Classification model analysis: Decision tree & Random Forest / Neural Network.

Juan Manuel Gonzalez Rincon | ID 23031523

Introduction: This document analyzes the performance of three classification models that uses different techniques; **Decision Three:** rule based, **Random Forest:** ensemble model and **Neural networks:** variable transformation based on layers on a census dataset.

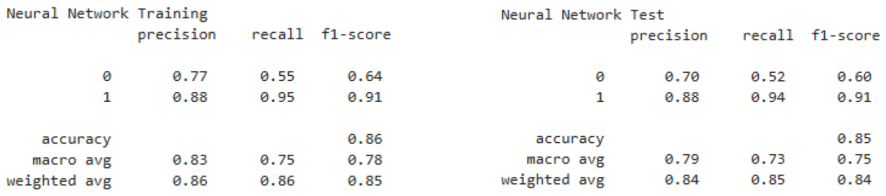
Subtopic1 | Neural Networks (Multilayer): Based on nodes organized in layers. Each layer receives inputs, applies unique weights, and uses an activation function to transform variables and compute classification probabilities. The model updates weights using gradient descent optimization, with learning rate and number of iterations as key parameters.

Subtopic2 | Decision Tree & Random Forest: Used rules at nodes based on a single attribute, forming branches that end in leaf nodes, where each class is pure. Node selection minimizes entropy. Random forests combine multiple decision trees trained on different data partitions, averaging their predictions for final classification.

Dataset: The Census Income Prediction dataset from UC Irvine predicts whether a person's income exceeds $50K/year based on 1994 census data. It includes 32,560 records with 14 input variables**,** variable types: **1) Nominal:** Marital Status, Occupation, Relationship, Work Class, Race, Sex and Country. **2) Ordinal:** Education and Education number. **3)Ratio:** Hours Per Week, Age, Fnlwgt, Capital Gain and Capital Loss. The base has 32,560 records and 15 columns; 1 the class variable (1 / >=$50K or 0 / <$50K) and the 14 variables mentioned before.

Data Preprocessing: 1) Duplicated values: First, the dataset had 24 duplicated records. **2)Outliers:** Normalized and filtered data within three standard deviations of the mean. it was 6%. **3) Missing Values:** They were present in three variables: Work Class, Occupation and Country, as the mode of Work Class and Country represented more than 70%, they were filled with it, for Occupation variable, which was more distributed, this variable was filled in proportion. **4) Discretization:** As the variables age and education manage an ordinal hierarchy, these variables were transformed into ordinal grouped variables using domain knowledge. **5) Aggregation:** The variables Marital Status, Work Class and Relationship were grouped in related variables but without losing the value added to the problem. **6) Binarization:** The Sex, Class and Country were represented by binary variables, Country had predominant mode with more than 90%. **7) Multicollinearity:** Remove highly correlated variable education / education number (0.81). **8) One-hot encoding:** The nominal data was finally transformed into binary classes using an extended representation with a column per value, this step is key for feeding both classification models. **9) Train-Test split:** Randomly split data into 70% training and 30% test sets.

Neural Network (Multilayer): The model used had one input and output layers, then two dense hidden layers with 128 and 64 nodes each one with the activation function Relu which is easily optimized with gradient descent and less prone to gradient vanishing, but to overcome this possible issue there were used two dropout layers with 20% each, finally the output layer has a sigmoid activation function due to the binary problem in case. The model was trained with 50 epochs, the results were:



```
Neural Network Training
              precision    recall  f1-score

           0       0.77      0.55      0.64
           1       0.88      0.95      0.91

    accuracy                           0.86
   macro avg       0.83      0.75      0.78
weighted avg       0.86      0.86      0.85
```

```
Neural Network Test
              precision    recall  f1-score

           0       0.70      0.52      0.60
           1       0.88      0.94      0.91

    accuracy                           0.85
   macro avg       0.79      0.73      0.75
weighted avg       0.84      0.85      0.84
```

Image 1. Neural Network classification results with training and test datasets.

The accuracy of the model was similar in both cases (0.85) and the precision of the class 1 (<=$50K) was better (0.88), due to the unbalance of the dataset (class 1 78%), the recall showed that the class 1 was predicted better (0.94) than class 0 (0.52) reflected in the recall.

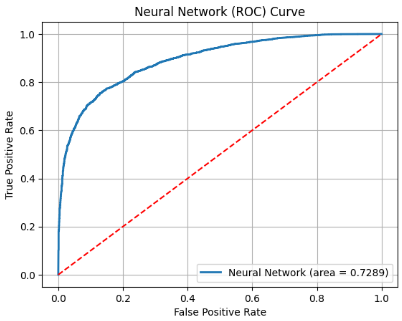
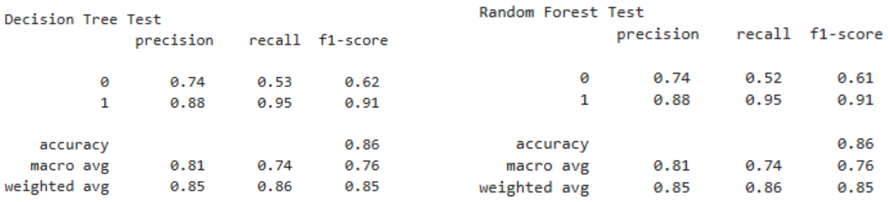


Chart 1. Neural Network ROC curve.

Although the model has better recall for class 1 (>=$50K), the ROC curve shows that the model differentiates well the classes, at least more than the random guess and with good accuracy.

Decision Tree & Random Forest: The hyperparameters were found trying different values and evaluating in the validation set, that happened with the ponderation of the number of leaves in the random selection and then the selection of number of leave nodes, the maximum number of nodes was 23, for the random forest, the number of trees was 200 and to find the maximum number of variables randomly selected to train different trees, the model was trained and validated with random number of this parameter, resulting in 26, that's what makes this model more robust than the decision three, the ponderation of different independent models.



```
Decision Tree Test
              precision    recall  f1-score

           0       0.74      0.53      0.62
           1       0.88      0.95      0.91

    accuracy                           0.86
   macro avg       0.81      0.74      0.76
weighted avg       0.85      0.86      0.85
```

```
Random Forest Test
              precision    recall  f1-score

           0       0.74      0.52      0.61
           1       0.88      0.95      0.91

    accuracy                           0.86
   macro avg       0.81      0.74      0.76
weighted avg       0.85      0.86      0.85
```

Image 2/3. Decision Tree & Random Forest classification results with training and test datasets.

The accuracy was similar to previous model (0.86), the proportion of class 0 predicted correct was better (0.53) than NN but lower than recall of class 1 (0.95), this issue due to the balance of the training data, random forest results were similar as well, that is because of the one-hot encoding and the way that random forest creates more veriety of trees by using different variables. The ROC curve for both models are the same, slighly higher than the NN performing better in the classification problem.

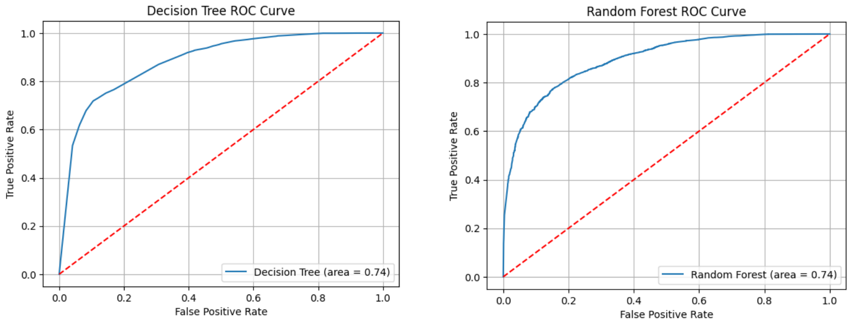


Chart 2/3. Decision Tree & Random Forest ROC curve.

Conclusion: All models performed comparably, with Decision Tree and Random Forest slightly outperforming Neural Networks in ROC and accuracy. These models are also more interpretable and practical for decision-making. To address class imbalance, sampling techniques could further enhance performance for Class 0 (<$50K), although decission tent to overfit, the results shows that validation test improves its generalization similar to random forest which was affected by the one hot encoding. Neutal Networks is also more computational demanding.

Bibliography: Tan, P., Steinbach, M., Karpatne, A. & Kumar, V. (2019). *Introduction to Data Mining*. 2nd ed. Boston: Pearson, Chapter 6, Section 7, pp. 451-464.

UCI Machine Learning Repository. "Census Income Dataset." Accedido el 5 de enero de 2025. https://archive.ics.uci.edu/dataset/20/census+income.