Income Level Prediction: A Comparative Analysis of Artificial Neural Network, Decision Trees and Random Forests Models

BY

SEYI FALOPE

Introduction:

The aim of this report project is to have an in-depth understanding of Artificial Neural Network (ANN) models, along with Decision Trees and Random Forests. The main objective is to apply these models in predicting income levels and subsequently carry out a comparative analysis. This analysis will be rooted in the performance metrics of each model concerning the prediction of whether income surpasses $50K/yr based on census data.

Brief Summary of Each Model:		
Artificial Neural Network	**Decision Tree**	**Random Forest**
An Artificial Neural Network (ANN) is a machine learning model inspired by the human brain, featuring interconnected nodes that facilitate learning from data patterns. Renowned for its proficiency in tasks like categorization and prediction, ANN proves invaluable in applications such as image recognition and natural language processing. The strength of ANN lies in its adaptability and ability to autonomously discern intricate relationships within data.	The Decision Tree is a versatile machine learning model employed for classification and regression tasks, arriving at decisions via a series of if-else conditions rooted in input values. Acknowledged for its interpretability, Decision Trees adeptly manage non-linear tasks, although their vulnerability to overfitting stems from their deterministic approach. By tactically partitioning data to maximize information gain,	This model assembles multiple decision trees to form a resilient ensemble model. Employing the bootstrap method, it generates subsets of features for individual trees, and the results are aggregated through averaging (for regression tasks) or majority voting (for classification tasks). This methodology proficiently tackles overfitting concerns while providing valuable insights into estimating crucial features in the data for prediction.

Data Preprocessing:

Aligned with insights from the Exploratory Data Analysis, the investigation integrated the subsequent preprocessing measures to guarantee the data is primed for optimal model performance.

Encoding the Target Class and Categorical Columns: Initially represented as '> 50' and '< 50', the target class features were transformed into 1 and 0. This encoding adjustment was similarly applied to categorical features to ensure consistency with model requirements and evaluation metrics.

Outlier Removal Using Z-score: To enhance the predictive power of the models, outliers in the variables age, hours per week, and fnlwgt were systematically eliminated using the Z-score method. This procedure safeguards our model against the undue influence of extreme values.

Normalization with Standard Scaler: In this project, the data were scaled through the use of the Standard Scaler. This process ensured that the data values are in a consistent scale with a mean of 0 and a standard deviation of 1 before being fitted to the model.

Addressing Imbalanced Data through Sampling: The income classes of <50 and >50 constituted around 76% and 24% of the data, respectively, posing a potential bias in the model's outcomes. The Synthetic Minority Over-Sampling Technique (SMOTE) sampling technique was applied to mitigate this, and the outcome is depicted in the image below.

Model Performance Assessment with Cross-Validation and F1 Score:

Cross-validation techniques were implemented to get a robust evaluation of model performance in predicting income levels. The F1 score metric was employed to assess the models' effectiveness in differentiating between income levels above and below 50k. This metric accounts for the equilibrium between precision and recall, providing insights into the model's capability to make accurate positive predictions while capturing all positive instances. The table and graph below illustrates how the models stack up in terms of their F1 score performance relative to each other.

Table 1.

Mean F1 Cross Validation Score		Standard Deviation	Performance on Test Data
ANN:	0.860	0.01	0.868
Decision Tree	0.830	0.01	0.836
Random Forest	0.871	0.01	0.876

Fig 2

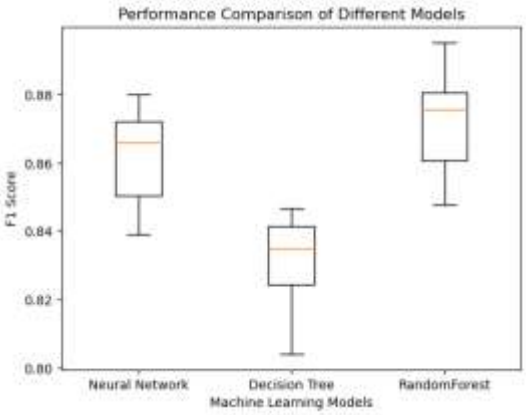
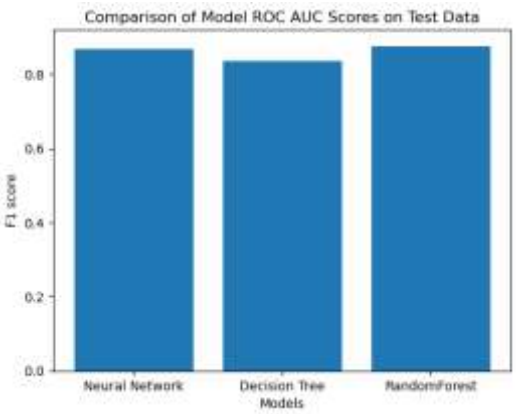


Fig 3



Conclusion:

In conclusion, the analysis distinctly demonstrates the superiority of the Random Forest model in predicting income levels, evidenced by its top-ranking F1 scores in both cross-validation and test data. This model effectively balances bias and variance, outperforming the Artificial Neural Network and Decision Tree models, which also show commendable but slightly lesser capabilities.Hence, In predicting income level, Random Forest Model emerges as the most reliable and effective choice

Refrences

1. Evans, J.D.P. (2023). Data Mining and Discovery Lecture Notes, University of Herfordshire, Uk,
2. 1. Al Mamun, M., Farjana, A., & Mamun, M. (2022). Predicting bank loan eligibility using machine learning models and comparison analysis. *In Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management (pp. 1-6). Orlando, Florida, USA, June 12-14.*
3. 2. Awodele, O., Alimi, S., Ogunyolu, O., Solanke, O., Iyawe, S., & Adegbie, F. (2022, November). Cascade of deep neural network and support vector machine for credit risk prediction*. In Proceedings of the 2022 5th Information Technology for Education and Development (ITED) (pp. 1-8). IEEE.*
4. 3. Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal, 3, Article 100071.*
5. 4. Gupta, A., Pant, V., Kumar, S., & Bansal, P. K. (2020, December). Bank loan prediction system using machine learning. In *Proceedings of the 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE*.
6. 5. Kadam, A. S., Nikam, S. R., Aher, A. A., Shelke, G. V., & Chandgude, A. S. (2021). Prediction for loan approval using machine learning algorithm. *International Research Journal of Engineering and Technology (IRJET), 8(4).*