CS 699 – DATA MINING

Project Report

INJURY SEVERITY PREDICTION

Submitted by,
Ratna Meena Shivakumar
Sneka Vetriappan

TABLE OF CONTENTS

1. STATEMENT:

The project aims to analyze and gain insights from the motor vehicle operator data involved in traffic collisions on county and local roadways within Montgomery County. The dataset contains details of all traffic collisions occurring on these roadways, including information about the drivers involved. By examining this dataset. The main **project goal** is to predict the severity of the injury. it helps emergency medical responders and hospital staff to prioritize treatment for those drivers who are most severely injured. This can help save lives and reduce the long-term impact of injuries and also predicting injury severity can help insurance companies and policymakers to better understand the cost and impact of car accidents. By analyzing data on injury severity, they can identify areas where improvements in road safety infrastructure, traffic regulations, or driver education are needed. Lastly, predicting injury severity can inform the design of new safety features in cars and other vehicles. By analyzing the types of injuries sustained in car accidents and the severity of those injuries, engineers and designers can develop new safety features to protect drivers and passengers in the event of a crash.

2. DESCRIPTION OF THE DATASET:

This dataset provides information on motor vehicle operators (drivers) involved in traffic collisions occurring on county and local roadways. The dataset reports details of all traffic collisions occurring on county and local roadways within Montgomery County, as collected via the Automated Crash Reporting System (ACRS) of the Maryland State Police, and reported by the Montgomery County Police, Gaithersburg Police, Rockville Police, or the Maryland-National Capital Park Police. This dataset shows each collision data recorded and the drivers involved. The dataset is made up of 157234 tuples and 43 attributes. Vehicle Damage Extent, Driver Distracted By, Surface condition, Vehicle movement, Driver substance abuse, Speed limit are some of the important factors to predict the details about a traffic collision.

FIELD NAMES AND DESCRIPTIONS:

S.No	Attribute	Description
1.	Report Number	A unique identification number assigned to each accident report.
2.	Local Case Number	A unique identification number assigned by the local law enforcement agency.
3.	Agency Name	The name of the law enforcement agency that reported the accident.
4.	ACRS Report Type	The type of Accident Reporting System used for the accident report.
5.	Crash Date/Time	The date and time at which the accident occurred.
6.	Route Type	The type of route where the accident occurred, such as interstate, state route, or local road.
7.	Road Name	The name of the road where the accident occurred.

8.	Cross-Street Type	The type of cross-street, such as an intersecting road or a driveway.
9.	Cross-Street Name	The name of the cross-street.
10.	Off-Road Description	Description of the accident location when it is off-road.
11.	Municipality	The name of the municipality where the accident occurred.
12.	Related NonMotorist	Any non-motorist involved in the accident, such as a pedestrian or bicyclist.
13.	Collision Type	The type of collision that occurred, such as rear-end, sideswipe, or head-on.
14.	Weather	The weather conditions at the time of the accident, such as clear, rain, or snow.
15.	Surface Condition	The road surface condition at the time of the accident, such as dry, wet, or icy.
16.	Light	The lighting conditions at the time of the accident, such as daylight, dark, or dusk/dawn.
17.	Traffic Control	The type of traffic control at the accident location, such as traffic signal, stop sign, or yield sign.
18.	Driver Substance Abuse	Whether the driver was under the influence of drugs or alcohol at the time of the accident.
19.	Non-Motorist Substance Abuse	Whether any non-motorist involved in the accident was under the influence of drugs or alcohol.
20.	Person ID	A unique identification number assigned to each person involved in the accident.
21.	Driver At Fault	Whether the driver was at fault for the accident.
22.	Injury Severity	The severity of any injuries sustained by those involved in the accident.
23.	Circumstance	Any special circumstances surrounding the accident, such as construction or a school zone.
24.	Driver Distracted By	The source of any distraction that caused the driver to lose focus, such as a phone, passenger, or outside event.
25.	Drivers License State	The state that issued the driver's license.
26.	Vehicle ID	A unique identification number assigned to each vehicle involved in the accident.

27.	Vehicle Damage Extent	The extent of the damage sustained by each vehicle involved in the accident.
28.	Vehicle First Impact Location	The location of the first point of impact on each vehicle involved in the accident.
29.	Vehicle Second Impact Location	The location of the second point of impact on each vehicle involved in the accident, if applicable.
30.	Vehicle Body Type	The body type of each vehicle involved in the accident, such as sedan, SUV, or truck.
31.	Vehicle Movement	The direction of travel of each vehicle involved in the accident, such as northbound or southbound.
32.	Vehicle Continuing Dir	The direction of travel of each vehicle involved in the accident after the collision, if applicable.
33.	Vehicle Going Dir	The direction the vehicle was going at the time of the accident, such as east or west.
34.	Speed Limit	The posted speed limit at the accident location.
35.	Driverless Vehicle	Whether one of the vehicles involved in the accident was driverless.
36.	Parked Vehicle	Whether one of the vehicles involved in the accident was parked.
37.	Vehicle Year	The year of manufacture of each vehicle involved in the accident.
38.	Vehicle Make	The make or manufacturer of each vehicle involved in the accident.
39.	Vehicle Model	The model of each vehicle involved in the accident.
40.	Equipment Problems	Any equipment problems reported for each vehicle involved in the accident.
41.	Latitude	The latitude of the accident location.
42.	Longitude	The longitude of the accident location.
43.	Location	A description of the accident location, such as an address or landmark.

Source : https://catalog.data.gov/dataset/crash-reporting-drivers-data

3. DESCRIPTION OF DATA MINING TOOL(S):

For processing and analyzing our enormous datasets, we have used data mining tools like Python, Excel, and RStudio.

Before being read into R, the data is preprocessed and encoded in Python. Python is a flexible programming language that offers a wide range of tools and frameworks for data mining and analysis. Label encoding is a technique used to convert categorical data into numerical data.

Python provides libraries such as scikit-learn and pandas, which include the LabelEncoder class and the factorize () method respectively, that can be used to perform label encoding.

We have utilized excel which is a widely used tool for data mining and analyzing because of its easy-to-use interface, powerful built-in functions, and its ability to handle and process large amounts of data. Excel allows you to quickly and easily clean and format data by using functions such as Find and Replace, Conditional Formatting, and Data Validation. This helps to ensure that the data is accurate and ready for analysis.

For Performing all the modeling and data visualization, we have utilized R programming language. R provides a wide range of classification algorithms, such as decision trees, random forests, and support vector machines, making it a powerful tool for classification tasks. In addition, R provides tools for evaluating the performance of classification models, such as confusion matrices and ROC curves. R provides a wide range of libraries, such as caret, which offers a unified interface for many different algorithms. These libraries make it easy to train and evaluate machine learning models. R studio provides a number of tools and packages for data visualization, data exploration, and data modeling, making it a useful tool for exploring complex dataset.

4. DESCRIPTION OF CLASSIFICATION ALGORITHM USED:

The classification algorithm that has been used in our project are:

(i) Decision tree using rpart.

The rpart package in R provides an implementation of the Classification and Regression Tree (CART) algorithm for building decision trees. Decision trees are a type of supervised learning algorithm that can be used for both classification and regression tasks. They are commonly used in data mining and machine learning applications due to their simplicity and interpretability. The tree structure allows users to see which variables are the most important predictors and how they interact with each other. Decision trees can be used for both classification and regression tasks and can handle both categorical and continuous data. The rpart algorithm can handle large datasets with many variables, making it a useful tool for analyzing complex data.

(ii) K-Nearest Neighbors

K-Nearest Neighbors (KNN) is an algorithm that is widely used for classification and regression tasks in R programming language. KNN is a non-parametric algorithm, which means that it does not make any assumptions about the distribution of the data. KNN is computationally efficient, and it can work well with large datasets.

(iii) SVM (Support Vector Machines)

Support Vector Machines (SVM) is a powerful and widely used machine learning algorithm in R programming language. SVM can achieve good generalization performance, which means that it can perform well on new, unseen data. Its ability to handle high-dimensional data, non-linearly separable data, and achieve good generalization performance make it a popular choice for data modeling.

(iv) Random Forest:

Random Forest is a method of collective learning that blends various decision trees to increase the model's robustness and accuracy. Using random subsets of the predictor variables and bootstrap samples of the training data, the algorithm builds numerous decision trees. To enhance the performance of the model, the number of trees and the quantity of variables employed at each split can be adjusted. Random Forest provides a measure of feature importance, which can help in feature selection and understanding the underlying data.

(v) Neural Network:

Neural networks can be used to model complex and non-linear relationships in data, making them useful for a wide range of problems. Neural networks can extract useful features from raw data, which can be used for downstream tasks such as classification. They can learn from unstructured data, such as images or text, which makes them useful for problems where the input data is not pre-processed or engineered.

5. ATTRIBUTE SELECTION METHODS:

(i) Recursive Feature Elimination (RFE): RFE is a feature selection algorithm used to select the most relevant subset of features from a given dataset. The weakest features are repeatedly eliminated using a backward selection method until the necessary number of features is obtained. The RFE algorithm functions by training a model on the full dataset—be it a support vector machine, a linear regression, or any other model—and then ranking the features according to their significance or contribution to the model. As soon as the target number of features is attained, the weakest feature or features are eliminated, and the procedure is then repeated with the remaining features.

(ii) Wrapper-based feature selection: Wrapper-based methods are used to select the most relevant subset of features from a given dataset. These methods select features based on the performance of a specific machine learning algorithm. The algorithm chooses a subset of features, trains the model using those features, then uses a validation set to assess the model's performance. Up until the necessary number of features is obtained, the procedure is repeated for various subsets of features. Wrapper-based approaches are distinct from filter-based methods in that when choosing features, they consider their interactions rather than only the correlation between each feature and the target variable.

(iii) Baruta: The Baruta method is an attribute selection technique that chooses the pertinent features from a subset of the dataset. The RFE method is used in the Baruta method to analyze a subset of the dataset that has been randomly chosen. To determine the most important traits, the procedure is performed several times, using a new subset each time. This method concentrates on the most consistent data, which can help to lessen the impact of noise and outliers in the dataset. The Baruta approach can reduce the computational complexity of the RFE method by restricting the amount of features that need to be assessed, making it particularly beneficial or datasets with a lot of features.

(iv) Filter Method: The filter method is an attribute selection strategy that chooses pertinent characteristics from a dataset using statistical metrics. The most significant features are chosen using this strategy based on their statistical characteristics after being analyzed independently of the learning process. The filter approach ranks the features in accordance with a predetermined criterion, such as their variance, correlation, or mutual information with the target variable. The model is then trained using the top-ranked features. As the features are chosen based on their statistical significance, it also makes it possible for the results to be easily interpreted.

(v) Chi Squared: A statistical method for selecting attributes is the chi-squared method. By assessing the statistical significance between each feature and the target variable, it is a filter approach that chooses the features that are most pertinent to the target variable. The input features for the chi-squared approach can be categorical or numerical, while the destination variable is categorical. Each feature's independence from the target variable is determined using the chi-squared test. The more dependent the feature is on the target variable, the higher the chi-square.

6. THE SET OF ATTRIBUTES SELECTED BY EACH ATTRIBUTE SELECTION METHOD:

(i)RecursiveFeatureElimination(RFE): Vehicle.Damage.Extent, Crash.Date.Time, Report.Number, Location, Collision.Type

(ii)Wrapper-basedfeatureselection: Vehicle.Damage.Extent,X,Driver.Distracted.By, Collision.Type, Road.Name, Vehicle.Movement, Equipment.Problems, Municipality, Agency.Name, Weather, Crash.Date.Time, Vehicle.Body.Type, Surface.Condition, Light , Route.Type, Parked.Vehicle, Vehicle.ID, Report.Number, Vehicle.Model, Traffic.Control

(iii) Baruta: X, Report.Number, Agency.Name, Crash.Date.Time, Collision.Type, Vehicle.Damage.Extent, Vehicle.Body.Type, Speed.Limit

(iv) Filter Method: Report.Number, Crash.Date.Time, Route.Type, Road.Name, Surface.Condition, Light, Traffic.Control, Driver.Substance.Abuse, Collision.Type, Driver.Distracted.By, Vehicle.ID, Vehicle.Damage.Extent, Vehicle.Body.Type, Vehicle.Model, Vehicle.Movement, Speed.Limit, Parked.Vehicle, Equipment.Problems, Location, class, Location

(v) Chi Squared: Report.Number, Crash.Date.Time, Municipality, Light, Traffic.Control, Collision.Type, Vehicle.Damage.Extent, Vehicle.Body.Type, Movement, Speed.Limit, Location, class

7. DETAILED DESCRIPTION OF DATA MINING PROCEDURE:

A data mining procedure is a set of steps that are followed to extract useful insights and knowledge from a large dataset. It involves various stages of processing, analysis, and interpretation of data to identify patterns, trends, and relationships.

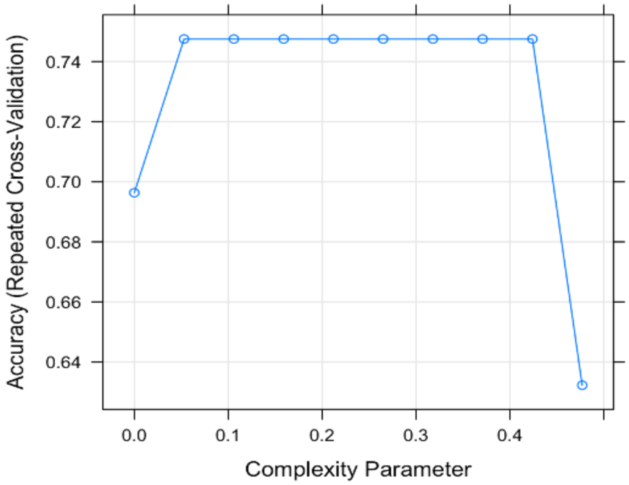
The dataset used in this procedure was obtained from Data.gov, which is the United States government's open data website. It provides access to datasets published by agencies across the federal government.

We utilized several data mining tools, including Python, Excel, and RStudio, to analyze our vast datasets. Python was used to preprocess the data and convert categorical data into numerical data using a technique called label encoding. Python's flexibility and diverse range of data mining and analysis tools, such as scikit-learn and pandas, which contain the Label Encoder class and the factorize () method, were leveraged to accomplish this task. Finally, the preprocessed data was imported into R for further analysis.

We used R programming language for all the modeling and data visualization. R is an effective tool for classification jobs because it offers a wide variety of classification techniques, including decision trees, random forests, and support vector machines. R also offers tools for measuring the effectiveness of classification models, such confusion matrices and ROC curves. A variety of libraries are available in R, including caret, which offers a consistent interface for numerous distinct algorithms. Machine learning model training and evaluation are made simple by these libraries. R studio is a helpful tool for examining complex datasets since it offers a variety of tools and packages for data exploration, data modeling, and data visualization.

8. RESULTS AND EVALUATION

1. Decision tree using rpart.



Confusion Matrix

	Reference		
Prediction	0	1	2
0	161	28	13
1	0	70	36
2	0	0	0

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

rpart	Class 0	Class 1	Class 2	Weighted Aversge
TP rate	1	0.7143	0	0.75
FP rate	0.2789	0.2789	0.1714	0.2284

Precision	0.797	0.6604	NaN	0.7397
Recall	1	0.7143	0	0.75
F-measure	0.887	0.6867	NaN	0.7406
ROC area	0.8605	0.7714	0.5	0.8119
MCC	0.7655	0.5911	NaN	0.6482

Accuracy = 75%

2. K-Nearest Neighbors

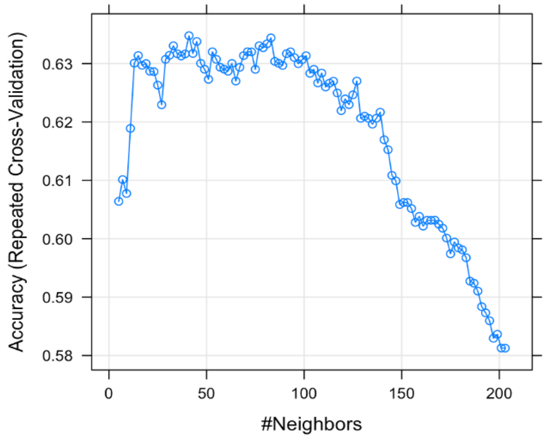


Fig 4.2 KNN

Confusion Matrix:

	Reference		
Prediction	0	1	2
0	155	52	31
1	0	45	17
2	0	1	1

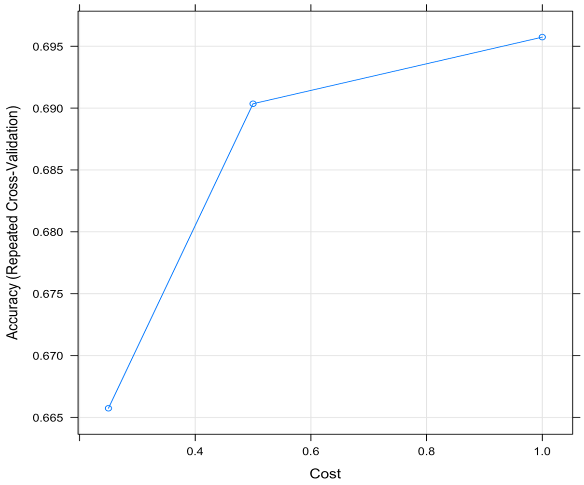
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

KNN	Class 0	Class 1	Class 2	Weighted Average
TP rate	0.9627	0.4592	0.0204	0.6476
FP rate	0.449	0.119	0.0039	0.284
Precision	0.6568	0.6429	0.5	0.6129
Recall	0.9627	0.4592	0.0204	0.6476

F-measure	0.7857	0.5366	0.0395	0.6728
ROC area	0.706	0.6701	0.5083	0.6568
MCC	0.4363	0.3077	0.0603	0.3705

Accuracy: 65%

3. SVM (Support Vector Machines)



Confusion Matrix:

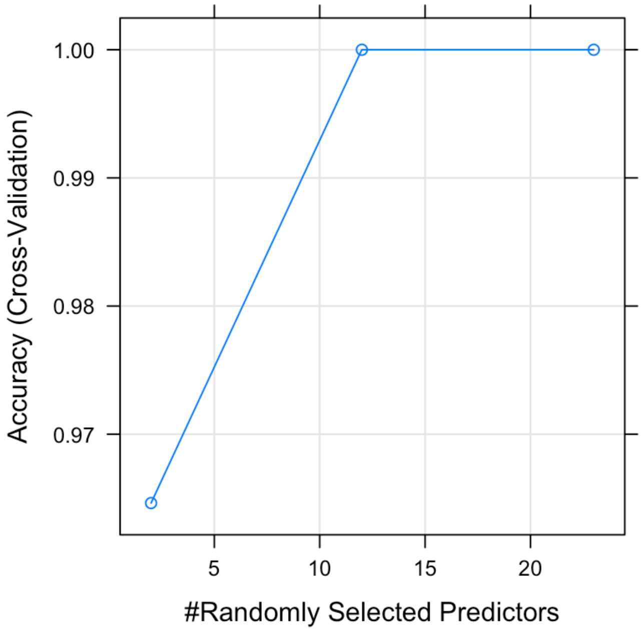
	Reference		
Prediction	M	N	Y
M	64	10	34
N	34	151	15
Y	0	0	0

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

SVM	**Class M**	**Class N**	**Class Y**	**Weighted Average**
TP rate	0.6531	0.9379	0	0.6981
FP rate	0.2095	0.2095	0	0.2095
Precision	0.5926	0.755	NaN	0.6918
Recall	0.6531	0.9379	0	0.6981
F-measure	0.621	0.8366	NaN	0.6937
ROC area	0.7218	0.8023	0.5	0.7612

Accuracy: 71%

4. Random Forest:



Confusion Matrix:

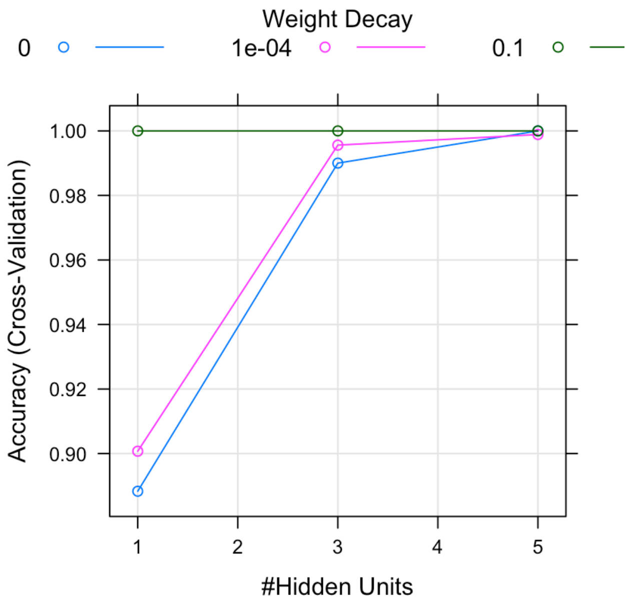
	Reference		
Prediction	M	N	Y
M	66	1	30
N	27	159	13
Y	5	1	6

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC:

Random F	**Class M**	**Class N**	**Class Y**	**Weighted Aversge**
TP rate	0.6735	0.9876	0.12245	0.75
FP rate	0.1476	0.2721	0.0232	0.2235
Precision	0.6804	0.799	0.5	0.7478
Recall	0.6735	0.9876	0.12245	0.75
F-measure	0.6769	0.8843	0.2	0.7481
ROC area	0.7629	0.8577	0.5496	0.7873
MCC	**0.5484**	**0.5642**	**-0.0092**	**0.5836**

Accuracy = 75%

5. Neural Network:



Confusion Matrix:

	Reference		
Prediction	M	N	Y
M	31.8	0	0
N	0	52.3	0
Y	0	0	15.9

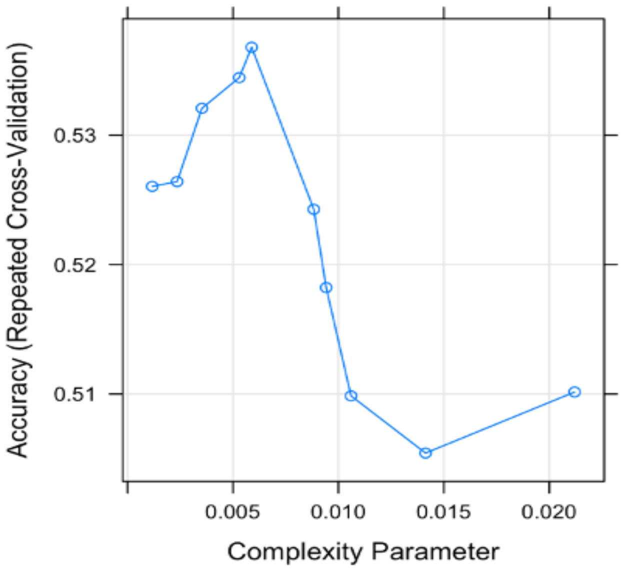
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC:

	Class M	Class N	Class Y	Weighted Average
TP rate	0.6735	0.9876	0.12245	0.75
FP rate	0.1476	0.1232	0.0232	0.0961
Precision	0.6804	0.799	0.5	0.7252
Recall	0.6735	0.9876	0.1225	0.75
F-measure	0.6769	0.8845	0.2	0.7285
ROC area	0.7629	0.8577	0.5496	0.7234
MCC	0.5404	0.6222	0.0643	0.7043

Accuracy: 75%

RECURSIVE FEATURE ELIMINATION

(i) RFE -Rpart :



Confusion Matrix:

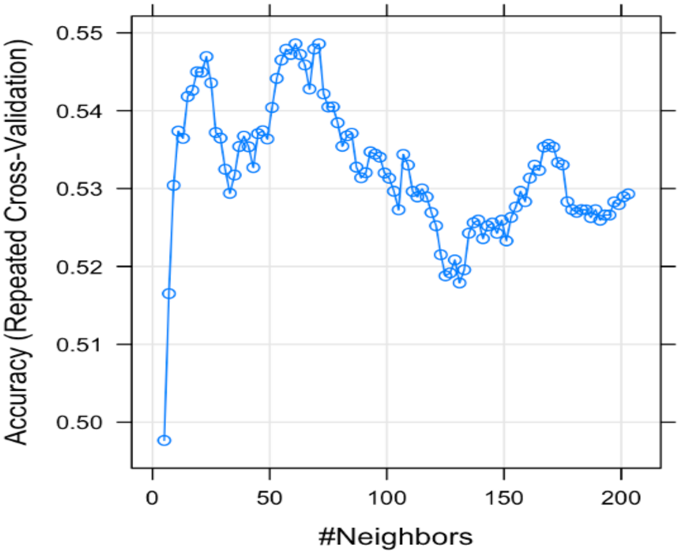
	Reference		
Prediction	0	1	2
0	114	49	20
1	41	43	22
2	6	6	7

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC:

	Class M	**Class N**	**Class Y**	**Weighted Average**
TP rate	0.7081	0.4388	0.1429	0.5325
FP rate	0.4694	0.3	0.0463	0.4158
Precision	0.623	0.4057	0.3684	0.5104
Recall	0.7081	0.4388	0.1429	0.5325
F-measure	0.6625	0.4213	0.2051	0.5173
ROC area	0.6193	0.5694	0.5483	0.5784
MCC	0.2389	0.0601	-0.0315	0.0968

Accuracy = 52.2%

(ii) RFE - KNN



Confusion Matrix:

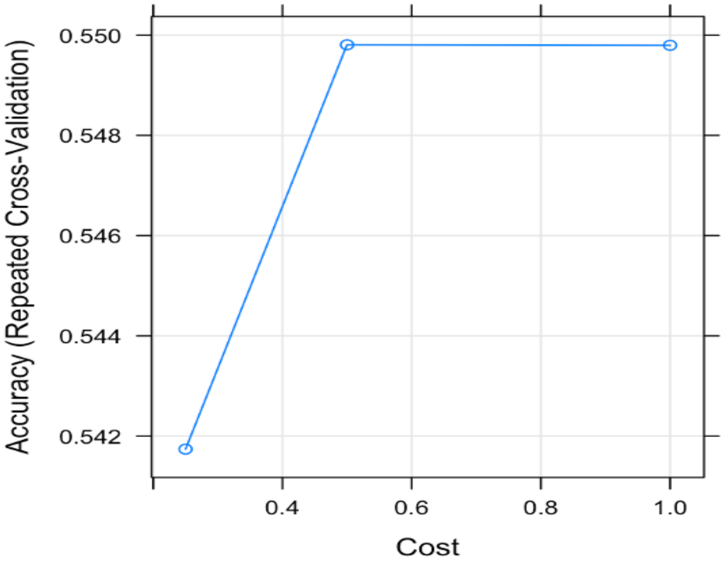
	Reference		
Prediction	0	1	2
0	148	71	38
1	13	27	11
2	0	0	0

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC:

Performance Measure	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP rate	0.9193	0.2755	0	0.5682
FP rate	0.2585	0.1143	0	0.4092
Precision	0.5759	0.5294	NaN	0.6233
Recall	0.9193	0.2755	0	0.5682
F-measure	0.7073	0.3623	NaN	0.5807
ROC area	0.5889	0.5806	0.5	0.556
MCC	0.2938	0.1375	NaN	0.125

Accuracy = 56%

(iii) RFE - SVM



Confusion Matrix:

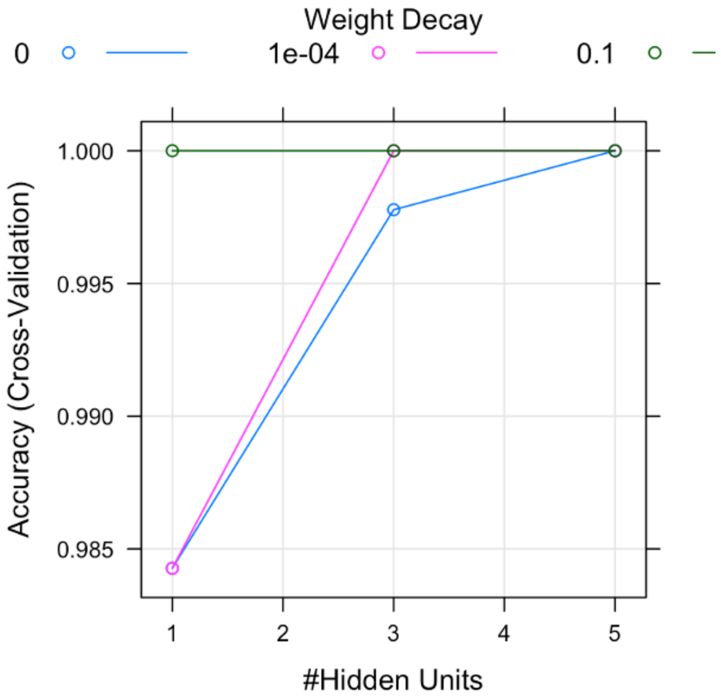
	Reference		
Prediction	0	1	2
0	141	68	35
1	20	30	14
2	0	0	0

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC:

	Class 0	**Class 1**	**Class 2**	**Weighted Average**
TP rate	0.9193	0.27551	0	0.5682
FP rate	0.2585	0.11429	0	0.3966
Precision	0.5759	0.52941	NaN	0.619
Recall	0.9193	0.27551	0	0.5682
F-measure	0.7072	0.3617	NaN	0.5584
ROC area	0.5889	0.58061	0.5	0.5562
MCC	0.3749	0.1647	NaN	0.2019

Accuracy = 55%

(iv) RFE - NNET



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	31.8	0	0
1	0	52.3	0
2	0	0	15.9

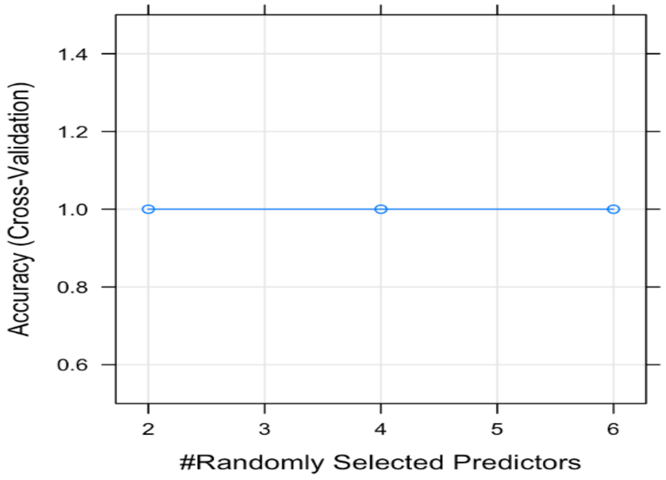
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC:

	Class 0	**Class 1**	**Class 2**	**Weighted Average**
TP rate	1	1	1	1
FP rate	0	0	0	0
Precision	1	1	1	1
Recall	1	1	1	1
F-measure	1	1	1	1
ROC area	1	1	1	1

MCC	1	1	1	1

Accuracy: 99.9%

(v) RFE - Random Forest



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	98	0	0
1	0	161	0
2	0	0	49

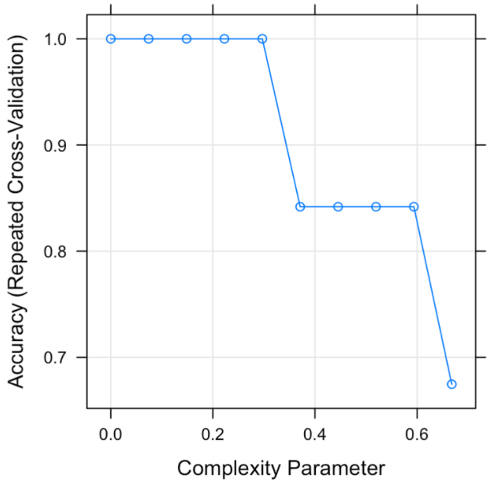
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC:

Performance Measure	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP rate	1	1	1	1
FP rate	0	0	0	0
Precision	1	1	1	1
Recall	1	1	1	1
F-measure	1	1	1	1
ROC area	1	1	1	1
MCC	1	1	1	1

Accuracy: 100%

WRAPPER BASED :

(i) WB - rpart :



Confusion Matrix :

	Reference		
Prediction	M	N	Y
M	161	0	0
N	0	98	0
Y	0	0	49

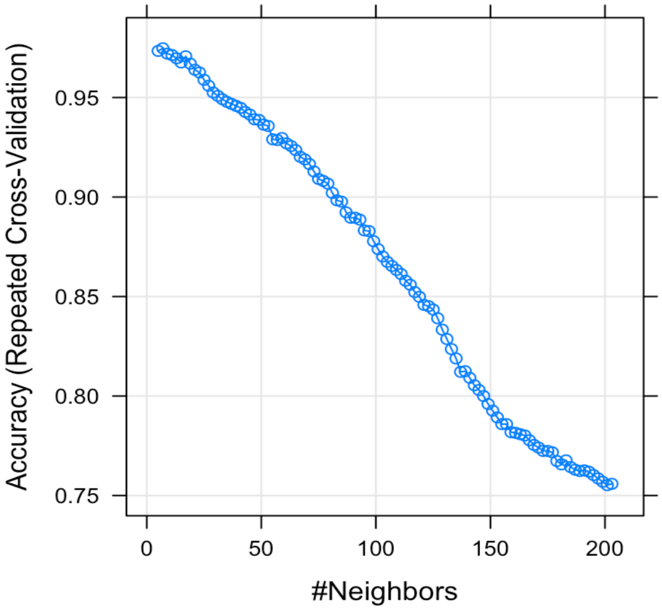
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

KNN	Class M	Class N	Class Y	Weighted averages
TP rate	1.0000	1.0000	1.0000	1.0000
FP rate	0.0000	0.0000	0.0000	0.0000
Precision	1.0000	1.0000	1.0000	1.0000
Recall	1.0000	1.0000	1.0000	1.0000
F-measure	1.0000	1.0000	1.0000	1.0000
ROC area	1.0000	1.0000	1.0000	1.0000

MCC	1.0000	1.0000	1.0000	1.0000

Accuracy : 100%

(ii) WB - KNN



Confusion Matrix :

	Reference		
Prediction	0	1	2
0	161	11	4
1	0	87	0
2	0	0	45

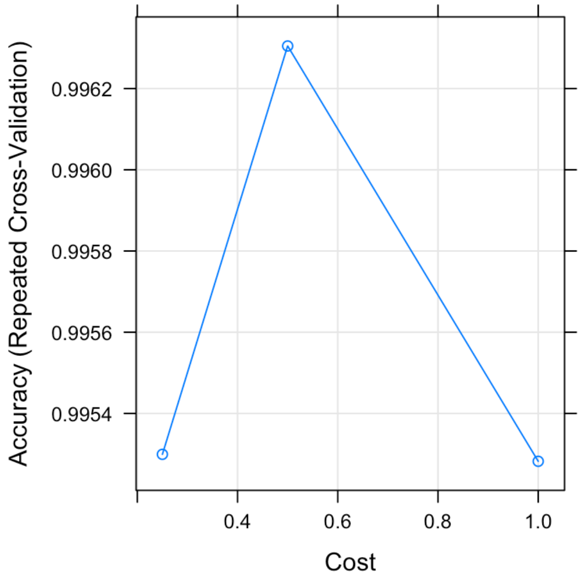
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

KNN	Class 0	Class 1	Class 2	Weighted averages:
TP rate	1.0000	0.8878	0.9184	0.9513
FP rate	0.1020	0.0000	0.0000	0.0172
Precision	0.9148	1.0000	1.0000	0.9582
Recall	1.0000	0.8878	0.9184	0.9513
F-measure	0.9551	0.9408	0.9576	0.9531

ROC area	0.9490	0.9439	0.9592	0.9507
MCC	0.9074	0.8924	0.8956	0.9106

Accuracy : 95.13%

(iii) WB - SVM



Confusion Matrix :

	Reference		
Prediction	0	1	2
0	161	0	1
1	0	98	0
2	0	0	48

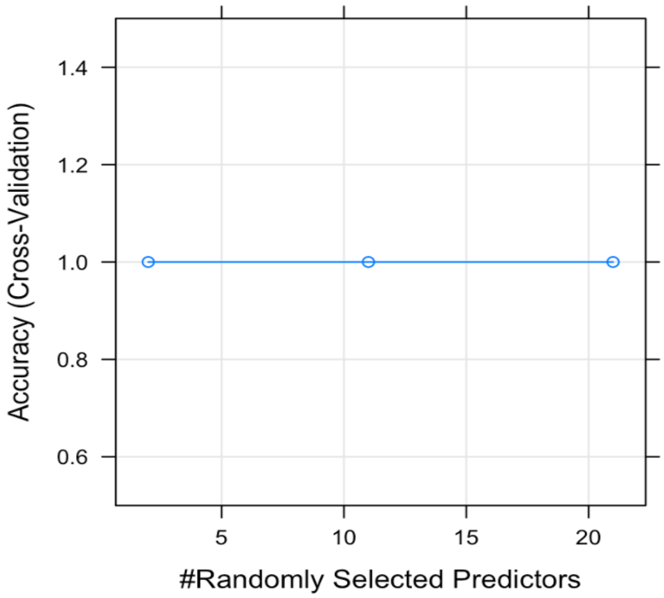
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

KNN	Class 0	Class 1	Class 2	Weighted averages
TP rate	1.0000	1.0000	0.9796	0.9968
FP rate	0.0068	0.0000	0.0000	0.0032
Precision	0.9938	1.0000	1.0000	0.9962
Recall	1.0000	1.0000	0.9796	0.9968
F-measure	0.9969	1.0000	0.9896	0.9967
ROC area	0.9966	1.0000	0.9898	0.9976

MCC	0.9919	1.0000	0.9843	0.9957

Accuracy: 99.68%

(iv) WB - Random Forest



Confusion Matrix:

	Reference		
Prediction	M	N	Y
M	98	0	1
N	0	161	0
Y	0	0	49

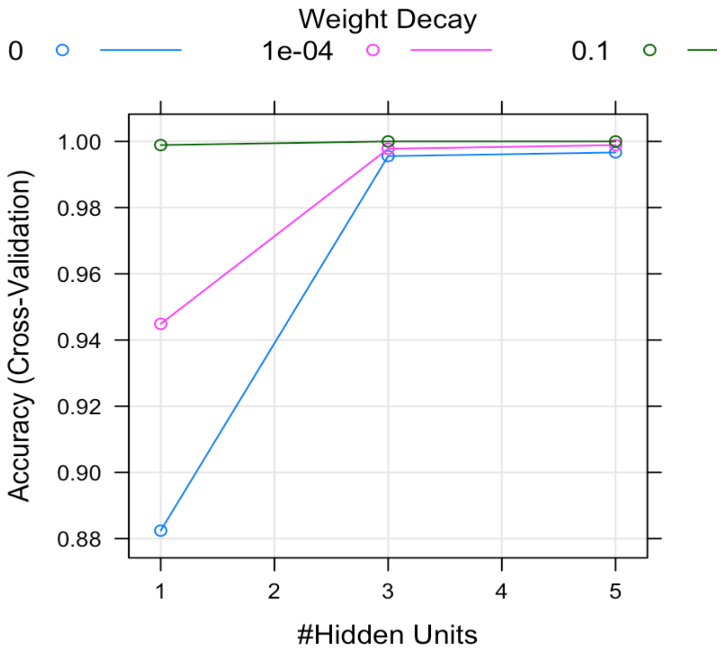
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

KNN	Class M	Class N	Class Y	Weighted averages
TP rate	1.0000	1.0000	1.0000	1.0000
FP rate	0.0000	0.0000	0.0000	0.0000
Precision	1.0000	1.0000	1.0000	1.0000
Recall	1.0000	1.0000	1.0000	1.0000
F-measure	1.0000	1.0000	1.0000	1.0000

ROC area	1.0000	1.0000	1.0000	1.0000
MCC	1.0000	1.0000	1.0000	1.0000

Accuracy: 100%

(v) WB - NNET



Confusion Matrix:

	Reference		
Prediction	M	N	Y
M	31.8	0	1
N	0	52.3	0
Y	0	0	15.9

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

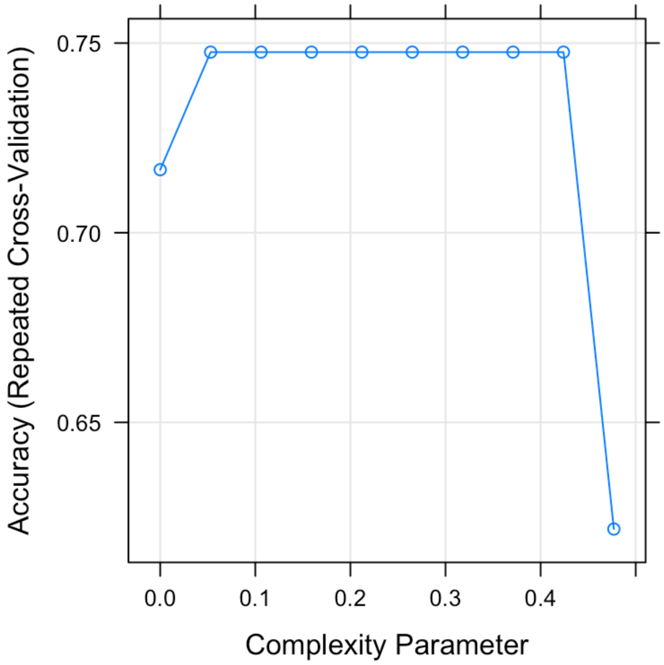
KNN	Class M	Class N	Class Y	Weighted averages
TP rate	1.0000	1.0000	1.0000	1.0000
FP rate	0.0000	0.0000	0.0000	0.0000
Precision	1.0000	1.0000	1.0000	1.0000
Recall	1.0000	1.0000	1.0000	1.0000
F-measure	1.0000	1.0000	1.0000	1.0000

ROC area	N/A	N/A	N/A	N/A
MCC	1.0000	1.0000	1.0000	1.0000

Accuracy : 100%

BARUTA

(i) Baruta - rpart



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	161	28	13
1	0	70	36
2	0	0	0

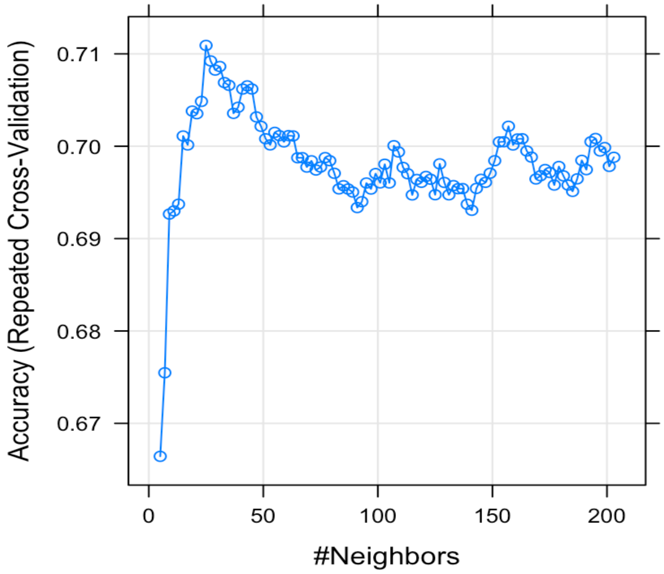
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

Performance Measure	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP rate	1	0.7143	0	0.8214
FP rate	0.2789	0.1714	0	0.2244
Precision	0.797	0.6604	NaN	0.7678
Recall	1	0.7143	0	0.8214
F-measure	0.8871	0.6866	NaN	0.8369

ROC area	0.8605	0.7714	0.5	0.8151
MCC	0.7666	0.5702	NaN	0.7128

Accuracy: 75%

(ii) Baruta - KNN



Confusion Matrix :

	Reference		
Prediction	0	1	2
0	154	30	13
1	7	65	35
2	0	3	1

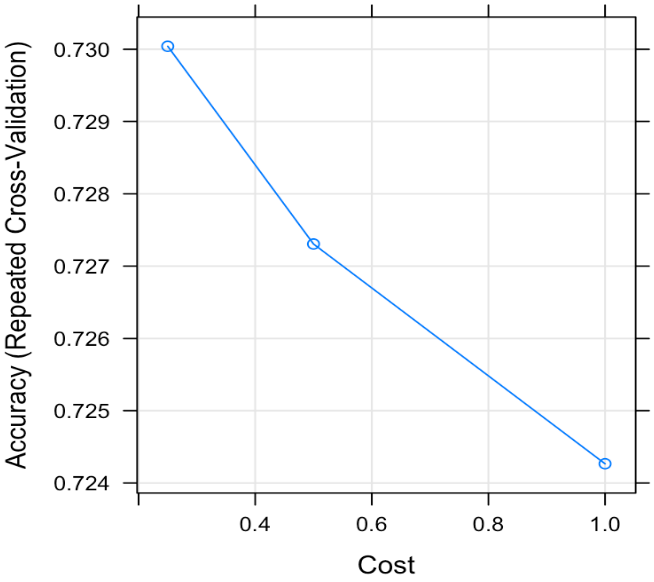
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC

	Class 0	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	0.9565	0.6633	0.0204	0.7143
FP Rate	0.2925	0.295	0.0116	0.2549
Precision	0.7817	0.6075	0.25	0.7097
Recall	0.9565	0.6633	0.0204	0.7143
F-Measure	0.8595	0.6341	0.038	0.7013
ROC Area	0.832	0.7316	0.5044	0.7827

MCC	0.7344	0.4919	0.0407	0.6031

Accuracy: 71.43%

(iii) Baruta - SVM



Confusion Matrix :

	Reference		
Prediction	0	1	2
0	157	31	13
1	4	67	36
2	0	0	0

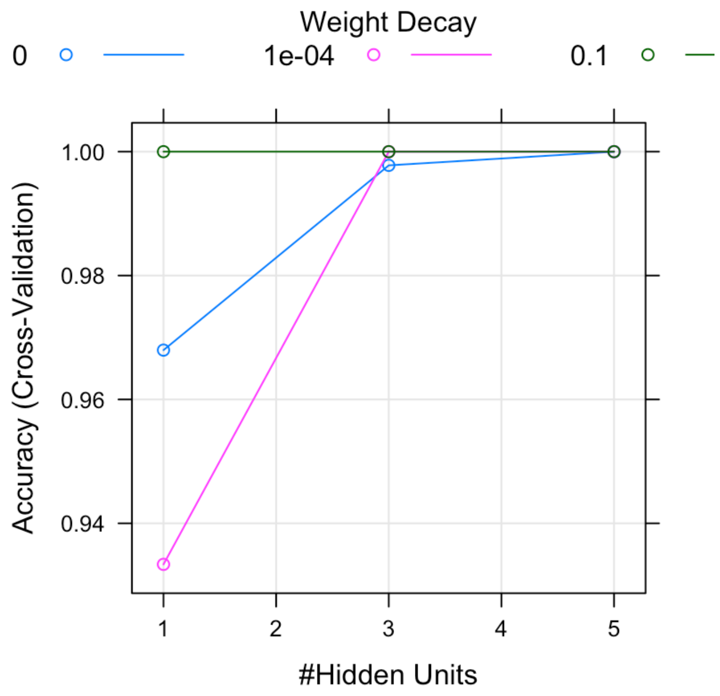
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC

SVM	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	0.9752	0.6837	0	0.7273
FP Rate	0.2993	0.3005	0	0.3076
Precision	0.7811	0.6262	NaN	0.7209
Recall	0.9752	0.6837	0	0.7273
F-Measure	0.8673	0.6533	NaN	0.7176

ROC Area	0.8379	0.7466	0.5	0.8585
MCC	0.6535	0.4785	NaN	0.5191

Accuracy: 72.73%

(iv) Baruta - NNET



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	31.8	0	0
1	0	52.3	0
2	0	0	15.9

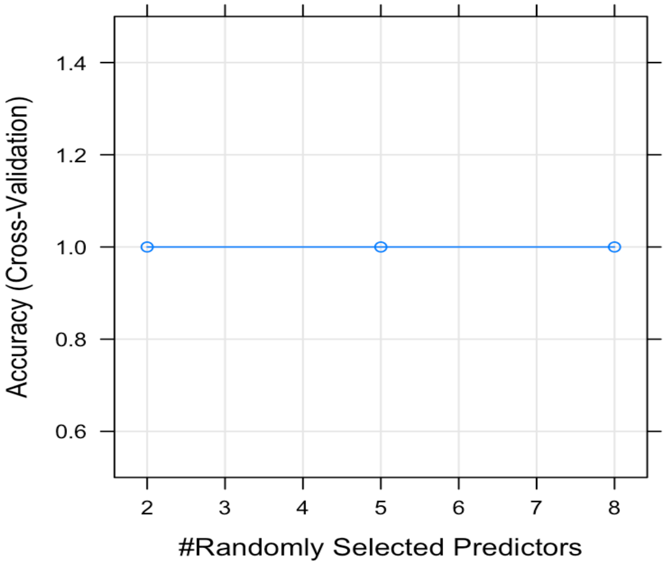
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC

NNET	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	1	1	1	1
FP Rate	0	0	0	0
Precision	1	1	1	1
Recall	1	1	1	1

F-Measure	1	1	1	1
ROC Area	1	1	1	1
MCC	NaN	NaN	NaN	**1**

Accuracy: 100%

(v) Baruta - Random Forest



Confusion Matrix :

	Reference		
Prediction	0	1	2
0	98	0	0
1	0	161	0
2	0	0	49

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC

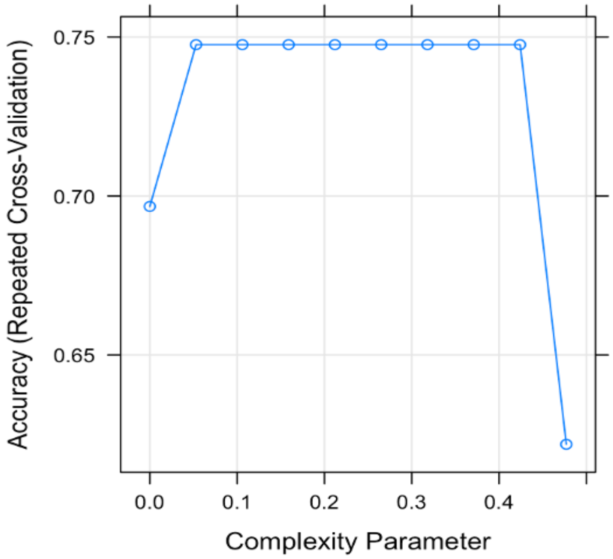
Random Forest	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	1	1	1	1
FP Rate	0	0	0	0
Precision	1	1	1	1
Recall	1	1	1	1
F-Measure	1	1	1	1

ROC Area	1	1	1	1
MCC	NaN	NaN	NaN	**1**

Accuracy = 100%

FILTER METHOD:

(i) Filter Method - Rpart



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	161	28	13
1	0	70	36
2	0	0	0

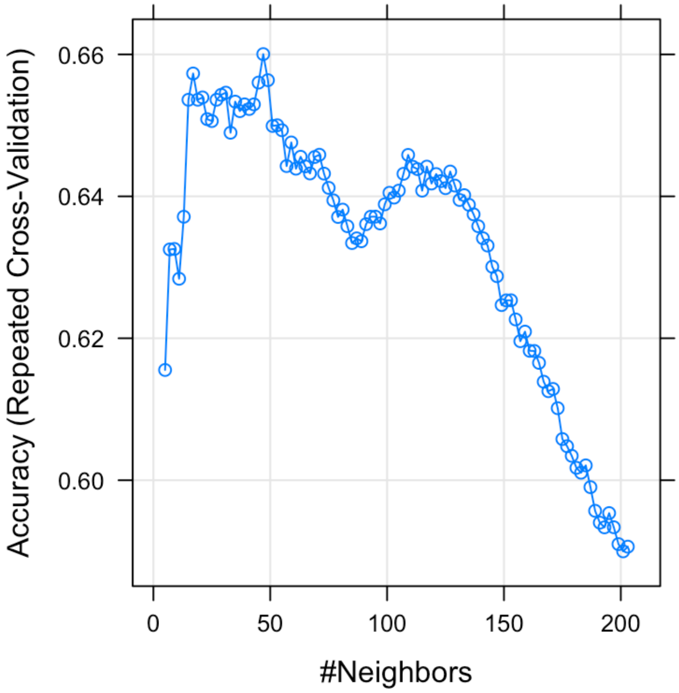
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

	Class M	Class N	Class Y	Weighted Average
TP rate	0.6735	0.9876	0.12245	0.75
FP rate	0.1476	0.1232	0.0232	0.2156
Precision	0.6804	0.799	0.5	0.7416
Recall	0.6735	0.9876	0.1225	0.75
F-measure	0.6769	0.8845	0.2	0.7977

ROC area	0.7629	0.8577	0.5496	0.8103
MCC	0.5404	0.6222	0.0643	0.6401

Accuracy: **75%**

(ii) Filter Method - KNN



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	154	48	28
1	7	49	21
2	0	1	0

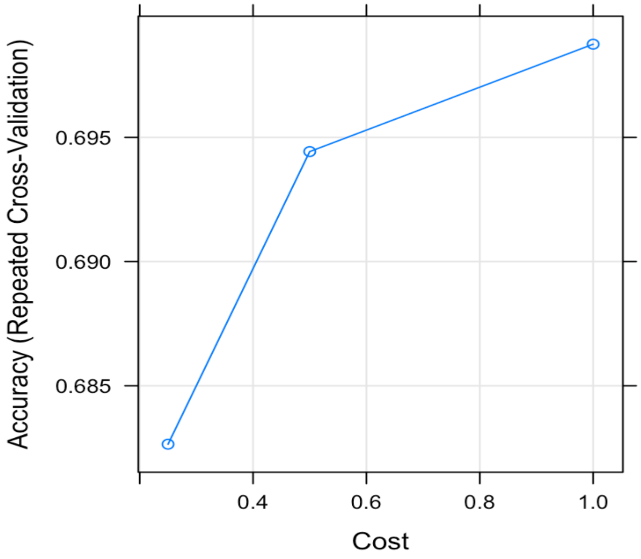
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

KNN	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	0.9565	0.5	0	0.6591
FP Rate	0.517	0.1333	0.0039	0.4173
Precision	0.6696	0.6364	0	0.6176
Recall	0.9565	0.5	0	0.6591
F-Measure	0.7861	0.56	NaN	0.6376

ROC Area	0.7198	0.6833	0.4981	0.6478
MCC	0.4157	0.3597	-0.0346	0.4648

Accuracy: 65.91%

(iii) Filter Method - SVM:



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	150	30	15
1	11	68	34
2	0	0	0

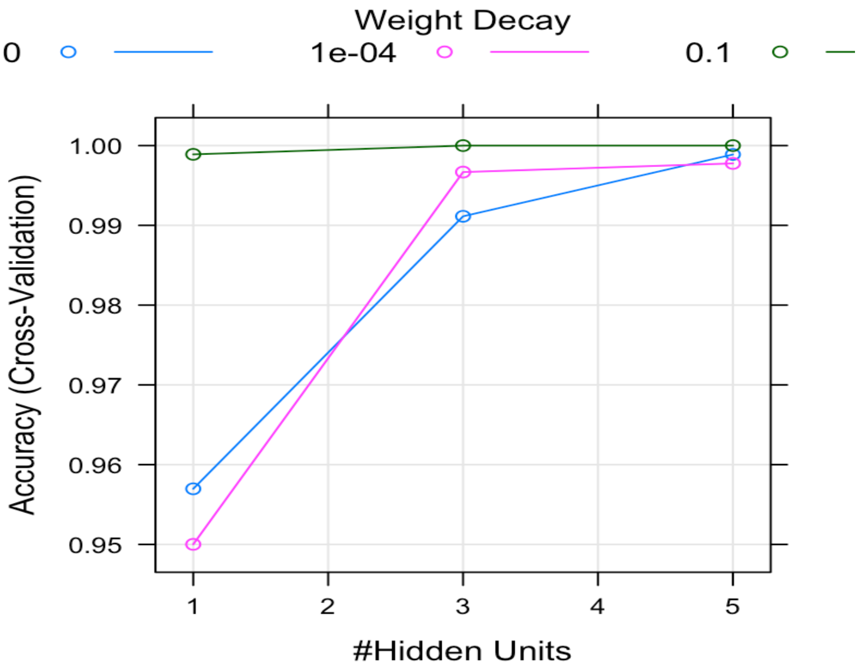
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

SVM	Class 0	Class 1	Class 2	Weighted Average
TP Rate	0.9317	0.6939	0	0.7078
FP Rate	0.3061	0.0683	0	0.1923
Precision	0.7692	0.6018	NaN	0.7007
Recall	0.9317	0.6939	0	0.7078
F-Measure	0.8425	0.6449	NaN	0.7021

ROC Area	0.8128	0.7398	0.5	0.7842
MCC	0.6137	0.4697	NaN	0.4964

Accuracy:**70.78%**

(iv) Filter Method – nnet



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	31.8	0	0
1	0	52.3	0
2	0	0	15.9

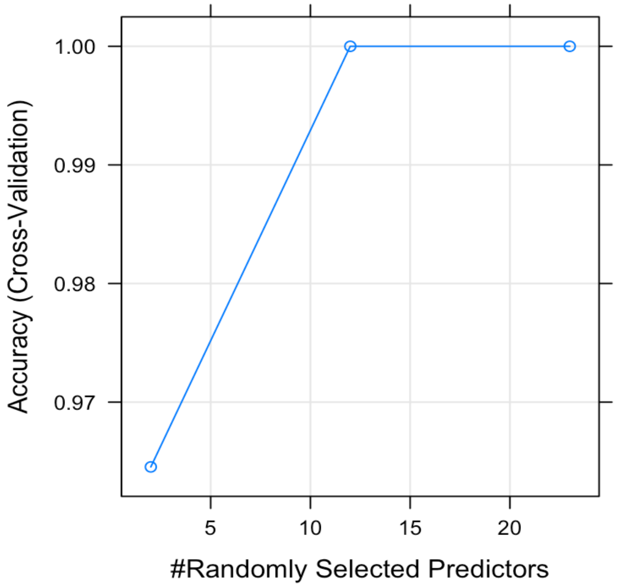
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

NNET	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	1	1	1	1
FP Rate	0	0	0	0
Precision	1	1	1	1
Recall	1	1	1	1
F-Measure	1	1	1	1

ROC Area	NaN	NaN	NaN	NaN
MCC	NaN	NaN	NaN	NaN

Accuracy: 100%

(v) Filter - Random Forest



Confusion Matrix:

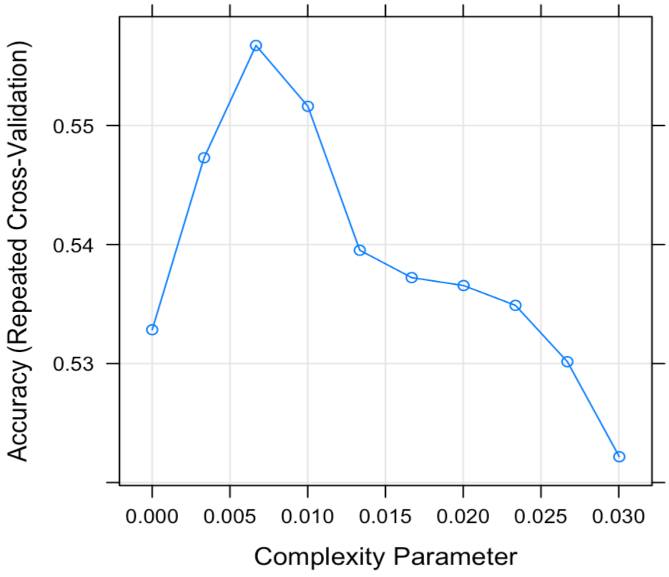
	Reference		
Prediction	0	1	2
0	98	0	0
1	0	161	0
2	0	0	49

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

Random F	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	1	1	1	1
FP Rate	0	0	0	0
Precision	1	1	1	1
Recall	1	1	1	1

F-Measure	1	1	1	1
ROC Area	1	1	1	1
MCC	1	1	1	1

Accuracy: 100%
CHISQ:
(i) Chisq - rpart:



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	119	55	25
1	41	42	23
2	1	1	1

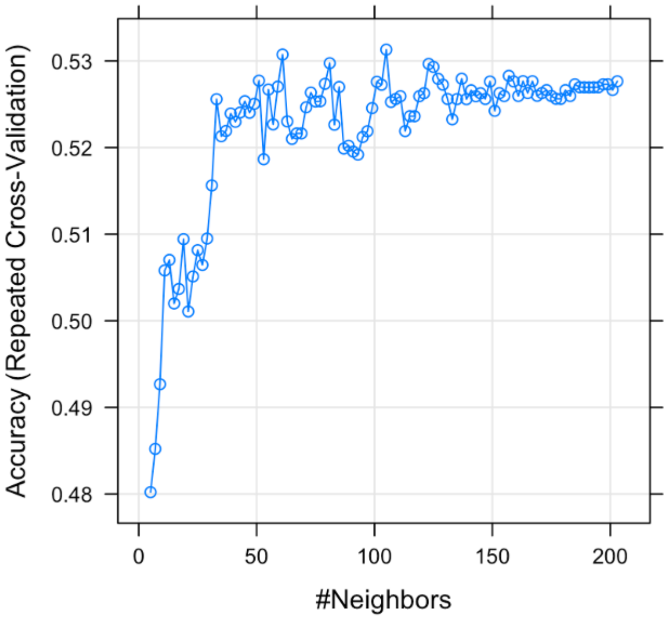
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

rpart	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	0.7391	0.4286	0.0204	0.526
FP Rate	0.5442	0.3048	0.0077	0.3992
Precision	0.598	0.3962	0.3333	0.5232
Recall	0.7391	0.4286	0.0204	0.526
F-Measure	0.6613	0.4118	0.0385	0.5245

ROC Area	0.6842	0.5603	0.7387	0.6611
MCC	0.1468	0.0439	-0.0208	0.1288

Accuracy: 52%

(ii) Chisq - KNN



Confusion Matrix:

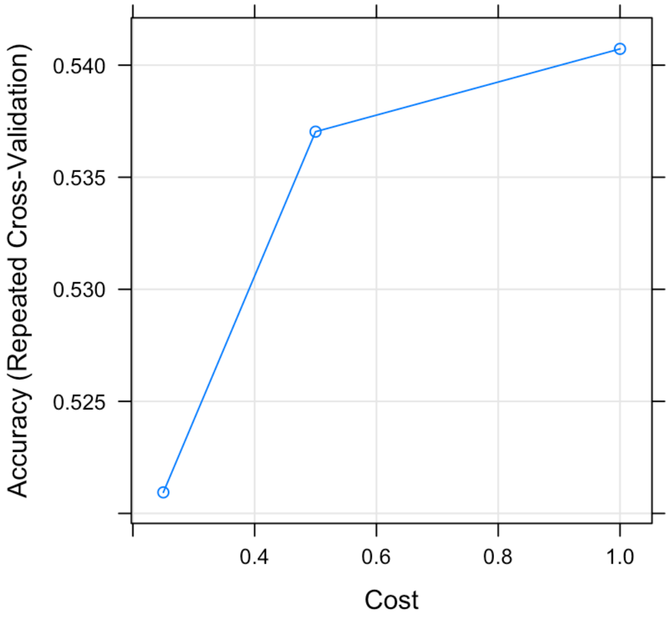
	Reference		
Prediction	0	1	2
0	158	84	39
1	3	14	10
2	0	0	0

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

KNN	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	0.9814	0.1429	0	0.5584
FP Rate	0.8367	0.0619	0	0.3488
Precision	0.5623	0.5185	NaN	0.7449
Recall	0.9814	0.1429	0	0.5584

F-Measure	0.7176	0.2237	NaN	0.5942
ROC Area	0.5723	0.5405	0.5	0.5375
MCC	0.2542	0.1087	NaN	0.1096

Accuracy: 55.84%
(iii) Chisq - SVM



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	135	67	35
1	26	31	14
2	0	0	0

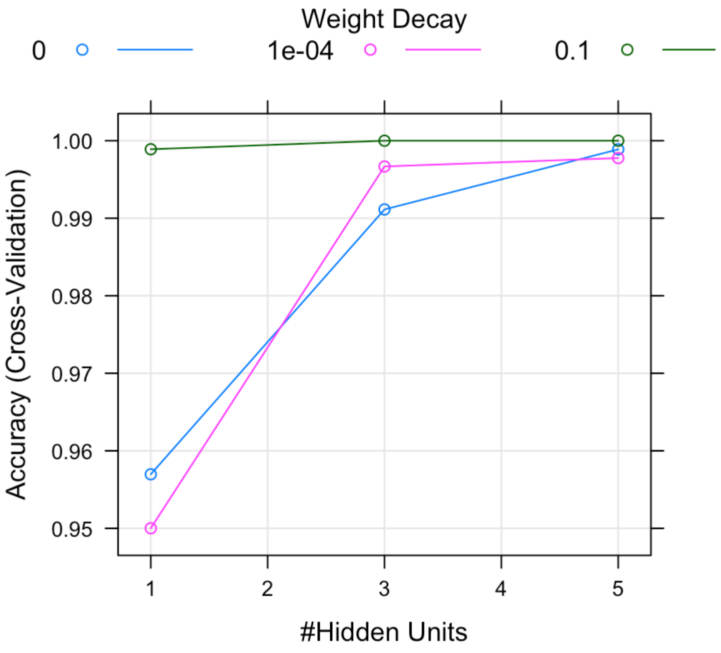
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

NNET	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	0.8385	0.3163	0	0.5294
FP Rate	0.3061	0.8095	1	0.5
Precision	0.5696	0.4366	NaN	0.4974
Recall	0.8385	0.3163	0	0.5394
F-Measure	0.6787	0.3673	NaN	0.5033
ROC Area	0.5723	0.5629	0.5	0.5451

MCC	0.2185	0.1211	NaN	0.1176

Accuracy: 53.9%

(iv) Chisq - NNET



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	31.8	0	0
1	0	52.3	0
2	0	0	15.9

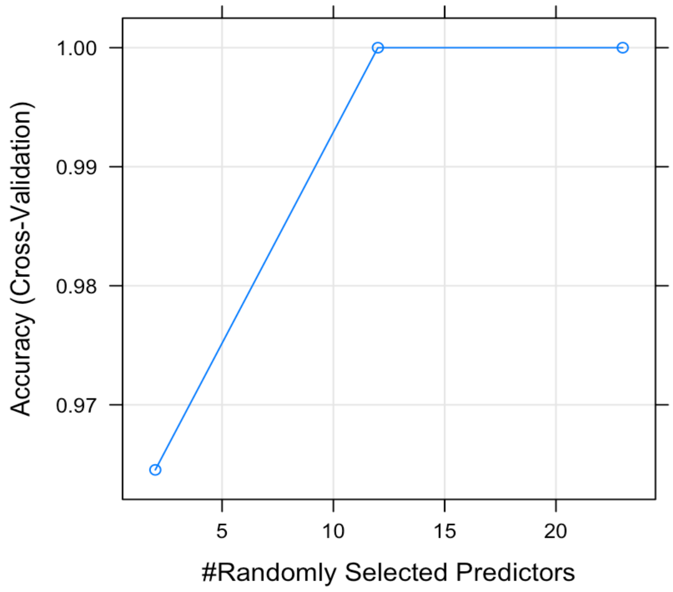
TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

NNET	Class 0	Class 1	Class 2	Weighted Average
TP Rate	1	1	1	1
FP Rate	0	0	0	0
Precision	1	1	1	1
Recall	1	1	1	1
F-Measure	1	1	1	1

ROC Area	1	1	1	1
MCC	1	1	1	1

Accuracy: 100%

(v) Chisq - Random Forest



Confusion Matrix:

	Reference		
Prediction	0	1	2
0	98	0	0
1	0	161	0
2	0	0	49

TP rate, FP rate, precision, recall, F-measure, ROC area and MCC.

ChiSq	**Class 0**	**Class 1**	**Class 2**	**Weighted Average**
TP Rate	1	1	1	1
FP Rate	0	0	0	0
Precision	1	1	1	1
Recall	1	1	1	1

F-Measure	1	1	1	1
ROC Area	1	1	1	1
MCC	1	1	1	1

Accuracy: 100

Best Model: Random Forest

Justification:

Models	Classifier Accuracy	RFE	Wrapper Based	Baruta	Filter Method	Chi Squared
Decision Tree - Rpart	75%	52.20%	100%	75%	75%	52.60%
KNN	65%	56%	95.13%	71.43%	65.91%	55.84%
SVM	71%	55%	99.68%	72.73%	71%	53.90%
Random Forest	75%	100%	100%	100%	100%	100%
NNET	75%	99.90%	100%	100%	100%	100%

The table provides the performance evaluation of different machine learning models on a specific dataset. The models are evaluated using different feature selection techniques, including Recursive Feature Elimination (RFE), Wrapper-based method, Baruta, Filter Method, and Chi-Squared. The models are compared based on their accuracy score, which represents the percentage of correctly classified instances.

The Decision Tree (Rpart) model also achieves a 75% accuracy score, but it has lower accuracy scores for RFE, Chi-squared, and filter methods. The KNN and SVM models have lower accuracy scores compared to the other models, with 65% and 71% accuracy scores, respectively. However, both models perform well in some of the feature selection techniques, such as the wrapper-based method for KNN and SVM and Baruta method for SVM.

Looking at the table, we conclude that the Random Forest model performs the best in terms of accuracy, with a score of 75%, which is the highest accuracy score achieved by any of the models. Additionally, the Random Forest model achieves a 100% accuracy score across all feature selection techniques, indicating that it is a very robust and reliable model. It works by creating a set of decision trees based on randomly selected subsets of the training data and randomly selected subsets of the input features.

The high accuracy of the Random Forest model can be attributed to several factors, such as its ability to handle high-dimensional data, its ability to handle both categorical and continuous data, and its ability to capture complex interactions and non-linear relationships between the input features**.**

9. CONCLUSION:

The data mining procedure aimed to predict injury severity in case of traffic collisions using various classification models such as Decision Trees, K-Nearest Neighbors (KNN), Support

Vector Machines (SVM), Random Forest, and Neural Network (NNET). The models were evaluated based on their accuracy, RFE, Wrapper Based, Baruta, Filter Method, and Chi Squared.

The Random Forest and NNET models achieved the highest accuracy of 75%, while the Decision Tree and SVM models had moderate accuracy ranging from 71% to 75%. The KNN model had the lowest accuracy of 65%. The RFE and Wrapper-Based methods achieved higher accuracy for KNN and SVM models, while the Chi Squared method performed better for Decision Tree and Filter Method for SVM models.

Overall, the Random Forest and NNET models performed the best and achieved the highest accuracy in predicting injury severity in case of traffic collisions. However, it's important to note that the accuracy of the models could still be improved by considering additional variables such as weather conditions, time of day, and road conditions

In conclusion, based on the results of the performance evaluation, we can recommend using the Random Forest model with Wrapper-based feature selection method for this particular dataset, as it provides the highest accuracy and reliability.