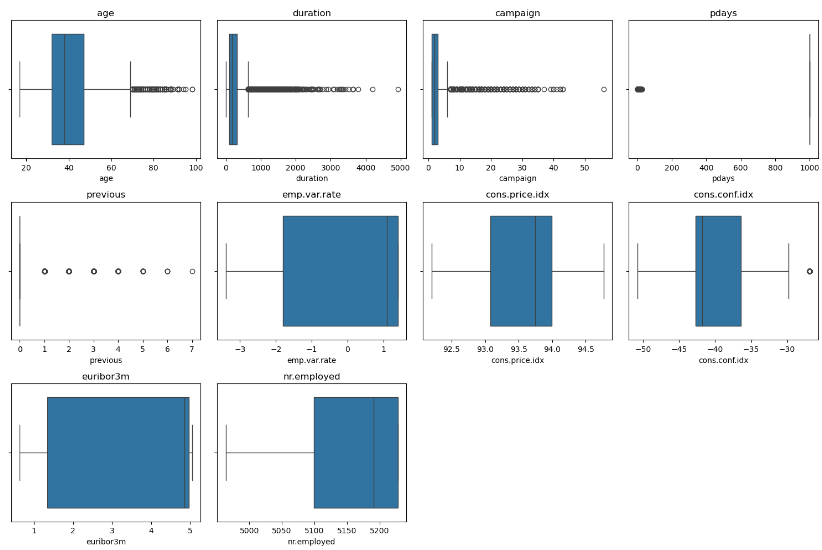
# **Decision Tree Classifier for Term Deposit Subscription Prediction Report**

## **Introduction**

This project aimed to develop a predictive model to classify whether a customer would subscribe to a term deposit using a decision tree classifier. The dataset was obtained from <https://archive.ics.uci.edu/dataset/222/bank+marketing.>

## **Data Exploration and Preprocessing**

* **Dataset Overview**: The dataset was loaded and explored, noting its structure, size, and data types for each feature. Initial inspection revealed both categorical and numerical features, with the target variable (y) indicating subscription status.
* **Outlier Detection and Handling**: Outliers in numerical columns were identified using boxplots. An interquartile range (IQR) technique was applied to mitigate their impact on the model, removing values that fell beyond 1.5 \* IQR from the 1st and 3rd quartiles.



* **Encoding Categorical Variables**: Categorical variables were transformed into a numerical format using **Label Encoding**, which maintains ordinal relationships. The target variable (y) was also label-encoded, with 1 representing subscription and 0 for non-subscription.

## **Data Splitting**

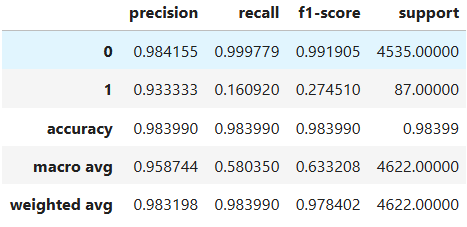
* **Training and Testing Split**: The dataset was split into training (80%) and testing (20%) sets to ensure model generalization.

## **Model Building and Tuning**

* **Algorithm Choice**: A **Decision Tree Classifier** was selected for this project.
* **Hyperparameter Tuning with GridSearchCV**: A 5-fold cross-validation was performed with various hyperparameters:
  1. max\_depth: None, 10, 20, 30, 40
  2. min\_samples\_split: 2, 10, 20
  3. min\_samples\_leaf: 1, 5, 10
  4. max\_features: 'sqrt', 'log2', None
  5. class\_weight: None, 'balanced'
* **Best Model Parameters**: The optimal parameters obtained were as follows:
  1. max\_depth: 10
  2. min\_samples\_split: 20
  3. max\_features: ‘sqrt’
  4. random\_state: 42

**Model Evaluation**

* **Accuracy**: The model achieved an accuracy of **98.4%** on the test set.
* **Classification Report**: The detailed classification report was as follows:



## **Findings**

**Overall Accuracy**: The model achieved an overall accuracy of approximately **98.40%**. This indicates that the model correctly classified around 98.40% of the test samples, which is an excellent performance.

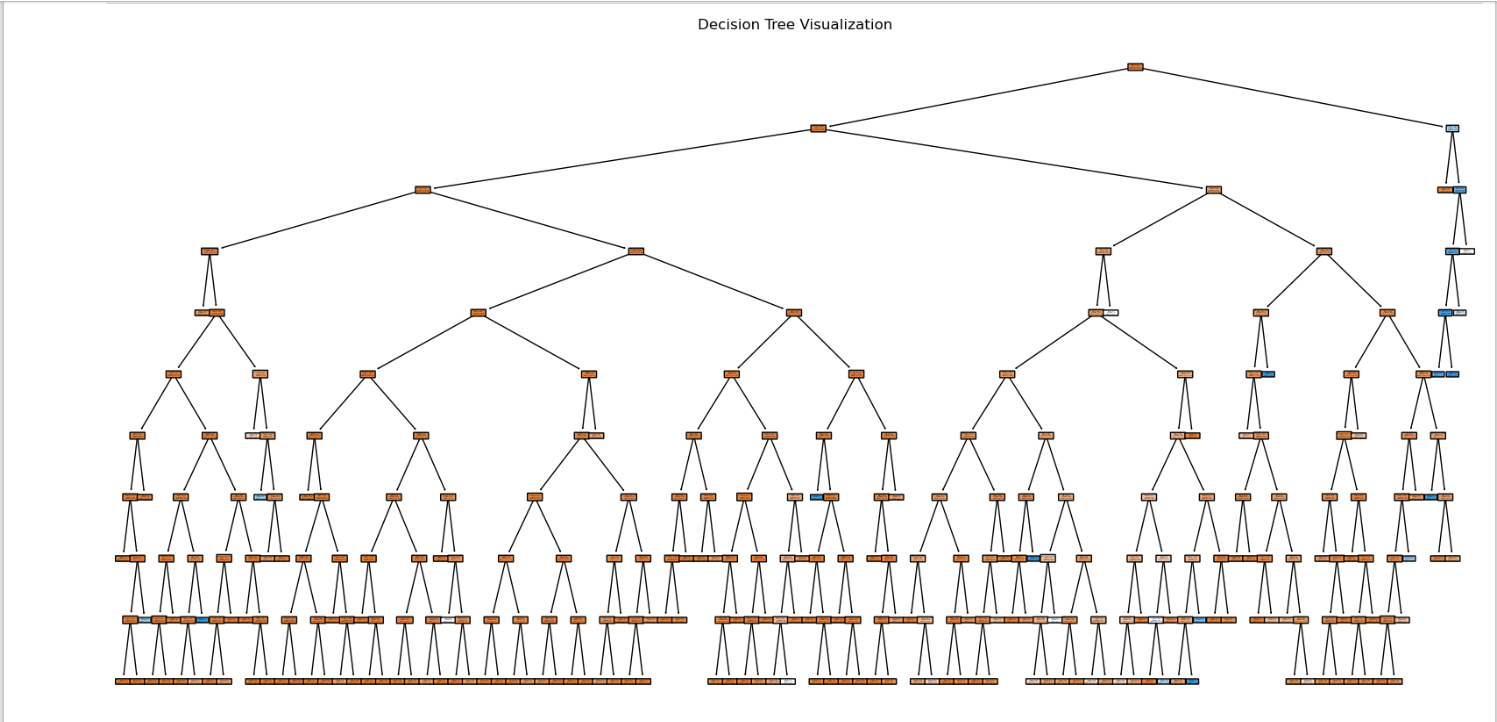
**Class Performance**:

* **Class 0 (Not Subscribed)**:
  1. **Precision**: 98.42% - The model is very precise in predicting non-subscribers, meaning that when it predicts a customer will not subscribe, it is correct about 98.42% of the time.
  2. **Recall**: 99.98% - The model successfully identifies nearly all actual non-subscribers, with a recall rate almost reaching 100%. This suggests that it has a low false negative rate for this class.
  3. **F1-score**: 99.19% - This harmonic mean of precision and recall indicates a strong balance between the two for class 0.
* **Class 1 (Subscribed)**:
  1. **Precision**: 93.33% - While still strong, the precision for predicting subscribers is lower compared to non-subscribers. This means that of all instances predicted to be subscribers, about 93.33% were indeed subscribers.
  2. **Recall**: 16.09% - The recall is significantly lower, indicating that the model is missing a large proportion of actual subscribers. This suggests a potential issue with the model's ability to recognize this minority class.
  3. **F1-score**: 27.45% - This low F1-score highlights the imbalance in performance, indicating that while the model is good at predicting non-subscribers, it struggles with identifying subscribers.

**Class Imbalance**:

* The support values reveal a significant class imbalance in the dataset, with **4535** samples of non-subscribers (class 0) and only **87** samples of subscribers (class 1). This imbalance could explain the high precision and recall for class 0 and the poor performance metrics for class 1

**Decision Tree Viz**:



## **Conclusion and Recommendations**

* The Decision Tree model successfully predicted term deposit subscriptions with notable accuracy, recall, and precision on balanced classes.
* While the model exhibits excellent overall accuracy and is very effective at identifying non-subscribers, it is significantly less effective at identifying subscribers. This indicates a need for exploring different algorithms that might better handle class imbalances, such as Random Forests or Gradient Boosting.