BRAC
UNIVERSITY

Inspiring Excellence

Predictive Modeling and Clustering for Travel Insurance Claims using Machine Learning

Group - 03

Submitted By

Name	ID
Sabbir Bin Abdul Latif	**23341056**
Rizwanul Islam	**21201129**

Submitted To

Mr. Rafeed Rahman
Senior Lecturer
Department of Computer Science and Engineering

Mahjabin Chowdhury
Adjunct Lecturer
Department of Computer Science and Engineering

1. Introduction

This project focuses on applying machine learning algorithms to an insurance dataset in order to predict claim occurrences. The problem is framed as a classification task, where the target variable indicates whether a claim was made or not. The motivation behind this work is to explore how data preprocessing, handling imbalance, and model selection affect prediction performance in real-world scenarios.

2. Dataset Description

- **Dataset Size:** 63,326 rows and 11 features
- **Target Variable:** **Claim** (binary: 0 = No claim, 1 = Claim)
- **Number of Features:** 10 input features + 1 output features
- **Feature Type:**
  - **Quantitative:** **Duration, Net Sales, Commission (in value), Age.**
  - **Categorical:** **Agency, Agency Type, Distribution Channel, Product Name, Destination, Gender.**
- **Encoding: Encoding:** Categorical variables were encoded numerically since ML models cannot process raw text data.
- **Correlation Heatmap:** We analyzed correlation between features and the target variable.

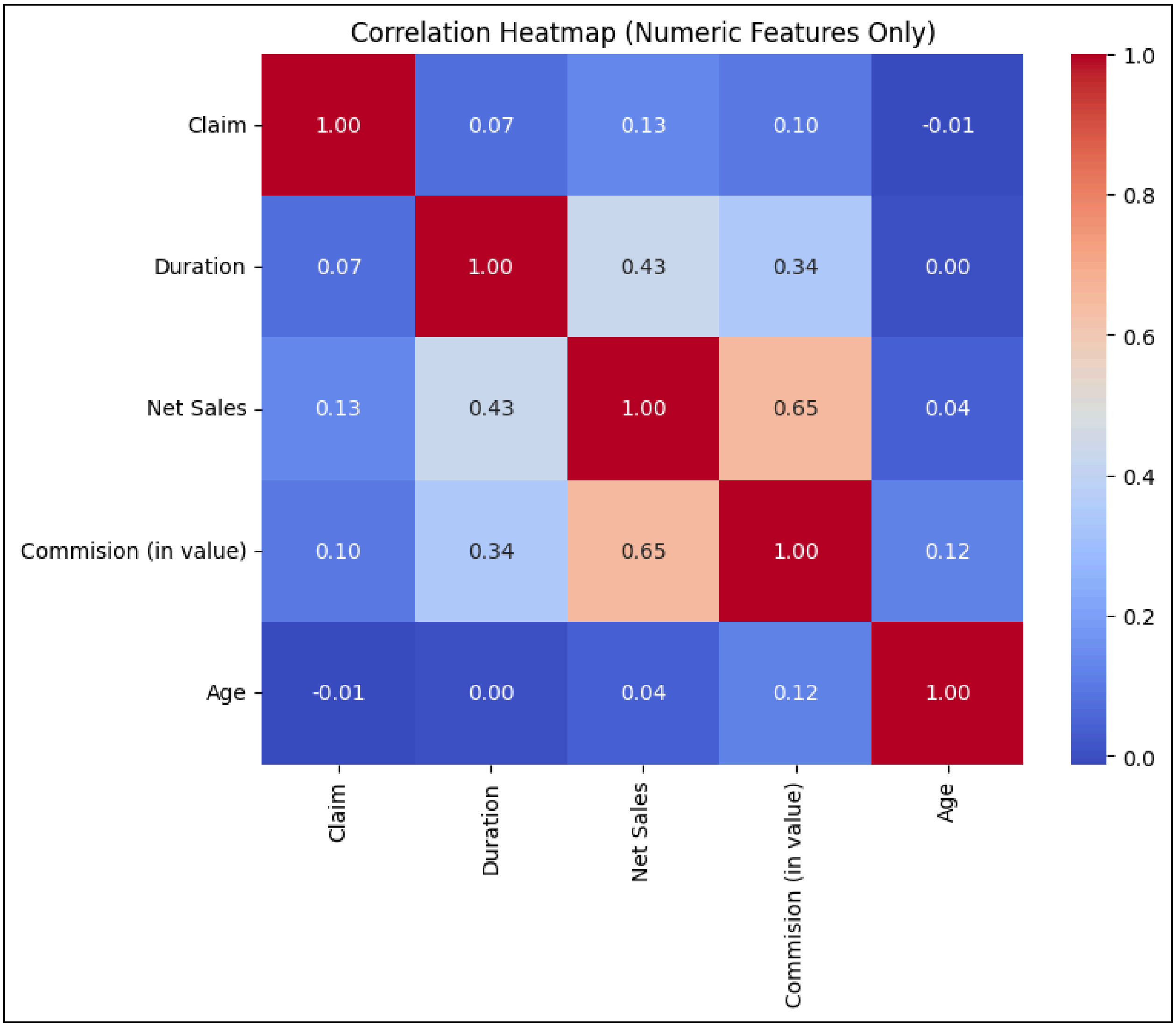

Figure 2.1: Correlation Heatmap

Understanding from Correlation Test: The heatmap revealed that most numerical features had weak correlations with the target variable, indicating that prediction relies on subtle patterns rather than strong linear relationships

3. Imbalanced Dataset & Exploratory Data Analysis

- **Claim Distribution:**
  - **No Claim (0):** 62,399 instances; **Claim (1):** 927 instances

This shows heavy imbalance between classes, which can bias model performance

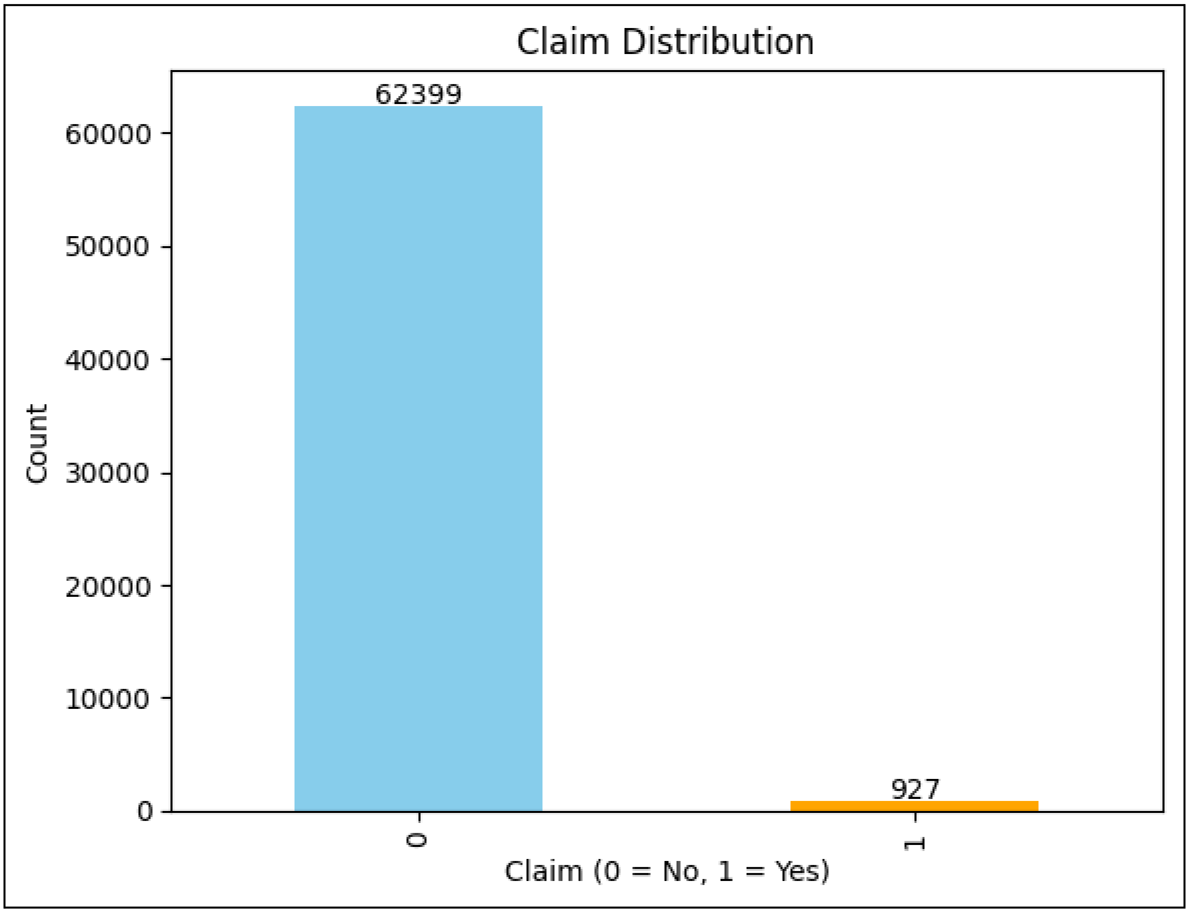

Figure 3.1: Claim Distribution

- **Claim Rate by Agency:** Some agencies showed higher claim tendencies, revealing potential business insights.

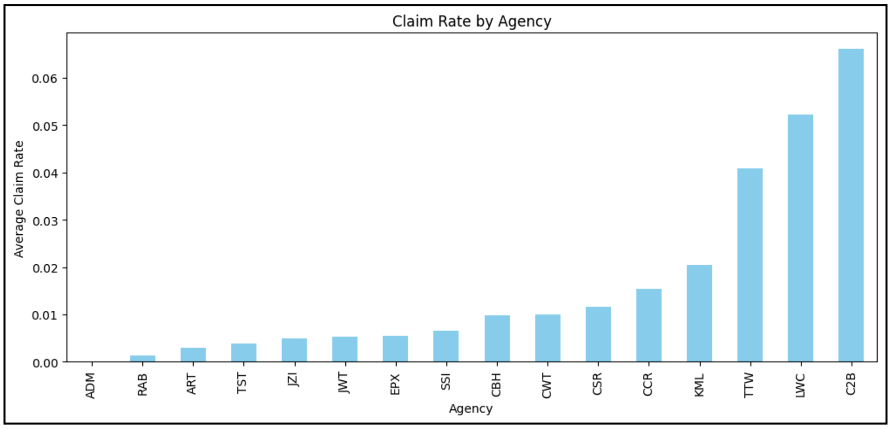

Figure 3.2: Claim Rate by Agency

- **Observation:** Since imbalance is extreme (only ~1.46% claims), models tend to predict the majority class (No Claim) more often.

4. Dataset Preprocessing

- **Missing Values:**
    - Only the Gender column had missing values (45,107 missing).
    - Handled by dropping/imputing as appropriate

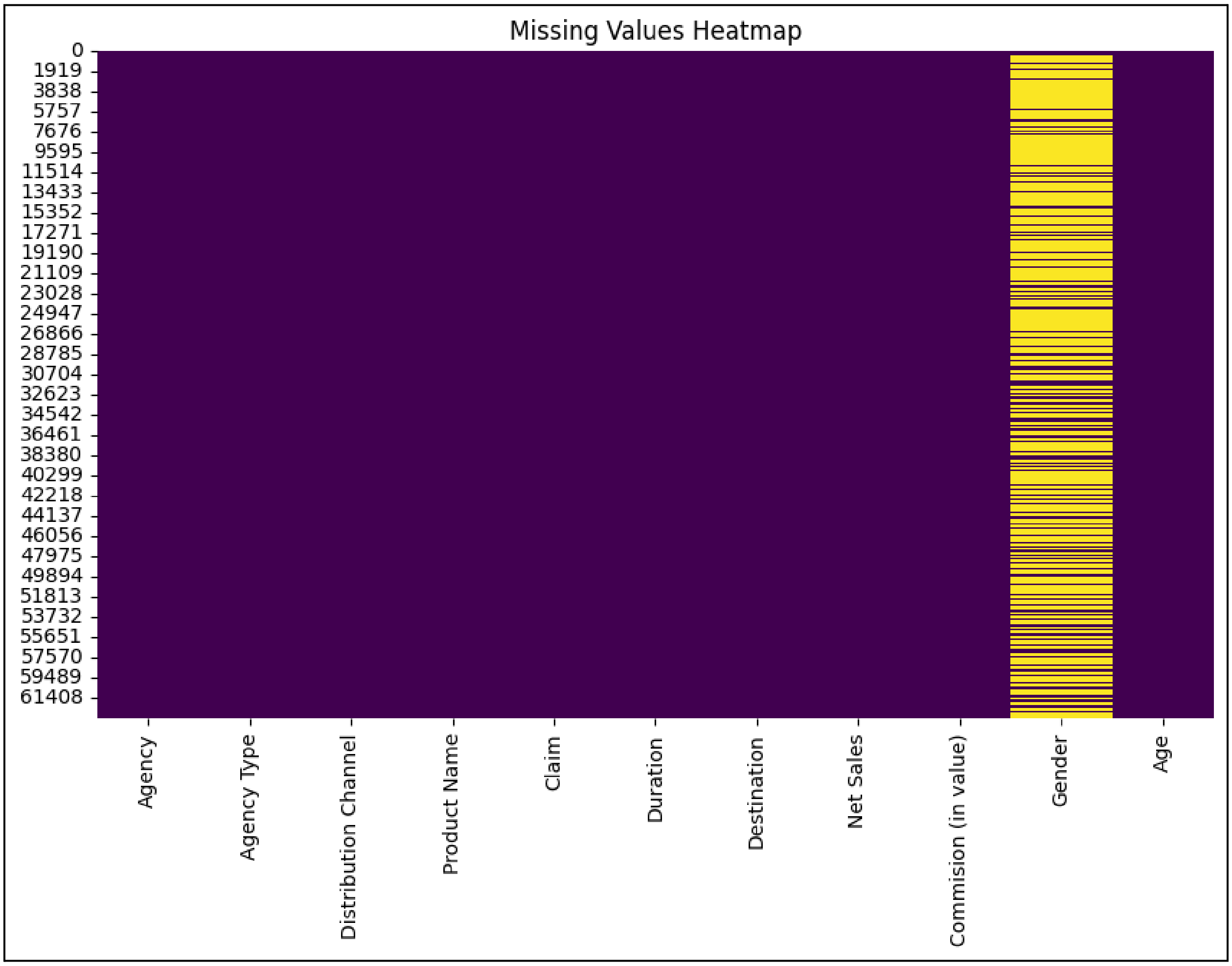

Figure 4.1: Missing Values Heatmap

- **Categorical Encoding:** Label encoding and one-hot encoding were applied where necessary.
- **Feature Scaling:** Continuous features (**Net Sales, Commision, Duration, Age**) were scaled to normalize ranges for distance-based and neural models.

5. Dataset Splitting

- **Train/Test Split:**
  - Training Set: 50,660 samples (80%)
  - Test Set: 12,666 samples (20%)
- **Stratified Sampling:** Ensured class imbalance ratio was preserved in both splits.

y_train distribution:

- No Claim: 49,918
- Claim: 742

y_test distribution:

- No Claim: 12,481
- Claim: 185

6. Model and Training & Testing

We applied three supervised models and one unsupervised method:

- **Logistic Regression**
  - **Accuracy: 0.7952**
  - **Precision: 0.0517**
  - **Recall: 0.7514**

- **F1-Score: 0.0968**

- **AUC: 0.8312**

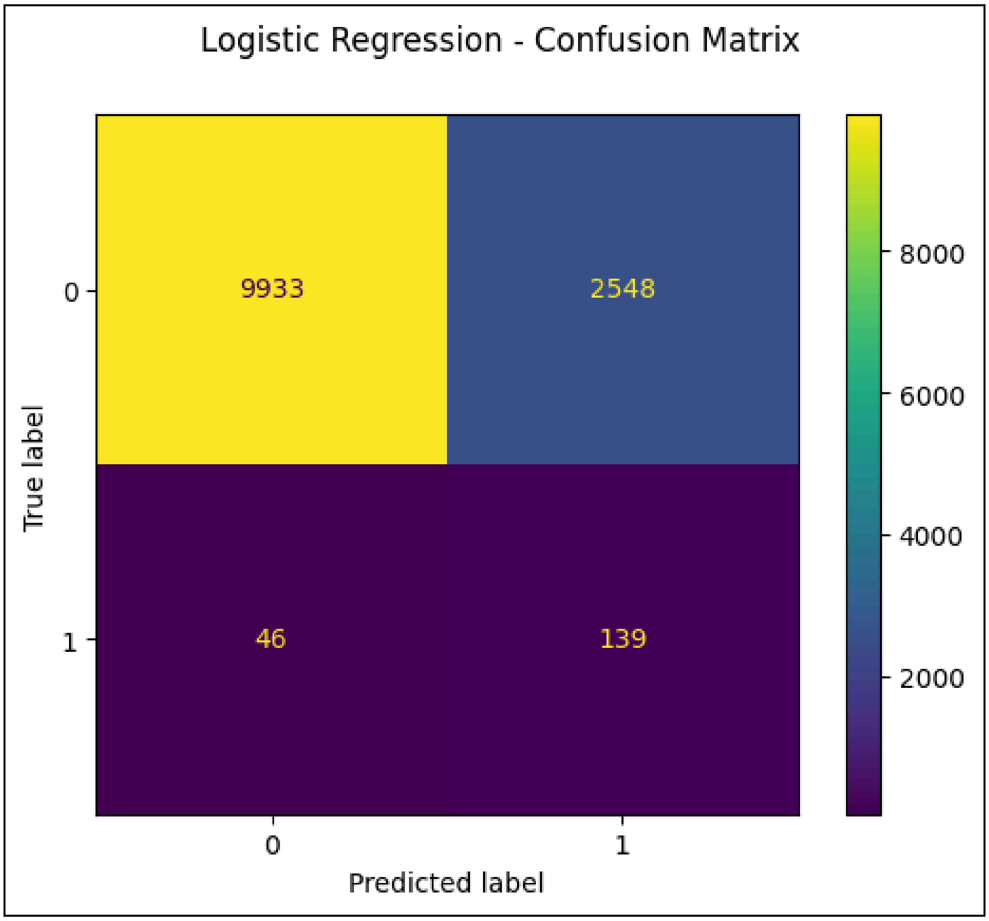

Figure 6.1: Logistic Regression Confusion Matrix

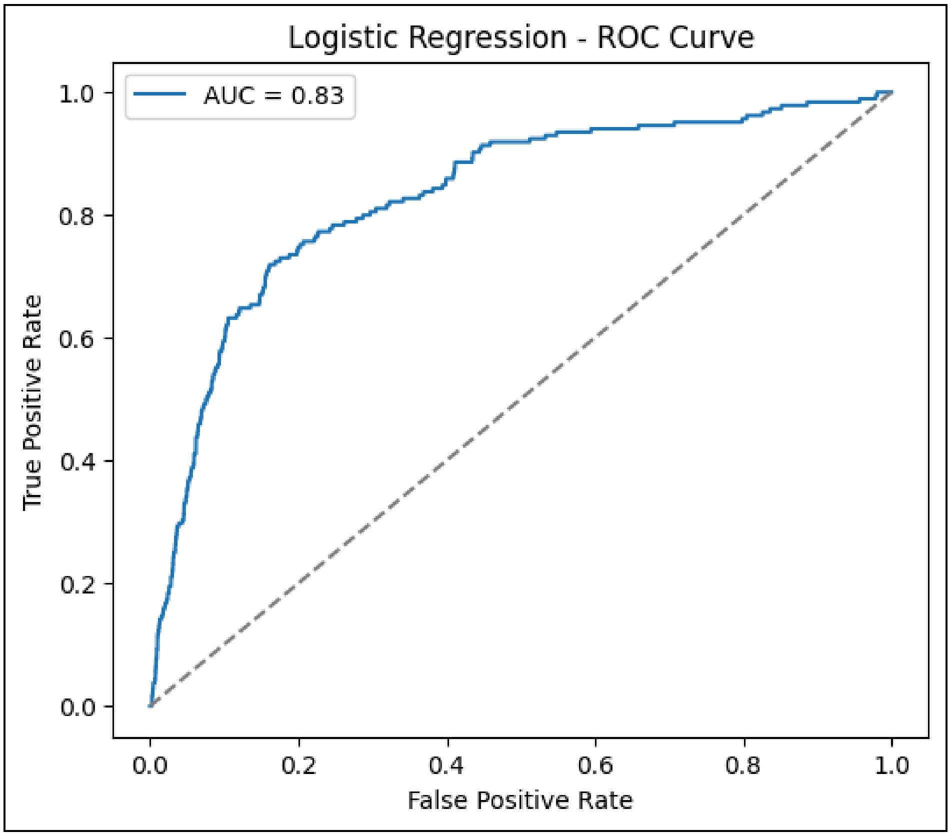

Figure 6.2: Logistic Regression ROC Curve

- **Decision Tree**
    - ○ **Accuracy: 0.9713**
    - ○ **Precision: 0.0680**
    - ○ **Recall: 0.0757**
    - ○ **F1-Score: 0.0716**
    - ○ **AUC: 0.5319**

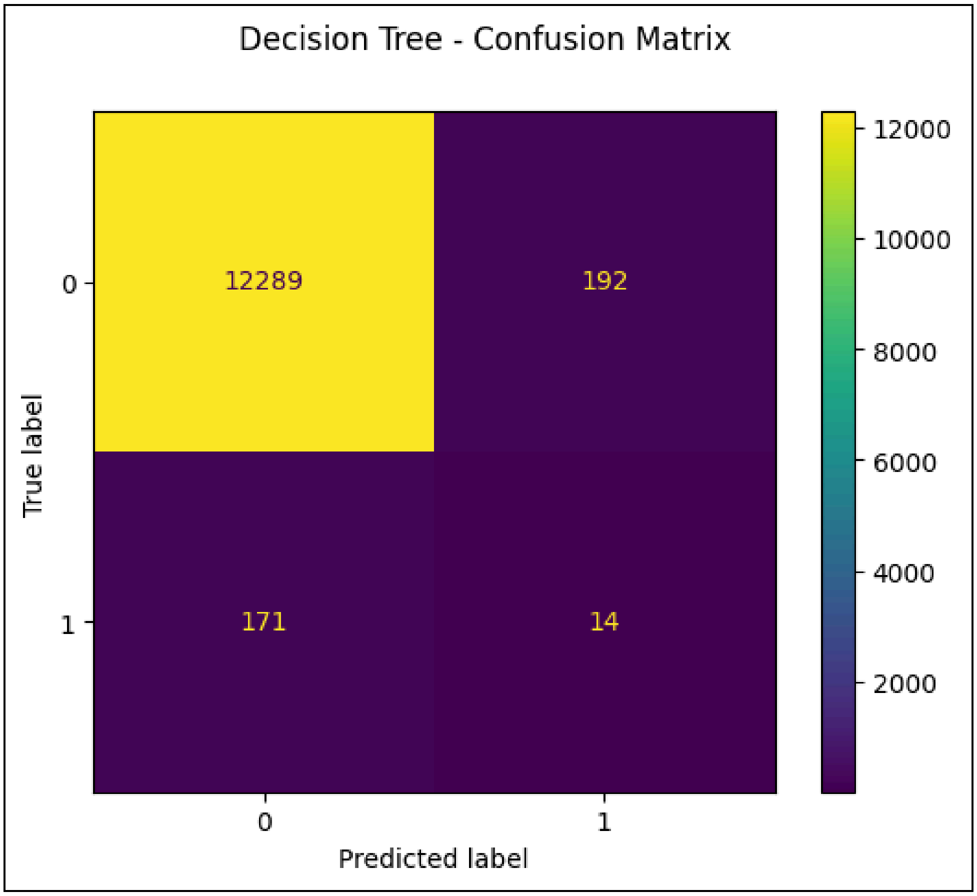

Figure 6.3: Decision Tree Confusion Matrix

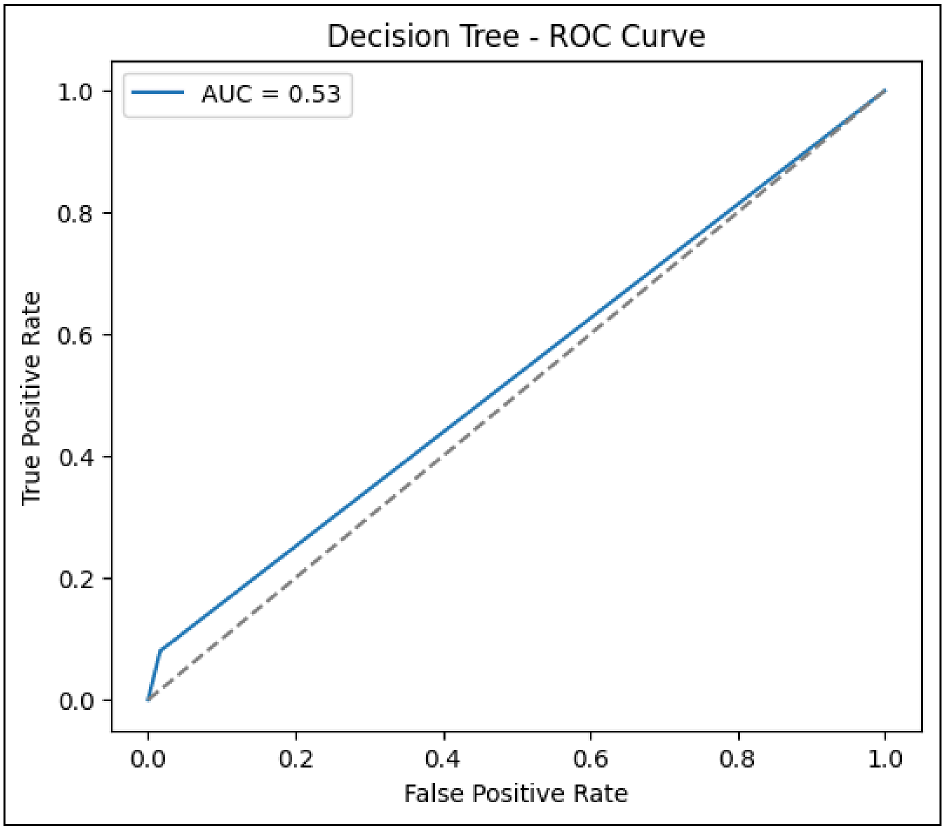

Figure 6.4: Decision Tree ROC Curve

- **Neural Network**
    - **Accuracy: 0.9854**
    - **Precision: 0.0000**
    - **Recall: 0.0000**
    - **F1-Score: 0.0000**
    - **AUC: 0.8383**

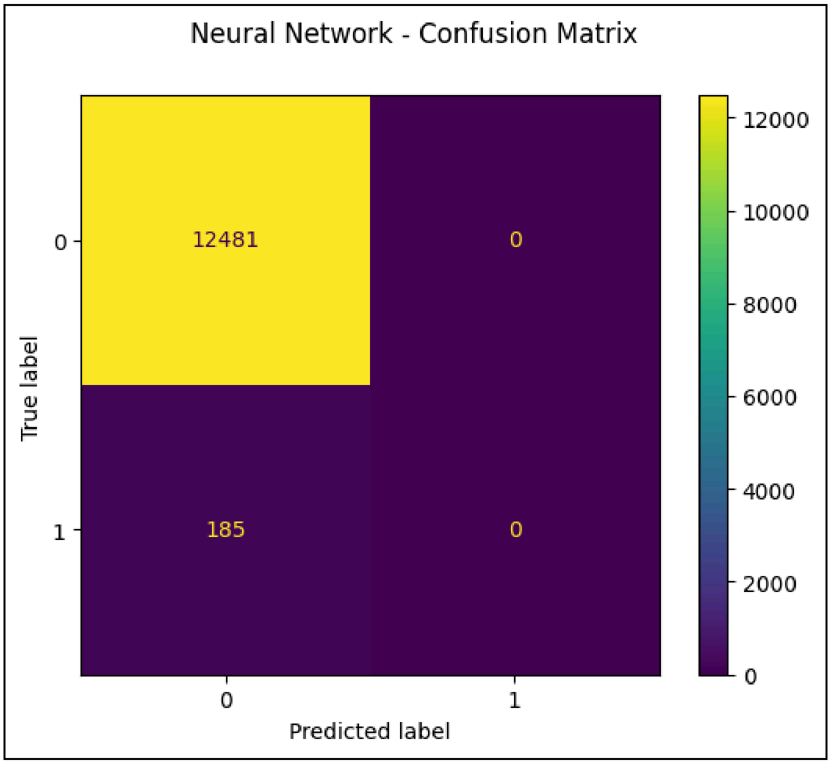

Figure 6.5: Neural Network Confusion Matrix

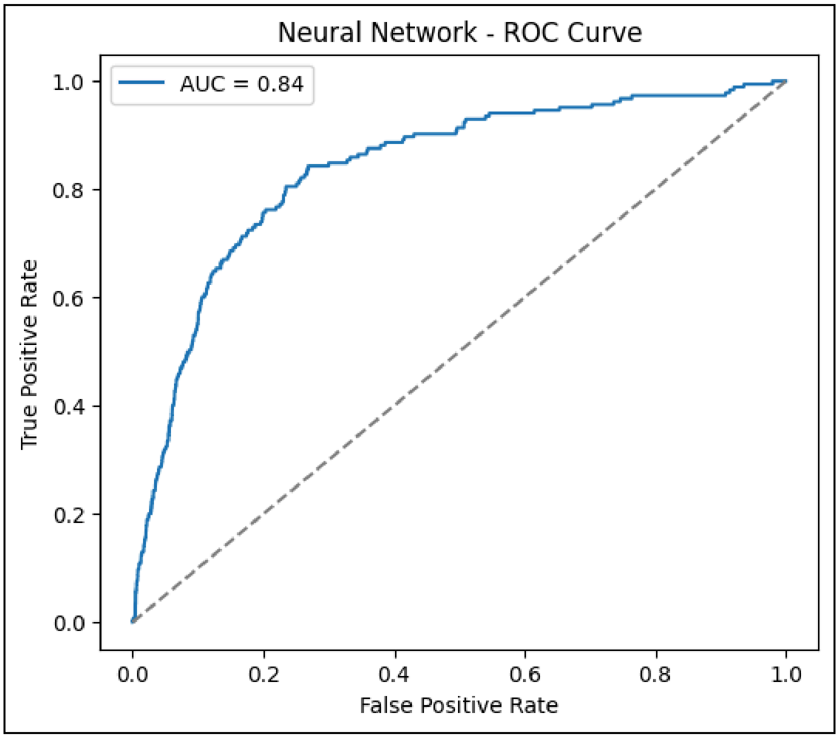

Figure 6.6: Neural Network ROC Curve

- K-Means Clustering (Unsupervised)
  - **Silhouette: 0.2835**
  - **Shows weak but visible cluster separation**

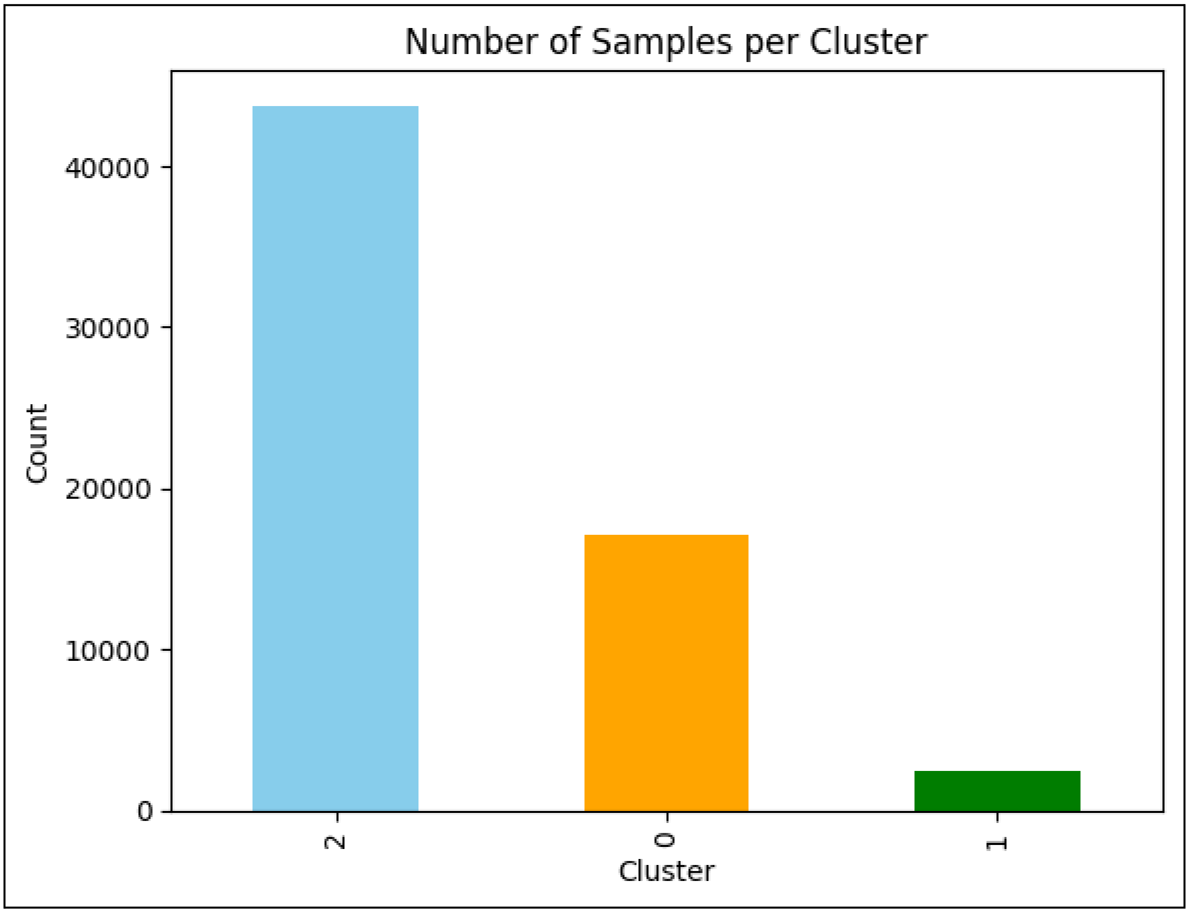

Figure 6.7: Number of Sample Per Cluster

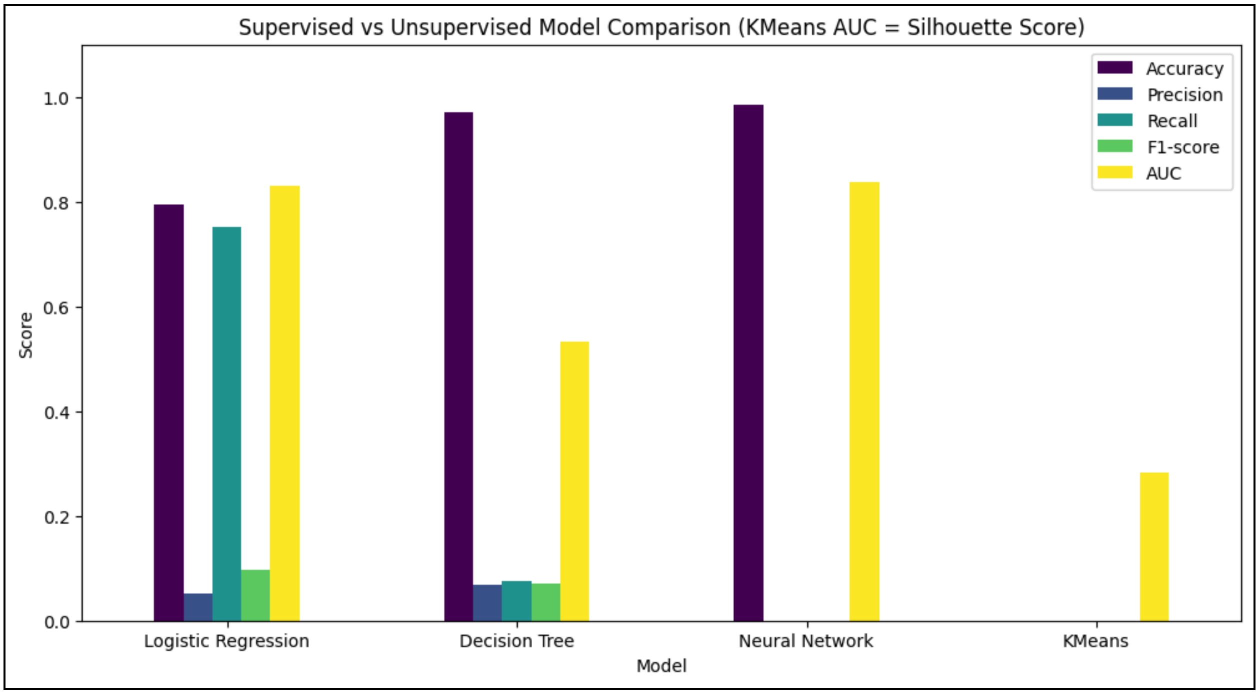

Figure 6.8: Supervised Vs Unsupervised Model Comparison

7. Model Selection / Comparison Analysis

Model comparison table:

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.7952	0.0517	0.7514	0.0968	0.8312
Decision Tree	0.9713	0.0680	0.0757	0.0716	0.5319
Neural Network	0.9854	0.0000	0.0000	0.0000	0.8383

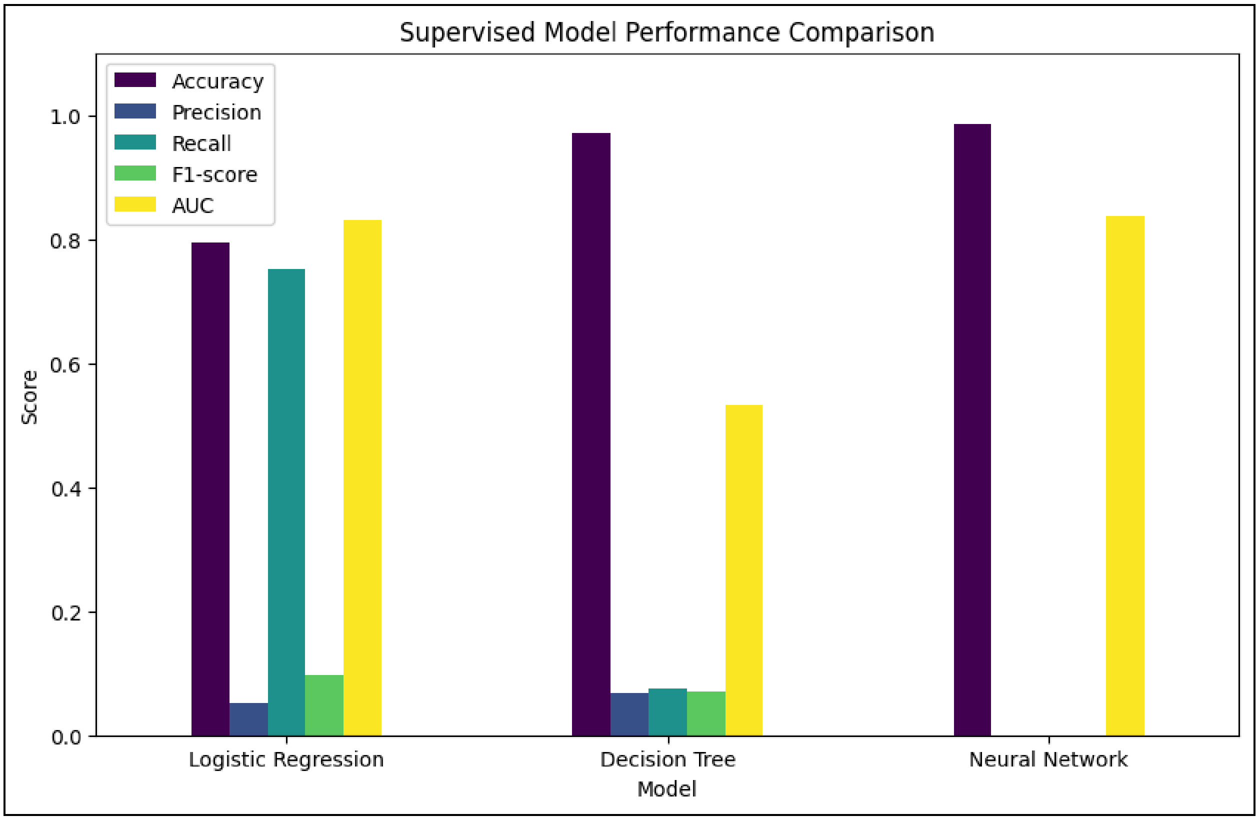

Figure 7.1: Supervised Model Performance Comparison

Key Findings:

- Logistic Regression performed best in terms of **recall** (capturing claims), but had poor precision.
- Decision Tree achieved high accuracy but very low AUC, indicating overfitting.
- Neural Network had the highest accuracy but completely failed to detect the minority class (claims).
- K-Means clustering gave moderate separation (Silhouette = 0.2835) but cannot replace supervised methods for prediction.

8. Conclusion

- **Overall Results:**

    - Neural Network gave the highest accuracy but failed in minority class detection.
    - Logistic Regression balanced recall and AUC but sacrificed precision.
    - Decision Tree overfitted, producing misleadingly high accuracy.
- **Challenges Faced:**
    - Heavy class imbalance severely reduced precision and F1-score.
    - Feature correlations were weak, making prediction harder.
    - Neural Network struggled without class balancing techniques like SMOTE.
- **Final Comment:**

  Imbalanced datasets require specialized handling. Without balancing, even high accuracy can be misleading. For future work, resampling strategies or anomaly detection models should be applied to better capture the minority "Claim" class.