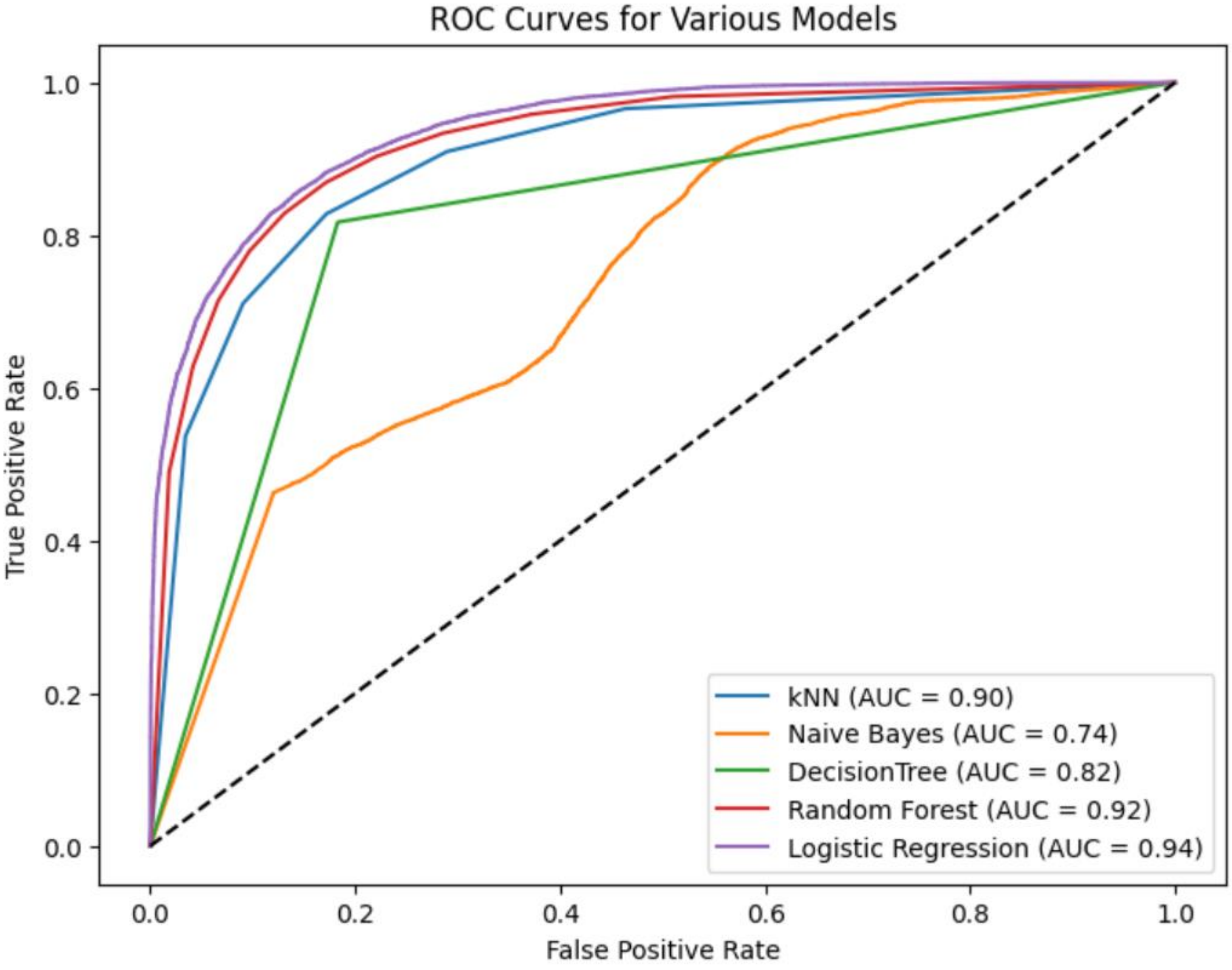
This plot is crucial for stakeholders to make decisions at a glance. The annotations on the bars display the exact accuracy values, enhancing the interpretability of the visualization.



3.3 ROC Curve Plotting

The ROC curves for each model are plotted by predicting probabilities on the test set. The ROC curve is a crucial diagnostic tool in binary and multiclass classification that shows the trade-off between sensitivity and specificity for different probability thresholds.

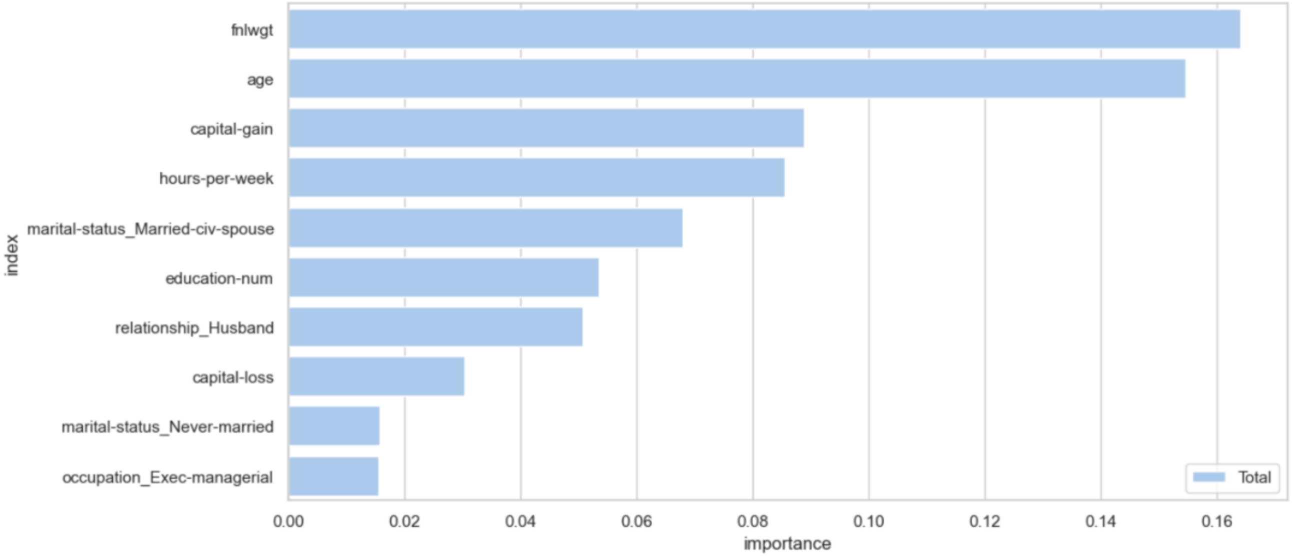
Analysis:
The Area Under the Curve (AUC) provides an aggregate measure of performance across all possible classification thresholds. A model with an AUC closer to 1 indicates a better model. Any error during the prediction or issues in model compatibility with multi-output predictions is caught and reported, ensuring robust evaluation.



3.4 Feature Importance Analysis

For the Random Forest model, feature importances are calculated. The feature importances indicate how valuable each feature was in the construction of the Random Forest trees.

Analysis:
The bar plot of feature importances helps identify which features are most influential in predicting the target variable. This can guide further model refinement and feature engineering to improve model performance. For instance, less important features can be removed to reduce model complexity and overfitting.



4. FUTURE WORK

In future work to enhance the predictive modeling, incorporating more sophisticated models such as Support Vector Machines, Gradient Boosting Machines, and neural networks could potentially yield better results due to their ability to capture more complex patterns in the data. Additionally, implementing k-fold cross-validation would offer a more robust estimate of model performance by mitigating the variance seen with a single train-test split. Addressing potential data imbalances through techniques like SMOTE or adjusting class weights during model training could also enhance fairness and accuracy. A deeper dive into feature engineering and exploring the interactions between features may reveal more intricate relationships and dependencies, further boosting model efficacy. Lastly, deploying the most effective model in a production environment, coupled with establishing a framework for continuous monitoring and regular model updates as new data emerges, would ensure sustained relevance and performance in real-world applications.

5. CONCLUSION

The analysis conducted through the series of Python scripts provides a thorough comparative evaluation of several machine learning models (k-Nearest Neighbors, Naive Bayes, Decision Trees, Random Forest, and Logistic Regression) applied to a dataset with the goal of predicting binary salary outcomes. By using metrics such as accuracy and ROC AUC, along with a detailed examination of feature importance for the Random Forest model, the study offers a robust framework for identifying the most effective model and understanding the contribution of different features to the predictive process.

5.1 Implications

The implications of a 1994 analysis focusing on adult income patterns can be wide-ranging and significant. Such an analysis might influence economic policies, particularly if it highlighted disparities in income based on factors like gender, race, or educational attainment. Governments and policymakers might respond by crafting initiatives aimed at reducing these disparities through targeted educational programs or labor market reforms. Additionally, findings indicating a correlation between higher education levels and increased income could encourage further investment in education and professional training programs. Moreover, if the analysis revealed a substantial portion of adults earning below the poverty line, it could lead to enhanced or expanded social welfare programs to support these individuals. Each of these potential outcomes underscores the importance of detailed income analysis in shaping effective and responsive public policies.

5.2 Human vs AI Contribution

In the data preprocessing stage, I meticulously cleaned the dataset by addressing missing values, stripping white spaces, and correcting erroneous entries, alongside implementing feature engineering where I devised new features like age groups and improved feature interpretability through normalization and standardization. I carefully selected and transformed relevant features for inclusion in my model, using techniques like one-hot encoding to suitably prepare the data for machine learning algorithms. For model selection, I demonstrated a nuanced understanding of various algorithms—kNN, Naive Bayes, Decision Tree, Logistic Regression, Random Forest—choosing them based on their distinct advantages and suitability for the dataset. I established an evaluation strategy, opting for metrics and methods like ROC curves and accuracy to effectively gauge model performance. Finally, my efforts culminated in the generation of insightful visualizations, using diverse plotting techniques to better understand data distributions and relationships between variables, thus aiding in comprehensive data analysis.

Once set up by humans, AI can significantly automate data preprocessing tasks like applying normalization, standardization, and encoding efficiently across large datasets. Through the use of machine learning libraries and tools, AI efficiently executes the algorithms chosen and configured by humans, handling computations for training, testing, and validating models. AI also optimizes models by adjusting weights, tuning hyperparameters through automated processes like grid search or random search and continuously improving learning algorithms based on feedback loops. Additionally, AI excels at identifying complex patterns in data that may not be easily visible to humans, such as detecting correlations, interactions among features, and predicting outcomes with high dimensional data.

5.3 Collaboration of Human and AI

The most effective data science workflows leverage both human intuition and creativity along with AI's computational power and pattern detection capabilities. Humans set strategic goals and design frameworks, while AI handles large-scale data manipulation and complex computations. The iterative feedback loop between human insights and AI outputs is crucial for refining models and achieving the most accurate and actionable results.

6. REFERENCES

i. https://archive.ics.uci.edu/dataset/2/adult

ii. https://www.kaggle.com/datasets/uciml/adult-census-income

iii. https://rpubs.com/Mr_President/income_prediction

iv. M. S. Acharya, A. Armaan and A. S. Antony, "A Comparison of Regression Models for Prediction of Graduate Admissions," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICCIDS.2019.8862140. keywords: {Linear regression;Measurement;Data models;Predictive models;Regression tree analysis;Linear Regression;Support Vector Regression;Decision Trees;Random Forest;Mean Squared Error},

v. S. Ansari and A. B. Nassif, "A Comprehensive Study of Regression Analysis and the Existing Techniques," 2022 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, 2022, pp. 1-10, doi: 10.1109/ASET53988.2022.9734973. keywords: {Measurement;Support vector machines;Analytical models;Machine learning algorithms;Atmospheric modeling;Simulation;Machine learning;data mining;machine learning;multicollinearity;regression analysis;statistical model},

vi. D. Kinaneva, G. Hristov, P. Kyuchukov, G. Georgiev, P. Zahariev and R. Daskalov, "Machine Learning Algorithms for Regression Analysis and Predictions of Numerical Data," 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2021, pp. 1-6, doi: 10.1109/HORA52670.2021.9461298. keywords: {Training;Machine learning algorithms;Machine learning;Predictive models;Prediction algorithms;Numerical models;Regression analysis;machine learning;normalization;globalization;generalization;dataset preparation},

vii. H. Luo, F. Lu and Q. Lu, "Regression Model in Selecting Network Information Technology Companies and Analyzing their Business Development Trends," 2010 First International Conference on Networking and Distributed Computing, Hangzhou, China, 2010, pp. 76-80, doi: 10.1109/ICNDC.2010.25. keywords: {Companies;Software;Information technology;Indexes;Industries;Radiofrequency identification;network information technology;regression models;method of factor analysis;business development trend},

viii. M. R. Putri, I. G. P. S. Wijaya, F. P. A. Praja, A. Hadi and F. Hamami, "The Comparison Study of Regression Models (Multiple Linear Regression, Ridge, Lasso, Random Forest, and Polynomial Regression) for House Price Prediction in West Nusa Tenggara," 2023 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS), Bali, Indonesia, 2023, pp. 1-6, doi: 10.1109/ICADEIS58666.2023.10270916. keywords: {Industries;Navigation;Linear regression;Buildings;Measurement uncertainty;Predictive models;Stakeholders;Comparison Study of Models;Regression;House Price Prediction;Accuracy Testing}