2021 2nd International Conference for Emerging Technology (INCET)
Belgaum, India. May 21-23, 2021

Predict the Performance Analysis of Supervised Learning Techniques Using Heart Disease Database

Dr S Govinda Rao
Professor in CSE
Gokaraju Rangaraju Institute of Engineering and Technology (GRIET),
Hyderabad, India
govindsampathirao@gmail.com

Dr P Chandrasekhar Reddy
Professor in CSE
GokarajuRangaraju Institute Of Engineering and Technology (GRIET),
Hyderabad, India

V Srinivas
Assistant Professor in CSE
Gokaraju Rangaraju Institute Of Engineering and Technology (GRIET),
Hyderabad, India

B S Anil Kumar
Assistant Professor in CSE
Gokaraju Rangaraju Institute Of Engineering and Technology (GRIET),
Hyderabad, India

***Abstract*- In this work, Prediction of heart attack is very uncontrolled problem faced by doctor abnormally in every hospitals they have to make a killing review feedbacks from their patients. The ability of the sentiment analysis is needful work for machine to get results in the form of feedback i.e. positive or Negative feedback. Machine learning techniques assess necessary job in this area and in this zone of research work to construct a product this can assists the machine learning computations to input the option choice with respect to the fore cast and assess the expectation. Heart disease commonly occurred disease and it is the critical speculation for unexpectedly death these days. Nearest Neighbour (KNN) featureless is the most well-known, viable and productive calculation utilized for acknowledgment .In this paper , evaluate heart disease occurrences prediction by using machine learning strategies. Which will gives comfortable heterogeneous forecast model and it aid to discover best symptomatic for medicinal services framework for early prediction of symptomatic heart disease.**

Keywords: Machine Learning, NLTK, Heart disease, KNN, Logistic Regression, Random Forest.

I. INTRODUCTION

Machine Learning is the unit of computer technology. It gives the "computer system ability to "Learn" and it progressively increase the performance of a assigned work on the heart disease dataset, without being especially programmed[2]. The machine learning name was invented in the year of 1959 by Arthur Samuel. It is embedded with Artificial intelligence and it Derive the study in machine learning with "Pattern Reorganization" and "Computational Learning" Theory in artificial intelligence .It scrutinize the study and built up the algorithm procedures that can be hustle the information and make prediction of data essentially given algorithms overcome following tightly with "Static Programs "those are also Static models are little easy to build and examine .The network incorporates for working towards "Data Breach", "optical character recognition", "Learning to rank" and "Computer vision". It closely related to "computational Statistics"[1]. (These are e also focus on prediction making through the use of computer).

- Data analytics
- Predictive analytics
- Data scientist

A. Supervisory Machine Learning

In SML the learning algorithm is granted with characterize. inputs, where the labels indicate the desired output.SML itself is consists below

- Classification: - where the output is qualitative, and
- Regression/Prediction: - where the output is quantitative.
- Medical diagnosis, Credit approval, mTarget marketing,, Treatment effectiveness analysis, Categorizing news stories as finance, weather, entertainment, sports, etc.

B. 1.2 Supervisory Learning Using Python

In this approach, we start with significant heart disease dataset and it having training attributes , target attributes. The Supervised Learning algorithm will known the differentiate among the training attributes and their required target variables and apply that learned relationship to classify whole new inputs (without targets).To demonstrate how supervised learning works, let's consider a supportive example of anticipate the marks of a student based on the number of hours he studied. Mathematically = f(X)+ C.

In this one , data has both input (training) values and output (target) values. in this dataset has continuous numerical values of attributes without any target labels, then it comes under Regression problem"[3].

C. Machine Learning:

The problem for machine learning is to identify information graphs in given data and then make prediction dependent on those regularly , large experiments to locate business optimistic queries , and aid them to take care of problems .Machine learning techniques determine your data and recognize patterns. In supervised learning, the model is Trained" with an enormous (large) volume of data and algorithms are used to anticipate a result from future sources of information[4]. Heart disease depicts a amplitude of

conditions that affect your heart. Under the Heart disease for suppose, Chest pain, chest tightness, Blood pressure and chest discomfort (angina) Spyder is an open source cross-platform unified growth environment (IDE) for technical programming in the Python language. Spyder support with different identified bundles in the logical Python stack, including NumPy, SciPy, Matplotlib, pandas, IPython, SymPy and Cython, just as other open source software.

II. RELATED WORK

- We have gathered the movie review data sets of various sizes and have chosen a portion of the broadly utilized and well-known administered AI calculations, for preparing the model. With the goal that the model will most likely sort the audit[5].
- These inputs can be gathered from various perspectives, for example, utilizing Web Applications, Mobile Applications, Feedback Systems, through mail, and so forth. Feedback also called audits are accumulated and distributed openly.
- So, that the clients can think about these audits while picking an administration or item and furthermore the associations can use these criticisms for giving better support of the clients.
- Heart disease is one of the important medical issue in the present life. In this paper, we are applied various supervisory learning/ data mining procedures applied on heart disease dataset to predict the coronary illness expectation are examined. Numerous creators researched on this region and connected different information mining strategies.
- Sentiment analysis alongside the AI procedures can bring about the structure of an elite canny framework and can confirmation its ability in the region of man-made consciousness. \
- But in some cases, it turns into an extremely intricate activity for the analysts to choose a suitable AI procedure as indicated by their prerequisite which leads them to inappropriate outcome with exceptionally poor exactness and execution of the model.

III. HEART DISEASE DATASET

The data set used for this work is from "Data. World" kept in the heart illness dataset is utilized. The dataset has 210 entries and 8 attributes. These 8 attributes are the most considered factors for the coronary ailment. Derelict of the way that it has 210 objects of data classes which only 10 are done and the remainder of the lines contain missing characteristics and removed from the analysis. The Data of patients recorded by number of times of chest misery and age in years. In dataset there are 8 characteristics used in this system, with 5 alphabetic and 3 numerical attributes. Approximately, the basic parameters are variable parameters and that ought to be analyzed for at regular intervals the lock (greatest pulse accomplished), circulatory strain (mm Hg), serum cholesterol in (mg/dl), and electrocardiographic results. In real world, data isn't continually completed and by virtue of the therapeutic data, it is for each situation veritable. To clear the amount of inconsistencies which are connected with data we use Data pre-taking care of.To predict heart disease the dataset holding 210 instances is collected from UCI repository Information about heart disease data set is shown in Table1

IV. METHODOLOGY

A. Introduction of K—Nearest Neighbour

I have assembled over 80% arrangement models and only 15-20% deterioration models. This measurement can be pretty much summing up all these through the business. The main agenda for this inclination towards identification models in that scientific issues include making a decision. In this Paper, we are discussed another broadly utilized grouping strategy called K-Nearest neighbours (KNN). Our spotlight will be essentially on how does the calculation work and how does the info parameter impact the output/prediction. The k-nearest neighbours (KNN) algorithm is a modern and very simple to use supervised machine learning algorithm that can be implemented to simplify both classification and regression problems[7].

TABLE I. HEART DISEASE DATASET

Age	**Chest Pain**	**BP**	**Blood (Sugar)**	**Rest Electro**	**Max (Heart Rate)**	**Exersice (angina)**	**Disease** (Positive (+), Negaticve (-))
43	Asympt	142	No	Normal	140	Y	+
39	Atypical Angina	125	No	Normal	165	Y	-
39	Non Anginal	161	Yes	Normal	165	N	-
42	Non Anginal	161	No	Normal	145	N	-
49	Asympt	142	No	Normal	135	N	-
50	Asympt	142	No	Normal	135	N	-
59	Asympt	142	Yes	Left_Vent_Hyper	120	Y	+
54	Asympt	200	No	Normal	139	Y	+
59	Asympt	135	No	Normal	126	N	+
56	Asympt	170	No	Abnormality	125	yes	+
52	Non Anginal	140	No	Abnormality	170	N	-
60	Asympt	100	No	Normal	125	N	+
55	Atypical Angina	160	Yes	Normal	143	Y	+
57	Atypical Angina	140	Yes	Normal	140	N	-
38	Asympt	110	No	Normal	166	N	+
60	Non Anginal	120	No	Left_Vent_Hyper	135	N	-
55	Atypical Angina	140	No	Normal	150	N	-
50	Asympt	140	No	Abnormality	140	Y	+

B. K—Nearest Neighbour (KNN)

In the acquirements step, the characterization shows the classifier by segregating the training set. In the characterization step, the class names for the given data are foreseen. The data set qualities and their related class labels examined inside part into a preparation set and test set. [10]

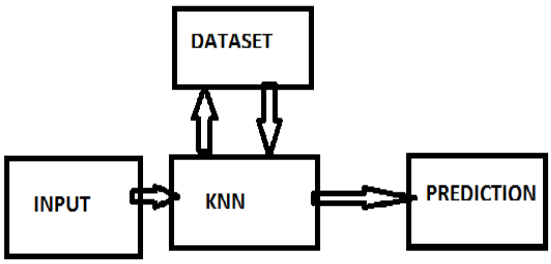


Fig. 1. Flow chart of KNN

Our proposed system intends to improve the execution of KNN classifier for disease conjecture. Count for our proposed strategy is showed up as Algorithm 1. [7]

Stage 1: Heart Disease informational index

Stage 2: Classification of informational

Collection into patients with coronary

Illness and typical.

Stage 3: Input the informational collection

Stage 4: Apply pre-preparing strategies Fill in

Missing qualities

Stage 5: Discard excess highlights

Stage 6: Apply (KNN) on Predominant highlights

Stage 7: Measure the execution of the KNN

Demonstrate.

Executing KNN in Scikit-Learn on Heart dataset to order the kind of Heart Disease dependent on the given data.

First steps, to apply our machine learning calculation we need to comprehend and investigate the given dataset. In this point of reference, we use Heart illness dataset which is imported from the scikit-learn group. [8]

Pip install ~m pandas

pip install ~m matplotlib

pip install ~m scikit-learn

C. K-Nearest Neighbours in scikit-learn

KNN relies upon information by correspondence issue, so that, by brilliance of a submitted test traits and making credits those are equivalent to that. The made properties are portrayed by n qualities. Each characteristic tends to a point in a n-dimensional space. Thus, every arranging traits are verified in n-dimensional point of reference space.At that moment when we give an ambiguous attribute, a k-Nearest Neighbour classifier checks the model space for the k created attributes those are closest to the ambiguous attributes. These k created attributes, that are the k nearest Neighbours of the ambiguous attributes. Directly, we give import KNN classifier from sklearn and apply to our data which by then gatherings the Heart Disease dataset [11].

D. Implementing KNN Algorithm with Scikit-Learn

In this part ,we will perceive how Python's Scikit-Learn library can be utilized to execute the KNN calculation in under 20 lines of code. The download and establishment directions for Scikit learn library are accessible at here.

The Dataset

We are using this heart disease dataset to utilize the well-known iris informational collection for our KNN model. The dataset comprises of four properties: sepal-width, sepal-length, petal-width and petal-length. These are the characteristics of explicit kinds of iris plant. The undertaking is to foresee the class to which these plants have a place. There are three classes in the dataset: Heart Disease, Heart Disease and Heart Disease.

Importing Libraries

import NumPy as np

import matplotlib.pyplot as plt

*import pandas as p*d

Importing the Dataset: the supervisory learning algorithms access the dataset and upload same it into our panda's data frame, then execute the following code:

```
data =
pd.read_csv('C:/Users/Desktop/Dataset/heartdataset.csv')
data.head()
print(data.head())
```

Advantages

- The algorithm is simple and simple to execute.
- There's no reason to construct a model, tune few parameters, or make extra suspicions.
- The algorithm is adaptable. It tends to be utilized for order, relapse, and search (as we will find in the following segment).
- As said before, it is sluggish learning calculation and, in this way, requires no preparation preceding making continuous expectations. It creates the KNN calculation very quicker than various other techniques calculations that require preparing e.g SVM, straight relapse, and so on.
- Since the calculation requires no preparation before making forecasts, new information can be included consistently.

V. RESULT AND DISCUSSION

There are 210 observations with 6 features each (age, chest_pain, rest_bpress, max_heart_rate, exercise_angina, disease).There are no null values, so we don't have to worry about that. There are 50 observations of each species.

A. Load the Dataset:

```
data=pd.read_csv('C:/Users/RonakModi/Desktop/17241D2510/heartdataset.csv')
```

TABLE II. AGE WISE HEART DISEASE PREDICTION

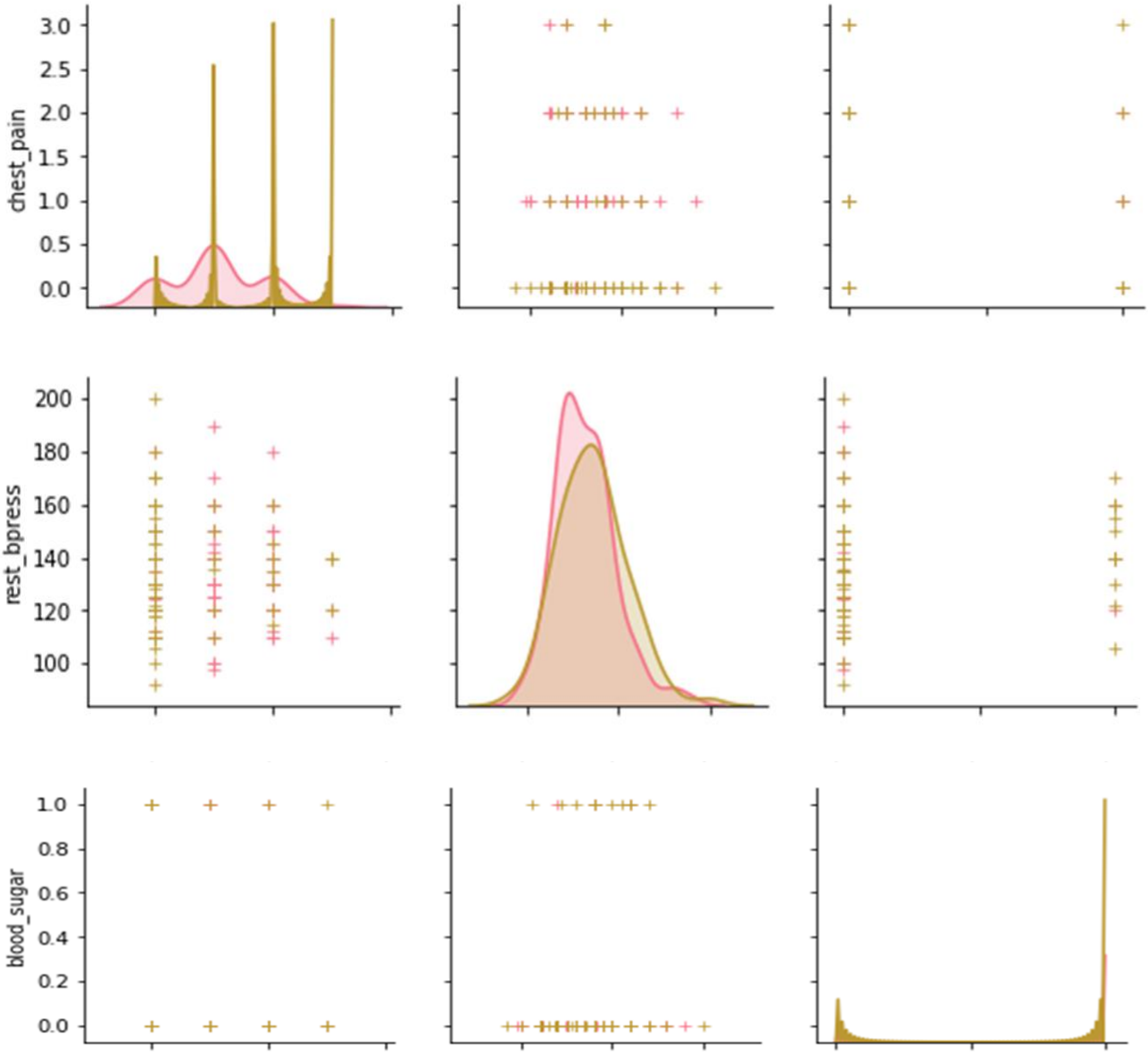
S. No	Age	Chest_Pain	Rest_Bpress	Max_Heart_Rate	Exercise_angina	Disease
0	43	Asympt	140	135	Yes	Positive
1	39	Atyp_angina	120	160	Yes	Negative
2	39	Non_anginal	160	160	No	negative
3	42	Non_anginal	160	146	No	Negative
4	49	Asympt	140	130	No	Negative

Data Transformation:

We found in our underlying investigation that the majority of the sections in our informational index are strings, however the calculations in scikit-learn see just numeric information. Fortunately, the scikit-learn library givesus numerous strategies for converting string data into numerical data.

TABLE III. PREPROCESSING HEART DISEASE DATA

S.No	Age	Chest_Pain	Exercise_angina	Disease
0	**43**	**0**	**1**	**1**
1	**39**	**1**	**1**	**0**
2	**39**	**2**	**0**	**0**
3	**42**	**2**	**0**	**0**
4	**49**	**0**	**0**	**0**



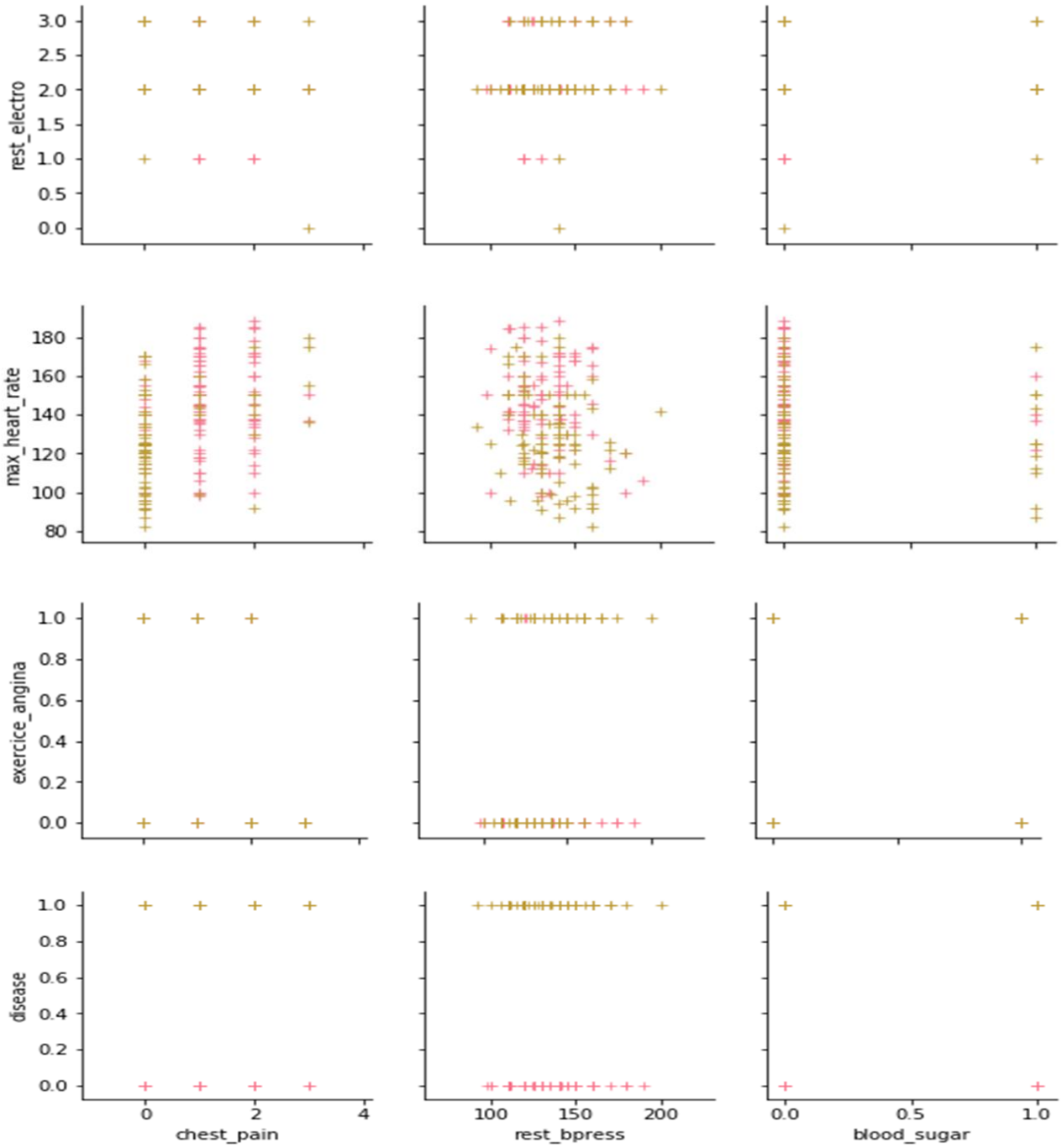


Fig. 2. Preprocessing of 50 observations of each species

Data Visualization:

We will utilize the violinplot() technique given by the Seaborn module to make the violin plot. Allows first import the seaborn module and utilize the set() technique to modify the size of our plot. We will consider the to be of the plot as - 1.0 by 2.0.

In its simple structure, a violin plot shows the appropriation of information crosswise over names. In the above plot we have marks 'rest_bpress' on the x-pivot and the estimations of 'infection' in the y-hub.

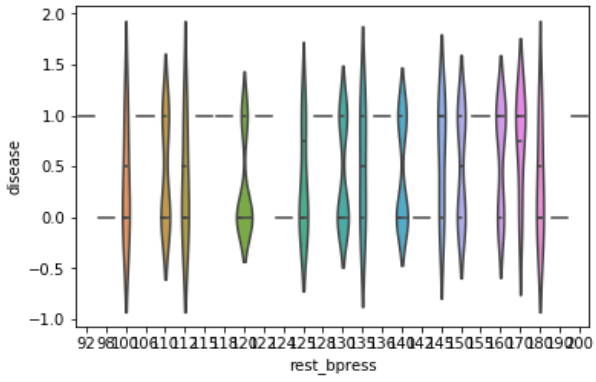


Fig. 3. Rest Bpass Vs estimation of Infection in disease

In its simple structure, a violin plot shows the conveyance of information crosswise over names. In the below plot we have marks 'max_heart_rate on the x-hub and the estimations of 'illness' in the y-pivot. The violin plot demonstrates to us that the biggest circulation of information is in the customer size '- 1.0', and the remainder of the customer size is 2.0.

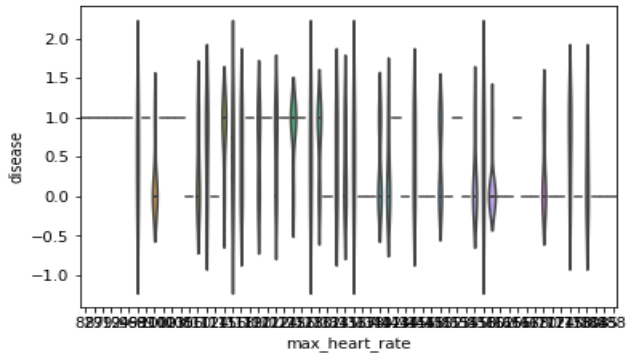


Fig. 4. Max Heart rate plots

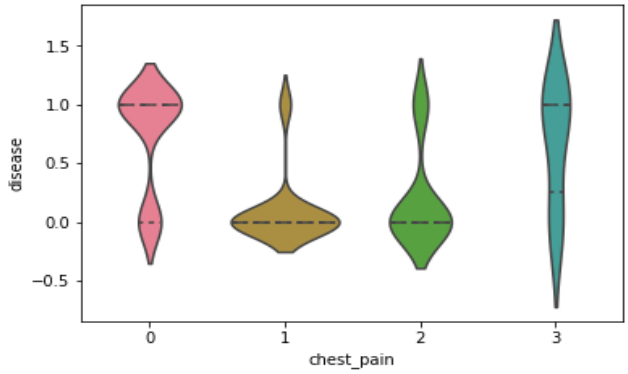


Fig. 5. Chest Pain Plots

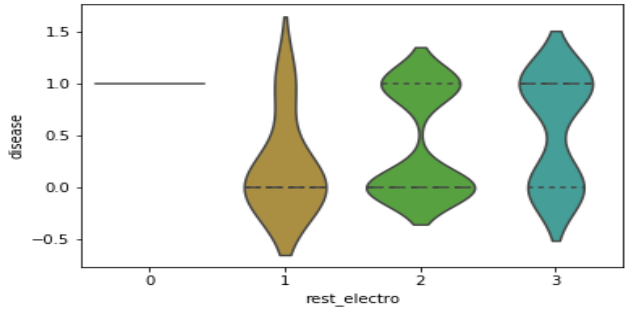


Fig. 6. Rest Elctro Plots

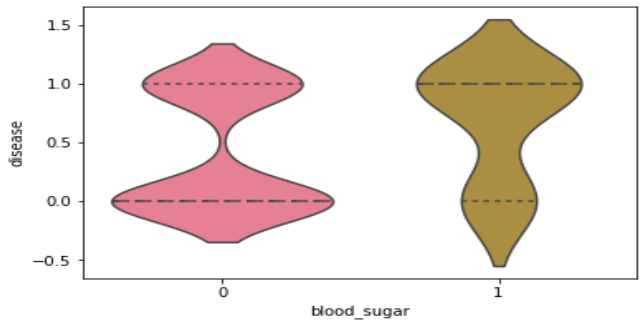


Fig. 7. Blood Sugar Plots

KNN Classifier/ Regression:

Here ,we have Final Accuracy Results of KNN classifier Algorithms with different N Values.

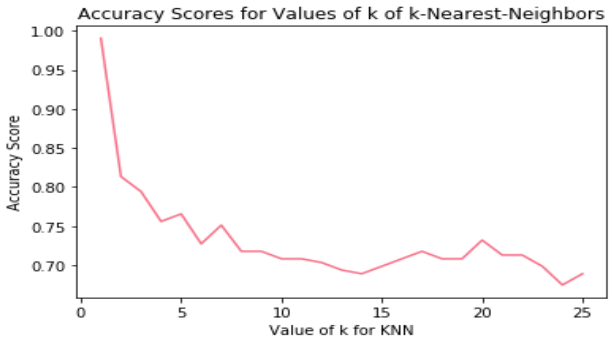


Fig. 8. Accuracy Results OF KNN Classifier

Here ,we have Final Accuracy Results of KNN Logistic Regression Algorithms with different N Values.

K-Neighbor Accuracy Score:
0.8452380952380952

Performance Comparison

K-Neighbor Accuracy Score:
0.7942583732057417

(125, 6)

In the past segments we have utilized the accuracy score() technique to gauge the precision of the KNN calculations. Presently, we will utilize the Random Forest Algorithm given by the RandomForestClassifier library to give us a visual report of how our models perform.

```
logreg = RandomForestClassifier()
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
print( RandomForestClassifier())
print(metrics.accuracy_score(y_test, y_pred))
```
Result : 0.7380952380952381

VI. CONCLUSION

In this paper watched out for the estimate of Heart Disease subject to KNN. Our system uses KNN as a classifier to diminish the misclassification rate. In this paper also investigates KNN based component decision measure to pick few features and to upgrade the request execution. The results prescribe that proposed system would altogether be able to improve the learning precision. From generation results, it is derived that KNN based segment assurance is imperative for course of action of Heart disease. This model helps the specialists in a gainful estimate of sicknesses with commanding features. In future, we have to fuse outfit classifiers with KNN to develop a decision sincerely strong system for early finding of Heart disease and besides should consider Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) for Heart Disease set.

REFERENCES

[1] BiswaRanjanSamal,MrutyunjayaPanda,"*Performanc e Analysis of Supervised MachineLearning Techniques for Sentiment Analysis*" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 14, No.5, May 2016

[2] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[3] TaiwoOladipupoAyodele "*Types of Machine Learning Algorithms*",NewAdvances in Machine Learning,YagangZhang(Ed.),InTech,2010,DOI: 10.5772/9385.

[4] M.A. Jabbar, "*Heart Disease Prediction System usingAssociativeClassification and GeneticAlgorithm*", ICECIT, pp 183-192, Elsevier, vol 1(2012).

[5] R. Chitra and Dr.V. Seenivasagam "*Heart Disease Prediction SystemUsing Supervised LearningClassifier"*Bonfring International Journal of SoftwareEngineering and Soft Computing, Vol. 3, No. 1, March 2013.

[6] Hardeep Singh Assistant Professor "*Performance analysis of unsupervised machine learning techniques for networktrafficclassification*"Fifth International Conference on Advanced Computing&Communication Technology 2015.

[7] MA Jabbar and Shirinasamreen "*Heart disease prediction system based on hidden naïve bayesclassifier"*Vardhaman college of Engineering Hyderabad, India.

[8] Bing Liu "Sentiment Analysis and Opinion Mining" Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012

[9] Siqian Chen, Jie Yang, Yun Gu "image sentiment analysis using supervised collective matrix factorization" 12th IEEE Conference on Industrial Electronics and Applications (ICIEA) Institute of Image Processing and Pattern Recognition Shanghai Jiao Tong University, Shanghai China,

[10] Jabbar MA "Prediction of heart disease using k-nearest neighbor and particleswarm optimization". Vardhaman College of Engineering,Hyderabad, India ISSN 0970-938X.

[11] Rao, S. Govinda, R. Rambabu, and P. VaraPrasada Rao. "Modified Hierarchical Clustering algorithms to Evaluate the Similarities of Growth Factor IR Inhibitors by Using Regression Analysis." 2018 4th International Conference on Computing Communication and Automation (ICCCA). IEEE, 2018.