Performance Analysis of Classification Algorithms

1.Introduction:

Data mining is the process to extract potentially valuable and relevant information from big amount of data sets. It includes a set of techniques, such as classification, clustering, association rule mining, anomaly detection, etc. Classification or prediction is the most commonly used technique of data mining. Classification is a kind of supervised learning technique that identifies the hidden relationships between dependent and independent variables. Supervised learning techniques extract certain important features from the training data and then use those features to test on unobserved data. A wide application of classification techniques is image classification, pattern recognition, medical disease diagnosis, fault detection, traffic accident severity analysis, and detecting financial trends.

In order to use the classification model for actual implementation, certain criteria are used to validate the performance of the model. Several types of classification techniques are existing, such as, NB, DT, K-Nearest Neighbor, RF, etc. The performance of all the classification algorithms is not similar on all data types. In other words, the performance of different classifiers is varied on different data sets. The data sets can have three basic types of attribute values: numeric, nominal or both. Therefore, the selection of any classification algorithms must utilize the knowledge about data and its attribute values. Wrong selection of classification algorithm will certainly lead to bad classification model and bad results. This motivates our study.

This report evaluates the performance of most popular classification algorithms, namely, NB, DT ,RF and KNN on three different types of data sets. The outcome of this study will certainly contributes in identifying if the different characteristics of the data affect the performance of classifiers. Also, we will identify that for what kind of data, which classification algorithms will be more suitable.

This study would be helpful for the beginners to choose among the set of classification techniques to perform on a variety of data set. We designed a GUI in which we can do analysis of datasets using classification techniques and implement the best classification technique among them in real time by giving values to dataset and doing the prediction

To identify which type of dataset and which classification technique is more suitable. There are three types of datasets (numerical, nominal, and mixed) and many types of classification algorithms. The wrong selection of classification algorithms will certainly lead to bad classification models and bad results. It is also very hard to write each and every time code to analyze the dataset and create a prediction model for it. So we developed a machine learning model to handle the above problems and also designed a GUI to make it user-friendly.

2.Goals of project :

- Performance analysis of classification algorithms on various type of datasets.
- Making the process of prediction and training the model with the help of a GUI.

3.METHODOLOGIES

Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. In a world where nearly all manual tasks are being automated, the definition of manual is changing. Machine Learning algorithms can help computers play chess, perform surgeries, and get smarter and more personal. With the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to decide

which algorithm is better and we can predict in real time too. Data mining has the ability to extract hidden knowledge from a huge amount of data. In this project we have focused on developing a system based on four classification methods namely Random Forest , Decision Tree Classifier [CART] , K-Nearest Neighbor and Naïve Bayes. We discuss about them briefly in below cases.

3.1. K-Nearest Neighbor

K-nearest neighbor algorithm is one of the classification algorithms. It is the simplest and easy than other data mining techniques.KNN is a non-parametric method used for classification and regression. It is a type of instance-based learning or lazy learning. This technique classifies new belongings based on similarity measure .the value of k always assign positive integer number .In this algorithm the training data are stored. Based on the neighbors or nearest prediction of test data is complete.

Phase I : Determine k which is the number of nearby neighbors.

Phase II : Estimate distance between the instance and training samples.

Phase III : The remoteness of the training samples are sorted and the closest neighbor based on the minimum the distance is determined in this step.

Phase IV : In this step we get all the classes of all the training data

Phase V : Use the majority of the class of closest neighbors as the prediction value of the query instance .

Advantages:

• KNN is pretty intuitive and simple.

• Very easy to implement for multi-class problem

• Can be used both for Classification and Regression

Disadvantages:

• KNN is computationally expensive.

• Variables should be normalized, or else higher range variables can bias the algorithm.

• Data still needs to be pre-processed.

3.2. Decision Tree Classifier (CART)

Decision tree is a Supervised machine learning algorithm used to solve classification problems. The main objective is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and

classification. The typical algorithms of decision tree are ID3,CART.The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes

Advantages:

• A major decision tree analysis advantages is its ability to assign specific values to problem.

• The decision tree model is transparent in nature.

• It allows for a comprehensive analysis of the consequences of each possible decision.

Disadvantages:

• May suffer from overfitting.

• Classifies by rectangular partitioning.

• Does not easily handle non-numeric data.

• Can be quite large- pruning is necessary.

3.3. Naive Bayes (NB)

Naive Bayes is a classification technique with a notion that defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. It is based on conditional probability. Naive Bayes is a machine learning classifier that employs the Bayes theorem. Naive Bayes classifiers assume attributes have independent distributions. It is considered to be fast and space-efficient. It also provides a simple approach, with clear semantics, representing and learning probabilistic knowledge. It is known as Naive because it relies on two important simplifying assumptions. The predictive attributes are conditionally independent and secondly, it assumes that no hidden attributes bias the prediction process. It is very fast to train and fast to classify.

Advantages:

• A Naive Bayesian model is easy to build and useful for massive datasets.
• It is simple and is known to outperform even highly sophisticated classification methods.
• Good results were obtained in most cases.

Disadvantages:

• Assumes class conditional independence, therefore loss of accuracy.
• Practically, dependencies exist among variables.

3.4. Random Forest (RF)

Random forest is a Supervised machine learning algorithm used to solve classification problems. It is a method that operates by constructing multiple decision trees during the training phase. The decision of the majority of the trees is taken as the final decision.

Advantages:

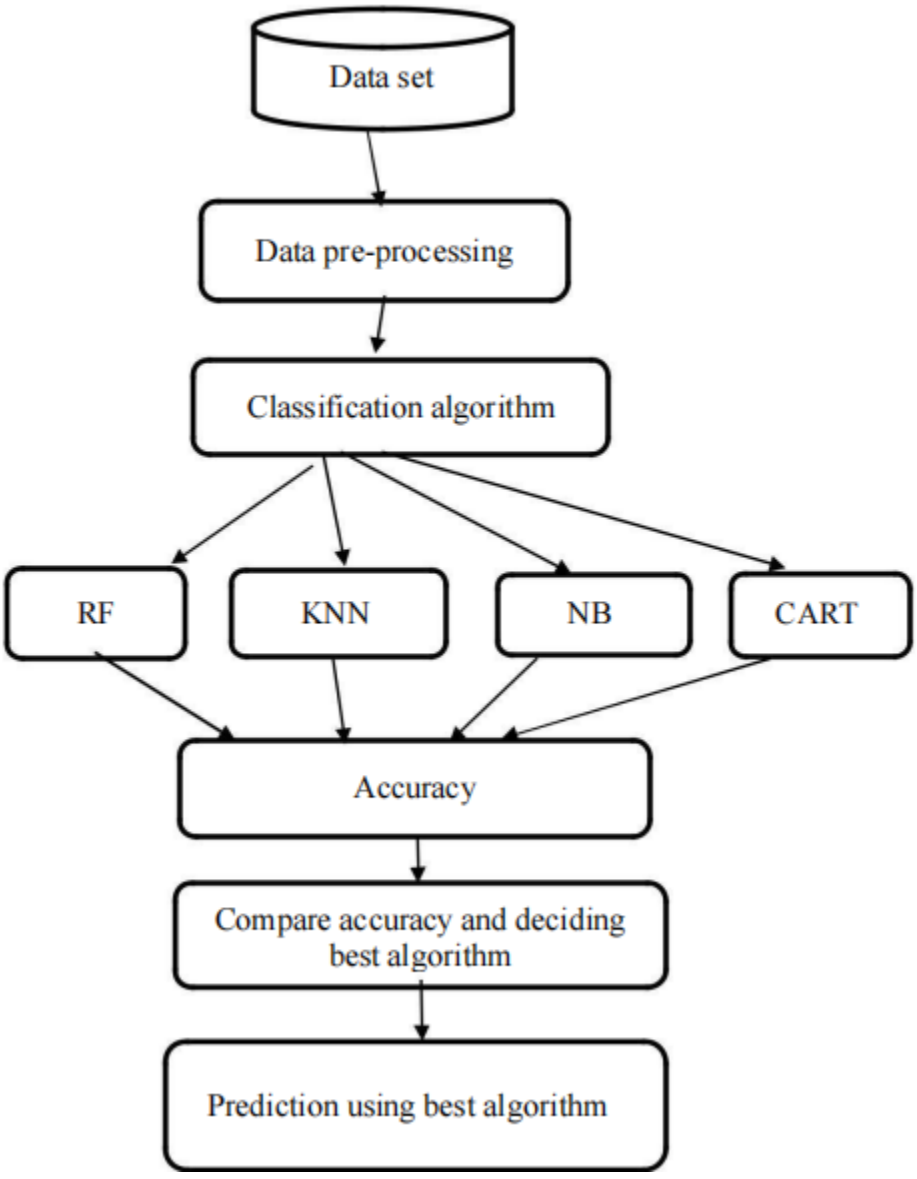
• As we mentioned earlier a single decision tree tends to overfit the data. The process of averaging or combining the results of different decision trees helps to overcome the problem of overfitting.
• Random forests also have less variance than a single decision tree. It means that it works correctly for a large range of data items than single decision trees.
• Random forests are extremely flexible and have very high accuracy.
• They also do not require the preparation of the input data. You do not have to scale the data.
• It also maintains accuracy even when a large proportion of the data are missing.

Disadvantages:

• The main disadvantage of Random forests is their complexity. They are much harder and more time-consuming to construct than decision trees.
• They also require more computational resources and are also less intuitive. When you have a large collection of decision trees it is hard to have an intuitive grasp of the relationship existing in the input data.

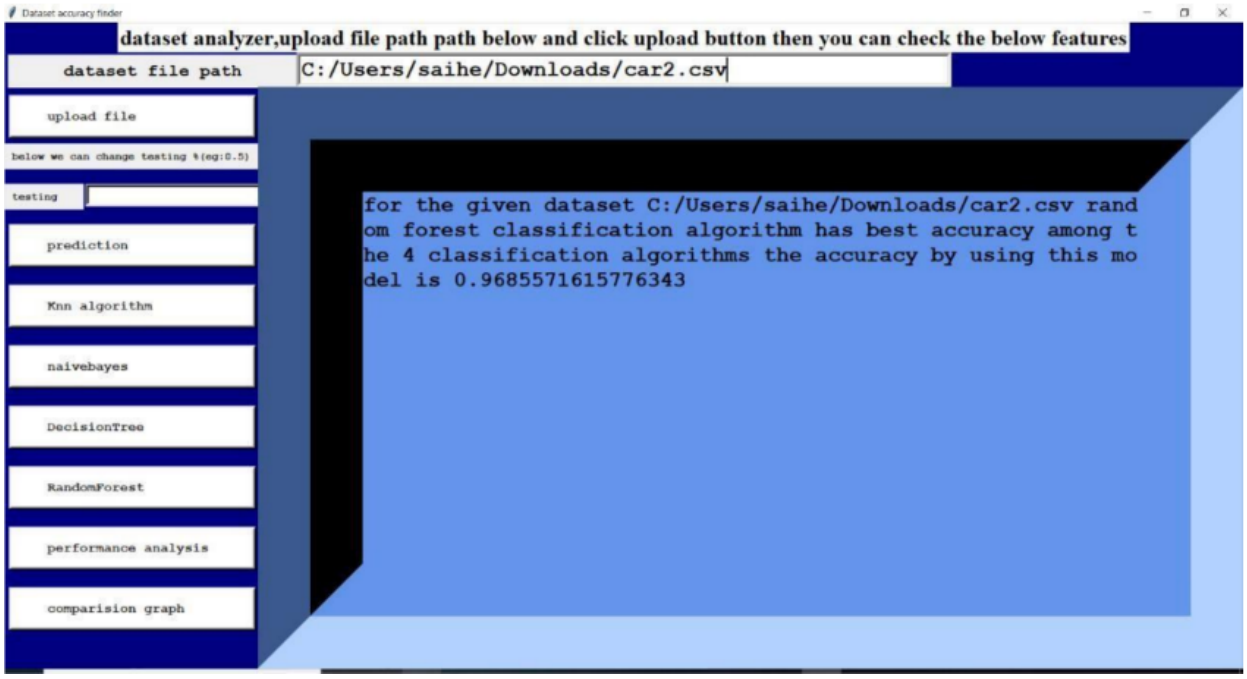
• In addition, the prediction process using random forests is more time-consuming than other algorithms.

4. High Level Design Solution

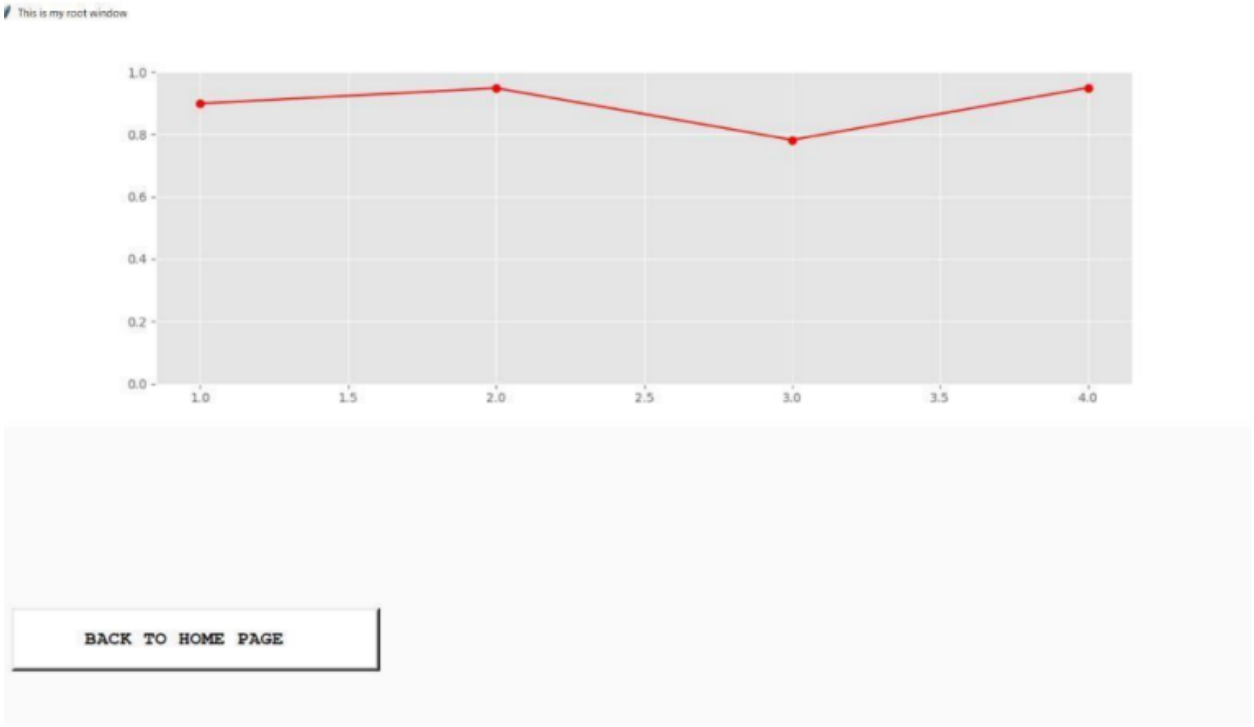


The developed system is in the form of a GUI and it is loaded from the command prompt by using python commands. Then dataset should be loaded and prediction part can be done .We can see them in below sections.

In this page the file is uploaded by giving file path and clicking upload file button. The default test size is 0.3 we can vary it by giving the value in testing textbox and clicking the upload file button . we can check accuracies of different algorithms by clicking on their respective buttons.

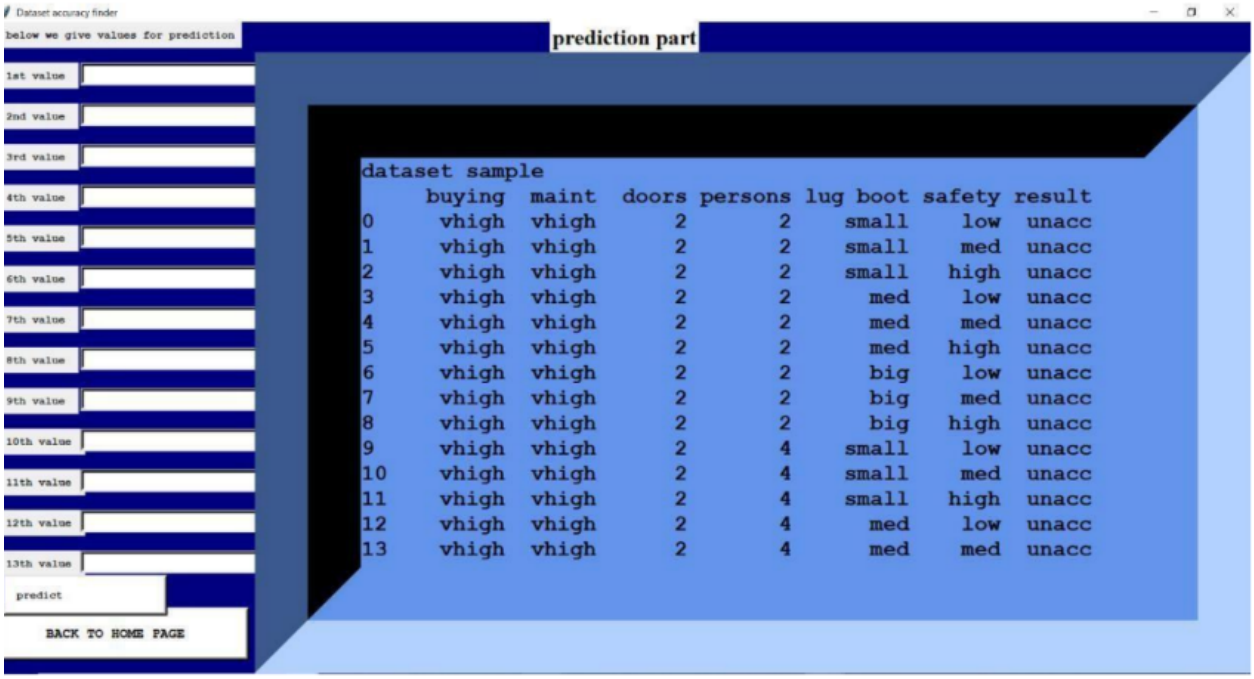


When we click on performance analysis button we can see which algorithm is best for that dataset and test size.

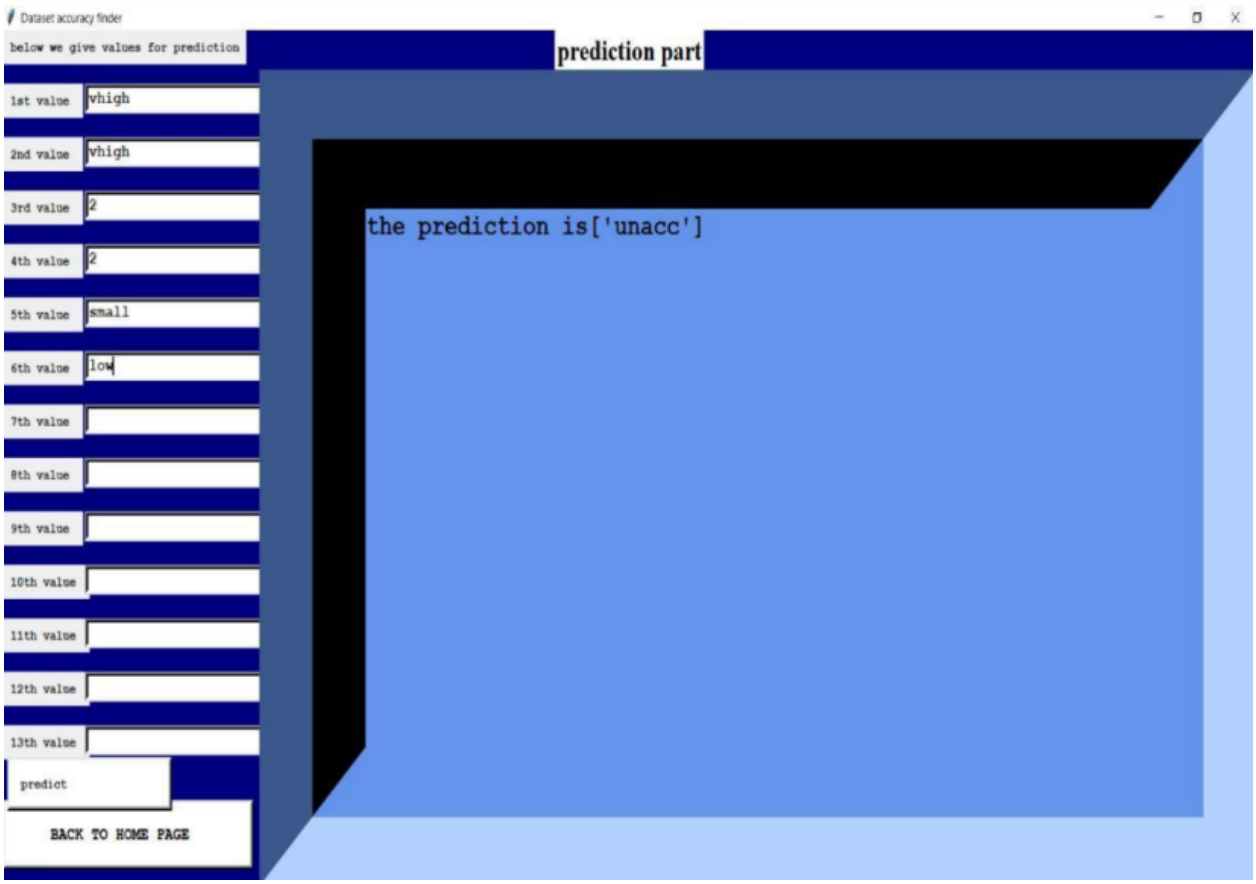


When we click on comparison graph button we can see graph four algorithms

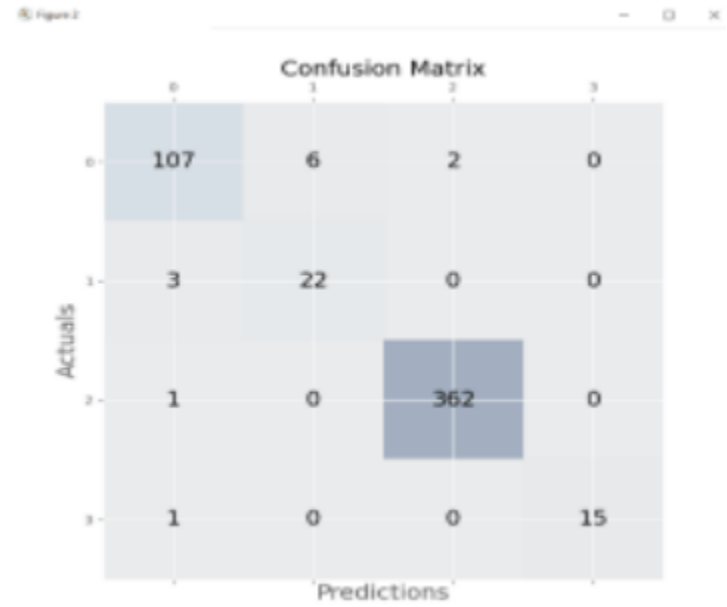
accuracy.



When we click on predict button it opens another window ,on the left side of the window we can view dataset which we used for performance analysis. Based on the dataset we must give values to right side to predict last column of the dataset.



When we give values and click on predict button we view prediction result on the right .



When we click on the confusion matrix we can view the confusion matrix of the best algorithm.

5. Tools and algorithms used to develop a solution:

- Jupyter Notebook: It is an open-source tool to support interactive data science and scientific computing across all programming languages.
- Python Libraries
    - **Matplotlib**: Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays. It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, and WxPythonotTkinter. It can be used in Python and IPython shells, Jupyter notebook, and web application servers also
    - **Sklearn**: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy, and Matplotlib.
    - **Tkinter**: Python has many options for developing Graphical User Interface. Of all methods, Tkinter is the most used method. Python with Tkinter will be very useful to create the GUI applications. Using Tkinter creating GUI is made easy. Tkinter calls will be translated into Tcl commands which are used in embedded interpreters and it made possible to use python and commands in one application. Tkinter provides various controls, such as labels, buttons, and text boxes used in a GUI. Controls are also, called widgets. Tkinter commonly comes bundled with Python, using Tk and is Python's standard GUI framework. It is open-source and available under the Python License.

- **Pandas**: The name is derived from the term “panel data”, an econometrics term for multidimensional structured data sets. It is a library for data manipulation and analysis. The library provides data structures and operations for manipulating numerical tables and time series. It is also known as the “Python Data Analysis Library”
- **Numpy**: NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object and tools for working with these arrays. This is a fundamental package for scientific computing with Python, adding support for large, multi-dimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on these arrays.

6.RESULTS W/ SOLUTION:

We can apply any dataset to our model, Now let us see few datasets used during the testing

Types of datasets:

We have divided the datasets based on the size and the type of contents.

Based on content

- Mixed (with text and numbers)
- Numeric (with only numbers)
- Nominal (with only text)

Based on size

- Small (<500)
- Medium (<10000)
- Large (>10000)

Based on the types we can form 9 different types of datasets.

6.1 Numeric

Diabetes Prediction- Numeric - Medium-786 Records

Diabetes dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

• Pregnancies: Number of times pregnancies.

• Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test.

• Blood Pressure: Diastolic blood pressure (mm Hg).

• Skin Thickness: Triceps skin fold thickness (mm).

• Insulin: 2-Hour serum insulin (mu U/ml).

• BMI: Body mass index (weight in kg/(height in m)^2).

• DiabetesPedigreeFunction: Diabetes pedigree function.

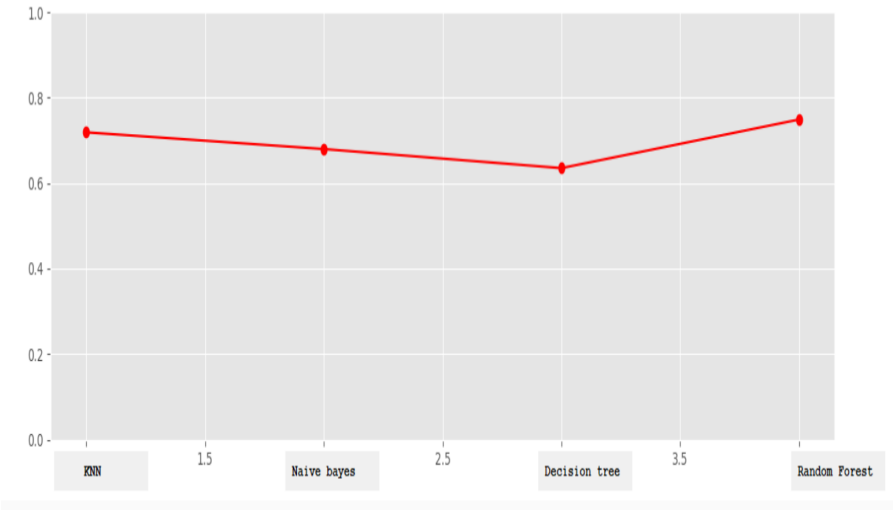
• Age: Age (years).

• Outcome: Class variable (0 or 1).

Dataset	No of attributes	No of Instances
PIDD	8	768

PIDD- Pima Indian Diabetes Dataset

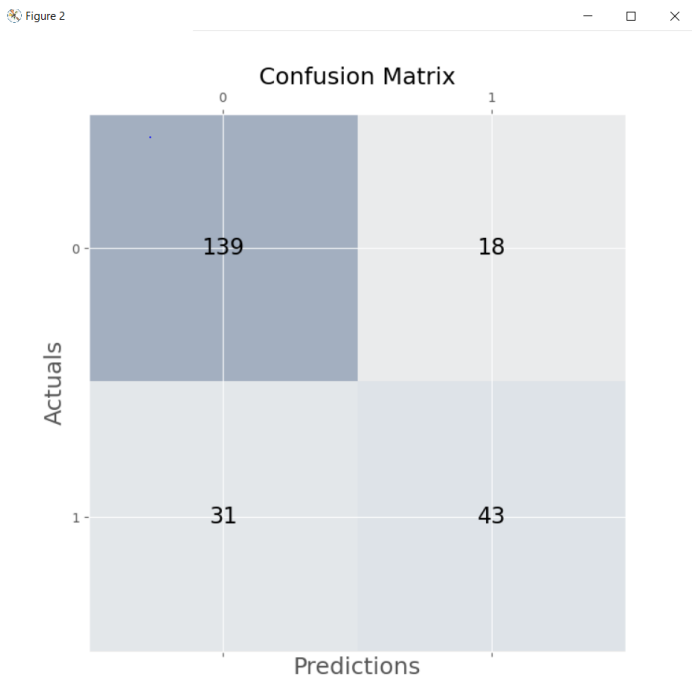
NO	NAME OF ATTRIBUTES ATTRIBUTES	TYPE
1	Number of pregnant	Numeric
2	Glucose	Numeric
3	Blood Pressure(mm HG)	Numeric
4	Skin Thickness	Numeric
5	Insulin	Numeric
6	Body Mass Index(BMI)	Numeric
7	Diabetes Pedigree function	Numeric
8	Age(years)	Numeric



Accuracies

- KNN : 0.718
- Naive Bayes : 0.679
- Decision Tree : 0.635
- Random Forest : 0.748

Random forest has the best accuracy



Confusion matrx for the random forest algorithm

True positive:43

True negative :139

False positive: 18

False negative: 31

Numeric - Other Datasets

- Prediction of sepsis - 110k Records - Large
  - Accuracies:
    - KNN : 0.9179
    - Naive Bayes : 0.9261
    - Decision Trees : 0.9266
    - Random Forest: 0.9263
- Heart - 300 Records - Small
  - Accuracies:
    - KNN : 0.806
    - Naive Bayes : 0.778
    - Decision Trees : 0.821
    - Random Forest: 0.816

For all the sizes we found random forest has improved the accuracies.

6.2 Nominal

Nominal - Car Evaluation - Medium -1800 Records

Car Evaluation is a UCI Machine Repository which has been derived from a simple hierarchical model where the database can be used for testing constructive induction and structure discovery methods. The inputs for the data set are lowercase. Apart from the basic idea, it has three moderate ideas which are PRICE, TECH, and COMFORT where each idea is in the first idea with its lower relatives.

The data set contains cases that have auxiliary data evacuated which is specifically related to the six input attributes: buying, input, maintenance, doors, persons, luggage, safety. Table1 shows how a car evaluation data set will evaluate the concept structure.

Car	Car acceptability
Price	Overall price
buying	Buying price
maint	Price of the maintenance
Tech	Technical characteristics
Comfort	Comfort
Doors	Number of doors
Persons	Capacity in terms of persons to carry
Lugg boot	The size of the luggage boot
Safety	Estimated safety of the car

Table CD model evaluation

The data for the data set is the data set characteristics are Multivariate, Number of Instances are 1728, the attribute characteristics are categorical, Number of instances are 6 , the associated tasks are Classification and

there are none missing values. The attribute information can be known in Table2.

buying	v-high,high, med, low
maintenance	v-high, high, med, low
doors	2, 3, 4,5
persons	2,4,5
luggage boot	small, med, med
safety	low, med, high

Table: CD Descriptor

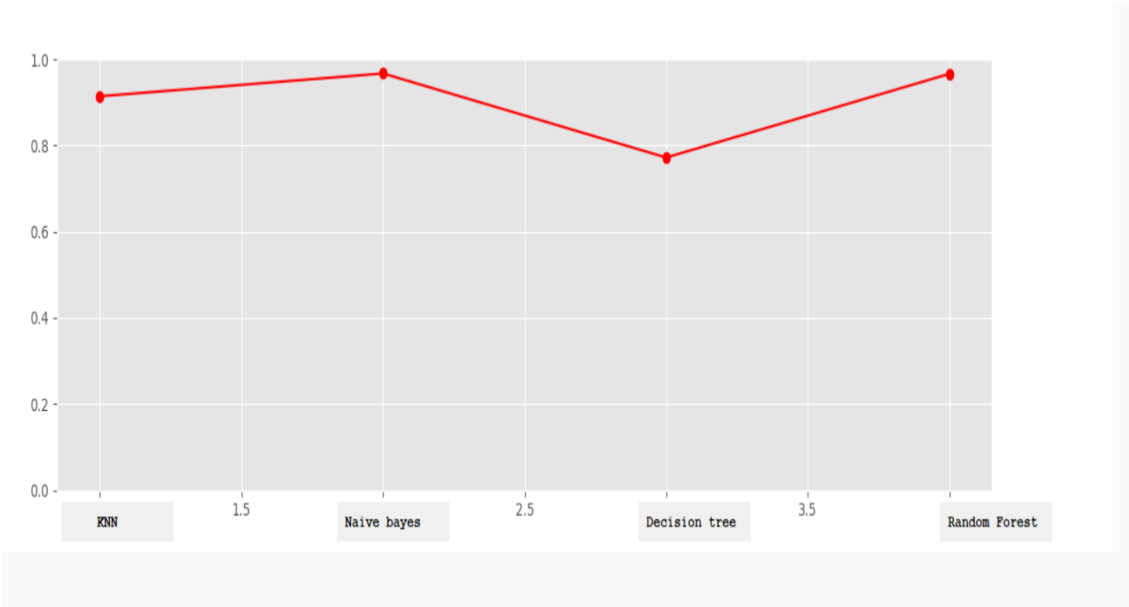
NO	NAME OF ATTRIBUTES	TYPE
1	Buying	Nominal
2	Maintenance	Nominal
3	Doors	Nominal
4	Persons	Nominal
5	Luggage Boot	Nominal
6	Safety	Nominal

Table : CD Attributes

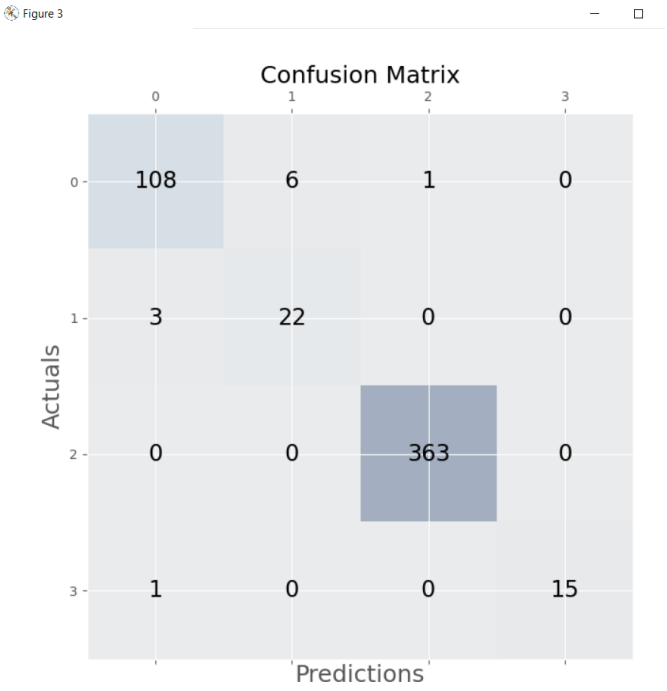
Accuracies:

- KNN : 0.913
- Naive Bayes : 0.966
- Decision Tree : 0.771
- Random Forest : 0.966

Random Forest has the best accuracy



Confusion matrix for Random Forest



For the other Nominal datasets

- Perform- 500 Records - small
  - Accuracies:
    - KNN : 0.634
    - Naive Bayes : 0.669
    - Decision Trees : 0.675
    - Random Forest: 0.722
- Animals - 100k Records - Large
  - Accuracies:
    - KNN : 0.971
    - Naive Bayes : 0.969
    - Decision Trees : 0.785
    - Random Forest : 0.972

For all the nominal cases we found Random Forest algorithm produces better accuracy.

6.3 Mixed

German Credit Data- Mixed- small-500 Records

German credit dataset is originally from the Kaggle website. The objective is to predict the purpose of the customer.

• Age: age (years)

• Sex: Gender (male or female)

• Job: number of jobs (1, 2,3 …)

• Housing: house details (own, rent, free ….).

• Saving account: status of saving account (little, moderate, quite rich …).

• Checking account: status of checking account (little, moderate).

• Duration: time.

• Credit amount: money

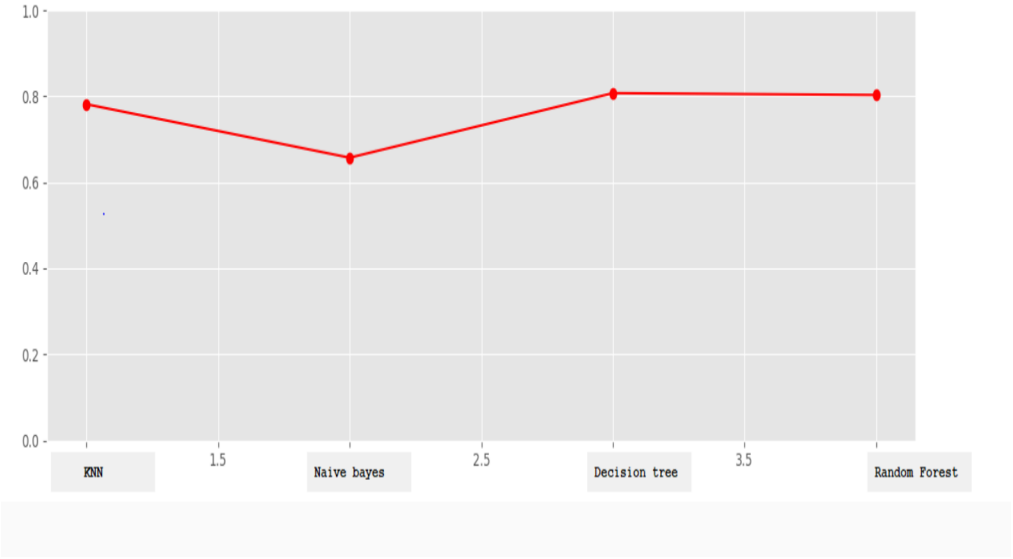
• Purpose (essential, non-essential)

Dataset	No of attributes	No of Instances
GCD	9	500

Table GCDDescriptor

NO	NAME OF ATTRIBUTES	TYPE
1	Age	Numeric
2	Sex	Nominal
3	Job	Numeric
4	Housing	Nominal
5	Saving account	Nominal
6	Checking account	Nominal
7	Credit amount	Numeric
8	Duration	Numeric
9	Purpose	Nominal

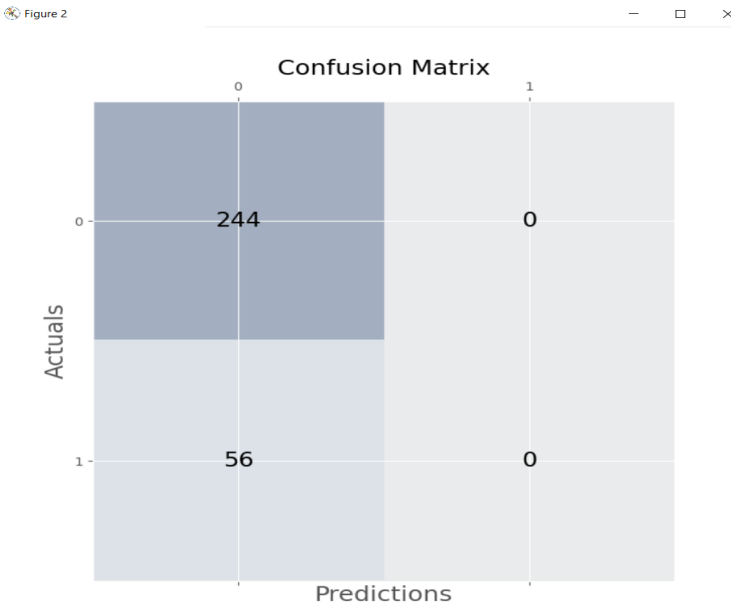
Table GCD Attributes



Accuracies:

- KNN : 0.781
- Naive Bayes : 0.657
- Decision Tree : 0.807
- Random Forest : 0.802

We found decision tree as the best accuracy algorithm, with random forest almost close to it.



Confusion matrix for the decision tree algorithm

True positives: 244

True negatives: 0

False positives: 56

False negatives: 0

Mixed Other datasets

- Power System - 11k Records - Large
  - Accuracies:
    - KNN : 0.928
    - Naive Bayes : 0.977
    - Decision Trees : 0.977
    - Random Forest : 0.989
- Abalone - 4k Records - Medium
  - Accuracies:
    - KNN : 0.523
    - Naive Bayes : 0.479
    - Decision Trees : 0.365
    - Random Forest : 0.540

For all types of Mixed datasets we found the random forest has the best accuracy.

7. Conclusion

	Small	Medium	Large
Numeric	Decision Tree/ Random Forest	Random Forest	Decision Tree/ Random Forest
Mixed	Decision Tree/ Random Forest	Random Forest	Naive Bayes/ Decision Tree
Nominal	Random Forest	Random Forest/ Decision Tree	Random Forest

- For all the types we found Random Forest provides the best accuracy.

- For large datasets Random forest takes a lot of resources and time, so for the huge datasets, it is better to opt for the alternative classification algorithm.

8.FUTURE WORK:

Our future work will consist of the selection of some real-world large data set and performing some suitable classification technique based on the nature and characteristics of the data and providing some important information about the data set. Other improvements can be adding more classification algorithms, considering more datasets, and making GUI more convenient for all the users.