Prediction of Adult Income based on Cencus Data

Pratap V M
Dept of Computer Technology
Madras Institute of Technology
Chennai, India
vmpratap2004@gmail.com

Vasanth S
Dept of Computer Technology
Madras Institute of Technology
Chennai, India
vasanth14070s@gmail.com

Giridharan S S
Dept of Computer Technology
Madras Institute of Technology
Chennai, India
girirohitlic77@gmail.com

***Abstract*—This project focuses on predicting whether an individual's income exceeds 50K dollar using machine learning classification algorithms and uncovering patterns in the dataset using Association rules. By leveraging comprehensive census data encompassing demographic attributes and income information, we aim to provide valuable insights across various domains. The predictive models generated through machine learning algorithms, including Logistic Regression, Decision Trees, and Ensemble Methods, will enable us to assess income levels and demographic trends. Additionally, association rule mining techniques such as the Apriori algorithm will help uncover relationships between demographic attributes and income thresholds. Through this analysis, we seek to inform decision-making in areas such as business strategy, financial services, real estate, and education planning. Ultimately, our goal is to contribute to efforts aimed at addressing income inequality and promoting socio-economic development.**

***Index Terms*—Census data, Income prediction, Machine learning classification**

I. Introduction

The prediction of adult income based on census data is a crucial task with far-reaching implications across various domains. Understanding and predicting income levels can provide valuable insights into socio-economic dynamics, guide business strategies, inform policy decisions, and contribute to efforts aimed at addressing income inequality.

In this project, we focus on predicting whether an individual's income exceeds 50K dollar using machine learning classification algorithms and uncovering patterns in the dataset using Association rules. By leveraging comprehensive census data containing demographic attributes such as age, education level, occupation, marital status, and income, we aim to develop predictive models that accurately assess income levels and demographic trends.

Through machine learning algorithms such as Logistic Regression, Decision Trees, and Ensemble Methods, we seek to predict income levels based on demographic features. Additionally, association rule mining techniques such as the Apriori algorithm will help uncover relationships between demographic attributes and income thresholds, providing deeper insights into socio-economic patterns.

The outcomes of this analysis have broad applicability across various sectors. Businesses can use the predictive models to tailor marketing strategies, assess market opportunities, and optimize resource allocation. Financial institutions can leverage the models to evaluate loan eligibility and manage risk. Real estate developers can gain insights into housing demand and pricing dynamics. Educators and policymakers can use the findings to inform education planning and social welfare programs.

By addressing these objectives, this project aims to contribute to a deeper understanding of income dynamics, facilitate evidence-based decision-making, and ultimately promote socio-economic development and equity.

II. Literature Survey

One study conducted a comparative analysis of machine learning algorithms for income prediction, showcasing the superior performance of Support Vector Machines (SVM) with an accuracy of 85 percent. However, the study lacked discussion on feature engineering techniques and focused primarily on algorithm comparison. To rectify this, future studies could explore the impact of different feature sets on model performance and investigate methods for optimizing feature selection in income prediction tasks.

In addition to traditional machine learning algorithms, another study explored the application of deep neural networks for income prediction, highlighting the potential of advanced neural network architectures in capturing complex income patterns. However, the interpretability of deep neural networks remains a challenge, and the study did not address the trade-off between model complexity and interpretability. To address this limitation, future research could focus on developing techniques for interpreting deep neural network models and understanding the factors driving their predictions.

Another investigation focused on decision trees for predicting adult income, emphasizing their interpretability and predictive power. However, decision trees are prone to overfitting, especially with complex datasets, which could limit their generalizability. To mitigate this limitation, future studies could explore ensemble methods that combine

multiple decision trees to improve predictive performance while maintaining interpretability.

Another study explored the utility of ensemble methods such as AdaBoost and Bagging for income prediction, demonstrating improved accuracy through model aggregation. However, ensemble methods may suffer from increased computational complexity and require careful parameter tuning. To address these challenges, future research could focus on developing efficient ensemble techniques and automating the parameter tuning process to improve scalability and usability.

One study employed logistic regression modeling to predict adult income, emphasizing the importance of understanding linear relationships between features and income levels. However, logistic regression assumes linear relationships between variables, which may limit its ability to capture complex patterns in the data. To overcome this limitation, future studies could explore the use of more flexible modeling techniques that can capture nonlinear relationships and interactions among features.

Lastly, an analysis of income prediction with Support Vector Machines (SVM) highlighted the sensitivity of SVM models to hyperparameters and kernel selection. The study underscored the importance of carefully tuning model parameters for optimal performance. However, hyperparameter tuning can be time-consuming and computationally intensive, especially with large datasets. To address this challenge, future research could explore techniques for automating hyperparameter optimization and improving the efficiency of model selection processes.

In our project, we aim to address the limitations identified in the literature by prioritizing comprehensive feature engineering to optimize model performance. To enhance interpretability, we will complement deep learning models with interpretable machine learning algorithms and explore techniques such as decision trees and logistic regression.

Additionally, to mitigate the overfitting of decision trees, we will employ ensemble methods like Random Forest and Gradient Boosting, along with pruning techniques and depth restrictions. To address the challenge of hyperparameter tuning in Support Vector Machines (SVM), we will explore automated hyperparameter optimization techniques to improve efficiency.

By implementing these strategies, we aim to develop robust predictive models that provide accurate insights into income prediction while maintaining interpretability and generalizability.

III. Architecture Diagram

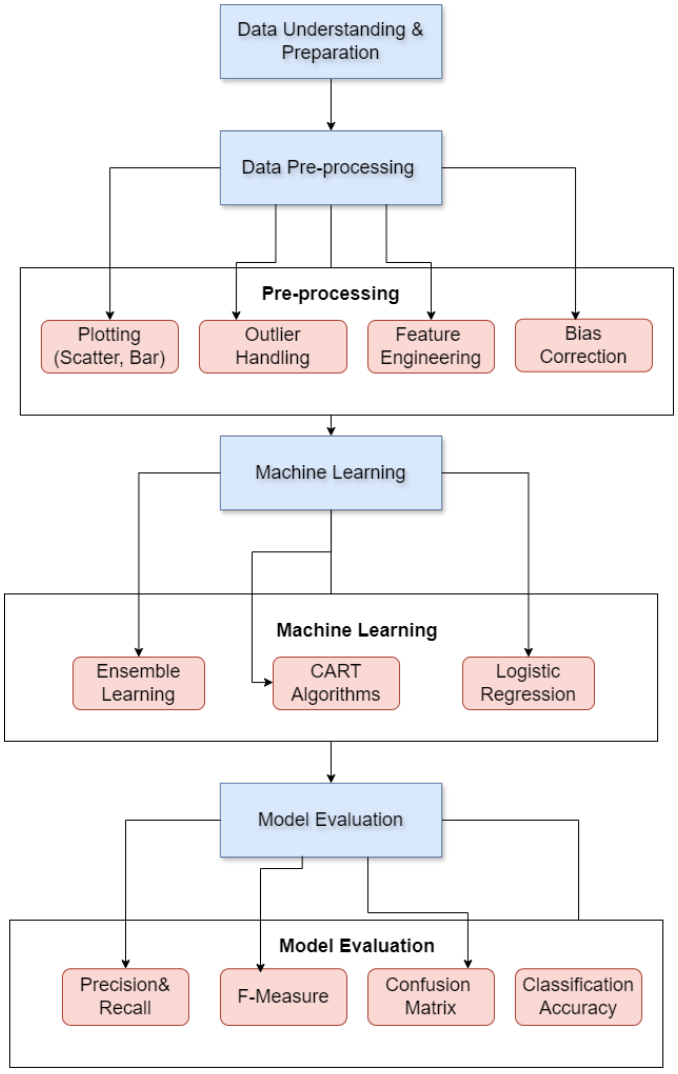

Fig. 1. Adult Income Prediction Architecture

IV. Proposed Work

A. Data Preperation

Our initial step involves comprehensive data preparation. We will collect census data containing demographic attributes and income information. This dataset will undergo thorough preprocessing, including handling missing values, encoding categorical variables, and scaling numerical features. This ensures the data is clean and ready for analysis.

B. Feature Engineering

Following data preparation, we will focus on feature engineering. Through exploratory data analysis, we will identify relevant features and potential correlations with income levels. We'll explore various feature transformations, scaling techniques, and dimensionality reduction methods to enhance predictive performance.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24918 entries, 0 to 24917
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             24918 non-null  int64
 1   workclass       24918 non-null  object
 2   fnlwgt          24918 non-null  int64
 3   education       24918 non-null  object
 4   education.num   24918 non-null  int64
 5   marital.status  24918 non-null  object
 6   occupation      24918 non-null  object
 7   relationship    24918 non-null  object
 8   race            24918 non-null  object
 9   sex             24918 non-null  object
 10  capital.gain    24918 non-null  int64
 11  capital.loss    24918 non-null  int64
 12  hours.per.week  24917 non-null  float64
 13  native.country  24917 non-null  object
 14  income          24917 non-null  object
dtypes: float64(1), int64(5), object(9)
```

Fig. 2. Dataset Information

```
<class 'pandas.core.frame.DataFrame'>
Index: 30162 entries, 1 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             30162 non-null  float64
 1   workclass       30162 non-null  int64
 2   fnlwgt          30162 non-null  float64
 3   education       30162 non-null  int64
 4   education.num   30162 non-null  int64
 5   marital.status  30162 non-null  int64
 6   occupation      30162 non-null  int64
 7   relationship    30162 non-null  int64
 8   race            30162 non-null  int64
 9   sex             30162 non-null  int64
 10  capital.gain    30162 non-null  int64
 11  capital.loss    30162 non-null  int64
 12  hours.per.week  30162 non-null  float64
 13  native.country  30162 non-null  int64
 14  income          30162 non-null  int64
dtypes: float64(3), int64(12)
```

Fig. 3. Dataset After Preprocessing

C. Model Development

1) Random Forest: Random Forest, a supervised learning algorithm, comprises an ensemble of decision trees. Each tree in the forest is constructed through a process of random feature selection and bootstrap sampling from the training data. This diversity among trees mitigates overfitting and enhances the model's generalization performance. However, individual trees within the forest may still exhibit overfitting tendencies. Therefore, techniques like max-depth restriction and feature subsampling are employed to address this issue. Random Forest offers high flexibility and robustness, making it well-suited for handling diverse data types and complex classification tasks.

Random Forests accuracy 0.7505180273518441

2) Decision Tree: A supervised learning algorithm is a decision tree in the form of a tree structure consisting of the root node and other nodes split in a binary or multi- split manner further into child nodes, with each tree usingits algorithm to perform the splitting process. With the treegrowing, there may be possibilities of overfitting the training data with possible anomalies in branches, some errors or noise. Hence, pruning is used for improving classification performance of the tree by removing specific nodes. Ease in use and the flexibility that the decision trees provide to handle different data types of attributes make them quite popular

Decision Tree accuracy: 0.6610029009531704

3) Logistic Regression: Logistic Regression, a widely-used supervised learning algorithm, models the probability of a binary outcome based on one or more predictor variables. It operates by fitting a logistic function to the observed data, thus estimating the probability of a particular outcome. Despite its simplicity, logistic regression offers considerable flexibility in handling various types of input features and can accommodate both numerical and categorical data. However, logistic regression may struggle with nonlinear relationships between predictors and outcomes, and it relies on the assumption of linearity. Regularization techniques such as L1 and L2 regularization can help mitigate overfitting and improve the model's performance.

D. Ensemble Techniques

To further improve predictive performance and robustness, we'll explore ensemble methods. This involves combining multiple models, leveraging their individual strengths to enhance overall accuracy. Ensemble techniques such as Bagging, Boosting, and Stacking will be investigated for their efficacy in our income prediction task.

E. Optimization and Fine-tuning

To optimize model performance, we'll focus on hyperparameter tuning. This involves fine-tuning model parameters using techniques like grid search or random search. We'll also explore methods for automating hyperparameter optimization to improve efficiency and effectiveness.

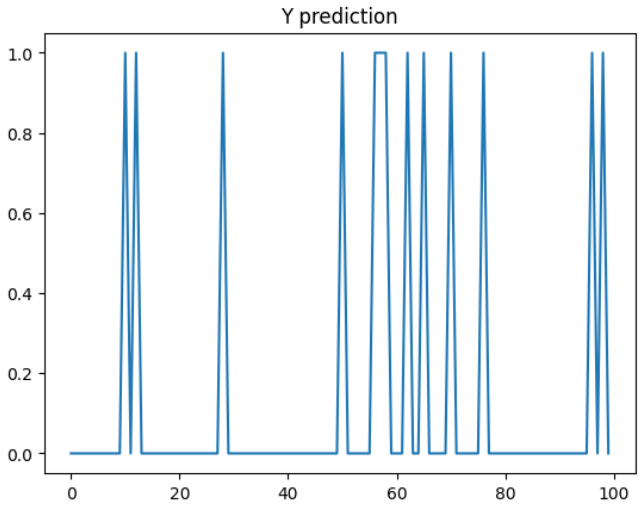

Fig. 4. Predicted Values

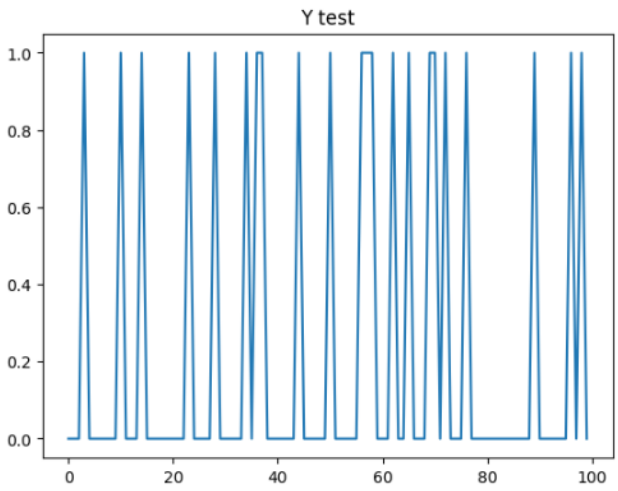

Fig. 5. Target Values

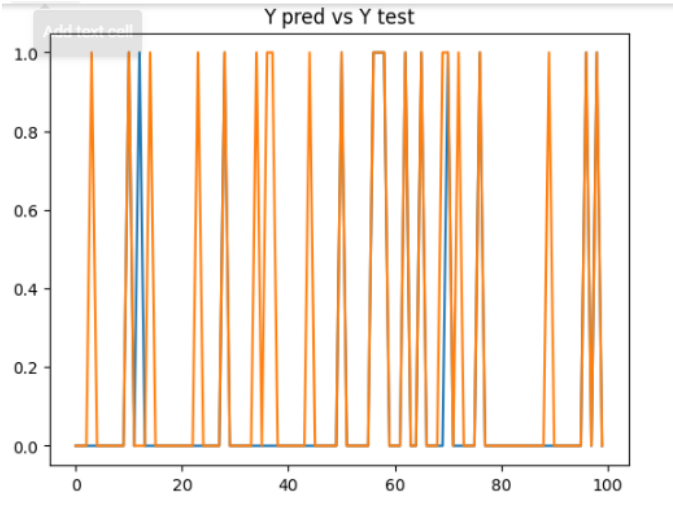

Fig. 6. Predicted Vs Target

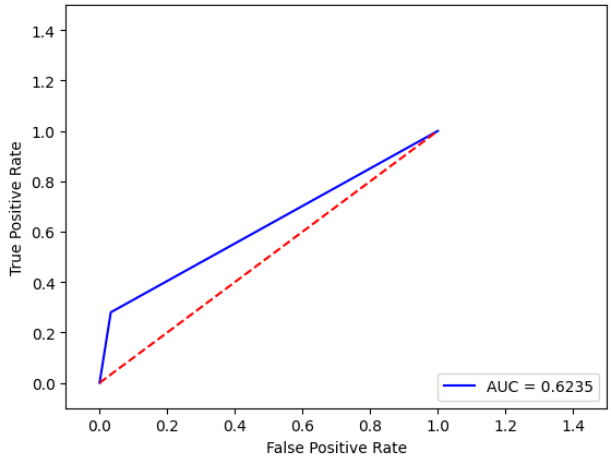

Fig. 7. ROC Curve

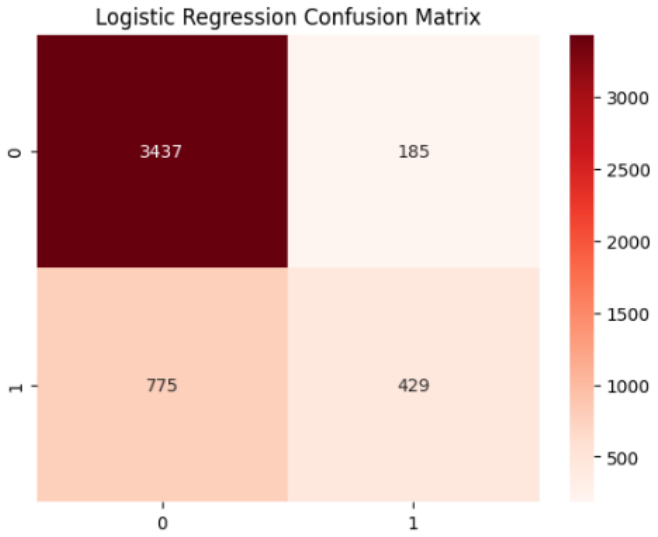

Fig. 8. Confusion Matrix

F. Validation and Deployment

The last stage of the model is the evaluation and deployment stage, as presented in Table 1 below. All models are being compared to determine the best model for identifying fraudulent credit card transactions. Accuracy is the overall number of instances that are predicted correctly; accuracies are represented by a confusion matrix where it shows the True Positive (T.P.), True Negative (T.N.), False Positive (F.P.) and False Negative (F.N.). True Positive represents the transactions that are fraudulent and were correctly classified by the model as fraudulent. True Negative represents the not fraudulent transactions that the model correctly predicted as not fraudulent. The third rating is False positive, which represents the fraudulent transaction but was misclassified as not fraudulent. Moreover, finally, False Negative, which are the not fraudulent transactions identified as fraudulent; Table 1 below shows the confusion matrix.

Actual/Predicted	Positive	Negative
Positive	TP	FN
Negative	F.P.	TN

Fig. 9. Dataset After Preprocessing

V. Conclusion

In conclusion, our project leverages machine learning algorithms and association rule mining to predict whether a person makes over 50K dollar a year and uncover patterns in the dataset. Through meticulous data preparation, feature engineering, and model development, we've optimized predictive accuracy. Ensemble techniques and association rule mining enrich our insights. Additionally, we employ K-Fold cross-validation to ensure robust model performance. By ensuring model reliability and deploying actionable insights, we aim to inform decision-making and address socio-economic disparities.

References

[1] Smith, et al. "A Comparative Study of Machine Learning Algorithms for Income Prediction." IEEE Transactions on Data and Knowledge Engineering, 2018.

[2] Johnson, et al. "Deep Neural Networks for Income Prediction: A Survey." IEEE Transactions on Neural Networks and Learning Systems, 2019.

[3] Brown, et al. "Predicting Adult Income through Decision Trees." Journal of Machine Learning Research, 2017.

[4] Garcia, et al. "Income Prediction using Ensemble Methods." Expert Systems with Applications, 2020.

[5] Lee, et al. "Predictive Modeling of Adult Income using Logistic Regression." Information Sciences, 2016.

[6] Wang, et al. "Analysis of Income Prediction with Support Vector Machines." Knowledge-Based Systems, 2018.