1. INTRODUCTION

Data mining is the process to extract potentially valuable and relevant information from big amount of data sets. Usually, it includes a set of technique, such as, classification , clustering, association rule mining, anomaly detection, etc. Data mining techniques have been widely used to analyze data from different domains such as business, medical and transportation. The wide applicability of these data mining techniques proved it as a reliable and result oriented in all real world domains. Classification or prediction is the most commonly used techniques of data mining. Classification is a kind of supervised learning techniques that identifies the hid- den relationships between dependent and independent variables. Supervised learning techniques extract certain important features from the training data and then it uses 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.

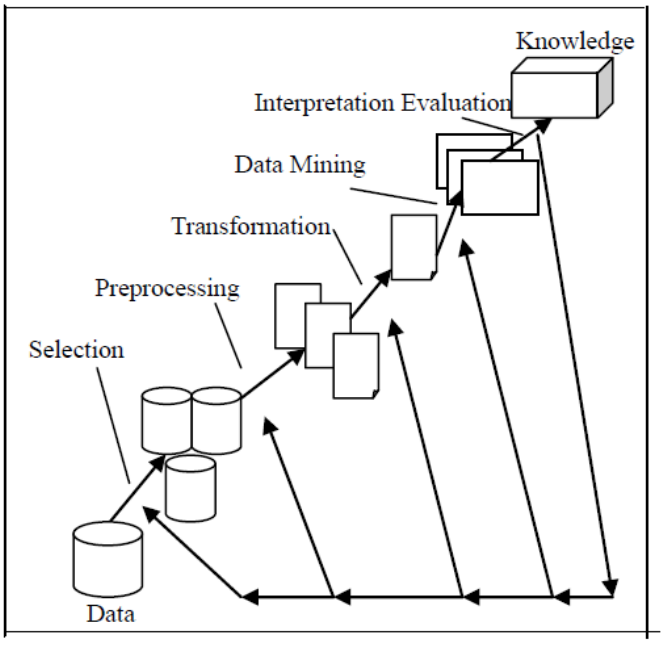

Fig 1.1 datamining process

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

1.1 Problem statement

To identify for which type of dataset which classification technique is more suitable. There are three types of datasets (numerical, nominal and mixed) and many types of classification algorithms. Wrong selection of classification algorithm will certainly lead to bad classification model and bad results. It is also very hard to write each and every time code to analyze the dataset and create prediction model for it. So we designed a GUI which handles all these issues.

1.2 Scope

- The scope of this project is to apply different type of datasets and come to a conclusion that for a particular type of dataset this algorithm is suitable ,It helps beginner to chose algorithm ,and we also want to decrease the work of a person by designing a GUI which analyses and creates a prediction model for any given dataset.

1.3 Objectives

- The objective is the extraction of examples and learning from huge measure of information, not the extraction of information itself. It likewise is a popular expression and is regularly connected to any type of extensive scale information or data handling (gathering, extraction, warehousing, investigation, and insights) and additionally any use of PC choice emotionally supportive network, including computerized reasoning, machine learning, and business knowledge. The real information mining errand is the programmed or self-loader examination of substantial amounts of information to concentrate already obscure fascinating examples, for example, gatherings of information records, uncommon records (oddity recognition) and conditions.

1.4 Proposed system

- The proposed performance analysis GUI is developed by using PYTHON and Tkinter package. We can upload datasets from our pc and check which algorithm is better for that dataset.Later we can create the prediction model for that dataset and apply it in real time.

2.LITERATURE REVIEW

2.1 Introducton to python

Python is a programming language, created by Guido van Rossum, and released in 1991. It is mainly used for

- Web development
- Software development
- Mathematics
- System Scripting

Python can be used in the following ways:

- Python can be used on a server to create web applications.
- Python can be used alongside software to create workflows.
- Python can connect to database systems. It can also read and modify files.
- Python can be used to handle big data and perform complex mathematics.
- Python can be used for rapid prototyping, or for production-ready software development.

Reasons to choose Python are as follows:

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
- Python has a simple syntax similar to the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
- Python can be treated in a procedural way, an object-orientated way or a functional way.
- Python has pre-built libraries.

Python syntax compared to other programming languages:

- Python was designed for readability, and has some similarities to the English language with influence from mathematics.
- Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.
- Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose.

2.2 Introduction to Numpy, Pandas and Sklearn

- NumPy is a library that allows you to efficiently load and work with large datasets and memory.It's free, open source, and widely used in real systems in Silicon Valley. It's the foundation on which many other machine learning libraries are built.
- Scikit-learn is a very popular machine learning library.Think of it as a Swiss army knife for machine learning.It provides easy-to-use implementations of many of the most popular machine learning algorithms.
- Pandas lets you represent your data as a virtual spreadsheet that you can control with code. It has many of the same features you find in Microsoft Excel for quickly editing your data and performing calculations

2.3 Introduction to Tkinter

Python offers multiple options for developing GUI (Graphical User Interface). Out of all the GUI methods, tkinter is most commonly used method. It is a standard Python interface to the Tk GUI toolkit shipped with Python. Python with tkinter outputs the fastest and easiest way to create the GUI applications. Creating a GUI using tkinter is an easy task.

To create a tkinter:

1. Importing the module – tkinter
2. Create the main window (container)
3. Add any number of widgets to the main window
4. Apply the event Trigger on the widgets.

2.4 Research done before

Many researchers have focused on developing new classification techniques, giving rise to the need to determine which technique to use in a given situation. According to the 'no free lunch' theory, no best technique exists for all situations (Pen, Wang, Kou, & Shi, 2011Peng , Y. , Wang, G. , Kou, G. , & Shi, Y.(2011). An empirical study of classification algorithm Evaluationforfinancial risk prediction.*Applied Soft Computing* . It is imperative to find an appropriate technique; therefore, the main Aim of our research is to define which technique to use in a specific situation. Existing approaches use the trial-and-error method, and there is a lack of systematic research concerning which classification technique should be used on a particular data-set, based on the characteristics of the data-set. Data-set characteristics are key in determining the classification algorithm's performance (Kwon & Sim, 2013Kwon, O. , & Sim, J. M. (2013). Bernado-Mansilla, E. , & Ho, T. K. (2005). Domain of competence of XCS classifier system in complexity measurement space. *IEEE Transactions onEvolutionary Comp utation* Chen & Shyu, 2011Chen, C. , & Shyu, M. L. (2011, August).

Clustering-based binary-class classification for imbalanced data sets. A comparative assessment of classification methods.*Decision Support Systems* , Kwon & Sim, 2013K won, O. , & Sim, J. M. (2013). Effects of data set features on the performances of classification algo rithms.*Expert Systems with Applications* Smith, Woo, Ciesielski, & Ibrahim, 2002Smith,

K. A. , Woo, F. , Ciesielski, V. ,& Ibrahim,R.(2002).Matching data mining algorithmsuitability to data characteristic using a self-organizingmap. In *Hybrid information systems* (pp. 169– 179) . Physica. Automatic recommendation of classification algorithms based on data set characteristics. *Pattern Recognition* ,Kiang (2003Kiang, M. Y. (2003)

A comparative assessment of classification methods.*Decision Support Systems* , points out that data characteristics affect the performance of classification methods. IEEE International Conference on(pp. 384–389). Las Vegas, NV: IEEE claim that the correlations among data-set characteristics affect algorithm performance,but few studies have analysed the influence of dataset characteristics on classification algorithm performance

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 over fitting.
- Classifies by rectangular partitioning.
- Does not easily handle nonnumeric data.
- Can be quite large- pruning is necessary.

3.3. Naive Bayes (NB)

Naive Bayes is a classification technique with a notion which defines all features are independent and unrelates 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 which employs the Bayes theorem. Naive Bayes classifiers assume attributes have independent distributions. It is considered to be fast and space efficient. It also provides 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 out perform even highly sophisticated classification methods.
- Good result 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 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 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 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 time-consuming than other algorithms.

3.5 DATASETS USED

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

3.5.1 Diabetes dataset (Numerical)

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

Table 3.1 PIDD Descriptor
PIDD- Pima Indian Diabetes Dataset

NO	NAME OF 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

Table 3.2 PIDD Attributes

3.5.2 Car Dataset (Nominal)

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, COMFORT where each idea is in the first idea with its lower relatives.

The data set contains cases which has 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 be evaluates 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 3.3 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 3.4 CDdescriptor

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 3.5 CD Attributes

We give these datasets to PAC GUI to analyze and provide best classification algorithm and we implement by by giving new values in prediction part

3.5.3 German credit dataset (Mixed)

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 (car, radio , business, tv ….)

Dataset	No of attributes	No of Instances
GCD	9	1001

Table 3.6 GCD Descriptor

GCD- German credit data

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 3.7 GCD Attributes

3.6 Accuracy Measures

The proposed methodology is evaluated on Datasets namely (PIDD, Car..), which are taken from UCI Repository. All models present some value of accuracy upon working on dataset. Each model shows different percentage of accuracy at the end and they are compared. ‘The model which presents highest accuracy is considered the best model’. Naive Bayes, K-Nearest Neighbor, Decision Tree [CART] , Random forest are used in this research work.

4. SYSTEM DESIGN AND IMPLEMENTATION

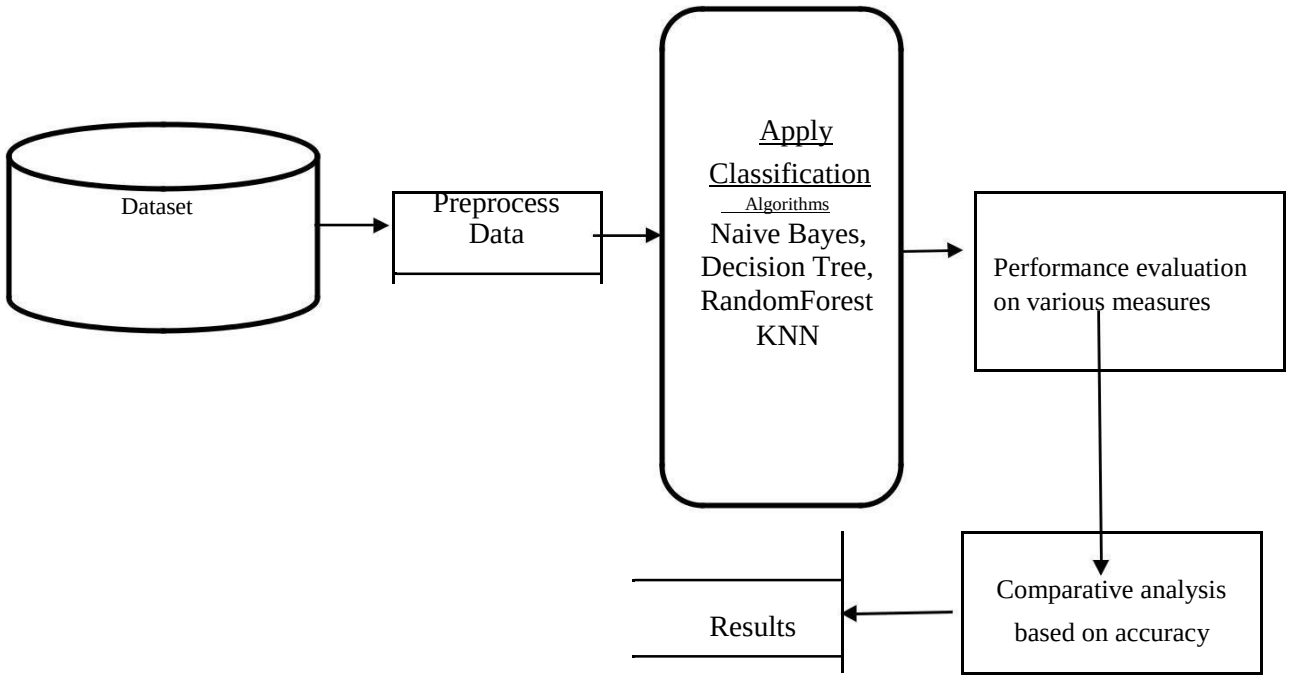

Fig:4.1: System Architecture

System design is the process of defining system architecture, modules and interfaces for the proposed system to satisfy specified requirements. The System architecture is shown in the Fig 4.1. The Data set is given as input to the system. The predictor helps in prediction of last attribute based on the attributes provided. The complete analysis of the attributes is done and then the evaluation is shown using graphs and textboxes . All of this work is done based on the code written backend in python and are represented in front end. A clear idea is obtained from graph and textboxes. Accuracy is predicted according to the models which are also coded backend in python. The backend is connected with Front end using Tkinter.

4.1 Data flow diagram

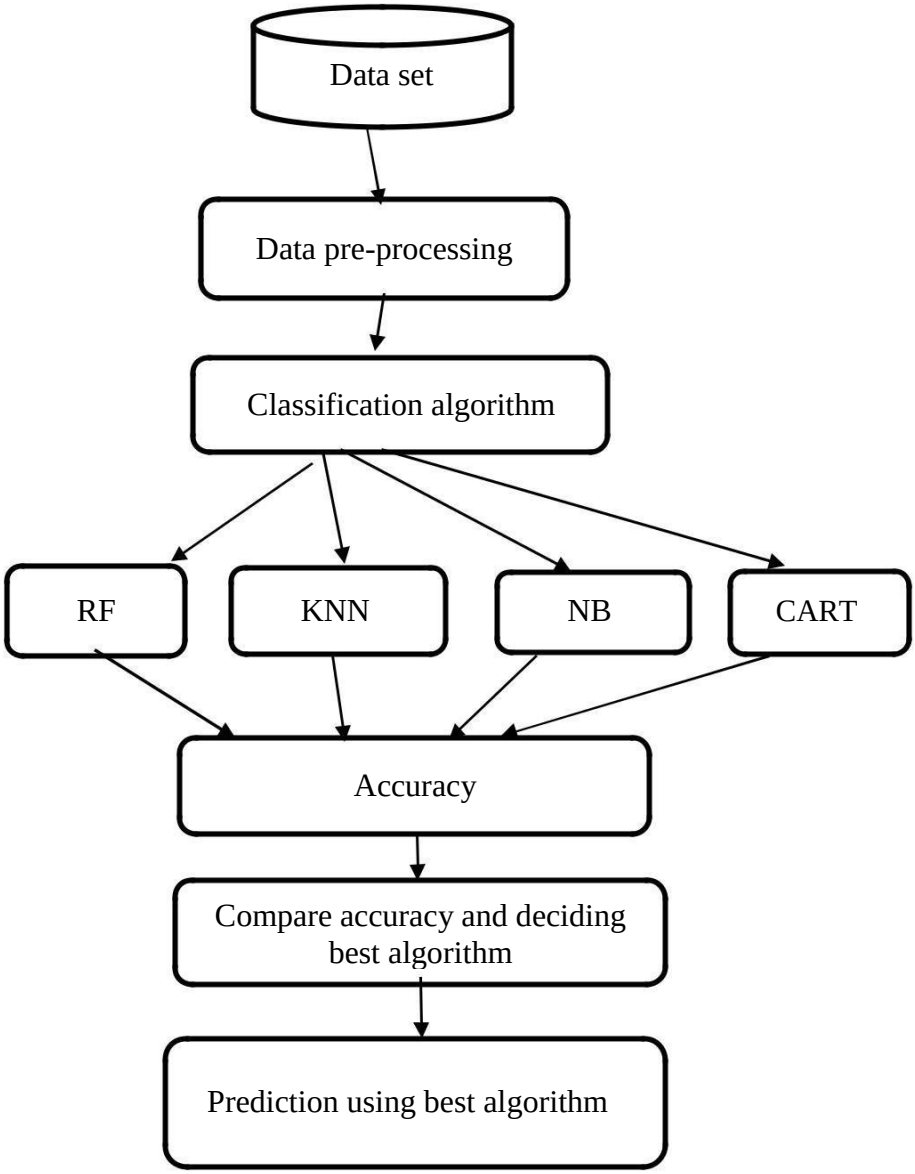

Fig 4.2 Dataflow diagram

In the above flow chart diagram the flow of operation starts by the selection of the dataset followed by the Model selection based on our requirement. Then we can check accuracy for different algorithms and know which is the best algorithm. Using that best algorithm we can do prediction part.

4.2 Class diagram

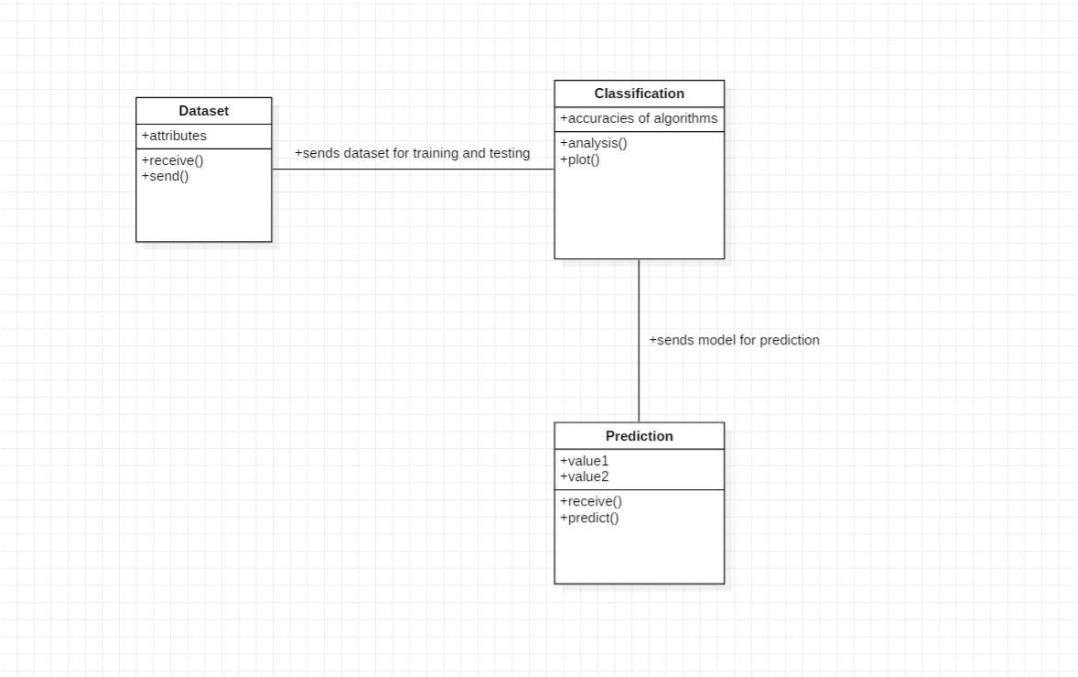

Fig 4.3 class diagram

The above figure depicts the class diagram containing three classes (Dataset , classification , prediction) and their associations.

4.3 Package diagram

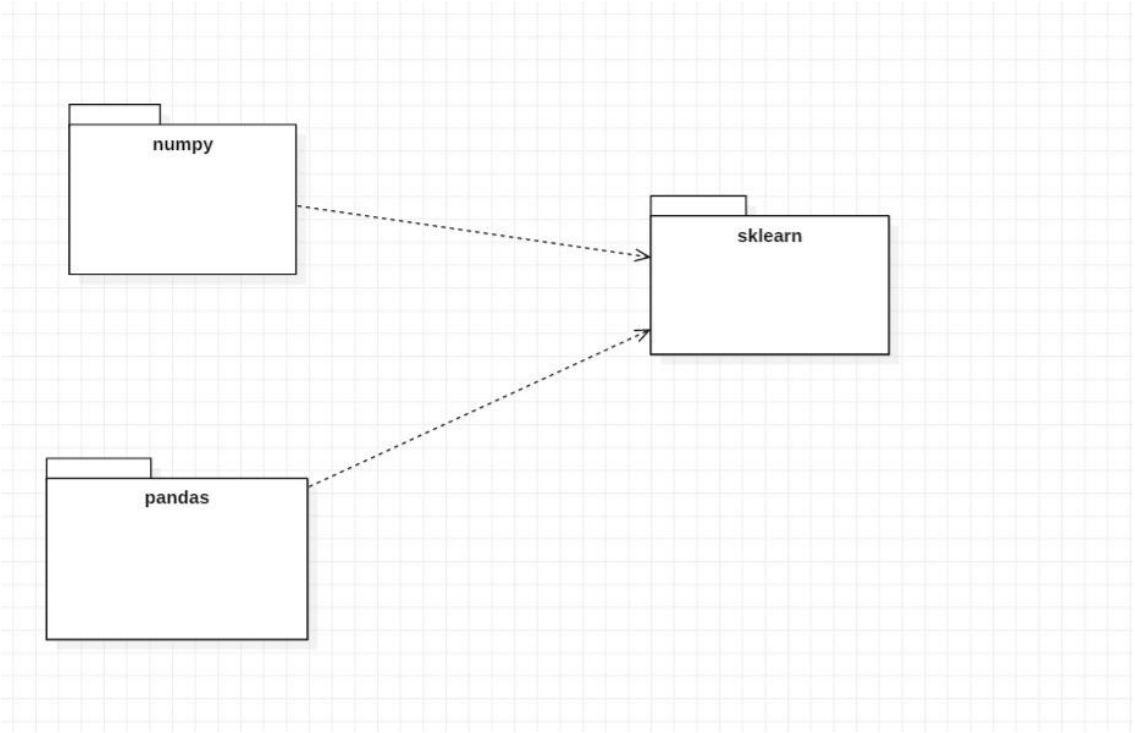

Fig 4.4 package diagram

The above figure depicts the package diagram containing three packages (numpy Sklearn , pandas) and their dependencies.

4.4 Hardware and Software requirements

- **Hardware:** Computer, Laptop.
- **Software:**
  1. Operating system: All Operating Systems
  2. Dataset: any Dataset.
  3. Front end: Tkinter .
  4. Back end: PYTHON.

- Output can be viewed in Computers, Laptops.

4.5 Implementation

- Open the command prompt and execute our python file myapp.py, after the execution GUI Prompts on the screen .
- In the GUI we can upload the datasets and see which classification Algorithm is better for that particular dataset.
- We can implement this model in real time by Providing few values and predicting the missing one
- In program we implement Sklearn , numpy , pandas and tkinter packages.
- Sklearn , numpy , pandas and tkinter are part of python and they are installed automatically during the Installation of python.
- If they are not installed, we can manually install them by using Install -pip- command in command prompt.

  Install -pip- numpy

  Install -pip- sklearn

  Install -pip- tkinter

  Install -pip- pandas

5. RESULT

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.

Fig: 5.1 File uploading

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.

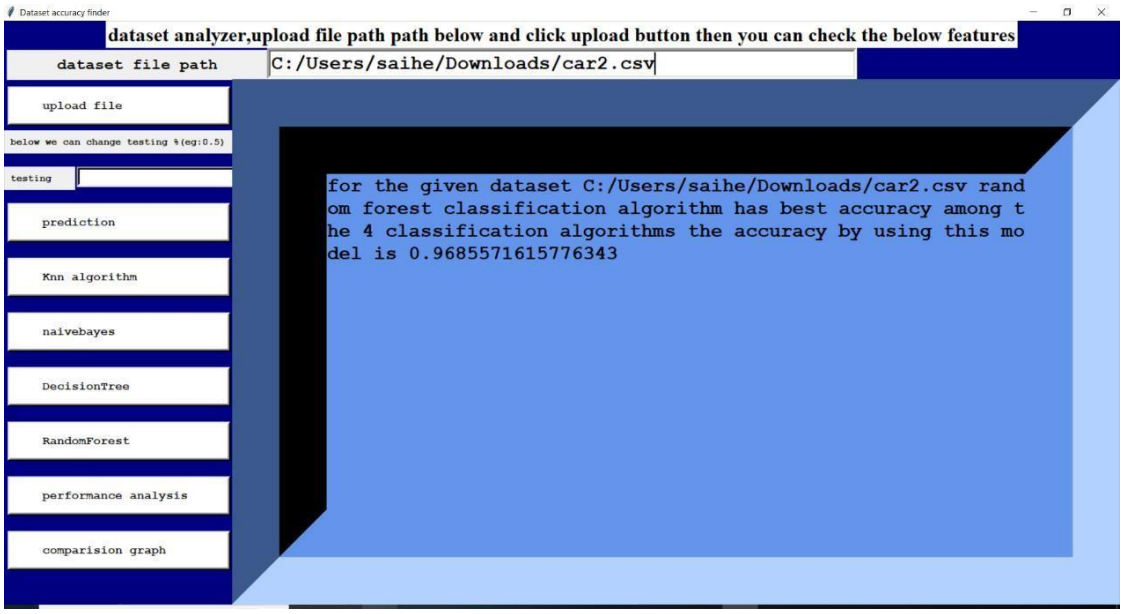

Fig: 5.2 performance analysis.

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

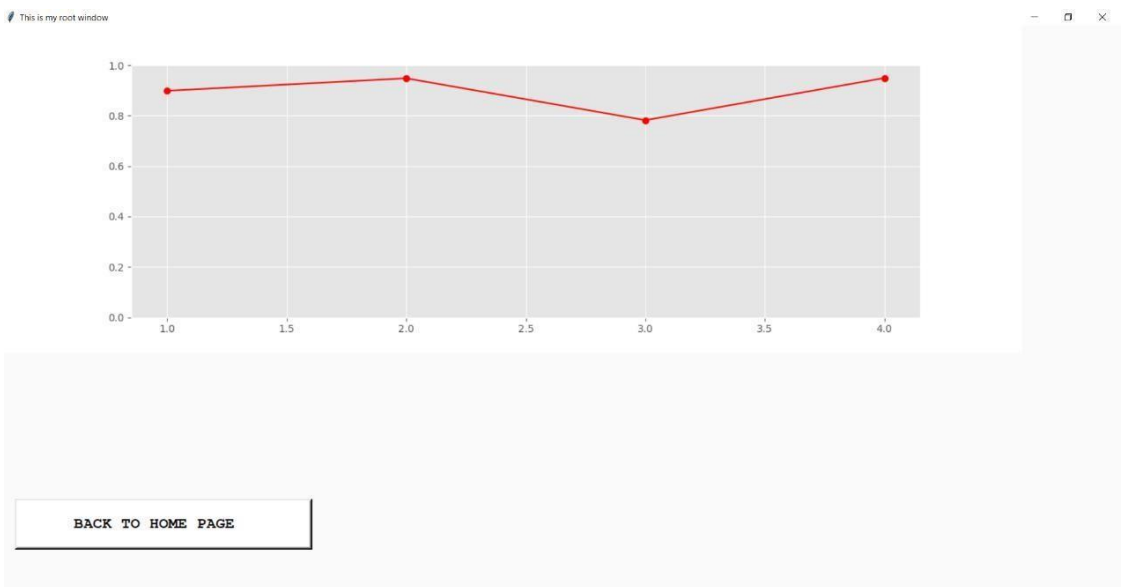

Fig: 5.3 comparing accuracy graphs

.

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

s

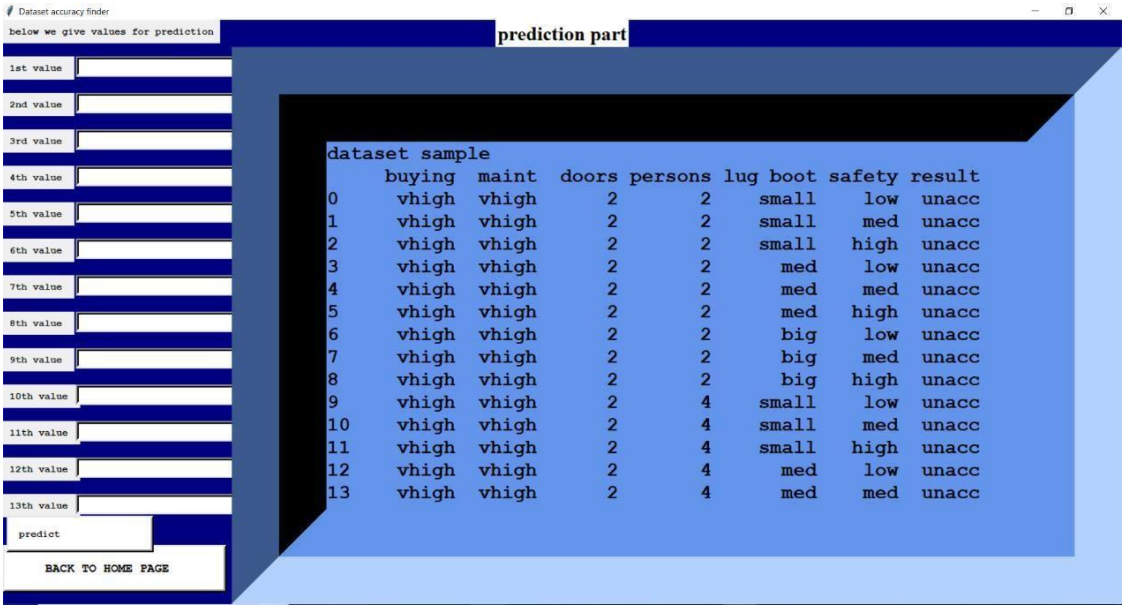

Fig: 5.4 prediction part.

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 dtaset we must give values to right side to predict last column of the dataset.

Fig: 5.5 Result of prediction.

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

6. CONCLUSION AND FUTURE WORK

In this study, we have used four popular classification techniques NB, decision tree , Knn and Random forest on different data sets. The purpose of this study is to check the performance of different classification algorithms on different data sets. In order to achieve this, we have used three different data sets i.e., numeric data (diabetes data set), nominal data (car data set) and mix attribute data (German credit card data set). Further, we performed all four classification techniques on these data sets and compared the result. The results illustrate that the performance of Random forest classifier on numeric and nominal data set is superior to the other three techniques. The result also revealed that on mixed attribute data set, the decision tree classification technique and Random Forest almost have same accuracy and they outperformed the other two techniques. Therefore, this study simply revealed the important information that the different classification algorithms have different accuracy and performance on different kind of data sets. Our study certainly helps the beginners to choose the best classification algorithm in order to apply different kind of data set. Our future work will consist of selection of some real world large data set and perform some suitable classification technique based on the nature and characteristics of the data and providing some important information out of the data set.

7. REFERENCES

- NBian. Available from: http://www.saedsayad.com/naive_ bayesian.htm
- Elkan C. Niave Baysian learning. Adapted from Technical Report No. CS97-557, Department of Computer Science and Engineering, University of California, San Diego; 1997. p. 1–11.
- Shahrukh T, Prashasti K. A survey on decision tree based approaches in data mining. International Journal of Advanced Research in Computer Science and Software Engineering. 2015; 5(4):25–71.
- Vijayarani S, Muthulakshmi M. Comparative analysis of bayes and lazy classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering. 2013; 2(8):3118–24.
- Durairaj M, Deepika R. Comparative analysis of classifi- cation algorithms for the prediction of leukemia cancer. International Journal of Advanced Research in Computer Science and Software Engineering. 2015; 5(8):787–91.
- k-fold cross validation. Available from: http://www.csie. ntu.edu.tw/~b92109/course/Machine %20Learning/Cross- Validation.pdf
- Rupali B, Sonia V. Implementation of ID3 algorithm. International Journal of Advanced Research in Computer Science and Software Engineering. 2013; 3(6):845–51.
- Measuringsearcheffectiveness.Availablefrom:https:// www.creighton.edu/fileadmin/user/HSL/ docs / ref/ Searching_-_Recall_Precision.pdf
- Baeza-Yates B, Ricardo R, Ribeiro-Neto RN, Berthier B. Modern information retrieval. New York, NY: ACM Press, Addison-Wesley; 1999. p. 1–103. ISBN: 0-201-39829-X.
- Zolfagharifar SA, Karamizadeh FV. Developing a hybrid intelligent classifier by using evolutionary learning (Genetic Algorithm and Decision Tree). Indian Journal of Science and Technology. 2016 May; 9(20):1–8.

- Kim M, Kim CJ. Factors associated with decision to partic- ipate in physical activity by people with spinal cord injury: An analysis using decision tree. Indian Journal of Science and Technology. 2016 Jul; 9(26):1–7.
- Azad C, Jha VK. Data mining based hybrid intrusion detec- tion system. Indian Journal of Science and Technology. 2014 Jan; 7(6):1–9.
- Rajalakshmi V, Mala GSA. Anonymization by data relo- cation using sub-clustering for privacy preserving data mining. Indian Journal of Science and Technology. 2014 Jan; 7(7):1–6.

APPENDIX

#importing packages

```
from tkinter import*
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import matplotlib.animation as animation
from matplotlib import style
from matplotlib.backends.backend_tkagg import FigureCanvasTkAgg
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import BernoulliNB
from sklearn.ensemble import RandomForestClassifier
```

 # creating window in Gui

```
me=Tk()
me.geometry("700x700")
me.title("Dataset accuracy finder")
melabel = Label(me,text="dataset analyzer,upload file path path below and click
upload button then you
can check the below features",bg='White',font=("Times",20,'bold'))
melabel.pack(side=TOP)
me.config(background='navy')
```

```
 #applying classification algorithms
  def dataset():
  global b
  global c
  global k
  global l
  c=b.get()

if(metext1.get()==''):
k=0.3
    else:
      k=float(metext1.get())

    global f
    global z
    global a1
    global a2
    global a3
    global a4

    dataframe = pd.read_csv(c)
    X = dataframe.iloc[:,:-1 ]
    y = dataframe.iloc[:, -1]
  #Encoding Categorical Data to Numerical Data

    for i in X.columns:
      if(X[i].dtype == 'object'):
          labelencoder_X = LabelEncoder()
          X[i] = labelencoder_X.fit_transform(X[i].astype(str))
    X.info()
    y.head()
  #splitting the data set

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = k, random_state = 0)
```

```
#knn classification
#from sklearn.preprocessing import StandardScaler
  sc = StandardScaler()
  X_train = sc.fit_transform(X_train)
  X_test = sc.transform(X_test)
#from sklearn.neighbors import KNeighborsClassifier
  classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
  classifier.fit(X_train, y_train)
  y_pred = classifier.predict(X_test)
#from sklearn.model_selection import cross_val_score
  accuracies1=cross_val_score(estimator=classifier,X=X_train,y=y_train,cv=10)
  print(accuracies1.mean())
  a1=accuracies1.mean()
#Decision Tree classification
#from sklearn.tree import DecisionTreeClassifier
  classifier = DecisionTreeClassifier()
  classifier.fit(X_train, y_train)
  y_pred = classifier.predict(X_test)
#from sklearn.model_selection import cross_val_score
  accuracies2=cross_val_score(estimator=classifier,X=X_train,y=y_train,cv=10)
  print(accuracies2.mean())
  a2=accuracies2.mean()


#naivebayes Classification
# Importing the dataset
  dataframe = pd.read_csv(c)
  X = dataframe.iloc[:,:-1 ]
  #X.iloc[0]=['vhigh','vhigh',2,4,'big','high']

  y = dataframe.iloc[:, -1]
#Encoding Categorical Data to Numerical Data
#from sklearn.preprocessing import LabelEncoder
  for i in X.columns:
    if(X[i].dtype == 'object'):
        labelencoder_X = LabelEncoder()
        X[i] = labelencoder_X.fit_transform(X[i].astype(str))
```

```
 #f=X.transform([['vhigh','med',2,4,'big','high']])
# l=X.iloc[0]
 X.info()

 y.head()
#splitting the data set
#from sklearn.model_selection import train_test_split
 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = k, random_state = 0)

#from sklearn.naive_bayes import BernoulliNB
 classifier = BernoulliNB()
 classifier.fit(X_train,y_train)
 y_pred=classifier.predict(X_test)
#from sklearn.model_selection import cross_val_score
 accuracies3=cross_val_score(estimator=classifier,X=X_train,y=y_train,cv=10)
 print(accuracies3.mean())
 a3=accuracies3.mean()

#Random Forest classification
#from sklearn.ensemble import RandomForestClassifier
 classifier = RandomForestClassifier(n_estimators=100,criterion='entropy',random_state=0)
 classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
#from sklearn.model_selection import cross_val_score
accuracies4=cross_val_score(estimator=classifier,X=X_train,y=y_train,cv=10)
print(accuracies4.mean())
a4=accuracies4.mean()
```

#Comparing accuracies to know best algorithm

```
def clrbut4():
 if((a1==a2)and(a1==a3)and(a1==a4)and(a1==0.0)):


 sentence="file is not uploaded"
 t.delete('0.0','end')
 t.insert(0.0,sentence)
elif((a2>a1)and(a2>a3)and(a2>a4)):


        sentence="for the given dataset "+str(c)+" NAIVE BAYES classification algorithm has best accuracy among the 4 classification algorithms the accuracy by using this model is "+str(a2)
 t.delete('0.0','end')
 t.insert(0.0,sentence)
elif((a1>a2)and(a1>a3)and(a1>a4)):
# sentence="for the given dataset  knn classification algorithm has best accuracy among the 4 classification algorithms the accuracy by using this model is "+a1
 sentence="for the given dataset "+str(c)+" KNN classification algorithm has best accuracy among the 4 classification algorithms the accuracy by using this model is "+str(a1)
 t.delete('0.0','end')
 t.insert(0.0,sentence)
elif((a3>a2)and(a3>a1)and(a3>a4)):
# sentence="for the given dataset  decision classification algorithm has best accuracy among the 4 classification algorithms the accuracy by using this model is "+a3
 sentence="for the given dataset "+str(c)+" DECISION TREE classification algorithm has best accuracy among the 4 classification algorithms the accuracy by using this model is "+str(a3)
 t.delete('0.0','end')
 t.insert(0.0,sentence)
elif((a4>a2)and(a4>a3)and(a4>a1)):


 sentence="for the given dataset "+str(c)+" random forest classification algorithm has best accuracy among the 4 classification algorithms the accuracy by using this model is "+str(a4)
 t.delete('0.0','end')
 t.insert(0.0,sentence)
```

#**prediction part**

```
  y = dataframe.iloc[:, -1]
#Encoding Categorical Data to Numerical Data
#from sklearn.preprocessing import LabelEncoder
     for i in X.columns:
       if(X[i].dtype == 'object'):
        labelencoder_X = LabelEncoder()
        X[i] = labelencoder_X.fit_transform(X[i].astype(str))
  #f=X.transform([['vhigh','med',2,4,'big','high']])
     l=X.iloc[0]
     X.info()


     y.head()
#splitting the data set
#from sklearn.model_selection import train_test_split
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = k, random_state = 0)


     if((a1==a2)and(a1==a3)and(a1==a4)and(a1==0.0)):


      sentence="file is not uploaded"
      t1.delete('0.0','end')
      t1.insert(0.0,sentence)
     elif((a2>a1)and(a2>a3)and(a2>a4)):
      sc = StandardScaler()
      X_train = sc.fit_transform(X_train)
      X_test = sc.transform(X_test)
      classifier = DecisionTreeClassifier()
      classifier.fit(X_train, y_train)
      y_pred = classifier.predict(X_test)
      f=classifier.predict([l])


      sentence="the prediction is"+str(f)
      t1.delete('0.0','end')
      t1.insert(0.0,sentence)
     elif((a1>a2)and(a1>a3)and(a1>a4)):
 # sentence="for the given dataset  knn classification algorithm has best accuracy among the 4 classification
```

```
algorithms the accuracy by using this model is "+a1
      sc = StandardScaler()
      X_train = sc.fit_transform(X_train)
      X_test = sc.transform(X_test)
#from sklearn.neighbors import KNeighborsClassifier
      classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
      classifier.fit(X_train, y_train)
      y_pred = classifier.predict(X_test)
      f=classifier.predict([l])

      sentence="the prediction is"+str(f)
      t1.delete('0.0','end')
      t1.insert(0.0,sentence)
     elif((a3>a2)and(a3>a1)and(a3>a4)):
      classifier = BernoulliNB()
      classifier.fit(X_train,y_train)
      y_pred=classifier.predict(X_test)

      f=classifier.predict([l])

      sentence="the prediction is"+str(f)
      t1.delete('0.0','end')
      t1.insert(0.0,sentence)
     elif((a4>a2)and(a4>a3)and(a4>a1)):

      classifier = RandomForestClassifier(n_estimators=100,criterion='entropy',random_state=0)
      classifier.fit(X_train, y_train)
      y_pred = classifier.predict(X_test)
      f=classifier.predict([l])

      sentence="the prediction is"+str(f)
      t1.delete('0.0','end')
      t1.insert(0.0,sentence)
    butequal=Button(me,padx=10,pady=10,bd=4,bg='white',command=predict,text="predict
",font=("Courier New",10,'bold'))
    butequal.place(x=5,y=680)
```

MVSR Engineering College, Dept. of CSE

COURSE
NAME:PROJECT
COURSE CODE:CS 414

Course Objectives:

1. Learn to survey the necessary domains for problem identification
2. Learn the process of planning the complete lifecycle of a project
3. Understand how to map requirements from a user into software specification.
4. Learn to apply concepts of software engineering for design of the identified real world problem
5. Improve the coding capabilities by implementing the various modules of project
6. Comprehend the suitable documentation procedure for a technical project.

Course Outcomes:

Code No.	Student will be able to
CS414.1	Summarize the survey of the recent advancements to infer the problem statement with applications towards society.
CS414.2	Design a software based solution within the scope of project.
CS414.3	Implement using contemporary technologies and tools.
CS414.4	Test and deploy the applications on real world environments.
CS414.5	Demonstrate qualities necessary for working in a team.
CS414.6	Generate a suitable technical document for the project.

Table 1: Relevance of CO-PO/PSO

CO	PO addressed	PSO addressed	Cognitive levels
CS414.1	1,2,3,4,5,6,7,8	1	Analyze, Evaluate
CS414.2	1,2,3,4,5,6,7,8	1	Analyze
CS414.3	1,2,3,4,5,6,7,8	1,2	Apply, Evaluate, Analyze
CS414.4	1,2,3,4,5,6,8,10	1,2	Apply, Evaluate, Analyze
CS414.5	8,9,10,11,12	1	Apply, Evaluate, Analyze
CS414.6	9,10,11,12	1	Apply, Evaluate, Analyze

Table 2: CO-PO/PSO matrix

CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CS414.1	3	2	3	2	2	2	2	1	-	-	-	-	1	-
CS414.2	3	3	3	2	2	2	2	2	-	-	-	-	3	-
CS414.3	3	3	3	2	2	2	2	2	-	-	-	-	3	3
CS414.4	3	3	3	2	2	2	-	1	-	1	-	-	2	3
CS414.5	-	-	-	-	-	-	-	2	2	2	2	2	2	-
CS414.6	-	-	-	-	-	-	-	-	3	2	2	2	2	-
CS414	3	3	3	2	2	2	1	2	1	1	1	1	3	2

Table 3: Justification for CO-PO/PSO Level – through number of sessions

CO	No of sessions	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2
CS414.1	6	4	4	4	2	2	2	2	1	-	-	-	-	1	-
CS414.2	12	6	7	7	4	4	4	4	4	-	-	-	-	7	-
CS414.3	18	8	8	8	7	7	6	5	6	-	-	-	-	10	8
CS414.4	12	6	5	5	4	4	4	-	2	-	1	-	-	4	8
CS414.5	6	-	-	-	-	-	-	-	3	2	2	2	2	2	-
CS414.6	6	-	-	-	-	-	-	-	-	4	2	2	2	2	-
Total	60	24	24	24	17	17	16	11	16	6	5	4	4	26	16

Table 5: % of classroom session and Correlation level

% of classroom sessions addressing a particular PO/ PSO	Level
>=40% of classroom sessions addressing a particular PO	3 : substantial(high)
25 to 40% of classroom sessions addressing a particular PO	2 : moderate(medium)
5 to 25% of classroom sessions addressing a particular PO	1 : slight (low)
< 5% of classroom sessions addressing a particular PO	- : no correlation

Table 6: PO/PSO addressed by the Project

Project Name	Domain	In-house/ Industry	PO/PSO addressed	Internal Guide

Table 7: Rubrics Evaluation

PO/PSO	PO1,PO2,PO6,PO7				PO3	PO4,PO5, PSO1	PO4,PO5, PSO2	PO8	PO9				PO10			PO11	PO12
Rubrics	R1				R2	R3	R4	R5	R6				R7			R8	R9
Roll. No.	CI	CII	CIII	Total	CIV	CV	CVI	CVII	CVIII	CIX	CX	Total	CXI	CXII	Total	CXIII	CIV
	4	4	4	12	4	4	4	4	4	4	4	12	4	4	8	4	4

Rubrics for project

Focus Areas:

1. Problem Formulation (PO1,PO2, PO6, PO7)
2. Project Design (PO3)
3. Build (PO4,PO5, PSO1)
4. Test & Deploy (PO4, PO5, PSO2)
5. Ethical responsibility (PO8)
6. Team Skills (PO9)
7. Project Presentation (P10)
8. Project management (PO11)
9. Lifelong Learning (PO12)

Focus Areas	Criterion [c]	Exemplary 4	Satisfactory 3	Developing 2	Unsatisfactory 1
Problem Formulation (PO1,PO2, PO6, PO7)	I - Identify/Define Problem Ability to identify a suitable problem and define the project objectives.	Demonstrates a skillful ability to identify / articulate a problem and the objectives are well defined and prioritized.	Demonstrates ability to Identify / articulate a problem and All major objectives are identified.	Demonstrates some ability to identify / articulate a problem that is partially connected to the issues and most major objectives are identified but one or two minor ones are missing or priorities are not established.	Demonstrates minimal or no ability to identify / articulate a problem and many major objectives are not identified.
	II - Collection of Background Information: Ability to gather background Information (existing knowledge, research, and/or indications of the problem)	Collects sufficient relevant background information from appropriate sources, and is able to identify pertinant/critical information;	Collects sufficient relevant background information from appropriate sources;	Collects some relevant background information from appropriate Sources.	Minimal or no ability to collect relevant background information
	III- Define scope of the problem Ability to identify problem scope suitable to the degree considering the impact on society and environment	Demonstrates a skillful ability to define the scope of problem accurately mentioning the relevant fields of engineering precisely. Considers, explains and evaluates the impact of engineering interventions on society and environment.	Demonstrates ability to define problem scope mentioning the relevant fields of engineering broadly. Considers and explains the impact of engineering interventions on society and environment	Demonstrates some ability to define problem scope mentioning some of the relevant fields. Some consideration of the impact of engineering interventions on society and environment.	Demonstrates minimal or no ability to define problem scope and fails to mention relevant fields of engineering. Minimal or no consideration of the impact of engineering interventions on society and environment
Project Design (PO3)	IV- Understanding the Design Process and Problem Solving: Ability to explain the design process including the importance of needs, specifications, concept generation and to develop an approach to solve a problem.	Demonstrates a comprehensive ability to understand and explain a design process. Considers multiple approaches to solving a problem, and can articulate reason for choosing solution	Demonstrates an ability to understand and explain a design process. Considers multiple approaches to solving a problem, which is justified and considers consequences.	Demonstrates some ability to understand and explain a design process. Considers a few approaches to solving a problem; doesn't always consider consequences.	Demonstrates minimal or no ability to understand and explain a design process. Considers a single approach to solving a problem. Does not consider consequences.

Build (PO4,PO5, PSO1)	V- Implementing Design Strategy: Ability to execute a solution taking into consideration design requirements using appropriate tool (software/hardware);	Demonstrates a skillful ability to execute a solution taking into consideration all design requirements using the most relevant tool.	Demonstrates an ability to execute a solution taking into consideration design requirements using relevant tool.	Demonstrates some ability to execute a solution but not using most relevant tool.	Demonstrates minimal or no ability to execute a solution. Solution does not directly attend to the problem.
Test & Deploy (PO4, PO5, PSO2)	VI- Evaluating Final Design: To evaluate/confirm the functioning of the final design. To deploy the project on the target environment	Demonstrates a skillful ability to evaluate/confirm the functioning of the final design skillfully, with deliberation for further Improvement after deployment.	Demonstrates an ability to evaluate/confirm the functioning of the final design. The evaluation is complete and has sufficient depth.	Ability to evaluate/confirm the functioning of the final design, but the evaluation lacks depth and/or is incomplete.	Demonstrates minimal or no ability to evaluate/confirm the functioning of the final design.
Ethical responsibility (PO8)	VII - Proper Use of Others' Work: Ability to recognize, understand and apply proper ethical use of intellectual property, copyrighted materials, and research.	Always recognizes and applies proper ethical use of intellectual property, copyrighted materials, and others' research.	Recognizes and applies proper ethical use of intellectual property, copyrighted materials, and others' research.	Some recognition and application of proper ethical use of intellectual property, copyrighted materials, and others' research.	Minimal or no recognition and/or application of proper ethical use of intellectual property, Copyrighted materials, or others' research.
Team Skills (PO9)	VIII - Individual Work Contributions and Time Management: Ability to carry out individual Responsibilities and manage time (estimate, prioritize, establish deadlines/ milestones, follow timeline, plan for contingencies, adapt to change).	Designated jobs are accomplished by deadline; completed work is carefully and meticulously prepared and meets all requirements.	Designated jobs are accomplished by deadline; completed work meets requirements.	Designated jobs are accomplished by deadline; completed work meets most requirements.	Some Designated jobs are accomplished by deadline; completed work meets some requirements.
	IX - Leadership Skills: Ability to lead a team. (i) Mentors and accepts mentoring from others. (ii) Demonstrates capacity for initiative while respecting others' roles. (iii) Facilitates others' involvment. (iv) Evaluates team Effectiveness and plans for improvements	Exemplifies leadership skills.	Demonstrates leadership skills.	Demonstrates some leadership skills at times.	Demonstrates minimal or no Leadership skills.
	X - Working with Others: Ability to listen to, collaborate with, and champion the efforts of others.	Skillfully listens to, collaborates with, and champions the efforts of others.	Listens to, collaborates with, and champions the efforts of others.	Sometimes listens to, collaborates with, and champions others' efforts.	Rarely listens to, collaborates with, or champions others' efforts.

Project Presentation (P10)	XI - Technical Writing Skills Ability to communicate the main idea with clarity. Ability to use illustrations properly to support ideas (citations, position on page etc)	Main idea is clearly and precisely stated. Materials are seamlessly arranged in a logical sequence Illustrations are skillfully used to support ideas	Main idea is understandable. Material moves logically forward, Illustrations are properly used to support ideas	Main idea is somewhat Understandable. Material has some logical order and is somewhat coherent or easy to follow. Illustrations are for the most part properly used to support ideas	Main idea is difficult to understand. Material has little logical order, and is often unclear, incoherent. Illustrations are used, but minimally support ideas. (not properly cited etc)
	XII - Communication Skills for Oral Reports Ability to present strong key ideas and supporting details with clarity and concision. Maintain contact with audience, and ability to complete in the allotted time	Presentation logically and skillfully structured. Key ideas are compelling, and articulated with exceptional clarity and concision. Introduction, supporting details and summary are clearly evident and memorable, and ascertain the credibility of the speaker Presentation fits perfectly within time constraint.	Presentation has clear structure and is easy to follow. Key ideas are clearly and concisely articulated, and are interesting. There is sufficient detail to ascertain speaker's authority, and presentation includes an introduction and summary. Presentation fits within time constraint, though presenter might have to subtly rush or slow down.	Presentation has some structure. Key ideas generally identifiable, although not very remarkable. Introduction, supporting details and/or summary may be too broad, too detailed or missing. Credibility of the speaker may be questionable at times. Presentation does not quite fit within time constraint; presenter has to rush or slow down at end	Presentation rambles. Not organized; key ideas are difficult to identify, and are unremarkable. No clear introduction, supporting details and summary. Speaker has no credibility. Presentation is unsuitablably short or unreasonably long.
Project management (PO11)	XIII - Monitoring and Controlling the Project	Monitors timelines and progress toward project goals on a daily basis. Provides accurate, complete reports of project progress.	Monitors timelines and progress toward project goals most of the time. Provides relatively accurate, complete reports of project progress with only minor errors or omissions	Seldom monitors timelines and progress toward project goals. Provides relatively accurate, yet clearly incomplete, reports of project progress	Does not monitor timelines and progress toward project goals. Provides inaccurate, incomplete reports of project progress
Lifelong Learning (PO12)	XI V - Extend Scope of Work: Ability to extend the project through implementation in other study areas	Demonstrates a skillful ability to explore a subject/topic thoroughly, discusses the road map to extend the project in other areas.	Demonstrates an ability to explore a subject/topic, and shows possible areas in which project can be extended	Demonstrates some ability to explore a subject/topic, providing some knowledge of areas in which project can be extended	Demonstrates minimal or no ability to explore a subject/topic, and does not discuss future work clearly mentioning other areas