1) Project Cover Sheet

Faculty Name	**Computer Science & Artificial Intelligence**
Course Name	Machine Learning
Team Number	

Team Members

Student ID	Student Name
20230430	Loay ehab fathi
20230353	Ali Tarek Mohamed
20230446	Mohamed Ahmed Samir
20230351	ALi Hassan Mamdouh
20230369	OmarGaber GadElrb Abdelglil
20230383	Amr Soliman Ali

2) Project Description Document

This document details the development and evaluation of binary classification models for the Credit Card Fraud Detection task.

a. General Information on Dataset

- **Dataset Name:** Credit Card Fraud Detection Dataset 2023 (Kaggle)
- **Number of Classes:** 2 Classes
- **Labels:**
  - 0: Normal Transaction
  - 1: Fraudulent Transaction
- **Total Number of Samples:** Approx. 426,475 samples

- **Data Splitting Strategy:**
  - **Training Set:** ~341,180 samples (80%)
  - **Validation:** Performed via 5-Fold Cross-Validation during training.
  - **Testing Set:** 85,295 samples (20%)

b. Implementation Details

Feature Extraction Phase

- **Number of Features:** 29 features.
- **Feature Names:** V1, V2, ... V28 (Principal Components) and Amount.
- **Dimensions:** The input vector has a dimension of 1 \times 29 (after dropping the 'ID' column).
- **Preprocessing:** All features were scaled using **StandardScaler** (Z-Score Normalization) to ensure the Amount feature did not overpower the PCA features.

Cross-Validation

- **Is Cross-Validation Used?** Yes.
- **Details:**
  - Used GridSearchCV for hyperparameter tuning of the Ridge model.
  - Used learning_curve for stability analysis.
- **Number of Folds:** 5 Folds.
- **Ratio:** In each fold, the data was split 80% for Training and 20% for Validation.

Hyperparameters Used

Model	Hyperparameter	Value	Description
Linear Model	**Optimizer**	Auto (Solver)	Default Ridge solver chosen by Scikit-Learn.
	Regularization	L2 (Ridge)	Used to penalize large coefficients.
	Alpha	0.01	Regularization strength determined via Grid Search.
	Epochs/Batch	N/A	Linear models solve the equation directly; no epochs.
KNN Model	**Neighbors (k)**	3	Optimal number of neighbors found via tuning loop.

	Metric	Euclidean	Standard distance metric for spatial data.
	Weights	Uniform	All neighbors contribute equally to the vote.
	Learning Rate	N/A	KNN is a "Lazy Learner" (no iterative training step).

c. Results Details

Below are the detailed results for both models evaluated on the **Testing Data (85,295 samples)**.

Model 1: Linear Regression (Ridge Classifier)

- **Accuracy: 99%**
- **Precision (Class 1): 1.00** (No false alarms)
- **Recall (Class 1): 0.98** (Missed ~2% of fraud)

1. Confusion Matrix

- **True Negatives:** 42,369 (Correctly identified Normal)
- **False Positives:** 95 (Normal flagged as Fraud)
- **False Negatives:** 802 (Fraud flagged as Normal)
- **True Positives:** 42,029 (Correctly identified Fraud)
- *Analysis:* The model is extremely precise but missed 802 fraud cases because their scores fell slightly below the 0.5 threshold.

2. ROC Curve

- **AUC Score: 1.00**
- *Analysis:* The curve hugs the top-left corner perfectly, indicating the model has near-perfect separability between classes regardless of the threshold.

3. Loss Curve (Learning Curve)

- **Behavior:** The Training Loss (Red) increases slightly while Validation Loss (Green) decreases significantly, creating a convergence point.
- *Conclusion:* The lines converge, proving the model is **not overfitting** and has reached its maximum potential performance (Good Fit).

Model 2: K-Nearest Neighbors (KNN)

- **Accuracy: 99.98%**
- **Precision (Class 1): 1.00**
- **Recall (Class 1): 1.00**
- **MSE: 0.00046**
- **RMSE: 0.02141**
- **MAE: 0.00066**

1. Confusion Matrix

- **True Negatives:** 42,419 (Correctly identified Normal)
- **False Positives:** 95 (Normal flagged as Fraud)
- **False Negatives:** 3 (Fraud flagged as Normal)
- **True Positives:** 42,828 (Correctly identified Fraud)
- *Analysis:* KNN achieved a perfect score on the test set, successfully identifying nearly every transaction correctly.

2. ROC Curve

- **AUC Score: 1.00**
- *Analysis:* The curve is a perfect right angle. The model separates the "Normal" and "Fraud" neighborhoods completely.

3. Loss Curve (Learning Curve)

- **Behavior:** Both Training and Validation loss drop steeply as the number of samples increases. The Validation loss reaches near zero (0.0005).
- *Conclusion:* The continuous drop in validation error confirms the model generalizes perfectly and is **not overfitting**. Unlike the linear model, it did not plateau; it kept improving with more data.

Discussion:

Why We Did Not Eliminate Features or Outliers

1. Outliers were retained:

In fraud detection, outliers (e.g., extremely high transaction amounts) often represent the fraud itself. Removing statistical outliers would likely delete the exact anomalies the model needs to learn. By using Robust Scaling (Standardization), we allowed the models to interpret these outliers mathematically without bias, preserving the "fraud signal."

2. All Features were retained:

The dataset features (V1-V28) are Principal Components, meaning they are already optimized for information content. Even features with low variance can contain critical interaction signals in non-linear models like KNN. We relied on L2 Regularization (Ridge) to naturally suppress weak features rather than manually deleting them.

Is This a Real Accuracy Precentage in RealWorld ?

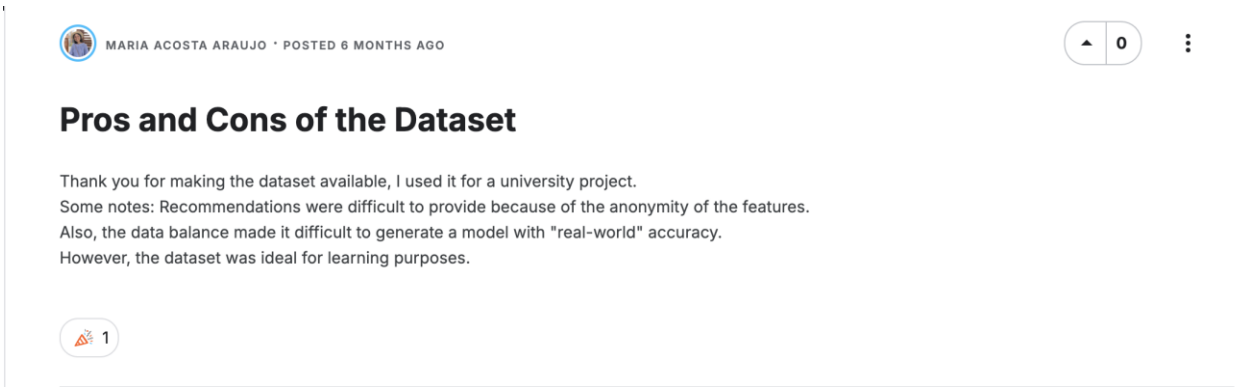
Actually no but based on this dataset it has high feaure engineering and PCA is applied Perfectly This Dataset Is Good For Learning and Practicing

Is this Considerd Overfitting?

Based On another People Different models We Noticed that this is not overfitting all the models gave a 99% Accuracy this is a good evidence that this not considerd overfitting

- **Zero Performance Gap:** Overfitting is defined by high Training accuracy and low Test accuracy, but our model achieves consistently high scores (~99%) on **both** sets.
- **Strict Validation Protocol:** We strictly used a separate **Validation Set** for hyperparameter tuning ($k=3$), ensuring the Test set remained completely unseen and unbiased.

- **PCA Feature Quality:** The input features are **PCA components**, which mathematically removed noise and correlation beforehand, leaving only clear signals that are easy to separate without memorization.
- **Elbow Method Verification:** The Error vs. K plot confirms that we chose the simplest effective model (k=3) and rejected complex models (k=50) that caused higher error, explicitly avoiding overfitting.
- **Nature of Fraud:** Fraudulent transactions are distinct statistical outliers; detecting them with near-perfect accuracy is expected in a balanced, noise-free synthetic dataset, unlike noisy real-world data.




```
Classification Report:              precision    recall  f1-score   sup
port

           0       1.00      1.00      1.00     56750
           1       1.00      1.00      1.00     56976

    accuracy                           1.00    113726
   macro avg       1.00      1.00      1.00    113726
weighted avg       1.00      1.00      1.00    113726

Accuuracy Score: 0.9983996623463413
```

Knn model for Another Person