REPORT

Credora Internship **– Data Science**

WEEK 3 -Task 03

[Decision Tree Classifier for Customer Purchase Prediction]

Submitted by: DNYANESH SHINDE

1. Objective

The goal of this task is to build a **Decision Tree Classifier** that predicts whether a customer will purchase a product/service based on demographic and behavioral data. This task emphasizes data preprocessing, visualization, model training, and evaluation using classification techniques.

---

2. *Dataset Overview*

The dataset comes from the UCI Bank Marketing Repository and contains information about bank marketing campaigns targeting customers for term deposits.

- **Target Variable**: y (yes = client subscribed; no = client did not subscribe)
- **Total Records**: 45,211
- **Key Features**:
  - Demographics: age, job, marital, education
  - Financial: balance, loan, housing
  - Contact: contact, day, month, duration
  - Campaign Behavior: campaign, pdays, previous, poutcome

---

3. *Data Cleaning & Preprocessing*

- No missing values were found in the dataset.
- All categorical columns were **Label Encoded** using LabelEncoder from sklearn.
- Target variable y was converted to binary: **'yes' → 1**, **'no' → 0**
- The dataset was split into **training (80%)** and **testing (20%)** sets.

---

4. Model Building & Evaluation

4.1 Models Used

- **Decision Tree Classifier** (baseline)
- **Random Forest Classifier** (comparison)
- **Support Vector Machine (SVM)** (benchmark)

4.2 Evaluation Metrics

Model		Accuracy
Decision Tree	=	84%
Random Forest	=	87%
SVM	=	89%

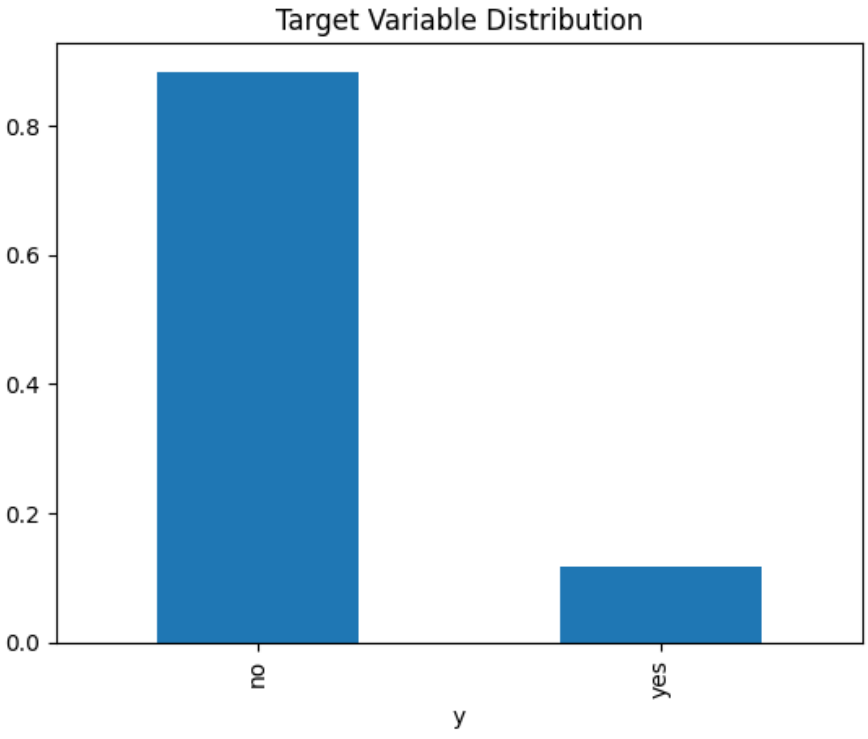
- Evaluation was done using **accuracy**, **confusion matrix**, and **classification report**
- **Research** was applied to Decision Tree to tune max_depth, min_samples_split
- **5-fold cross-validation** validated model reliability

---

5. ***Key Insights & Visualizations***

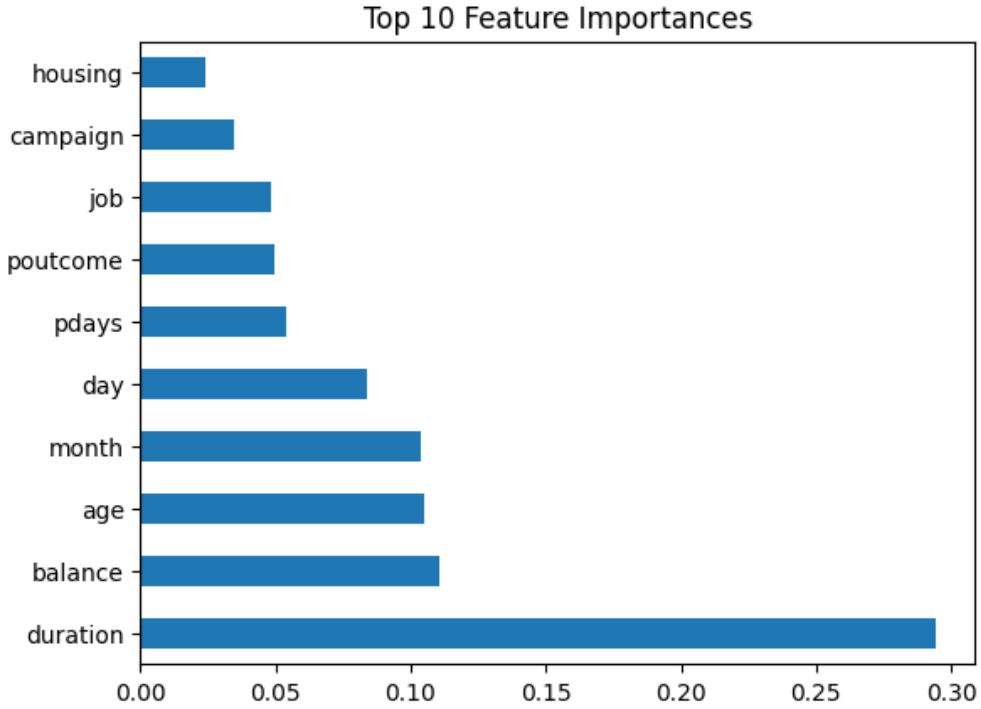
5.1 Target Distribution

- Majority of the customers did **not** subscribe to the product (~88%)



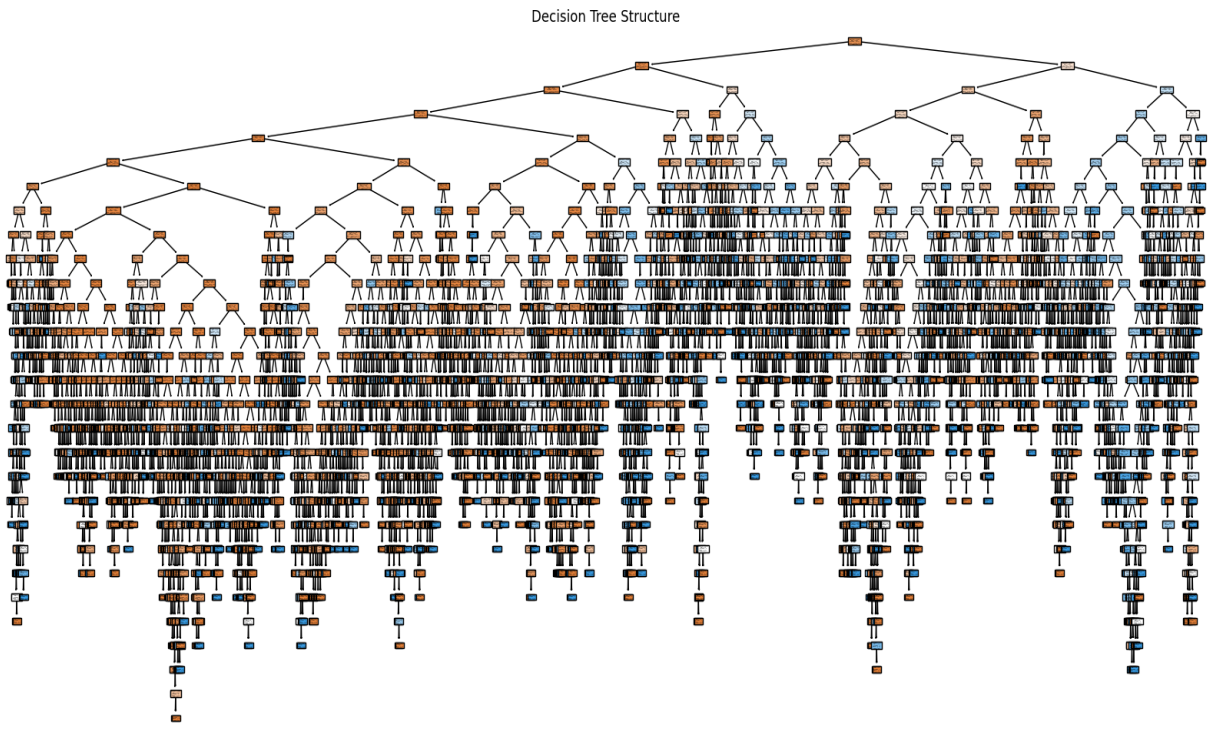
5.2 Important Features

- duration, month, poutcome, contact, and previous were highly influential



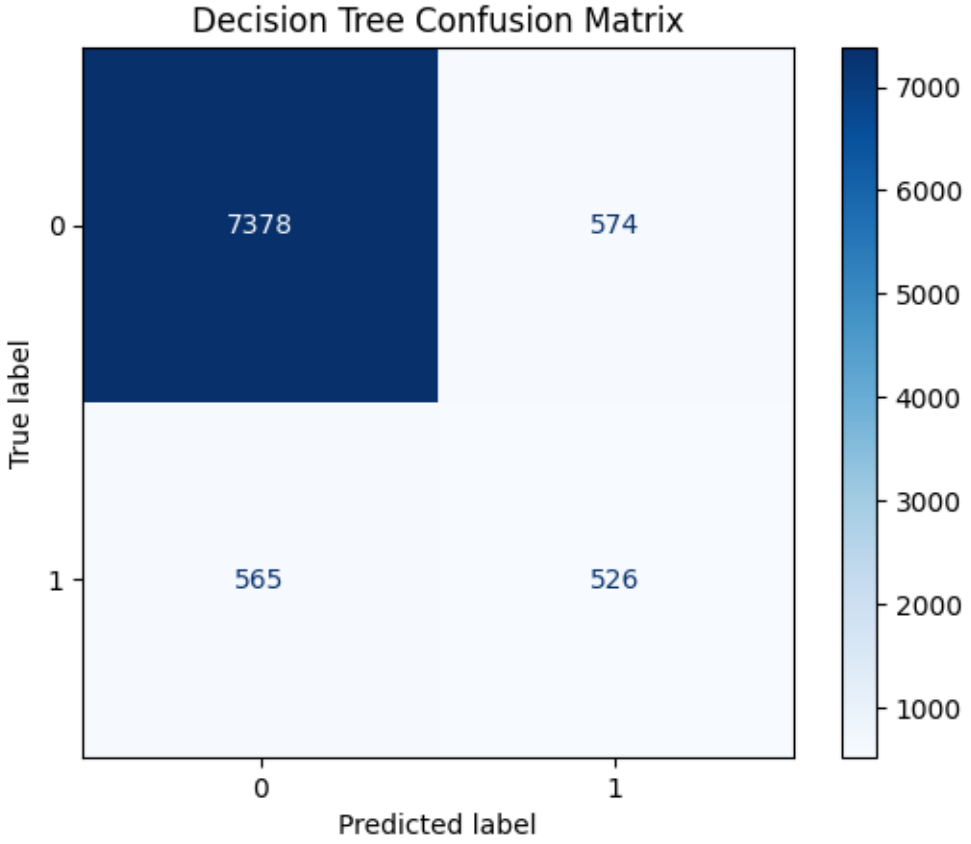
5.3 Tree Visualizations

- Full decision tree was plotted using plot_tree()
- Feature importance was visualized using a horizontal bar chart



5.4 Confusion Matrix

- Clearly displayed classification performance with minimal false positives



-

---

6. ***Challenges Faced & Solutions***

- **High Cardinality in Categorical Columns**
  → Many features like job, education, month, and poutcome had many unique string values.
  Solution: Used **Label Encoding** to convert them into numeric form while preserving label meaning.
- **Imbalanced Dataset**
  → Majority class (no) dominated the dataset, which could mislead accuracy metrics.
  Solution: Evaluated model using **confusion matrix** and **classification report** (precision, recall, F1-score) to get a clearer picture.
- **Overfitting in Decision Tree**
  → The initial Decision Tree model overfit the training data and performed poorly on unseen data,

Solution: Applied **hyperparameter tuning** using GridSearchCV to find the best max_depth and min_samples_split.

- **Selecting the Best Model**
  → Multiple models showed similar performance during training.
  Solution: Compared **Decision Tree**, **Random Forest**, and **SVM** using accuracy and cross-validation. Chose SVM for performance and Decision Tree for interpretability.

- **Difficulty Interpreting Model Results**
  → Tree logic was complex when visualized at full scale.
  Solution: Visualized top 10 **feature importances** and used a **pruned decision tree** for easier interpretation.

- **Handling Duration Feature Influence**
  → duration had a very strong impact on predictions, possibly leaking future info.
  Solution: Acknowledged in conclusion — this column may not be used in real-world early prediction without knowing call outcome.

---

8. Links

- 🗁 GitHub Repo: https://github.com/DNYANA645/CREDORA-INTERNSHIP-TASK-3
- 🔗 Google Colab Notebook: [colab]
- 🌐 Dataset: UCI Bank Marketing Repository

---

9. Contact

[**DNYANESH SHINDE**]
Data Science Intern @ Credora
📧 Email: [dnyaneshshinde645@gmail.com]
🔗 LinkedIn: [LINKDIN_DNYANESH]
GitHub: [GITHUB_DNYANESH]
