MET CS 699

Data Mining Analysis

Fraud Detections
&
Warranty Claims

Presented By: Tejaswi Lnu and Raghav Jindal

INTRODUCTION:

Data mining is the process of discovering patterns, correlations, and trends in large datasets and using this information to make informed decisions. One area where data mining can be particularly useful is fraud detection, where it can help identify patterns of behavior that are indicative of fraudulent activity. In this project, we apply data mining techniques to the problem of fraud detection in a company that provides warranty services to customers.

The goal of this project is to develop a machine-learning model that can accurately classify warranty claims as genuine or fraudulent, based on historical data. The dataset used in this project consists of many warranty claims submitted by customers, along with various attributes that describe the claims, such as the type of product, the date of purchase, and the amount of the claim. A significant portion of these claims are suspected to be fraudulent, making it difficult for the company to accurately identify and reject them.

To address this challenge, we apply a variety of data mining techniques to the dataset, including Generalized Linear Models, Support Vector Machines, Gradient Boosting Machines, Classification and Regression Trees, Lasso Regression, Ridge Regression, Recursive Feature Elimination, and Principal Component Analysis. These techniques are used to identify the most important features in the dataset and build an accurate predictive model that can classify warranty claims as genuine or fraudulent.

The project also discusses the limitations and challenges of using data mining techniques for fraud detection and provides recommendations for future research in this area. By applying data mining techniques to the problem of fraud detection in warranty claims, this project aims to help the company identify and reject fraudulent claims, thereby improving its financial stability and maintaining the trust of its genuine customers.

DATA MINING GOAL:

This project aims to develop a machine-learning model that can accurately classify warranty claims into genuine and fraudulent categories. The model should be able to analyze various features associated with each claim, such as product information, customer information, and claim details, to identify patterns that indicate potential fraud.

DETAILED DESCRIPTION OF DATASET:

The company has a large dataset of historical warranty claims, which will be used to train and test the machine-learning model. The dataset includes information about the product, customer, and the type of claim filed. The model should be able to learn from this dataset and accurately classify new warranty claims as genuine or fraudulent.

The success of this project will be measured by the accuracy of the model in correctly classifying warranty claims. A high-performing model will help the company detect and reject fraudulent claims, resulting in significant cost savings and improved customer satisfaction
Description of each column in the dataset:

1. **Region**: The geographic region where the warranty claim was filed.
2. **Area**: The specific area within the region where the warranty claim was filed.
3. **State**: The state where the warranty claim was filed.
4. **City**: The city where the warranty claim was filed.
5. **Consumer profile**: Information about the consumer who filed the warranty claim, such as age, gender, and occupation.
6. **Product category**: The category of the product for which the warranty claim was filed, such as electronics or appliances.
7. **Product type**: The specific type of product for which the warranty claim was filed, such as a smartphone or a refrigerator.
8. **AC1001 issue**: The specific issue reported for an air conditioners product, such as a faulty compressor or poor cooling.
9. **AC1002 issue**: The specific issue reported for an air conditioner product, such as faulty fan or noise issue.
10. **AC1003 issue**: The specific issue reported for an air conditioner product, such as gas leakage or water dripping.
11. **TV2001 issue**: The specific issue reported for a TV product, such as a blurry picture or sound issue.
12. **TV2002 issue**: The specific issue reported for a TV product, such as no power or poor reception.
13. **TV2003 issue**: The specific issue reported for a TV product, such as remote control not working or distorted sound.
14. **Claim value**: The amount claimed by the consumer for the warranty service.
15. **Service center**: The service center where the warranty claim was filed.
16. **Product age**: The age of the product for which the warranty claim was filed.
17. **Purchased from**: The source from where the product was purchased.
18. **Call details**: Information about the call made to the service center for filing the warranty claim.
19. **Purpose**: The purpose of the warranty claim, such as repair or replacement.
20. **Fraud**: A binary column indicating whether the warranty claim was fraudulent or genuine.

The dataset consists of many warranty claims filed over a certain period. The dataset includes information about the product, the customer, the type of claim filed, and the specific issues reported for air conditioner and TV products.The dataset also includes information about the location of the claim, the amount claimed, the age of the product, and other relevant details. The target variable is the binary column 'Fraud', indicating whether the warranty claim was fraudulent or genuine.

CLEANING OF DATA

A pristine data collection is required for all data analyses. The procedure is outlined in the four stages below:
A crucial stage in the processing of data is data cleansing, which involves reviewing and examining the data for mistakes, inconsistencies, and inaccuracies. Data cleaning makes sure that

data is dependable, precise, and comprehensive. Here are a few typical techniques for cleansing data:

1. **Handling Missing Values**: When cleansing data, missing data can be a big issue. Imputing missing values or removing the records with missing data are two ways to deal with missing values.
2. **Standardizing Data:** When standardizing data, it is important to make sure that the formats of the data values are uniform. For instance, making sure that all times are formatted the same.

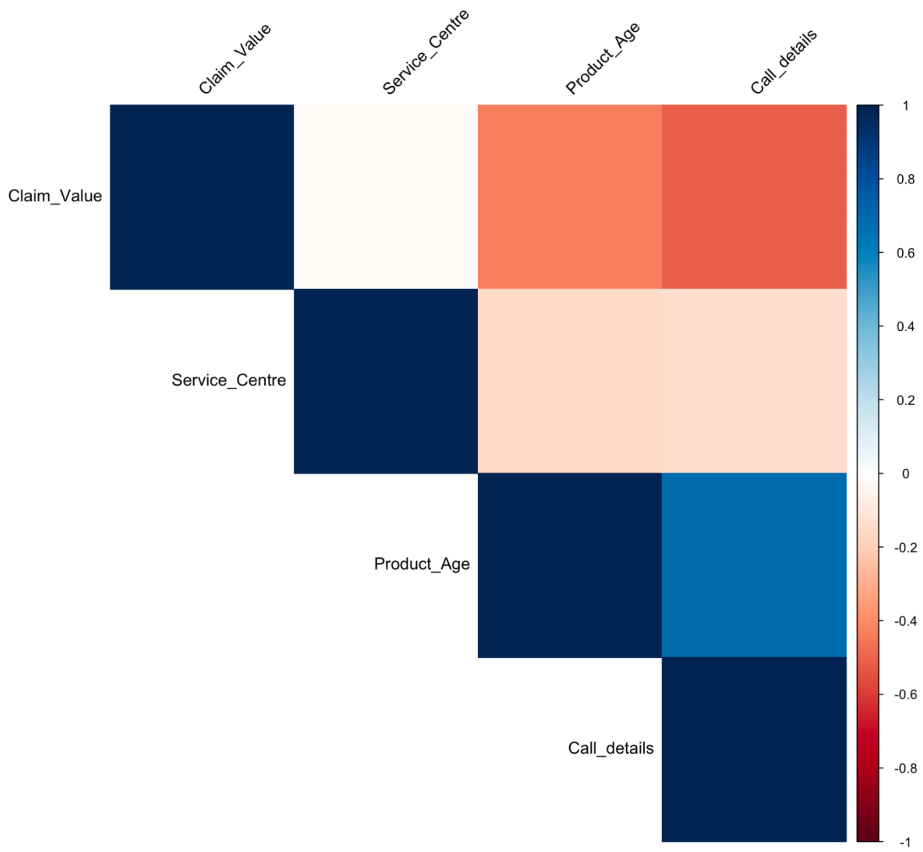
DETAILED DESCRIPTION DATA MINING TOOLS:

1. **caret** - A package that provides a unified interface for training and testing many different models in R.
2. **ggplot2** - A popular package for data visualization in R, which can help with exploratory data analysis and model interpretation.
3. **dplyr** - A package for data manipulation in R that can be used to clean and transform the dataset.
4. random Forest - A package for building random forests, a type of ensemble machine learning model, in R.
5. **glmnet** - A package for fitting Lasso and Ridge regression models in R.
6. **rpart** - A package for building classification and regression trees in R.
7. **e1071** - A package that provides SVM (Support Vector Machine) algorithms for classification and regression in R.
8. **ROCR** - A package for evaluating the performance of binary classifiers, such as those used in fraud detection.
9. **mlr** - A package that provides a unified interface for building and evaluating machine learning models in R.
10. **party** - A package for building decision trees and random forests in R, with support for categorical and continuous variables.
11. **randomForest** - RandomForest is an R package that implements the ensemble learning method for decision trees for classification and regression.
12. **naiveBayes** - naiveBayes is an R package that provides an implementation of the Naive Bayes algorithm for classification tasks.
13. **nnet** - nnet is an R package that provides functionality for building and training neural networks for classification and regression tasks.
14. **gbm** - gbm is an R package that provides an implementation of gradient boosting machines for regression and classification tasks.
15. **kknn** - kknn is an R package that provides an implementation of k-nearest neighbors algorithm for classification and regression tasks.
16. **Rweka** - RWeka is an R package that provides an interface to Weka machine learning algorithms for classification, regression, and clustering tasks.
17. **pROC** - pROC is an R package that provides tools for evaluating binary classifiers using ROC curves and related metrics.
18. **corrplot** -corrplot is an R package that provides tools for visualizing correlation matrices using various plot types and customization options.

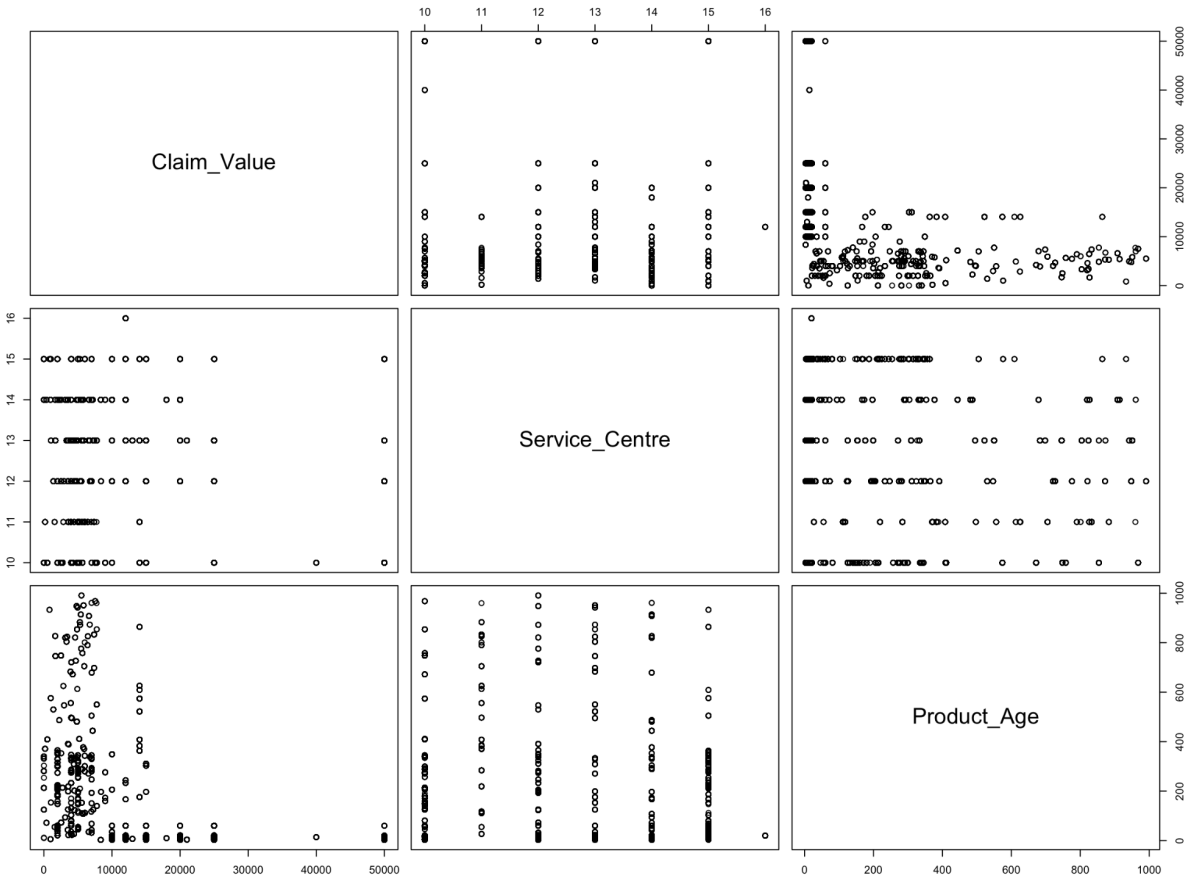
DATA VISUALIZATION:

HEAT PLOT

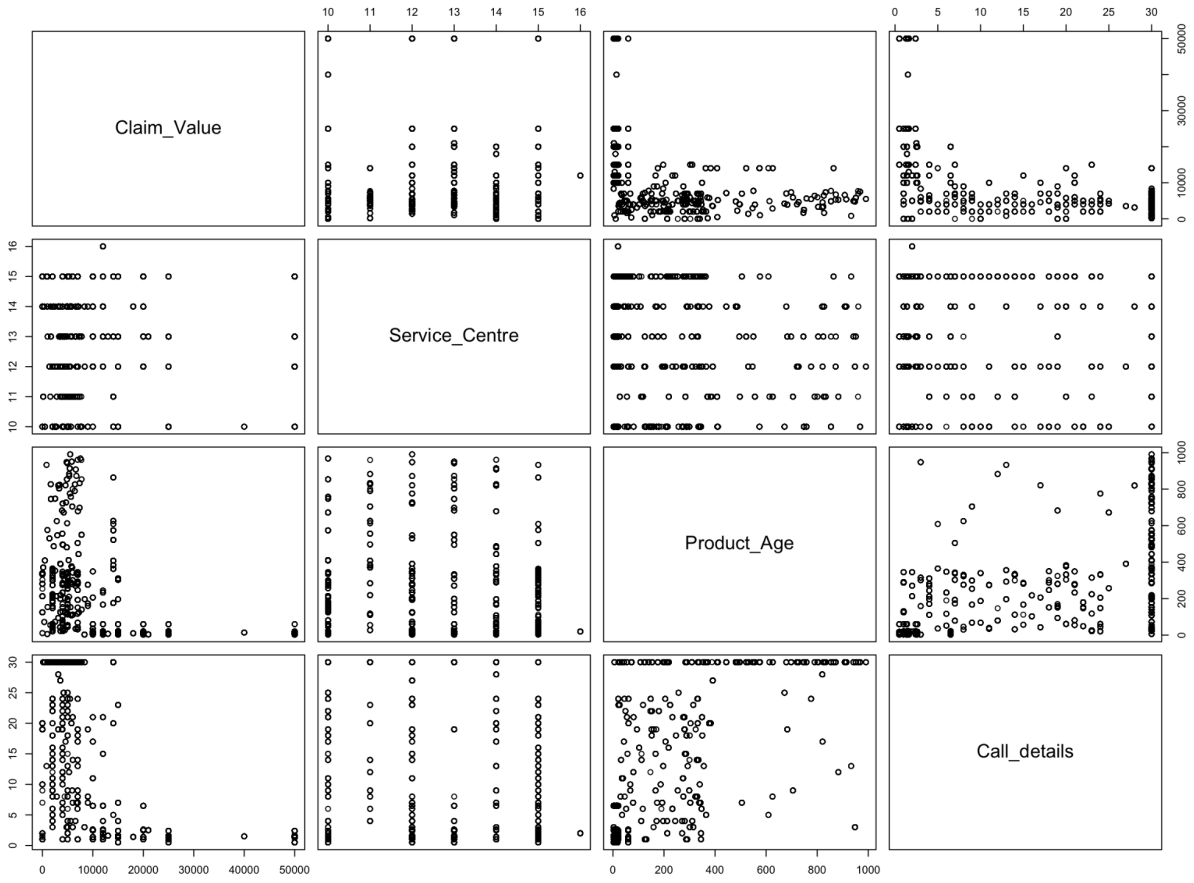
Heat plot calculating correlation btw 4 attributes claim value, service center, product age and call details



Scatter plot BTW Claim value, Service_centre, and Product age:



Scatter plot BTW Claim value, Service Centre, Product age &Call_details:



BRIEF DESCRIPTION OF CLASSIFICATION ALGORITHMS USED

The project utilizes six different classification algorithms. These algorithms are listed as follows:

1. **Generalized Linear Models (GLMs):** are a class of statistical models that extend the linear regression framework to handle response variables that are not normally distributed, such as counts or binary outcomes. GLMs are particularly useful for modeling the relationship between a response variable and a set of predictor variables, or covariates when the response variable is continuous, binary, or count data.GLMs use a link function to transform the response variable so that it is linearly related to the predictor variables. The link function is often chosen based on the nature of the response variable, and there are several commonly used link functions, such as the logarithmic link function for count data or the logistic link function for binary data. In the context of fraud detection for warranty claims, GLMs could be used to model the probability of a warranty claim being fraudulent, based on a set of predictor variables

such as the type of product, the age of the product, and the customer's location. The GLM would then provide a predicted probability of fraud for each warranty claim, which could be used to flag claims that are more likely to be fraudulent for further investigation. The choice of link function and variance function would depend on the nature of the response variable, such as whether it is binary or continuous.

2. **Support Vector Machines (SVMs):** are a class of supervised machine learning algorithms that are commonly used for classification and regression analysis. SVMs work by finding the optimal boundary or hyperplane that separates the data into different classes, with the goal of maximizing the margin between the classes. In the context of fraud detection for warranty claims, SVMs can be used to build a model that classifies warranty claims as either genuine or fraudulent. The SVM would learn from historical data, which would consist of a set of labeled warranty claims, with each claim assigned to either the genuine or fraudulent class. To build an SVM model, the warranty claims data would be transformed into a high-dimensional feature space, with each feature representing a different aspect of the claim, such as the product type, the location of the customer, and the time of the claim. The SVM would then try to find the hyperplane that best separates the genuine and fraudulent claims based on these features. One of the key advantages of SVMs is that they can handle complex, nonlinear relationships between the predictor variables and the response variable by using kernel functions. Kernel functions transform the data into a higher-dimensional space, where it may be easier to find a linear boundary or hyperplane that separates the data into different classifications and also have a regularization parameter, which helps to prevent overfitting by controlling the complexity of the model. Overfitting occurs when the model fits the training data too closely, which can result in poor performance on new, unseen data. In addition, SVMs can handle unbalanced data sets, where one class has significantly fewer observations than the other. This is important in the context of warranty claims, where fraudulent claims are likely to be much less frequent than genuine claims. Overall, SVMs are a powerful tool for fraud detection in warranty claims, as they can handle complex relationships between the predictor variables and the response variable, and can handle unbalanced data sets.

3. **Random Forest:** A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient-boosted trees. However, data characteristics can affect their performance. The first algorithm for random decision forests was created in 1995 by Tin Kam Ho using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to the classification proposed by Eugene Kleinberg.

4. **K-nearest neighbors (KNN):** KNN is a non-parametric estimation method which can be used for both regression and classification problems. It can be utilized in a motley of pattern recognition and estimation problems. KNN algorithm evaluates the output value of any given input vector by examining the output values in the vicinity of similar k neighbors. The measure of similarity is usually determined using a function that determines the distance. Various such distance calculating functions have been deployed to evolve KNN algorithms, such as Chebyshev distance, Mahalanobis distance, Euclidean distance, etc**.** The output of any given sample is evaluated by weighted averaging or normal averaging of its k nearest neighbors. The optimal value of k can be obtained with the help of a validation error curve. The algorithm iterates from 1 to the total number of points in the training data and calculates the distance between each point in the training dataset and the given test data. The distances calculated by the algorithm are then sorted in ascending order out of which the most frequent class in the top k rows of the sorted array is then returned as the predicted class.

5. **Classification and Regression Trees (CART):** It is a non-parametric, decision tree-based algorithm used for both regression and classification tasks. CART is a popular machine learning algorithm that works by recursively splitting the data into smaller subsets based on the value of a selected predictor variable, to minimize the impurity or variance within each subset. In the context of fraud detection for warranty claims, CART can be used to build a decision tree model that classifies warranty claims as either genuine or fraudulent. The CART model would learn from historical data, which would consist of a set of labeled warranty claims, with each claim assigned to either the genuine or fraudulent class. To build a CART model, the warranty claims data would be split into smaller subsets based on the value of a selected predictor variable. This process is repeated recursively for each subset, creating a tree-like structure that represents the decision rules used to classify the warranty claims. At each node of the tree, the algorithm selects the predictor variable that best splits the data into subsets that are more homogeneous in terms of the response variable, which in this case is the genuine or fraudulent label of the warranty claim. One of the key advantages of CART is that it can handle non-linear relationships between the predictor variables and the response variable, and can handle both continuous and categorical predictor variables. CART models are also easy to interpret, as the decision rules used to classify the data are represented by a tree-like structure that can be visualized and understood by humans. However, CART models are prone to overfitting if the tree is too complex, which can result in poor performance on new, unseen data. To address this issue, techniques such as pruning can be used to simplify the tree and prevent overfitting. In addition, ensemble techniques such as Random Forest can be used to improve the performance of the CART model by aggregating the predictions of multiple trees.

6. **Gradient Boosting Machines (GBMs):** are a type of ensemble learning algorithm that combines multiple weak models, such as decision trees, to create a stronger, more accurate model. GBMs are a popular and powerful machine learning algorithm that can

be used for both regression and classification tasks.In the context of fraud detection for warranty claims, GBMs can be used to build a model that classifies warranty claims as either genuine or fraudulent. The GBM would learn from historical data, which would consist of a set of labeled warranty claims, with each claim assigned to either the genuine or fraudulent class. To build a GBM model, the warranty claims data would be split into training and testing sets. The algorithm would then iteratively build decision trees on the training set, with each subsequent tree attempting to correct the errors made by the previous tree. The algorithm would assign higher weights to the observations that were misclassified by the previous tree, and lower weights to the correctly classified observations. This iterative process continues until predefined stopping criteria are met, such as a maximum number of trees or a minimum improvement in the error rate.One of the key advantages of GBMs is that they can handle nonlinear relationships between the predictor variables and the response variable. GBMs are also robust to outliers and can handle missing values. In addition, GBMs can be easily tuned to prevent overfitting by adjusting hyperparameters such as the learning rate, the number of trees, and the depth of the trees are also known for their high predictive accuracy, and they have been used successfully in a variety of applications, including fraud detection, credit risk assessment, and customer churn prediction. Overall, GBMs are a powerful tool for fraud detection in warranty claims. They can handle complex, nonlinear relationships between the predictor variables and the response variable and generate accurate predictions. However, GBMs can be computationally expensive and require careful tuning to prevent overfitting.

7. **Neural networks:** also known as artificial neural networks (ANNs), are a type of machine learning algorithm inspired by the structure and function of biological neurons in the human brain. Neural networks can be used for a wide range of tasks, including classification, regression, and pattern recognition. In the context of fraud detection for warranty claims, neural networks can be used to build a model that classifies warranty claims as either genuine or fraudulent. The neural network would learn from historical data, which would consist of a set of labeled warranty claims, with each claim assigned to either the genuine or fraudulent class. To build a neural network model, the warranty claims data would be split into training and testing sets. The neural network architecture would then be designed, which involves specifying the number of layers, the number of nodes in each layer, the activation functions used, and the optimization algorithm used for training the network. The neural network would then be trained on the training set using a backpropagation algorithm, which involves iteratively adjusting the weights and biases of the neurons in the network to minimize the error between the predicted and actual labels of the warranty claims. Once the network is trained, it can be used to classify new, unseen warranty claims as either genuine or fraudulent. One of the key advantages of neural networks is their ability to learn complex, nonlinear relationships between the predictor variables and the response variable. Neural networks are also robust to noisy data and can handle missing values. In addition, neural networks can be easily tuned by adjusting hyperparameters such as the number of layers, the number of nodes in each layer, and the learning rate. However, neural networks can be computationally expensive to train and require large amounts of data to achieve good performance. Neural networks are also known to be black-box models, meaning that it

can be difficult to understand how the network is making its predictions. Overall, neural networks are a powerful tool for fraud detection in warranty claims. They can handle complex, nonlinear relationships between the predictor variables and the response variable and generate accurate predictions. However, they can be computationally expensive and may require careful tuning to prevent overfitting. Additionally, the interpretability of neural networks may be a concern in some applications.

8. **Weka:** (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. It provides a graphical user interface to easily build and evaluate classification models. Weka includes various classifiers such as decision trees, Naive Bayes, k- nearest neighbor, support vector machines, and neural networks. Weka is an open-source software tool and supports many standard data formats. It can be used for both supervised and unsupervised learning tasks. Weka also offers feature selection techniques to identify the most relevant features in a dataset. It has a vast community of developers and users who continuously contribute to the development of new algorithms and extensions. Weka is widely used in academic research and in industry for various applications, including bioinformatics, finance, text mining, and image processing. It is a powerful and versatile tool for data mining and machine learning, and its user-friendly interface makes it accessible to users with little or no programming experience.

BRIEF DESCRIPTION OF ATTRIBUTE SELECTION METHODS:

1. **Random Forest:** Random forest attribute selection is a method used to identify the most important variables in a dataset. It is based on the random forest algorithm, which builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. During the process, each decision tree is trained on a randomly selected subset of the features, and the importance of each feature is calculated based on how much it contributes to the overall prediction accuracy. The feature importance scores are then used to rank the variables in order of significance, and the top-ranked variables are selected for further analysis or modeling. Random forest attribute selection is a useful technique for reducing the dimensionality of a dataset and improving model performance by focusing on the most informative features.

2. **Correlation-based:** Correlation-based attribute selection is a technique used in machine learning and data mining to select a subset of relevant features from a larger set of features or attributes. The technique involves calculating the correlation between each feature and the target variable (i.e., the variable to be predicted), and selecting the features that have the highest correlation with the target variable. In this technique, features with low correlation are discarded, as they are deemed to have little impact on the target variable. This helps to reduce the dimensionality of the data and improve the accuracy and efficiency of the predictive model. Correlation-based attribute selection can be applied to both numerical and categorical data, and there are several metrics used to calculate correlation, such as Pearson correlation coefficient, Spearman's rank correlation, and Kendall's tau correlation. The choice of metric depends on the type of data and the research question being addressed.

3. **Recursive Feature Elimination (RFE):** is a feature selection method that uses a model to recursively remove attributes from the dataset until a desired number of attributes is achieved.In RFE, a machine learning model is trained on the entire set of features and the importance of each feature is evaluated. The least important feature is then removed, and the model is retrained on the remaining features. This process is repeated until the desired number of features is reached. The importance of each feature is determined by the model's coefficient or a feature's influence on the model's performance. The features with the lowest coefficient or least influence on the model's performance are removed first. This process of removing features is called backward selection.RFE is a powerful feature selection method because it considers the interactions between features in the model. As the algorithm removes features, the remaining features may become more important and gain a higher weight in the model. By considering the interactions between features, RFE can help identify the most informative subset of features.

4. **Recursive Feature Elimination with Cross-Validation (RFECV):** It is a feature selection method that uses cross-validation to determine the optimal number of features to retain in a model. Like standard RFE, RFECV works by iteratively removing the least important features from a model until a desired number of features is reached. However, unlike standard RFE, RFECV uses cross-validation to evaluate the performance of the model at each iteration and to determine the optimal number of features. In RFECV, the dataset is split into k-folds, and the RFE algorithm is applied to each fold. The performance of the model is then evaluated using cross-validation, and the number of features that provide the best cross-validation score is retained. The optimal number of features is determined by evaluating the cross-validation score for each iteration of the RFE algorithm. The cross-validation score is plotted against the number of features, and the point at which the score begins to plateau is considered the optimal number of features. By using cross-validation to evaluate the performance of the model at each iteration, RFECV can help prevent overfitting and select the optimal number of features for the model.

5. **Principal Component Analysis (PCA):** It is a statistical technique used for dimensionality reduction and feature extraction. PCA is a powerful attribute selection method that can be used to identify the most important features in a dataset. In PCA, a linear transformation is applied to the dataset to transform the original features into a set of new features called principal components. These principal components are uncorrelated and represent the directions of maximum variance in the dataset. The first principal component captures the most variance in the dataset, while each subsequent principal component captures less variance. By selecting the top k principal components, PCA can reduce the dimensionality of the dataset while retaining the most important information.PCA can be used for attribute selection in machine learning models by selecting the top k principal components and using them as input features for the model. By selecting only the most important features, PCA can improve the performance of the model while reducing the risk of overfitting.

6. **Recursive Feature Elimination with Support Vector Machine (RFE-SVM)**: It is a feature selection method that combines the RFE algorithm with SVM.Like standard RFE, RFE-SVM works by iteratively removing the least important features from a model until a desired number of features is reached. However, instead of using a linear regression model as in standard RFE, RFE-SVM uses an SVM model to evaluate the importance of each feature. In RFE-SVM, the SVM model is trained on the full feature set and the importance of each feature is determined based on the weights assigned to them by the SVM. The feature with the lowest weight is removed and the model is retrained on the remaining features. This process is repeated until the desired number of features is reached. By using SVM to evaluate the importance of each feature, RFE-SVM can handle non-linear relationships between the features and the target variable, making it useful for datasets with complex relationships.

7. **Filter-based method using the chi-square**: attribute selection method is a statistical technique used for feature selection in data mining and machine learning. It is based on the chi-square statistic, which measures the independence between two categorical variables. In the context of feature selection, the chi-square test is used to determine whether the presence of a particular feature is significantly associated with the target variable. To apply the chi-square attribute selection method, the data is first organized into a contingency table, which shows the frequency counts of the different feature-target variable combinations. The chi-square test is then performed on each feature to determine its significance level, which indicates how likely it is that the presence of the feature is associated with the target variable. Features with high chi-square statistics and low p-values are considered to be highly correlated with the target variable and are selected for further analysis, while features with low chi-square statistics and high p-values are discarded. Chi-square attribute selection is a relatively simple and fast method for feature selection, but it assumes that the data is categorical and that the relationship between the features and the target variable is linear. It may not be appropriate for datasets with continuous variables or complex nonlinear relationships between the features and the target variable.

THE SET OF ATTRIBUTE SELECTED BY EACH ATTRIBUTE SELECTION METHOD:

1. **Recursive Feature Elimination (RFE):**
   Fraud,Purchased_from,City,Product_Age,Call_details,Claim_Value,State,Service_Centre,Purpose
2. **Recursive Feature Elimination with Cross-Validation (RFECV):**
   Fraud,Purchased_from,City,Product_Age,Call_details,Claim_Value,State,Service_Centre,Purpose,TV_2002_Issue,Consumer_profile,Area,TV_2003_Issue,AC_1001_Issue,AC_1002_Issue,Product_type,Product_category,TV_2001_Issue,AC_1003_Issue
3. **Recursive Feature Elimination with Support Vector Machine (RFE-SVM):**
   Fraud,Purchased_from,City,Product_Age,Call_details,Claim_Value,State,Service_Centre,Purpose,TV_2002_Issue,Consumer_profile,Area,TV_2003_Issue,AC_1001_Issue,AC_1002_Issue,Product_type,Product_category,TV_2001_Issue,AC_1003_Issue
4. **Principal Component Analysis (PCA):**

State,Area,City,Consumer_profile,Product_category,Product_type,Purchased_from,Purpose,Fraud

5. **Random Forest:**
   Fraud,Product_Age,Claim_Value,City,Call_details,Purchased_from,State,Service_Centre,Purpose
6. **Filter-based method using the chi-square:**
   Fraud,Product_Age,Claim_Value,Call_details,Purchased_from,City,State,Service_Centre,AC_1002_Issue
7. **Correlation-based:**
   Claim_Value,TV_2002_Issue,TV_2003_Issue,Service_Centre,Fraud

PROCEDURE:

1. Defining the categorization methods we would use, such as Naive-Bayes, Weka, Neural Net, Support Machine Vector, and Random Forest, was the first stage in the Analysis portion of the R code. The next stage entailed applying 10-fold cross-validation using these eight classification algorithms to the full dataset. Then, in order to assess the efficiency of the categorization algorithms, we gathered the average measures for accuracy, true positive rate, false positive rate, precision, recall, F1-measure, Matthew's correlation coefficient (MCC), and ROC AUC for the 10 folds. After data pre-processing, the clean dataset only had rows and the original dataset had about 8000 rows and 20 characteristics. This suggests that the original sample contained a sizable number of missing values and/or anomalies.
2. By eliminating anomalies and null values during the pre-processing stages, the machine learning models that were taught on the data performed better. This indicates that these actions were successful in raising the dataset's quality.
3. To determine the probability that a person will go back to jail after being released, eight machine learning models were chosen: Weka, Random Forest, Support Vector Machine (SVM), Neural Nets, and Naive Bayes. This suggests that the issue being addressed was a categorization issue.
4. Using a 66-33 divide to separate the dataset into training and test data, the models were trained on the training data and assessed on the test data using a confusion matrix. This indicates that a holdout technique was used to correctly validate the models.
5. Of all the models,Random Forest did the best at predicting the probability that a person would go back to prison after being released, with an accuracy rate of 97.6% across the entire dataset.
6. To extract sections of the most crucial characteristics from the original dataset, several attribute selection methods were used:Correlation Based Feature Selection, Random Forest Feature Selection, and Recursive Feature Selection and few more.The most pertinent characteristics to include in the models were determined using these attribute selection methods, which can enhance the performance of the models.

The data were divided into training and testing groups for each of the segments derived from the attribute selection methods, and the eight machine learning models were trained on the newly created training dataset. 64 models in total were taught and assessed.

RESULT AND PERFORMANCE:

PERFORMANCE MEASURES OF ALL MODELS

(SO AS TO FIT EASILY CHANGED ORIENTATION)

	MODEL	TP	TN	FP	FN	Accuracy (R)	TPR	FPR	Precision	Recall	F-Measure
WHOLE	GLM	2577	150	32	76	0.9358	0.9713531851	0.1758241758	0.9877347643	0.9713531851	0.9794754846
	SVM	2606	159	3	67	0.9429	0.9749345305	0.01851851852	0.9988501342	0.9749345305	0.9867474441
	RF	2587	46	22	180	0.976	0.9349475967	0.3235294118	0.9915676504	0.9349475967	0.9624255952
	KNN	2566	49	43	177	0.9675	0.9354721108	0.4673913043	0.9835185895	0.9354721108	0.9588938714
	CART	2570	129	39	97	0.9407	0.9636295463	0.2321428571	0.985051744	0.9636295463	0.9742228961
	NNET	2609	214	0	12	0.9245	0.9954215948	0	1	0.9954215948	0.9977055449
	GBM	2596	81	13	145	0.9668	0.9470995987	0.1382978723	0.995017248	0.9470995987	0.9704672897
	J48	2586	46	23	180	0.9757	0.9349240781	0.3333333333	0.9911843618	0.9349240781	0.9622325581
RFE	GLM1	2559	175	50	51	0.9206	0.9804597701	0.2222222222	0.9808355692	0.9804597701	0.9806476336
	SVM1	2609	226	0	0	0.9203	1	0	1	1	1
	RF1	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
	KNN1	2567	53	42	173	0.9665	0.9368613139	0.4421052632	0.9839018781	0.9368613139	0.9598055711
	CART1	2609	201	0	25	0.9291	0.990508732	0	1	0.990508732	0.9952317376
	NNET1	2609	226	0	0	0.9203	1	0	1	1	1
	GBM1	2587	78	22	148	0.9647	0.9458866545	0.22	0.9915676504	0.9458866545	0.9681886228
	J48-1	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
RFE-CV	GLM2	2577	150	32	76	0.9358	0.9713531851	0.1758241758	0.9877347643	0.9713531851	0.9794754846
	SVM2	2606	159	3	67	0.9429	0.9749345305	0.01851851852	0.9988501342	0.9749345305	0.9867474441

	RF2	2587	46	22	180	0.976	0.9349475967	0.3235294118	0.9915676504	0.9349475967	0.9624255952
	KNN2	2586	53	23	173	0.9732	0.9372961218	0.3026315789	0.9911843618	0.9372961218	0.9634873323
	CART2	2570	129	39	97	0.9407	0.9636295463	0.2321428571	0.985051744	0.9636295463	0.9742228961
	NNET2	2609	214	0	12	0.9245	0.9954215948	0	1	0.9954215948	0.9977055449
	GBM2	2595	81	14	145	0.9665	0.947080292	0.1473684211	0.9946339594	0.947080292	0.9702748177
	J48-2	2586	46	23	180	0.9757	0.9349240781	0.3333333333	0.9911843618	0.9349240781	0.9622325581
RFE-SVM	GLM3	2574	175	35	51	0.9259	0.9805714286	0.1666666667	0.9865848984	0.9805714286	0.9835689721
	SVM3	2609	226	0	0	0.9203	1	0	1	1	1
	RF3	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
	KNN3	2567	48	42	178	0.9683	0.935154827	0.4666666667	0.9839018781	0.935154827	0.9589092267
	CART3	2609	201	0	25	0.9291	0.990508732	0	1	0.990508732	0.9952317376
	NNET3	2609	214	0	12	0.9245	0.9954215948	0	1	0.9954215948	0.9977055449
	GBM3	2587	78	22	148	0.9647	0.9458866545	0.22	0.9915676504	0.9458866545	0.9681886228
	J48-3	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
PCA	GLM4	2572	150	37	76	0.934	0.9712990937	0.1978609626	0.9858183212	0.9712990937	0.9785048507
	SVM4	2600	174	9	52	0.9354	0.9803921569	0.04918032787	0.9965504025	0.9803921569	0.9884052462
	RF4	2587	46	22	180	0.976	0.9349475967	0.3235294118	0.9915676504	0.9349475967	0.9624255952
	KNN4	2586	53	23	173	0.9732	0.9372961218	0.3026315789	0.9911843618	0.9372961218	0.9634873323
	CART4	2609	141	0	85	0.9503	0.9684484039	0	1	0.9684484039	0.983971337
	NNET4	2587	53	22	173	0.9735	0.9373188406	0.2933333333	0.9915676504	0.9373188406	0.9636803874
	GBM4	2596	56	13	170	0.9757	0.9385394071	0.1884057971	0.995017248	0.9385394071	0.9659534884
	J48-4	2577	43	32	183	0.9735	0.9336956522	0.4266666667	0.9877347643	0.9336956522	0.9599552989
RANDOM FOREST	GLM5	2574	175	35	51	1	0.9805714286	0.1666666667	0.9865848984	0.9805714286	0.9835689721
	SVM5	2609	226	0	0	0.9243	1	0	1	1	1

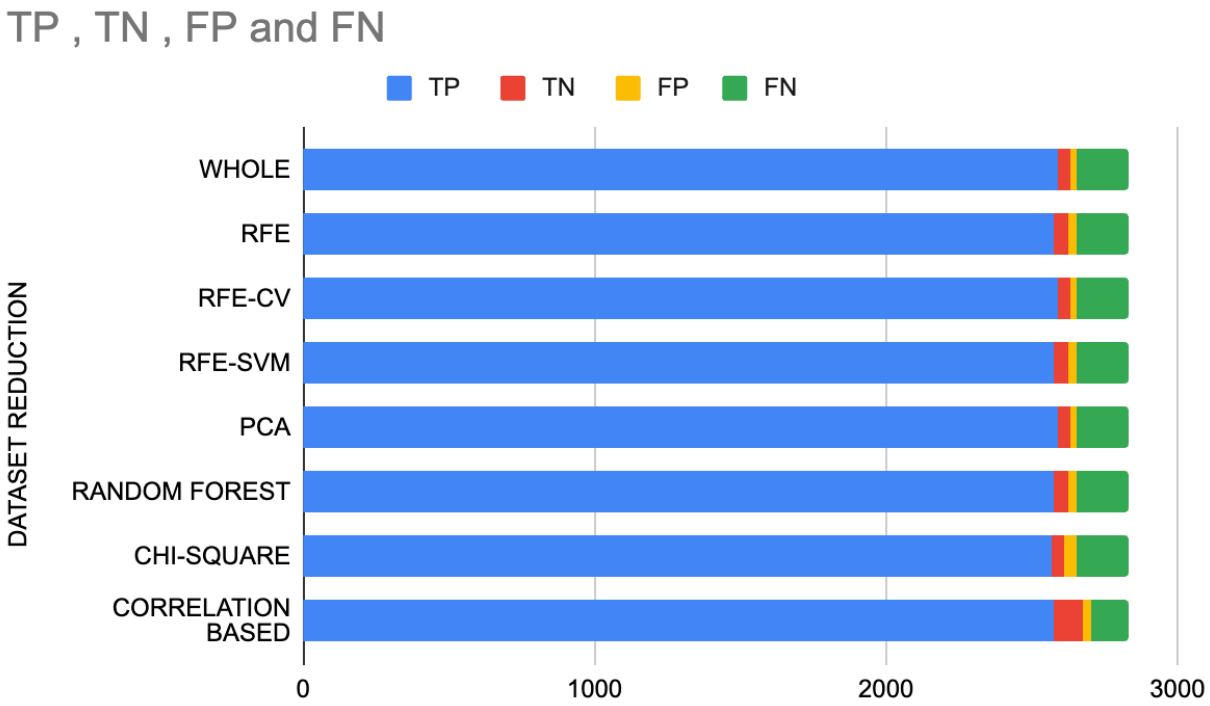
	RF5	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
	KNN5	2567	48	42	178	0.9683	0.935154827	0.4666666667	0.9839018781	0.935154827	0.9589092267
	CART5	2609	201	0	25	0.9291	0.990508732	0	1	0.990508732	0.9952317376
	NNET5	2518	132	91	94	0.9213	0.9640122511	0.4080717489	0.9651207359	0.9640122511	0.9645661751
	GBM5	2578	70	31	156	0.9644	0.9429407462	0.3069306931	0.9881180529	0.9429407462	0.9650009358
	J48-5	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
CHI-SQUARE	GLM6	2575	176	34	50	0.9259	0.980952381	0.1619047619	0.986968187	0.980952381	0.983951089
	SVM6	2609	226	0	0	0.9203	1	0	1	1	1
	RF6	2567	45	42	181	0.9693	0.9341339156	0.4827586207	0.9839018781	0.9341339156	0.9583722233
	KNN6	2585	73	24	153	0.9658	0.9441197955	0.2474226804	0.9908010732	0.9441197955	0.9668973256
	CART6	2609	201	0	25	0.9291	0.990508732	0	1	0.990508732	0.9952317376
	NNET6	2598	209	11	17	0.9224	0.993499044	0.05	0.9957838252	0.993499044	0.9946401225
	GBM6	2573	78	36	148	0.9598	0.9456082323	0.3157894737	0.9862016098	0.9456082323	0.965478424
	J48-6	2567	48	42	178	0.9683	0.935154827	0.4666666667	0.9839018781	0.935154827	0.9589092267
CORRELATION BASED	GLM7	2609	226	0	0	0.9203	1	0	1	1	1
	SVM7	2609	226	0	0	0.9203	1	0	1	1	1
	RF7	2578	98	31	128	0.9545	0.9526977088	0.2403100775	0.9881180529	0.9526977088	0.970084666
	KNN7	2574	108	35	118	0.9496	0.956166419	0.2447552448	0.9865848984	0.956166419	0.9711375212
	CART7	2609	200	0	26	0.9295	0.9901328273	0	1	0.9901328273	0.9950419527
	NNET7	2609	226	0	0	0.9203	1	0	1	1	1
	GBM7	2582	144	27	82	0.9397	0.9692192192	0.1578947368	0.9896512074	0.9692192192	0.9793286554
	J48-7	2578	101	31	125	0.9534	0.9537550869	0.2348484848	0.9881180529	0.9537550869	0.9706325301

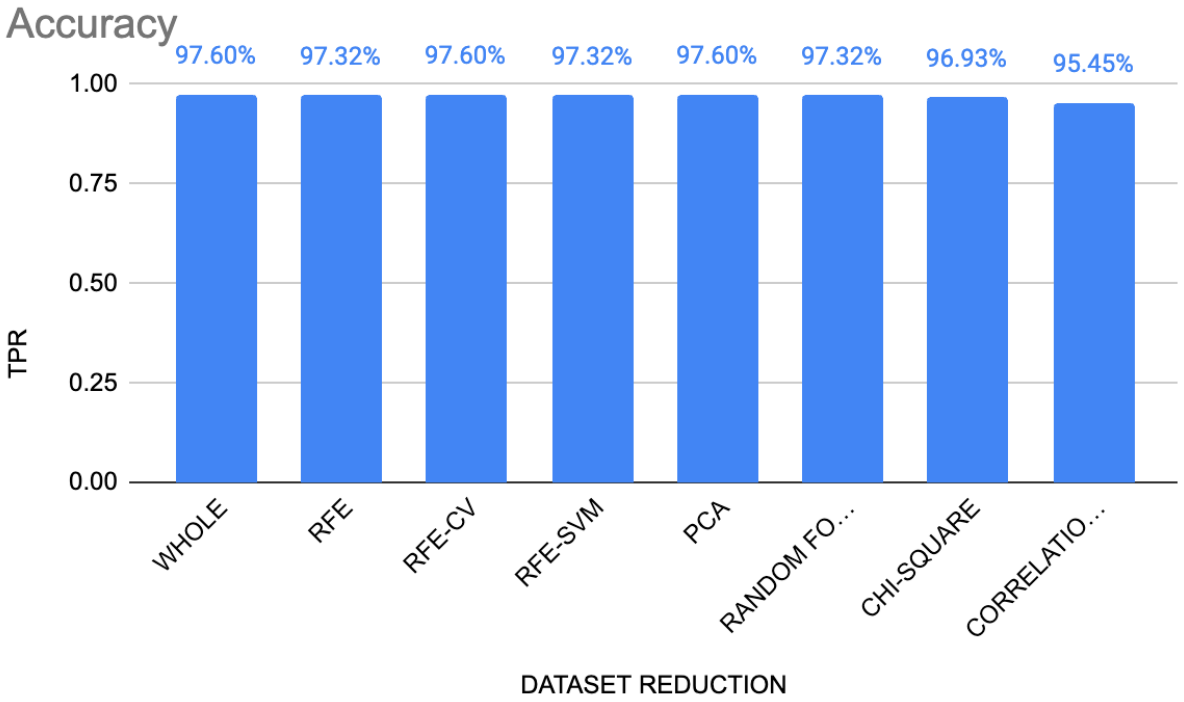
Best Model: Random Forest Classifier with 97.6% for reduced dataset according to Recursive Feature Elimination with Cross Validation, Principal Component Analysis feature reduction

BEST PERFORMANCE MEASURES WITH ACCORDANCE TO EACH REDUCTION ATTRIBUTE SELECTION

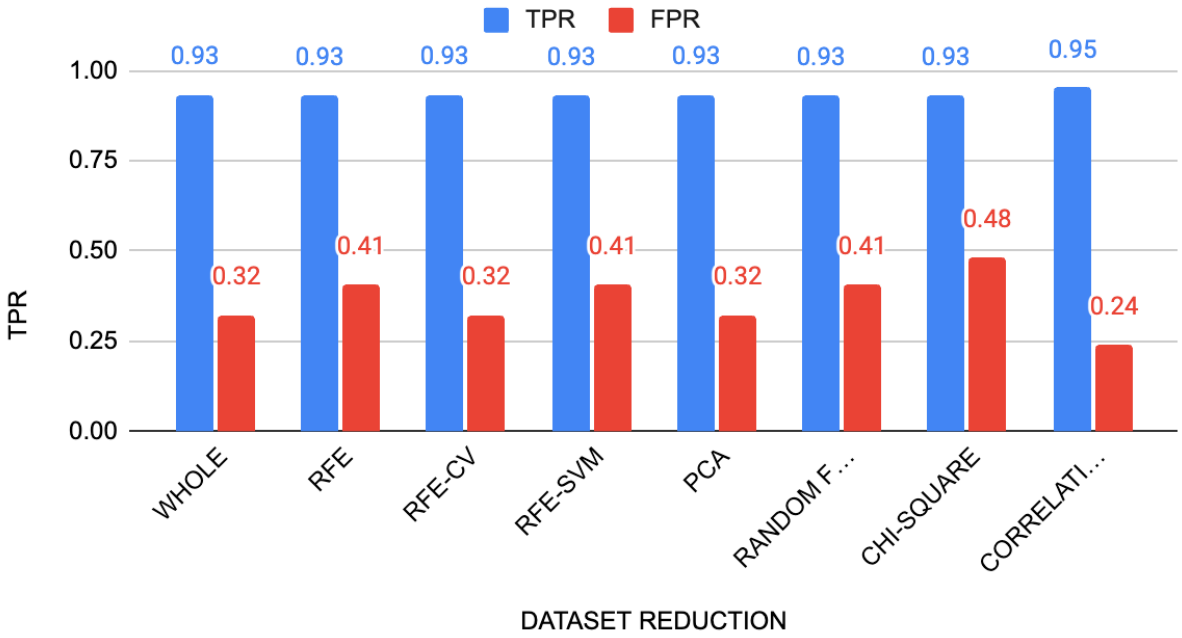
DATASET REDUCTION	MODEL	TP	TN	FP	FN	Accuracy (R)	TPR	FPR	Precision	Recall	F-Measure
WHOLE	RF	2587	46	22	180	0.976	0.9349475967	0.3235294118	0.9915676504	0.9349475967	0.9624255952
RFE	RF1	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
RFE-CV	RF2	2587	46	22	180	0.976	0.9349475967	0.3235294118	0.9915676504	0.9349475967	0.9624255952
RFE-SVM	RF3	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
PCA	RF4	2587	46	22	180	0.976	0.9349475967	0.3235294118	0.9915676504	0.9349475967	0.9624255952
RANDOM FOREST	RF5	2578	45	31	181	0.9732	0.9343965205	0.4078947368	0.9881180529	0.9343965205	0.9605067064
CHI-SQUARE	RF6	2567	45	42	181	0.9693	0.9341339156	0.4827586207	0.9839018781	0.9341339156	0.9583722233
CORRELATION BASED	RF7	2578	98	31	128	0.9545	0.9526977088	0.2403100775	0.9881180529	0.9526977088	0.970084666

GRAPHS

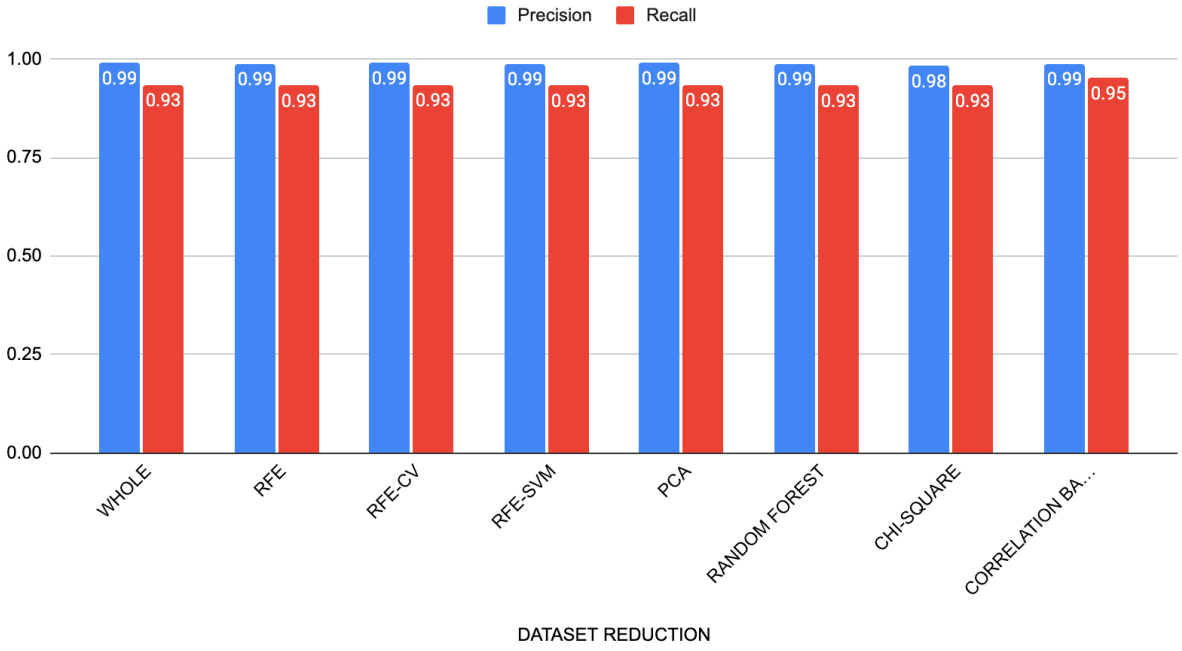

TP , TN , FP and FN
TP
TN
FP
FN
WHOLE
RFE
RFE-CV
RFE-SVM
PCA
RANDOM FOREST
CHI-SQUARE
CORRELATION BASED
DATASET REDUCTION
0
1000
2000
3000

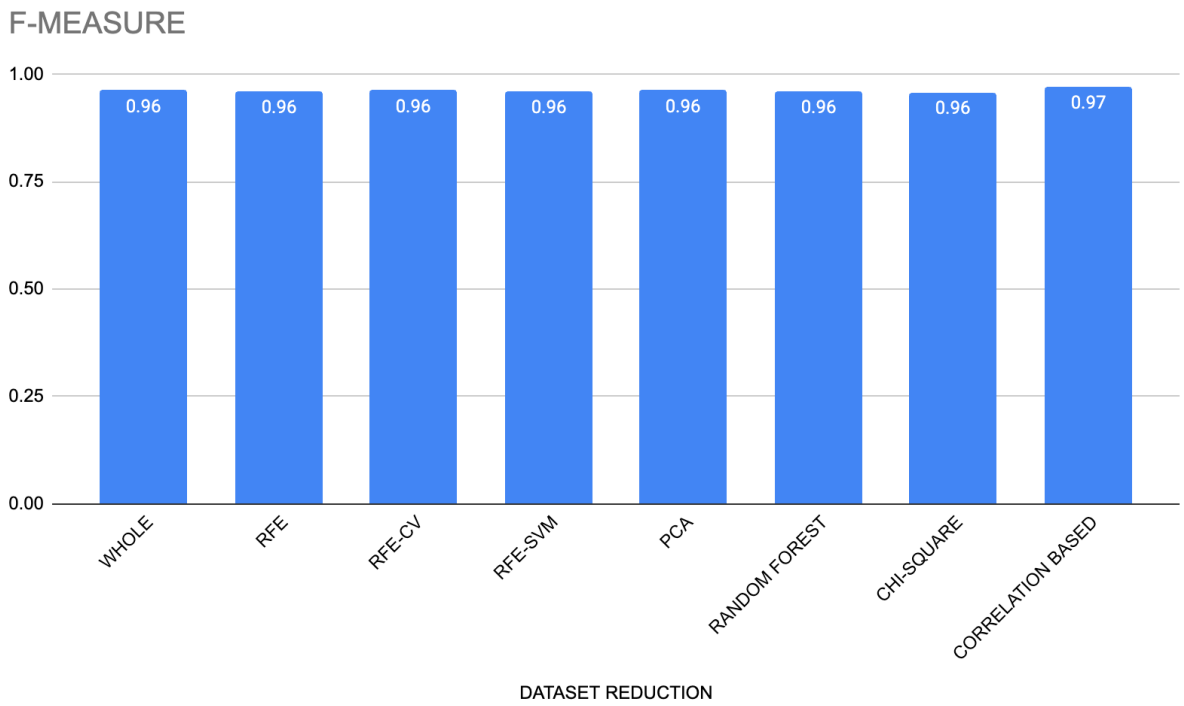

Accuracy
97.60%
97.32%
97.60%
97.32%
97.60%
97.32%
96.93%
95.45%
1.00
0.75
0.50
0.25
0.00
TPR
WHOLE
RFE
RFE-CV
RFE-SVM
PCA
RANDOM FO…
CHI-SQUARE
CORRELATIO…
DATASET REDUCTION

TPR vs FPR



Precision and Recall





LEARNINGS FROM THE PROJECT:

Feature selection is an essential stage in machine learning, and effective feature selection methods can speed up model training.
The quality and amount of the data used to teach machine learning classifiers have a significant impact on the accuracy of those systems.The initiative increased knowledge of using the R computer language to train machine learning models.Learned by doing in the areas of feature selection, model selection, and assessment.A better comprehension of how to improve the performance of various classification algorithms as well as their advantages and disadvantages.
Future machine learning initiatives in different domains may benefit from experience.

CONCLUSION:

The conclusion drawn from the project highlights several important aspects of machine learning. Firstly, the significance of feature selection in machine learning is emphasized. Feature selection techniques enable the identification of relevant features that contribute most to the prediction task, which can reduce the time taken to train machine learning models, while improving their performance. This is crucial as the quality and quantity of data used for model training heavily influence the accuracy of the classifiers.

The project also demonstrates how proficiency in training machine learning models can be improved using the R programming language. R is a widely used programming language for data analysis and statistical computing, and its popularity is due to its ease of use and powerful

analytical capabilities. The project's experience in using R for machine learning can be useful for future projects, as it enables efficient data manipulation, feature selection, and model development, which can significantly enhance the performance of machine learning models.

The project also provided hands-on experience in data pre-processing, model selection, and evaluation. Data preprocessing is a critical step in machine learning, as it ensures that the data is consistent, accurate, and complete, thereby improving the performance of machine learning models. Model selection and evaluation are also crucial, as it involves selecting the appropriate algorithm for the task at hand and assessing its performance. Through the project, the team has gained a practical approach.

Furthermore, the project has provided a deeper understanding of the strengths and limitations of different classification models and how to optimize their performance. By comparing and contrasting the performance of eight different machine learning algorithms, including KNN, Filter-based method using the chi-squar**e,** Random Forest, Support Vector Machine (SVM), Neural Nets, and the team gained insights into the trade-offs between different algorithms and how to choose the most suitable algorithm for a specific task.

In conclusion, the project has provided valuable practical experience in various aspects of machine learning, including data pre-processing, feature selection, model selection, and evaluation, using the R programming language. The team gained a deeper understanding of the strengths and limitations of different classification models and how to optimize their performance, which can be useful for future machine-learning projects in diverse domains.

Work done by team members:

1. Tejaswi Lnu:

a) Made the Project Proposal
b) Found Dataset to be used
c) Developed R Program to clean the dataset obtained
d) Developed 16 models using 2 attribute selection methods.
e) Made Final Project Presentation

2. Raghav Jindal:

a) Developed an R Program to save these datasets and call all required libraries.
b) Found out which classification models.
c) Developed 40 models using 5 attribute selection methods and for the whole dataset.
d) Performed Data Visualization on pre-processed data
e) Made Final Project Report