**🧾 Project Title: Credit Card Fraud Detection Using Machine Learning**

**👩‍💻 Submitted by: Bhargavi Sai Jalluri**

📅 **Date**: 05/06/2025

**1️⃣ Introduction**

Credit card fraud is a growing concern in the digital financial world. The goal of this project is to build a **machine learning model** that can effectively detect fraudulent transactions based on historical data.

Due to the **imbalance** in data (very few fraudulent cases), this project emphasizes choosing appropriate algorithms and evaluation metrics that handle such data distributions.

**2️⃣ Objective**

To develop a model that:

* Accurately classifies credit card transactions as **fraudulent (1)** or **genuine (0)**.
* Minimizes **false positives** (flagging genuine transactions as fraud).
* Handles **imbalanced data** efficiently.

**3️⃣ Dataset Description**

📁 **Source**: Kaggle - Credit Card Fraud Detection  
📌 **Total Records**: 284,807 transactions  
📌 **Fraud Cases**: 492 (0.17% of total)

**🔍 Features:**

* **V1 to V28**: Anonymized PCA components
* **Amount**: Transaction amount (scaled during preprocessing)
* **Time**: Time elapsed since first transaction
* **Class**: Target variable (0 = Genuine, 1 = Fraudulent)

**4️⃣ Technologies Used**

* 🐍 **Python**
* 📚 **Pandas**, **NumPy**
* ⚙️ **Scikit-learn**
* 📈 **Matplotlib / Seaborn (optional for visuals)**

**5️⃣ Data Preprocessing Steps**

* 🔄 **Drop ‘Time’** as it adds little predictive power.
* 💸 **Normalize 'Amount'** using StandardScaler.
* 📊 **Split** data into features (X) and target (y).
* 🔀 **Train-test split** with 70% training and 30% testing.
* ⚖️ Use **undersampling** or **SMOTE** (optional) to deal with class imbalance.

**6️⃣ Model Selection**

🎯 Chosen Model: **Random Forest Classifier**

**Why Random Forest?**

* Handles **imbalanced datasets** well.
* Provides **feature importance**.
* High **precision and recall** in binary classification.

**7️⃣ Model Training & Evaluation**

**✅ Evaluation Metrics:**

* **Accuracy**: Overall correctness
* **Precision**: % of predicted frauds that are correct
* **Recall (Sensitivity)**: % of actual frauds detected
* **F1 Score**: Balance between precision and recall

**📊 Example Result:**

* Accuracy: ~99%
* Precision: 92%
* Recall: 86%
* F1 Score: 89%

**Confusion Matrix:**

|  | **Predicted: No Fraud** | **Predicted: Fraud** |
| --- | --- | --- |
| Actual: No | 85288 | 15 |
| Actual: Fraud | 12 | 148 |

**8️⃣ Conclusion**

* 📌 Machine learning models can **significantly reduce** undetected fraud.
* 📌 **Random Forest** is a reliable, interpretable model for this task.
* 📌 For better accuracy, further improvements like **real-time detection**, **deep learning**, or **ensemble models** can be explored.

**9️⃣ Future Scope**

* Deploy the model as an **API** in real-time banking systems.
* Integrate **deep learning** (LSTM/Autoencoders) for higher precision.
* Perform **continuous learning** with real-time transaction data.

**🔟 References**

* Kaggle Dataset: https://www.kaggle.com/mlg-ulb/creditcardfraud
* Scikit-learn Documentation
* Research papers on fraud detection and data imbalance