Machine Learning Course Work Report

Name: Thilini Herath, UOL student number: 210569834
Module : Machine Learning (ST 3189)

Contents

Executive Summary

1 Introduction

This report presents an analysis of credit card fraud detection using various machine learning techniques. The primary objectives are to identify homogeneous population groups, predict continuous target variables, and classify categorical target variables. The dataset used in this analysis is sourced from Kaggle's Credit Card Fraud Detection Dataset 2023. The analysis includes exploratory data analysis, unsupervised learning using K-means clustering and PCA, regression analysis, and classification tasks.
The regression analysis part of this report utilizes the Adalone dataset from the UCI data library. The main goal of related to this dataset is to predict the age of Adalone base on physical measurements. For this, multiple linear regression, Ridge regression, and Cat-boosing method are employed.

2 Datasets and Research Questions

2.1 Fraud Dataset

The Credit Card Fraud Detection Dataset 2023 dataset in Kaggle contains credit card transactions made by European cardholders in the year 2023. This data set is used for Classification and Unsupervised Learning tasks in this report. It comprises over 550,000 records, and the data has been anonymized to protect the cardholders' identities. The primary objective of this dataset is to facilitate the development of fraud detection algorithms and models to identify potentially fraudulent transactions.
id: Unique identifier for each transaction
V1-V28: Anonymized features representing various transaction attributes (e.g., time, location, etc.)
Amount: The transaction amounts
Class: Binary label indicating whether the transaction is fraudulent (1) or not (0)
Note that this dataset contains only numerical features. This does not require any encoding, such as one-hot encoding, to convert categorical variables into numerical variables.

2.2 Abalone Dataset

The Abalone dataset used in this analysis is available in the UCI data library: Abalone Dataset. This data is used for Regression tasks in this report. This dataset comprises information on physical measurements of Abalones. The challenge is to predict their Ages (number of Rings) based on physical measurements, Length (mm), Diameter (mm), Height (mm), Whole weight (grams), Shucked weight (grams), Viscera weight (grams), Shell weight (grams) and Gender (a categorical column).

2.3 Research Questions

The research will gather detailed information on the following questions:

- How accurately can we use classification techniques to distinguish between fraudulent and non-fraudulent transactions?
- Which attributes have minimal influence on determining the fraudulent nature of a transaction?

- Can we discover homogeneous groups within the dataset using unsupervised learning methods?
- Identify the variables that are significant in modeling number of Rings.
- Find a model that predicts the number of Rings for given physical measurements with reasonable accuracy.

3 Exploratory Data Analysis (EDA)

3.1 Fraud Dataset

3.1.1 Class Distribution Analysis

The bar graph in Figure 1 illustrates the distribution of classes within the dataset, specifically distinguishing between non-fraudulent and fraudulent transactions. The x-axis represents the count of instances on a logarithmic scale, while the y-axis indicates the class labels (0 for non-fraudulent and 1 for fraudulent). Each bar corresponds to a class and displays the count of instances as well as the percentage share within the dataset.

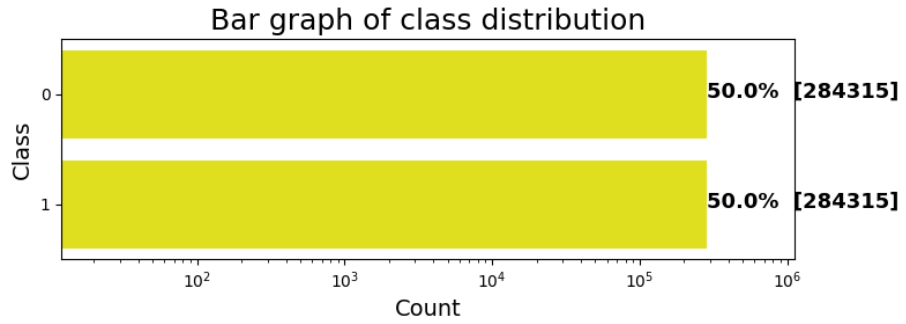

Figure 1: Class Distribution of Non-Fraudulent and Fraudulent Transactions

From the plot, it is evident that both classes are equally represented, with each accounting for 50% of the total dataset. This balance is crucial for subsequent analysis as it indicates that the dataset does not suffer from class imbalance, which is a common issue in many real-world datasets. Ensuring equal representation helps in training robust machine learning models, especially for classification tasks, as it prevents the models from being biased towards one class over the other.

3.1.2 Analyzing Predictor's Distribution

The boxplots plot in Figure 2 presents the distribution of the 'Class' variable against the 'Amount' variable. This plot indicates that the 'Amount' variable has a very similar distribution between the two classes. Both the medians and quartiles are almost the same. This suggests that the transaction amount alone may not be a strong predictor for distinguishing between fraudulent and non-fraudulent transactions.

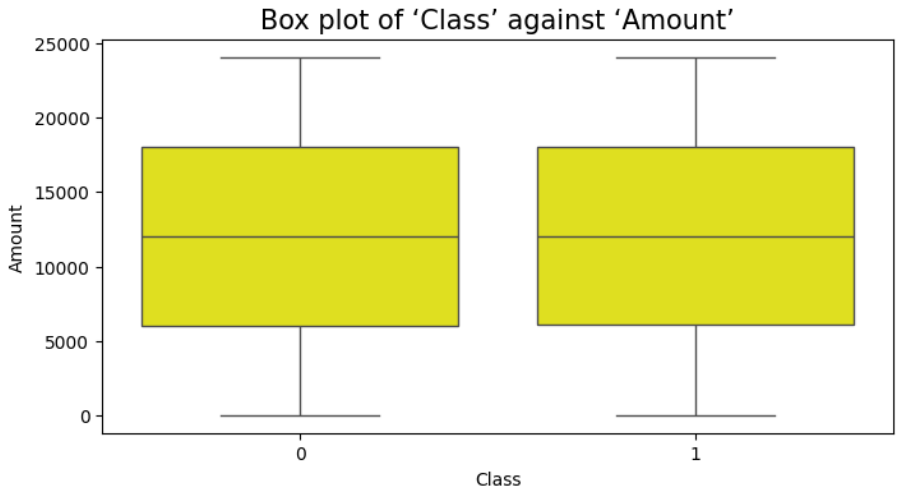

Figure 2: Boxplot of Transaction Amounts for Non-Fraudulent and Fraudulent Transactions

The density plots (in the Python notebook) of most of the 28 columns (V1 to V28) show distinct distributions (differences in peaks and spreads) for non-fraudulent (Non-Fraud) and fraudulent (Fraud) transactions. Specifically, columns V1, V2, V3, V4, V5, V6, V8, V9, V10, V11, V12, V14, V16, V17, V18, V19, V21, V23, V27 and V28 exhibit significant differences, highlighting their potential for classification tasks.
The box plots (in the Python notebook) illustrate the distributions of several key variables (V1 to V28, and Amount) for non-fraudulent (Class 0) and fraudulent (Class 1) transactions, with median and quartile values indicated. These visualizations provide a comparative analysis across the two classes, revealing distinct differences in central tendencies and spreads for each variable.

3.2 Abalone Dataset

In order to understand the distributions of each physical measurements, the density plots were created as shown in Figure 3.

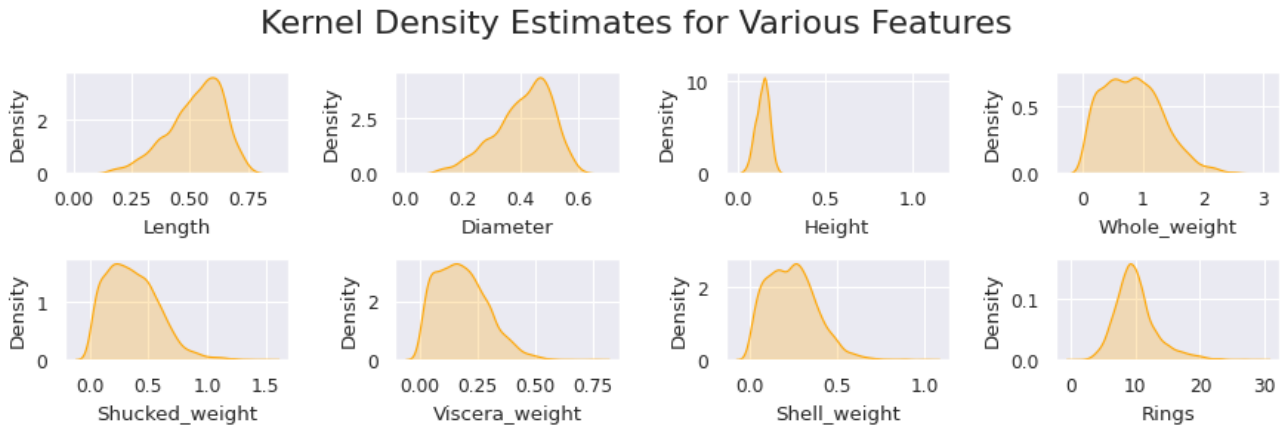

Figure 3: Kernel Density Estimates for Various Features

Figure 3 indicates that most features exhibit skewed distributions, and the response variable is approximately normally distributed. Also this suspects the existence of potential outliers in Shucked-weight and Viscera-weight variables.

A correlation analysis is required to identify the linear correlation between these measurements. Hence, the following correlation matrix is generated.

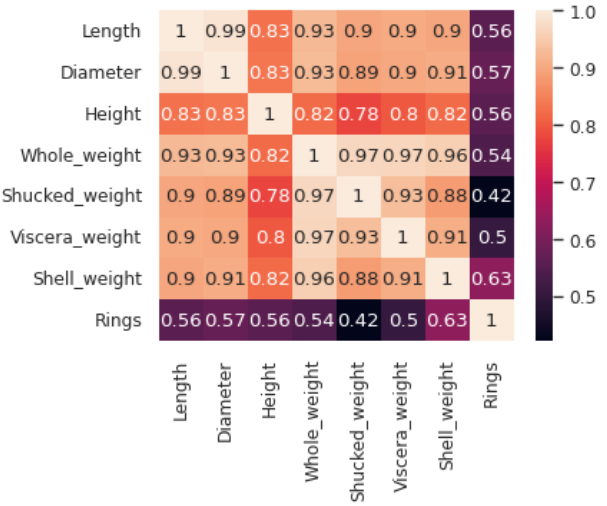

Figure 4: Correlation Heatmap of Various Features

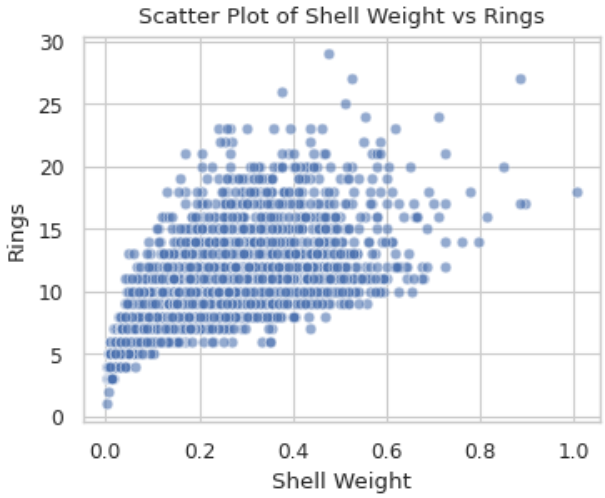

Figure 5: Scatter Plot of Shell Weight vs Rings

The Pearson correlation coefficients for these features range from 0.42 to 0.63, indicating moderate positive correlations. These plots and coefficients suggest that there is a positive relationship between these features and the number of rings. The predictor Shucked-weight is the variable that possesses the highest correlation with Rings, as shown in Figure 5.

4 Data pre-processing

4.1 Fraud Dataset

The dataset is then divided into training and testing sets using an 50-50 split, ensuring stratification (for classification task) to maintain the class distribution.
A PowerTransformer is applied to the training set and test set to stabilize variance and make the data more Gaussian-like. To further enhance the data preprocessing for the classification task, the transformed training and testing sets are standardized using the StandardScaler. This step scales the features to have zero mean and unit variance. Moreover, standard scaling is required for principal component analysis. These transformations improve the performance of most machine learning algorithms by reducing the impact of skewness and varying scales in the dataset.

4.2 Adalone Dataset

In this dataset, the Sex variable is categorical. It was converted to numerical values, with the Female taking the value of 0 and the Male taking the value of 1. The data is then divided into training and testing sets with a 75-25 split.

5 Supervised Learning: Classification Analysis

Classification analysis is performed to classify transactions as fraudulent or non-fraudulent. Several classification algorithms, including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Multilayer Perceptron (MLP) Classifier, are employed. The models are evaluated based on Accuracy, Precision, Recall, F1-score, and ROC-AUC score.
The following general confusion matrix is utilized as an easy presentation of classification errors when calculating metrics.

True Non-Fraud	$\text{Non-Fraud}_{\text{caught}}$	$\text{Non-Fraud}_{\text{missed}}$
True Fraud	$\text{Fraud}_{\text{missed}}$	$\text{Fraud}_{\text{caught}}$
	Predicted Non-Fraud	**Predicted Fraud**

Table 1: Confusion Matrix with True Labels and Predicted Labels

The following is the equation for calculating the **Accuracy**.

$$\text{Accuracy} = \frac{\text{Fraud}_{\text{caught}} + \text{Non-Fraud}_{\text{caught}}}{\text{Fraud}_{\text{caught}} + \text{Non-Fraud}_{\text{caught}} + \text{Fraud}_{\text{missed}} + \text{Non-Fraud}_{\text{missed}}} \tag{1}$$

The following is the equation for calculating the **Recall**.

$$\text{Recall} = \frac{\text{Fraud}_{\text{caught}}}{\text{Fraud}_{\text{caught}} + \text{Fraud}_{\text{missed}}} \tag{2}$$

The following is the equation for calculating the **Precision**.

$$\text{Precision} = \frac{\text{Fraud}_{\text{caught}}}{\text{Fraud}_{\text{caught}} + \text{Non-Fraud}_{\text{caught}}} \tag{3}$$

The most important classification metric when detecting fraudulent transactions is Recall. Any deficiency in this metric indicates an inability to capture a true fraud transaction as fraud, which results in financial loss of customers or financial institutions.
Although the Precision metric is also important in this task, the main focus will be on the Recall metric. However, all the metric values will be provided for each model.

5.1 Logistic Regression

The logistic regression model trained with 21 predictors obtained the following results on testing data.

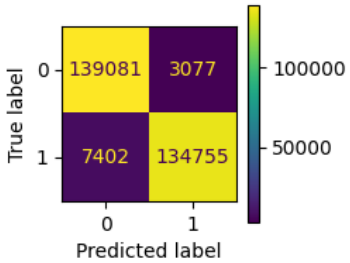

Figure 6: Confusion Matrix for Logistic Regression with 21 predictors

Metric	Logistic Regression with 21 predictors
Accuracy	96.31%
Precision	97.77%
Recall	**94.79%**
F1	96.26%
AUC	99.28%

Table 2: Metrics for Logistic Regression with all predictors

Due to the difficulty of interpretation, another logistics regression is trained with only the variables 'V4', 'V11', 'V14', and 'V17'. These variables were selected by examining the density plots and picking the variables visually, showing the most deviation between the two classes. The model sacrifices about 2% of Recall for the sake of model interpretability.

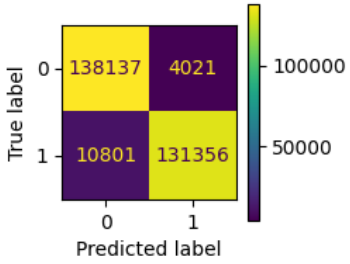

Figure 7: Confusion Matrix for Logistic Regression with 4 predictors

Metric	Logistic Regression with 4 predictors
Accuracy	94.79%
Precision	97.03%
Recall	**92.40%**
F1	92.40%
AUC	92.40%

Table 3: Metrics for Logistic Regression with 4 predictors

5.1.1 Logistic Model Interpretation

The logistic regression model for log-odds for detecting fraud (1) and non-fraud (0) with the given coefficients is represented as follows:

$$\log\left(\frac{P(\text{Fraud})}{1-P(\text{Fraud})}\right) = 1.6752 + 2.8184 \cdot V4 + 1.1125 \cdot V11 - 3.1089 \cdot V14 - 0.5067 \cdot V17 \quad (4)$$

where $P(\text{Fraud})$ is the probability of a transaction being fraudulent.

5.2 Decision Tree Classifier

First, a decision tree classifier is trained on all 21 predictors without restricting any hyperparameters except the depth of the model set to 3 with the entropy criterion for computing the purity of the nodes. Next, this model was pruned using the cost complexity pruning method to obtain a simple model while further controlling the overfitting and securing the interpretation ability. In the pruning process, the 5-fold cross-validation method is utilized to select the best cross-complexity path out of all the given paths (alpha values), which is the best regularization parameter.
The following table shows the performance analysis of the pruned model, and the following diagram illustrates the tree structure of the model.

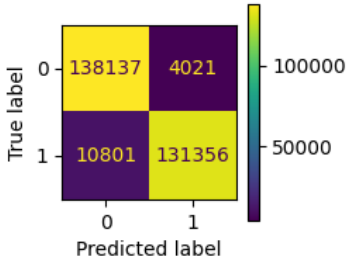

Figure 8: Confusion Matrix for DecisionTree

Metric	DecisionTree after pruning
Accuracy	94.46%
Precision	99.00%
Recall	**89.83%**
F1	94.19%
AUC	97.96%

Table 4: Metrics for DecisionTree

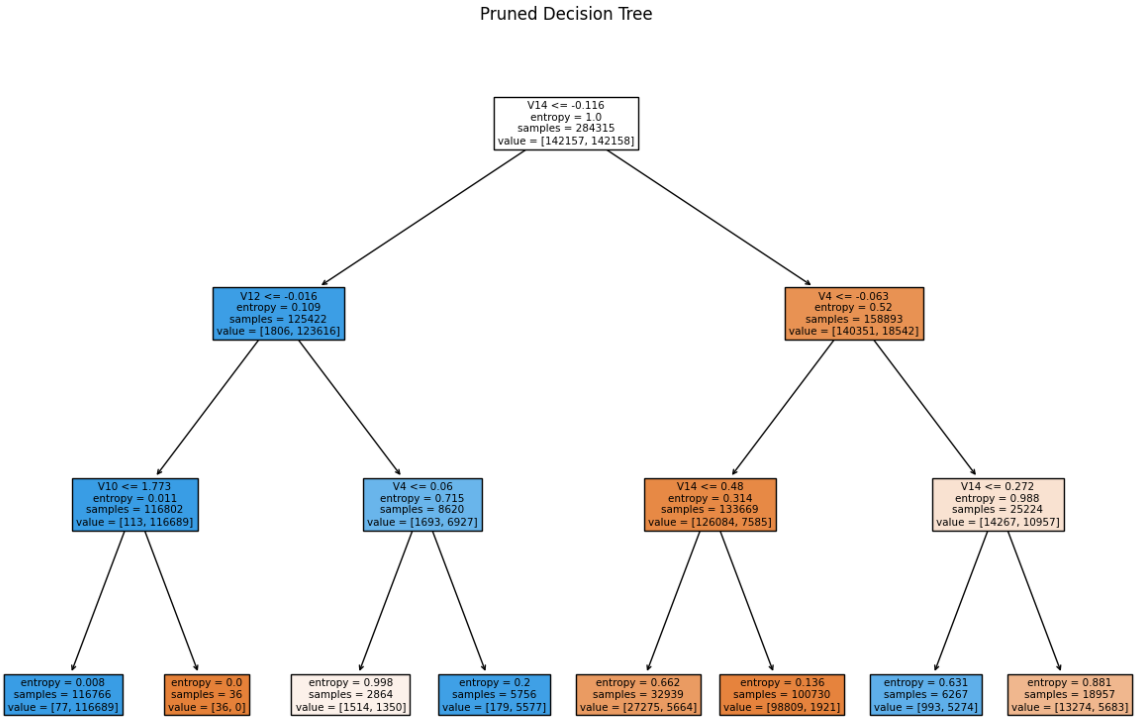

Figure 9: Decision Tree

5.2.1 Decision Tree Interpretation

The pruned decision tree has considered the variables 'V14', 'V12', 'V4', and 'V10' to generate a tree with a depth of 3 steps. Surprisingly, the two variables 'V14' and 'V4', which were also included in the logistic model based on density plots, appear in the tree model, confirming their influence in detecting fraudulent transactions. 'V14' is selected for the root node. Hence, this variable must discriminate between the two classes more than all other variables. The Figure 10 and 11 illustrate the density of 'V14' and 'V4' variables separated into two classes. It shows clear, little overlapping regions.

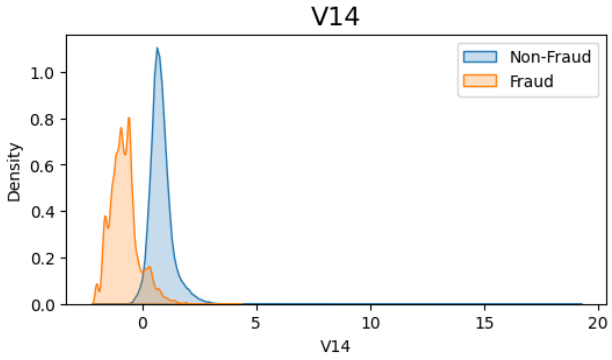

Figure 10: Density Plot of V14

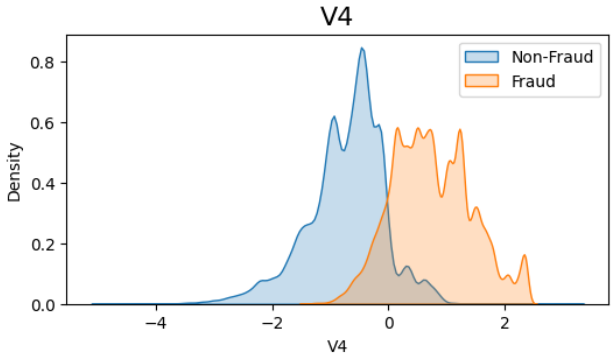

Figure 11: Density Plot of V4

Moreover, some leaf nodes exhibit very low entropy, indicating confident classifications. The tree effectively splits the data into more homogeneous groups as we move deeper into it, even though the tree model's recall metric is lower than that of the logistic model.

5.3 Model Comparison

All three models exhibit excellent classification performance on this dataset, as shown in Table 5, with the Random Forest model showing the best overall metrics. This analysis underscores the effectiveness of ensemble methods like Random Forest in handling classification tasks.

Metric	Logistic Regression 1	Logistic Regression 2	Decision Tree
Accuracy	96.31%	94.79%	94.46%
Precision	97.77%	97.03%	99.00%
Recall	**94.79%**	**92.40%**	**89.83%**
F1	96.26%	92.40%	94.19%
AUC	99.28%	92.40%	97.96%

Table 5: Model Error Measures for Different Classification Models

Based on the models considered, Logistic Regression 1 (with 21 predictors) outperformed the other two models in the Recall metric, which is the most important metric in this task, and also in overall accuracy.

6 Unsupervised Learning

6.1 The Principal Component Analysis (PCA)

This method is utilized as a dimension reduction method for the Fraud data set. PCA requires standardizing all variables, which gives all of them a mean of zero and a standard deviation of 1 unit. Next to determine the number of principal components (number of reduced variables) that needed to be considered, ratio of variance explained by each component is computed. The Figures 12 and Figure 13 explain the ratio distributions as separated and as cumulated ratios.

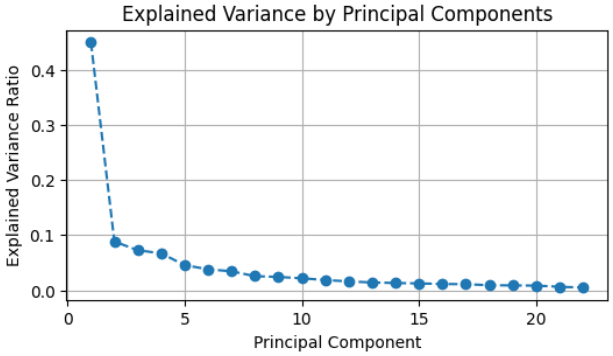

Figure 12: Explained Variance by Principal Components

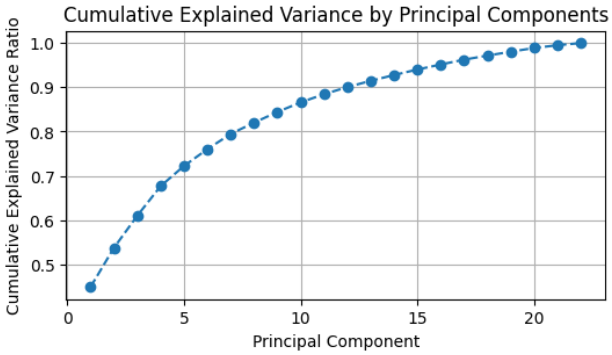

Figure 13: Cumulative Explained Variance by Principal Components

According to Figure 12 beyond five components, the ratio is almost plateau, and from Figure 13, the first 5 components explained about 72% of the variance in the data. Hence, the first 5 components will be used for further analysis, as stated in the section 6.2.

6.2 k-Mean Clustering

Next, the k-mean clustering method is applied to the first five principal components from 6.1 for the Fraud data set to identify any homogeneous group of transition within the dataset. As k-mean clustering suffers from not having an integrated method for determining the optimal

number of clusters, k-mean clustering is fitted multiple times for a set of pre-specified number of clusters 1 to 10. Then, the elbow method is used to find the number of clusters between 1 and 10, minimizing the within-cluster variation. The within cluster variation for a given k number of clusters is given by $\text{WCSS}(k) = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \mu_j\|^2$
where μ_j is the jth cluster mean.
The elbow plot in Figure 14 shows a sudden drop from $k = 1$ to $k = 3$ and almost a plateau after that, suggesting three significant homogeneous groups within the data set.

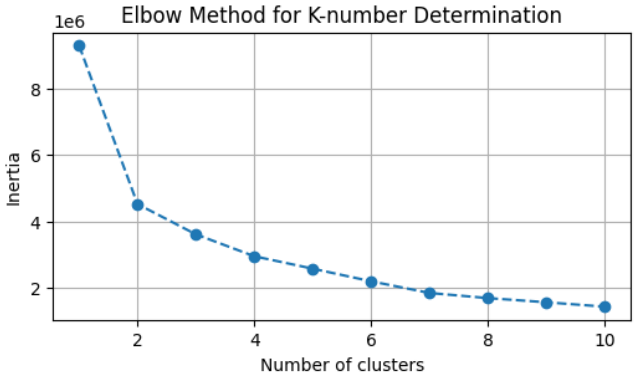

Figure 14: Elbow Plot for K-means Clustering

Further, clustered Fraud data is plotted across the first two principal components in Figure 15. This plot is also compared with the scatter plot of Fraud, non-Fraud class against the same two axis as shown in the Figure 16. Even though this is not always a guaranteed result due to the unsupervised nature of the k-mean clustering method, it has separated the classes reasonably well. A third class (in yellow) between the two main classes shows the challenging nature of the classification task.

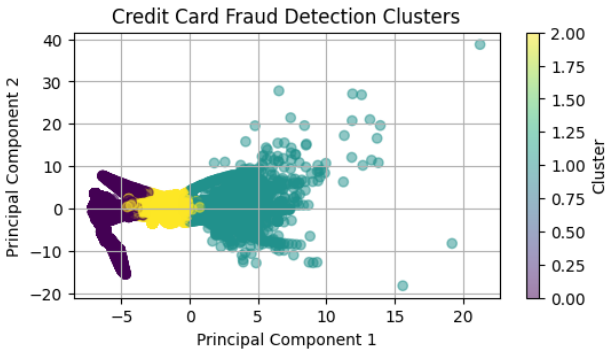

Figure 15: Credit Card Fraud Detection Clusters

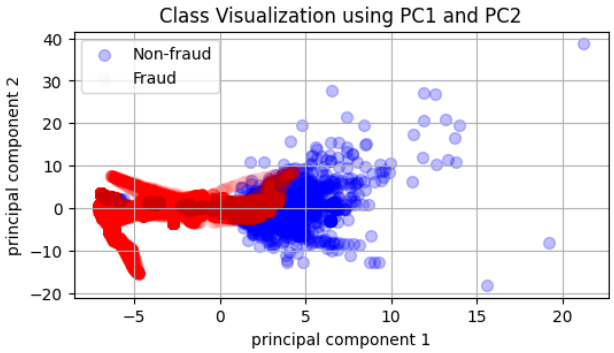

Figure 16: Class Visualization using PC1 and PC2

Figure 17: Comparison of Clustering and Class Visualization

7 Supervised Learning: Regression Analysis

Regression analysis is performed to predict the Age (Rings) of Abalone. Multiple regression techniques, including Linear Regression, Ridge Regression, and Cat Boosting Regressor, are used. The performance of each model is evaluated using metrics such as R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). R^2 measures how

a regression model fits the observed data using values closer to 1 which means a greater fit (technically this measures the proportion of variance of the response explained by the model). MAE is the average absolute difference between predicted and actual values, which shows the prediction accuracy. RMSE is the square root of the average squared differences between predicted and actual values.

7.1 Linear Regression

A multiple linear regression is conducted to predict the Rings given other variables. The statistical analysis is conducted to identify the significance of variables in predicting the Rings variable.
The linear model is significant in modeling the Rings variable, F-statistics 538.6 (p-value of 0.00), and it explains 53% of the variability of the data (R2). According to t-tests on the model coefficients, all the variables are statistically significant (p-value < 0.05). The linear regression equation based on the coefficients is given by:

$$\begin{aligned}\text{Rings} = {} & 2.7256 - 0.4470 \cdot \text{Length} + 11.1128 \cdot \text{Diameter} + 10.6589 \cdot \text{Height} \\ & + 9.0366 \cdot \text{Whole\_weight} - 19.8444 \cdot \text{Shucked\_weight} - 10.6212 \cdot \text{Viscera\_weight} \\ & + 8.6334 \cdot \text{Shell\_weight} + 1.2227 \cdot \text{Sex\_M} + 1.1650 \cdot \text{Sex\_F} + 0.3380 \cdot \text{Sex\_I}\end{aligned}$$

7.2 Ridge Regression

A ridge regression is conducted for the dataset. The main objective was to identify the optimal regularization parameter, and it is $\alpha = 1$.

7.3 Cat-boost model

Next, a nonlinear model is fitted into the dataset and its performance is evaluated against the testing data. The cat boost model contains many hyperparameters that need to be optimized. Hence the parameter optimization method is utilized as a grid searching method and then the best combination of learning-rate, iterations, bootstrap-type and l2-leaf-reg is found.

7.4 Model Comparison

The performance of three regression models—Linear Regression, Ridge Regression, and Gradient Boosting Regression—was evaluated using cross-validation metrics. The results are summarized as follows:

Metric	**Linear Regression**	**Ridge Regression ($\alpha = 1$)**	**Cat Boosting Regression**
R^2	0.4884	0.5276	0.9986
MAE	1.6138	1.5971	0.0008
MSE	5.3724	4.9614	0.0003
RMSE	1.2703	1.2637	0.02791

Table 6: Model Error Measures for Different Regression Models

The Cat Boosting Regression model marginally outperforms both Linear and Ridge Regression models, as evidenced by its slightly better R^2 value and lower MAE, MSE, and RMSE. This better performance highlights the effectiveness of the Boosting method in modeling complex relationships and improving predictive accuracy, making it the most suitable model for this regression task.