MA 212 Regression Analysis

Credit Card Fraud Detection Using Machine Learning

TEAM MEMBERS

- VEERAMALLA AKSHAY (230041039)
- K. VIVEK TEJ (230041014)
- MEDA SRINIVAS TEJA (230041022)

PROJECT GOALS

To detect fraudulent credit card transactions,
as it is essential to figure out the fraudulent transactions so that customers do not get charged for the purchase of products that they did not buy.

DATA COLLECTION

Data of transactions such as:

- **Time** elapsed seconds between the first and other transactions
- **Amount** of each transcation
- **Class** fraudulent or not a fraudulent(1 or 0)

KAGGLE DATASET

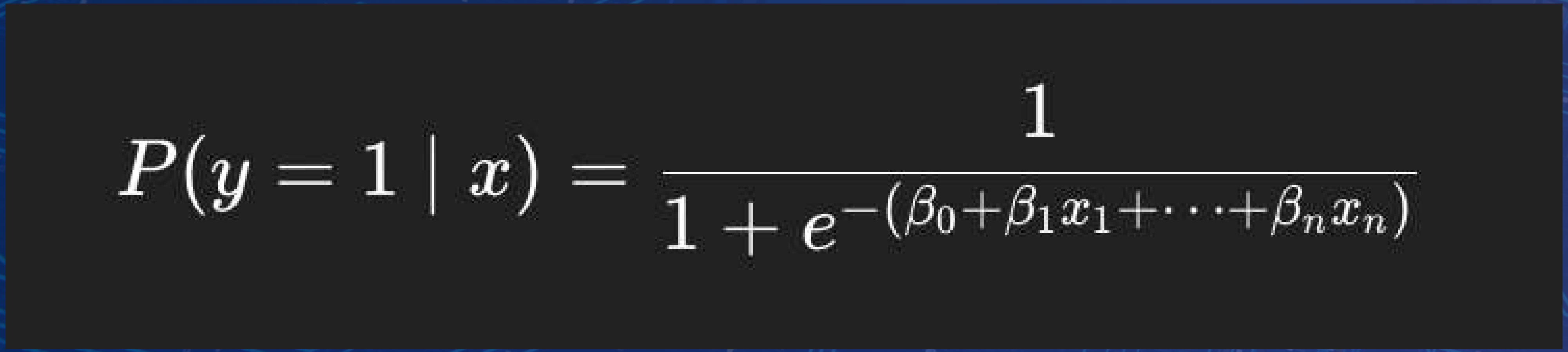
MACHINE LEARNING MODELS USED

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machine
- Decision Tree

MODEL 1: LOGISTIC REGRESSION

A linear, probabilistic model used for binary classification. It predicts the probability of a transaction being fraudulent (1) or not (0) using the sigmoid function.

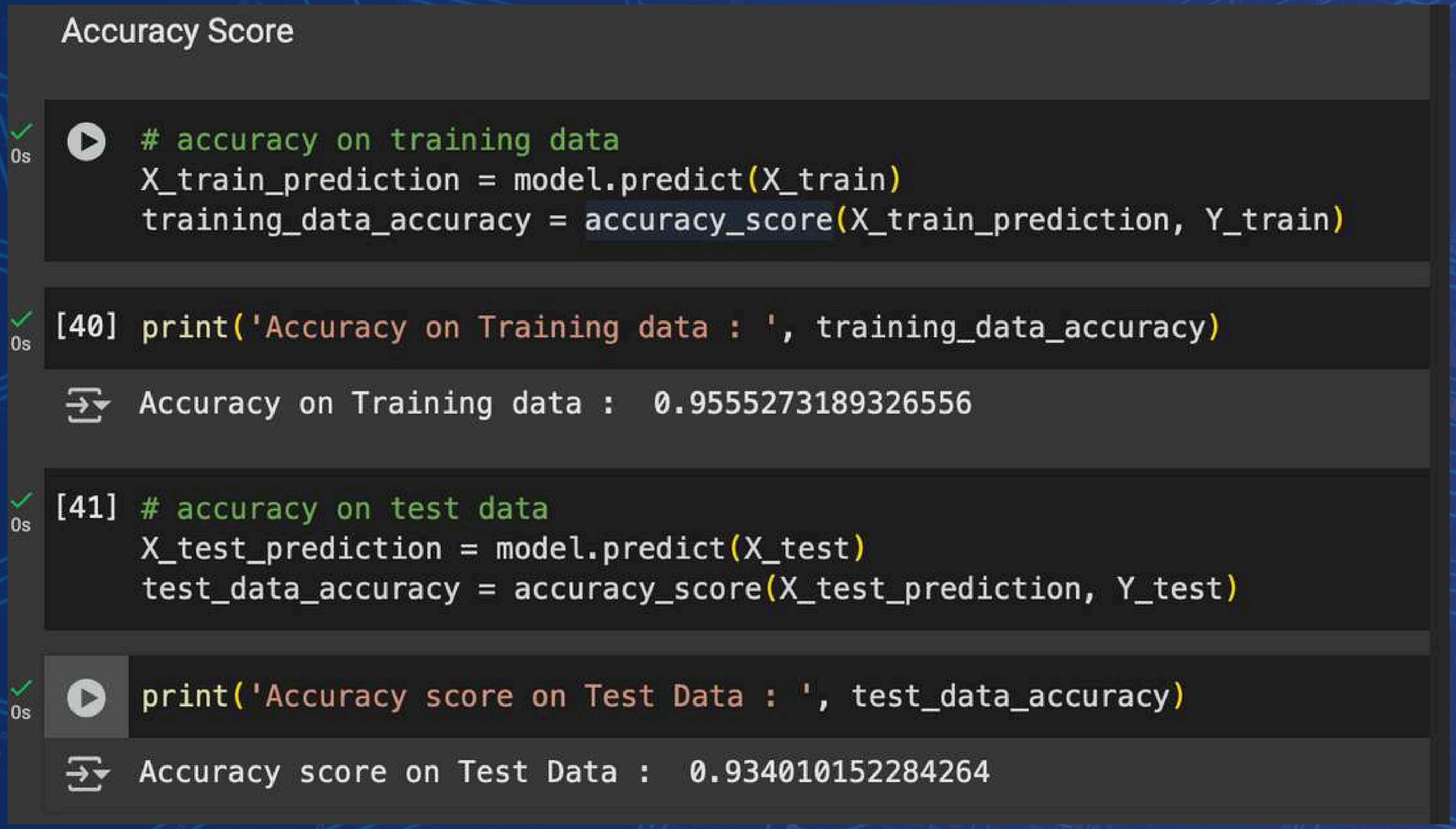
Binary classifier using sigmoid function:



$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

- Fraud implies ‘1’ and Non-Fraud implies ‘0’

- In this project we used 70% of the data for training and 30% of the data for testing
- The model managed to score an Accuracy on Training data of 95.52% , while it scored an Accuracy score on Test Data of 93.40%.



Accuracy Score

```
# accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
[40] print('Accuracy on Training data : ', training_data_accuracy)
```

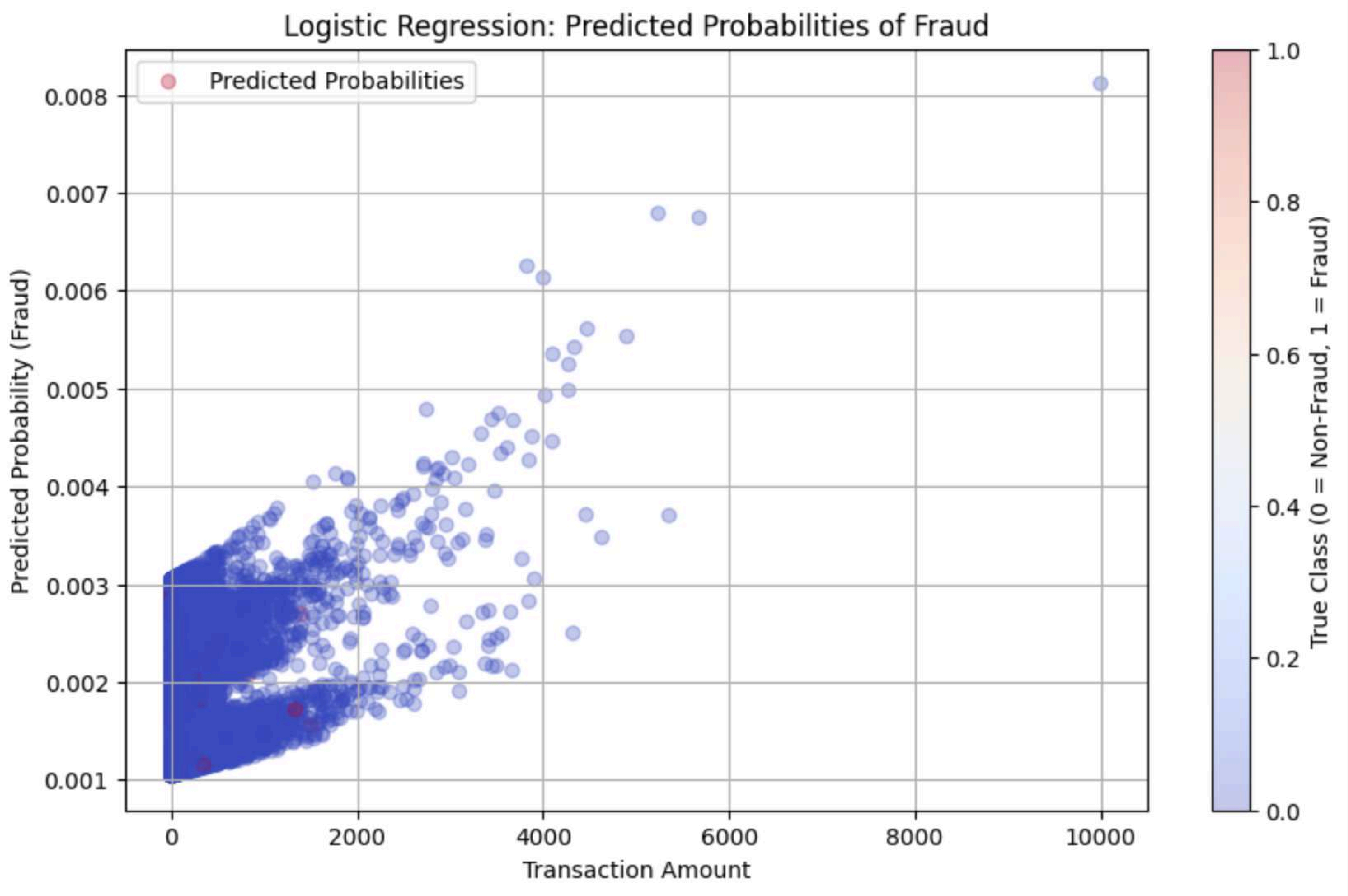
```
Accuracy on Training data :  0.9555273189326556
```

```
[41] # accuracy on test data
     X_test_prediction = model.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
print('Accuracy score on Test Data : ', test_data_accuracy)
```

```
Accuracy score on Test Data :  0.934010152284264
```

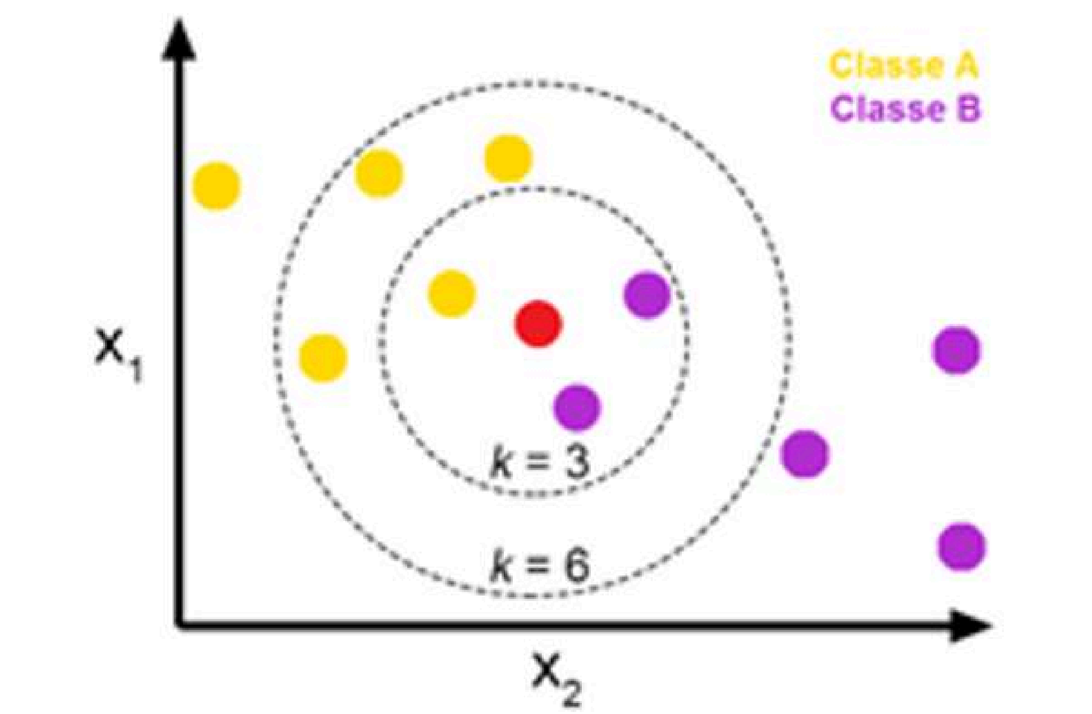
- So,If we consider high amount transaction at a very odd time then z gives a high value then probability nears 1, implies that it is fraud


Logistic Regression: Predicted Probabilities of Fraud
● Predicted Probabilities
Predicted Probability (Fraud)
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
2000
4000
6000
8000
10000
Transaction Amount
True Class (0 = Non-Fraud, 1 = Fraud)
1.0
0.8
0.6
0.4
0.2
0.0

MODEL 2: K-NEAREST NEIGHBORS (KNN)

KNN is a non-parametric, lazy learning algorithm. It classifies a transaction based on the majority class among its k-nearest neighbors using distance metrics like Euclidean distanceh text

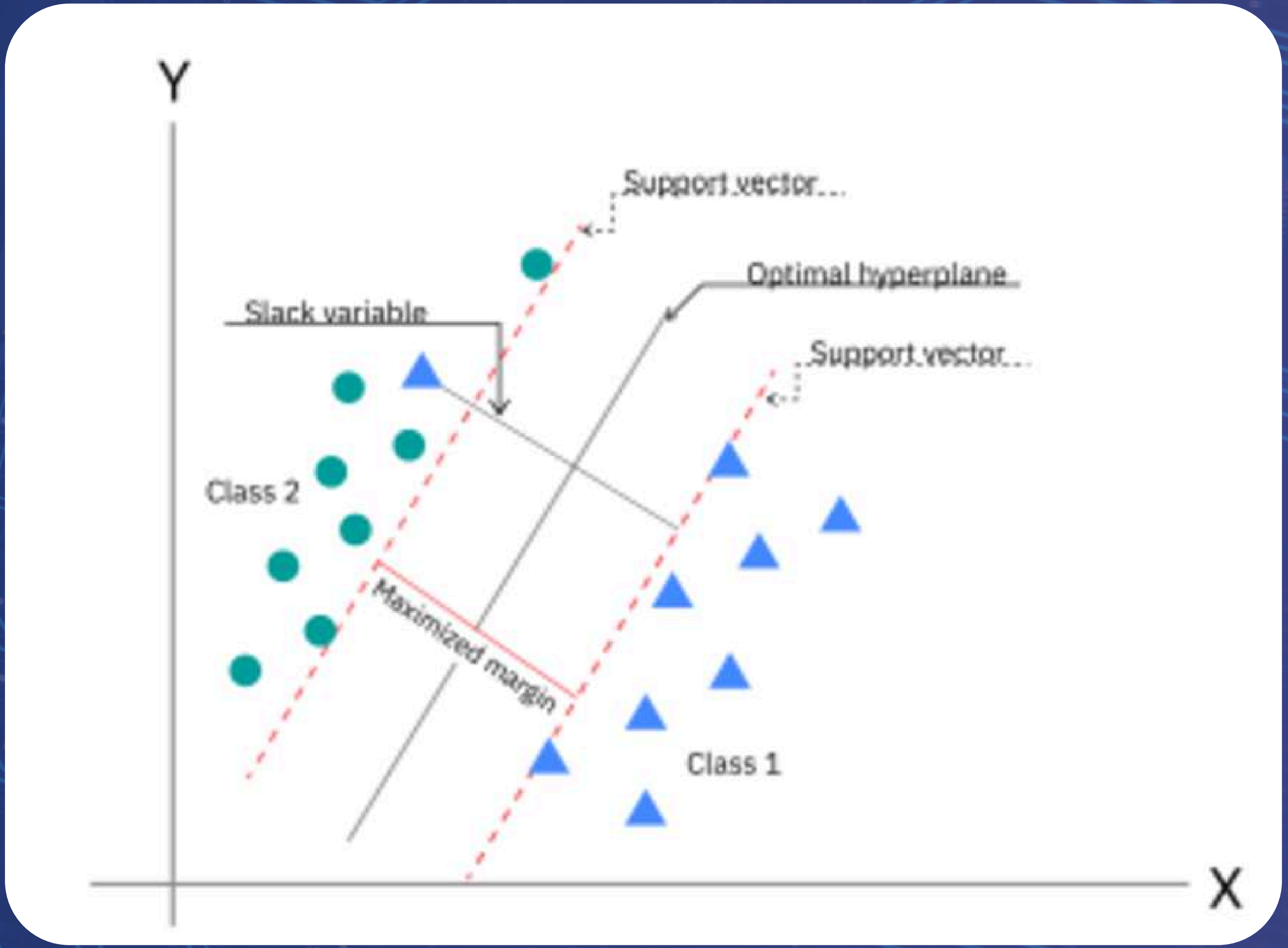
- Advantage: Extremely high accuracy
- Limitation: Computationally expensive for large datasets



MODEL 3: SUPPORT VECTOR MACHINE (SVM)

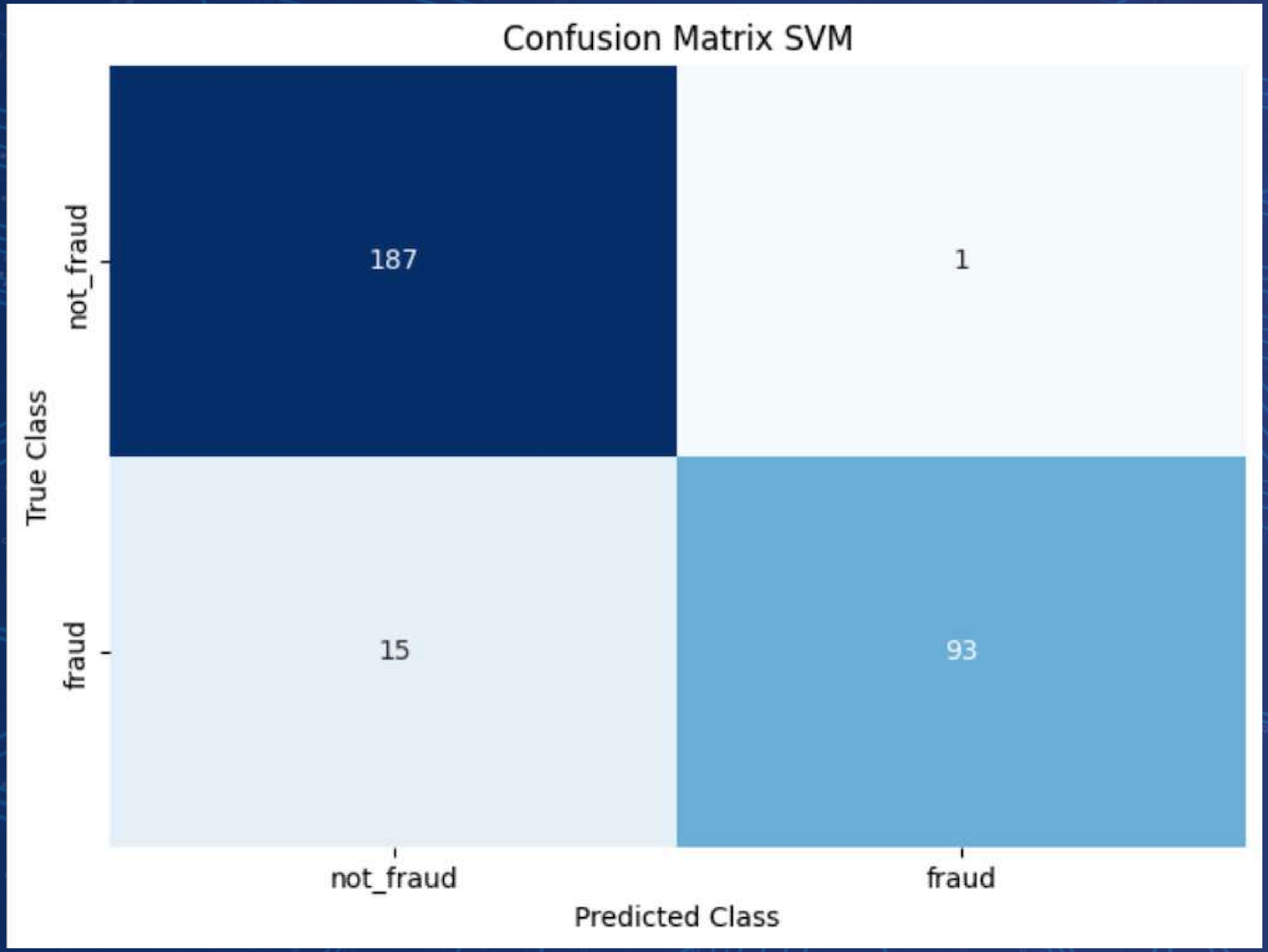
A Support Vector Machine (SVM)is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space

- SVM tries to find the optimal hyperplane that maximizes the margin between two classes.
- It works well in high-dimensional spaces.
- SVM works well when there is a clear margin between fraudulent and non-fraudulent classes.



CONFUSION MATRIX

A confusion matrix is a table that compares predicted values to actual values for a classification model.



Confusion matrix for our model

Metrics for confusion matrix are:

- Precision = TP / (TP + FP)
- Recall (Sensitivity) = TP / (TP + FN)
- F1 Score = 2 × (Precision × Recall) / (Precision + Recall)
- Accuracy = (TP + TN) .

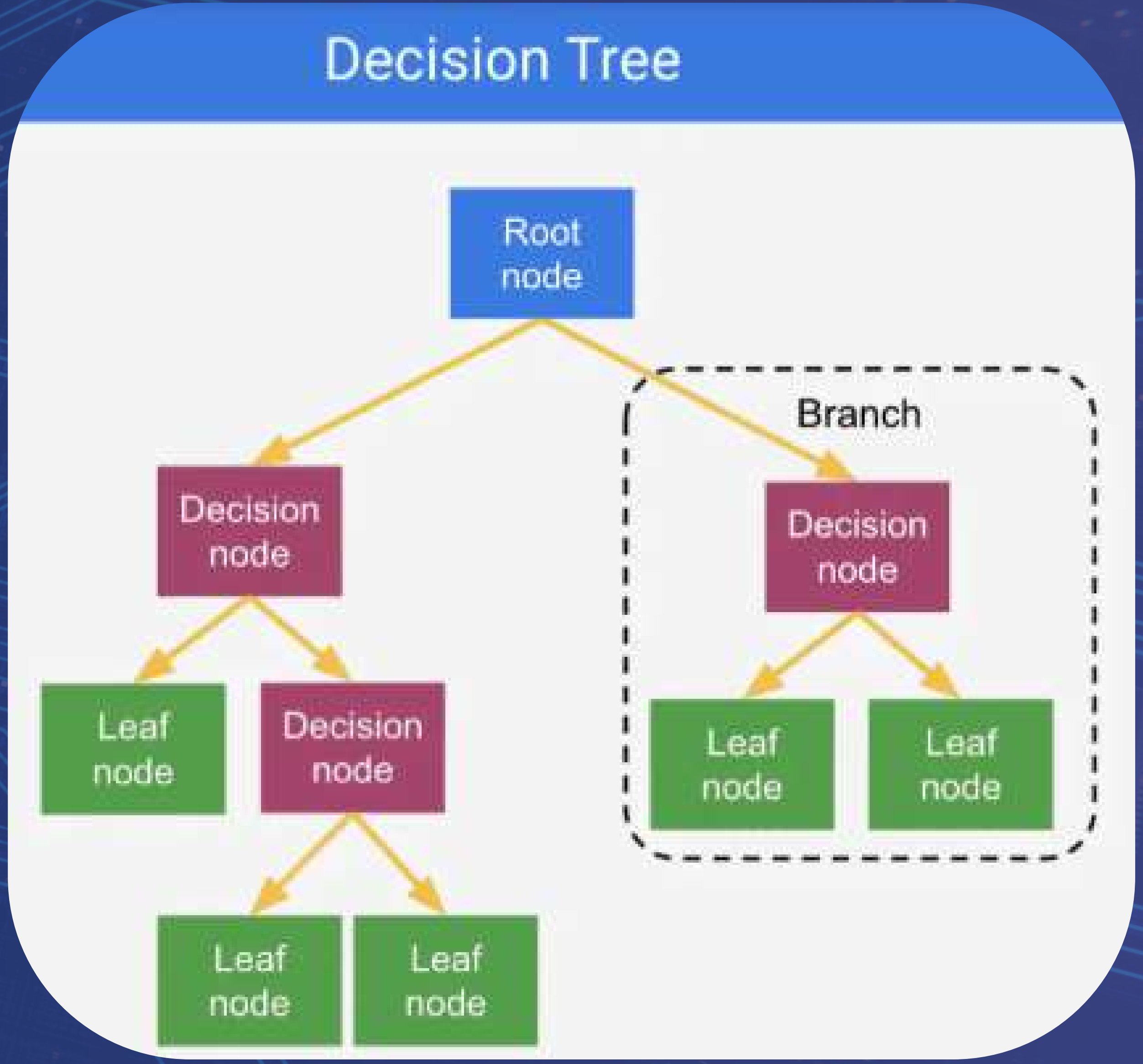
From Confusion matrix, we can derive that

Precision	**98.9%**
Recall	**86.1%**
F1 Score	**92.1%**
Accuracy	**94.6%**

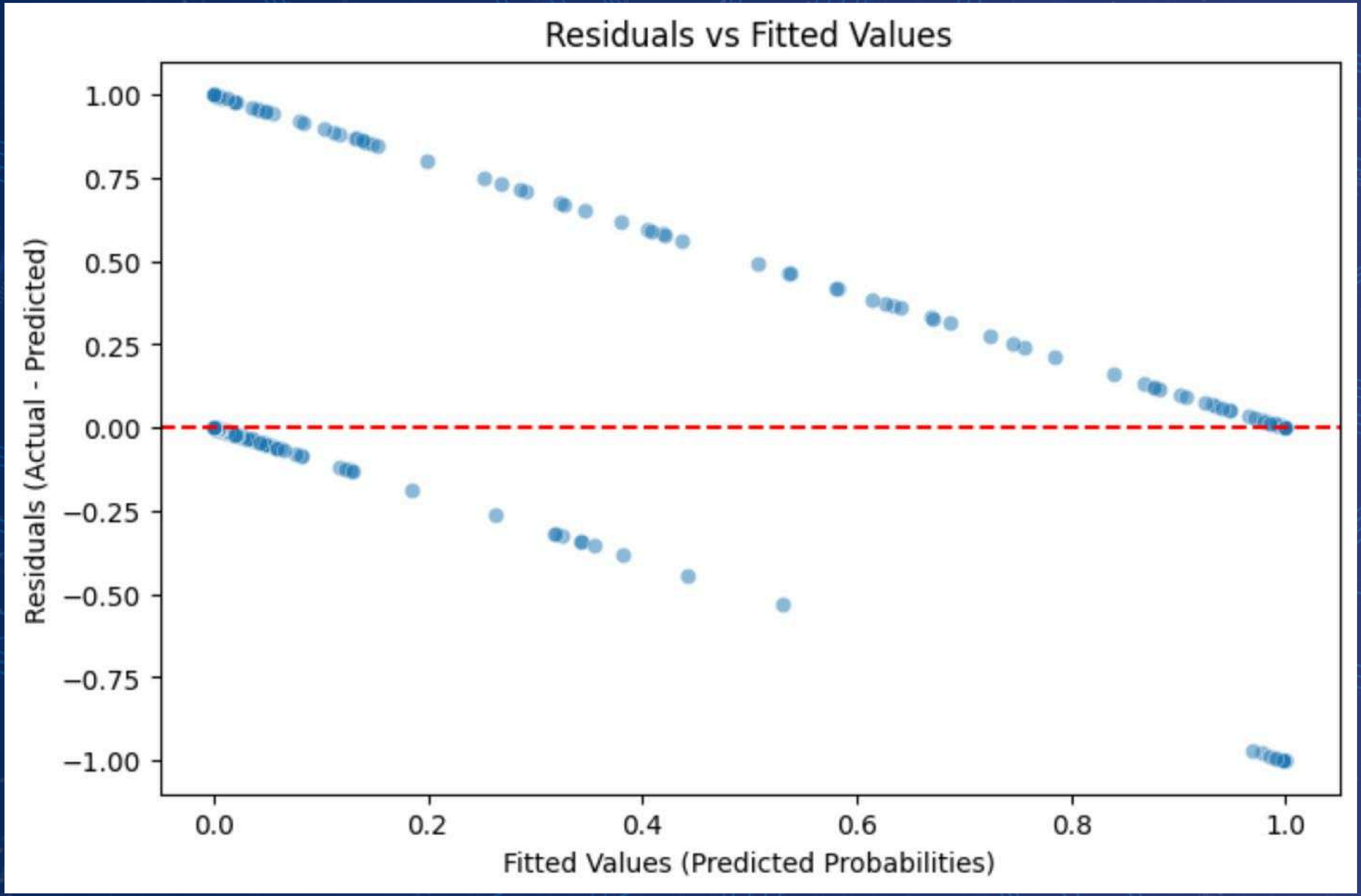
From these values we conclude that these are excellent scores especially precision

MODEL 4: DECISION TREE (DT)

A Decision Tree splits data into subsets based on feature values. It is easy to visualize and interpret. Decision Trees can easily capture non-linear patterns, which is useful in complex fraud scenarios



RESIDUAL PLOT ANALYSIS

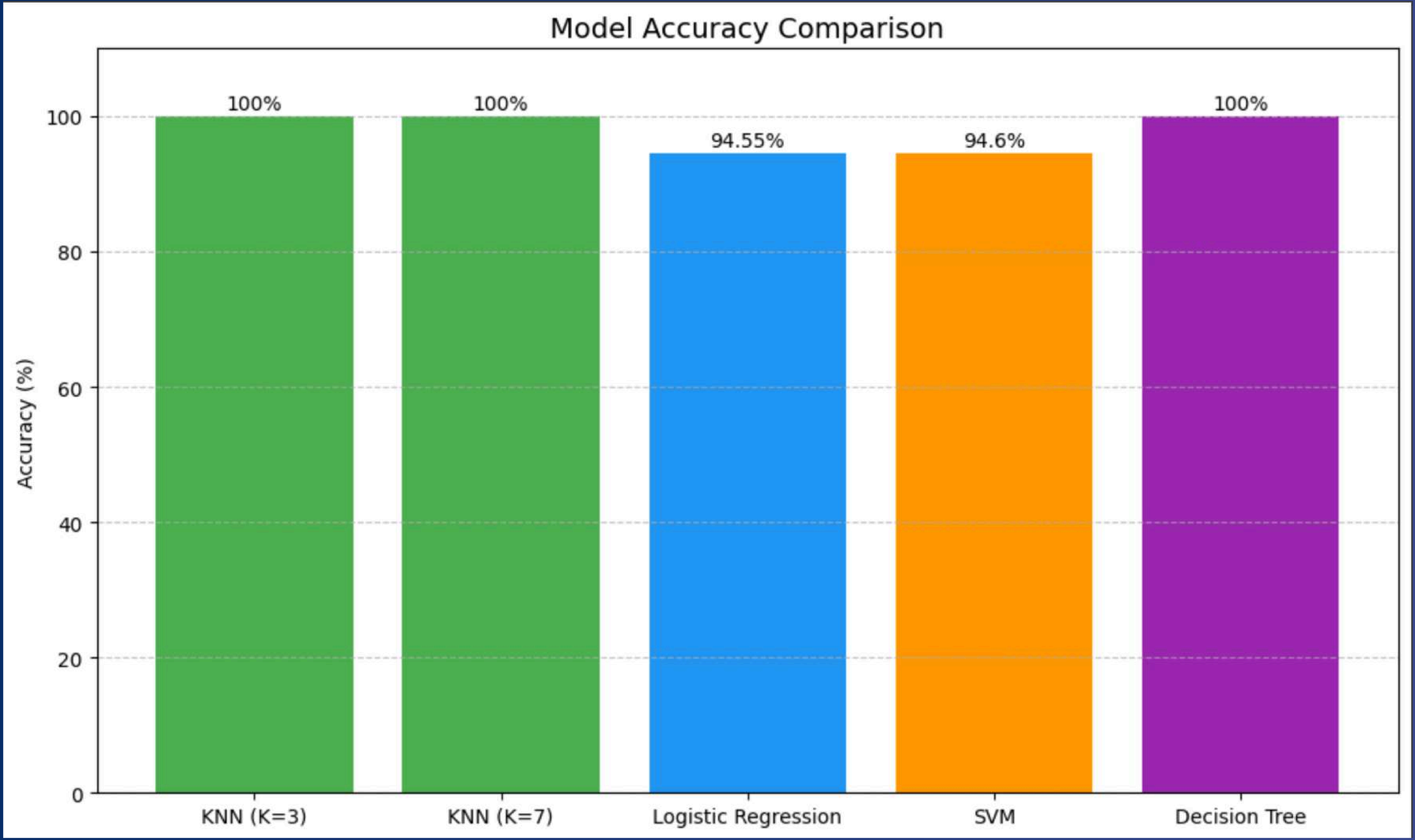


KEY OBSERVATIONS:

- We see a clear V-shaped pattern.
- No increasing or decreasing spread of residuals, which suggests homoscedasticity.
- Points are clustering near predicted probabilities of 0 and 1.
- That means this model is making confident predictions (close to 0 or 1), which is common in imbalanced datasets like fraud detection.

RESULTS:

Model	Accuracy(%)
KNN(K=3)	100
KNN(K=7)	100
Logistic Regression	94.55
Support Vector Machine	94.60
Decision tree	100


Model Accuracy Comparison
Accuracy (%)
0
20
40
60
80
100
100%
100%
94.55%
94.6%
100%
KNN (K=3)
KNN (K=7)
Logistic Regression
SVM
Decision Tree

CONCLUSION:

- KNN and Decision Tree models achieved the highest accuracy but may suffer from scalability and overfitting, respectively.
- Logistic Regression, while slightly less accurate, offers excellent interpretability.
- SVM balances performance and complexity well.
- In future, advanced methods like Random forest ,Gradient boosting should be explored.

ACKNOWLEDGMENT

WE WOULD LIKE TO EXPRESS OUR PROFOUND GRATITUDE TO DR. ARSHAD

THANK YOU !