IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DISEASE DIAGNOSIS PREDICTION FROM PATIENT SYMPTOMS USING MACHINE LEARNING

[1]Sindhu J, [2]B M Bhavya

[1]Research Scholar, Dept. of MCA, PES College of Engineering Mandya, Karnataka

[2]Assistant Professor, Dept. of MCA, PES College of Engineering Mandya, Karnataka

Abstract: This project is primarily focused on the creation of a machine learning algorithm-based system for disease prediction in a variety of diseases. Medical facilities must improve in order to make better decisions about patient diagnostic and treatment alternatives. In healthcare, machine learning allows humans to handle large, complicated medical information and then analyse them for clinical insights. Physicians can then use this information to provide medical care. As a result, when machine learning is used in healthcare, it can improve patient satisfaction. In this project, we want to combine machine learning capabilities in healthcare into a single system. Instead of diagnosis, healthcare may be made smart by implementing disease prediction utilising machine learning predictive algorithms.

Index Terms **-K-Nearest Neighbor, DCT, Naïve Bayes, Disease Prediction.**

I. INTRODUCTION

Artificial intelligence has made computers smarter and has given them the ability to reason. Machine learning is considered a subfield of AI research in a number of studies. Different experts believe that insight cannot be gained without learning. Unsupervised, Semi-Supervised, Supervised, Reinforcement, Evolutionary Learning, and Deep Learning are examples of Machine Learning Techniques. These learnings are used to quickly classify large amounts of data. For quick categorization of massive data and accurate illness prediction, we employ K-Nearest Neighbor (KNN) and related machine learning algorithms such as knn, naive bayes, and decision tree machine learning algorithms. Because medical data is growing at an exponential rate, using it to predict the correct disease is becoming increasingly important. However, because processing big data is becoming increasingly important in general, data mining plays an increasingly important role, and classification of large datasets using machine learning is becoming increasingly simple.

II. LITERATURE SURVEY

Konstantina Kourou and colleagues [1] proposed a research on machine learning applications in cancer prognosis and prediction in 2015. They offered an overview of many recent machine learning algorithms for cancer detection prediction in this work. They have offered a review of freshly released content for cancer detection research thus far. In 2015, P.Swathi Baby et al [2] suggested a project to develop a predictive mining-based diagnosis and prediction system. Weka and Orange software are used to analyse the illness data set. Machine learning techniques such as AD Trees, J48, K star, Nave Bayes, and Random Forest are used to investigate the performance of each algorithm, which provides statistical analysis and disease prediction. Their findings suggest that the best algorithms for the used dataset are K-Star and Random Forest, where building models takes less time (0 sec and 0.6 sec) and the ROC values are 1. [3] On the basis of accuracy, accuracy, and execution time for CKD prediction, the performance of Bayesian classifier, support vector machine (SVM) classifier, and Knearest neighbour is compared. 7(1), 42254, International Journal of Advances in Engineering and Technology (IJAET). [4] The goal of the project was to use vector-based vector machines (SVM) and artificial neural networks to forecast sickness (ANN). The purpose of this research is to compare the accuracy and run-time of the two methods. The experimental results reveal that RNA has a higher yield than other algorithms.

III. METHODOLOGY

1 Collection of Datasets

The datasets were obtained from kaggle.com, which is a website that provides raw datasets. The datasets are in csv format, and Pandas Libraries are used to read them.

2 Data Preprocessing

- Finding missing values
  Check for missing values like null, nan etc, in datasets
- Fill missing values
  Fill missing values by applying different statistical methods like
  Mean, median, mode an and fillna method etc. pandas and numpy are used for filling values in datasets

3 Data Analysis

- Analysis of all the data columns present in the datasets. Seaborn technique is used to analyses the datasets
- Each column is compare with column attrition in the datasets to check the important of each attribute

4 Splitting data into train and test

- Split data into train and test using sklearn libraries
- The data is splitted into 80% for training and 20% for testing
- The trained data are passed to the algorithms

5 Apply Classification Algorithms

- The sklearn library is used to deploy algorithms into our project
- The trained data are passed to the knn, naïve bayes decision tree
- The data are trained and checked for validation

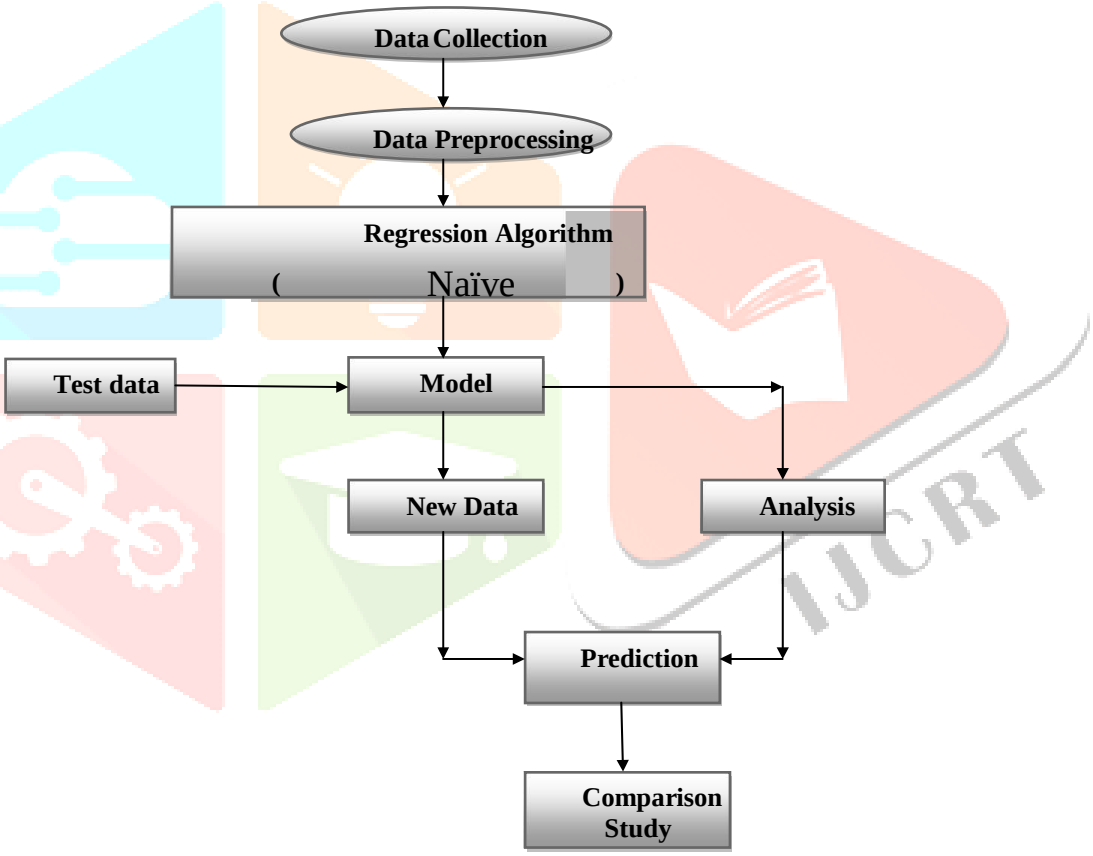

Figure 1 Proposed Methodology

3.1 KNN Algorithm Pseudocode

1. Calculate "d(x, x_i)" i =1, 2, ….., **n**; where **d** denotes the Euclidean distance between the points.
2. def euclidean_distance(x,y):
3. return sqrt(sum(pow(a-b,2) for a, b in zip(x, y)))
4. Arrange the calculated **n** Euclidean distances in non-decreasing order.
5. Let **k** be a +ve integer, take the first **k** distances from this sorted list.
6. Find those **k**-points corresponding to these **k**-distances.
7. Let $\mathbf{k}_i$ denotes the number of points belonging to the i^{th} class among **k** points i.e. $k \geq 0$
8. If $k_i > k_j \; \forall \; i \neq j$ then put x in class i.

3.2 Naïve Bayes

- Derivation:
- D : Set of tuples
- Each Tuple is an 'n' dimensional attribute vector
- X : (x1,x2,x3,…. xn)
- Let there be 'm' Classes : C1,C2,C3…Cm
- Naïve Bayes classifier predicts X belongs to Class Ci iff
- P (Ci/X) > P(Cj/X) for 1<= j <= m , j <> i Maximum Posteriori Hypothesis
- P(Ci/X) = P(X/Ci) P(Ci) / P(X)
- Maximize P(X/Ci) P(Ci) as P(X) is constant With many attributes, it is computationally expensive to evaluate P(X/Ci). Naïve Assumption of "class conditional independence"
- ∏= = n k P X Ci P xk Ci 1 (xk./ Ci)
- P(X/Ci) = P(x1/Ci) * P(x2/Ci) *…* P(xn/ Ci)

P(A|B) = Fraction of worlds in which B is true that also have A true

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Corollary:

$$P(A \wedge B) = P(A|B)\,P(B)$$

$$P(A|B)+P(\neg A|B) = 1$$

3.3 Decision Tree Code

Information Gain

```
infoGain(examples, attribute, entropyOfSet)
  gain = entropyOfSet
  for value in attributeValues(examples, attribute):
    sub = subset(examples, attribute, value)
    gain -=  (number in sub)/(total number of examples) * entropy(sub)
  return gain
```

Entropy

```
entropy(examples)
  '''
  log2(x) = log(x)/log(2)
  '''
  result = 0
  # handle target attributes with arbitrary labels
  dictionary = summarizeExamples(examples, targetAttribute)
  for key in dictionary:
    proportion = dictionary[key]/total number of examples
    result -= proportion * log2(proportion)
  return result
```

IV. RESULTS AND DISCUSSION

K-Nearest Neighbor (KNN), Naive Bayes, and Decision Tree Machine Learning are examples of machine learning techniques. Algorithms are used to classify large amounts of data quickly and accurately forecast disease.The KNN Algorithm displays 61.63 percent, the Naive Bayes algorithm shows 99.61 percent, and the Decision Tree algorithm shows 100 percent. In comparison to KNN and Nave Bayes, Decision Tree produces the best results.

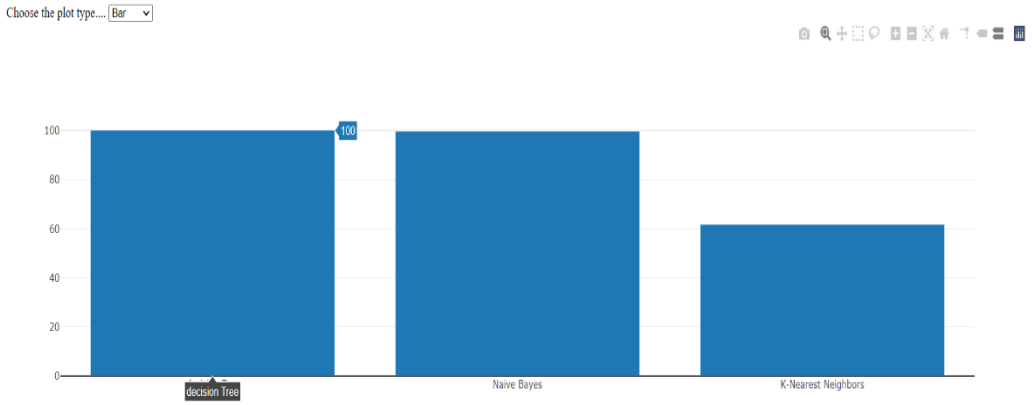

Figure 2:Accuracy of the Algorithm

V. CONCLUSION

In the health field, the use of data mining techniques for predictive analysis is critical since it allows us to detect diseases sooner and thereby save people's lives by anticipating cures. In this application, Data Mining algorithms such as Logistic Regression, KNN, and Naive Bayes are used to forecast healthy people and patients with health care data. The Naive Bayes classifier proved its performance in predicting with the best outcomes in terms of shortest execution time, according to simulation findings.

REFERENCES

[1] Konstantina Kourou et.al, "Machine learning applications in cancer prognosis and prediction" Computational and structural bitechnology Journal, Elsevier.

[2] P.Swathi Baby, T. Panduranga Vital ,"Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms" International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-018, Vol. 4 Issue 07, July-2015,206-210.

[3] K.R.Lakshmi1, Y.Nagesh2 and M.VeeraKrishna3, "Performance Comparision of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability", International Journal of Advances in Engineering & Technology, Mar. 2014, Vol. 7, Issue 1, pp. 242-254.

[4] Dr. S. Vijayarani, Mr.S.Dhayanand, "Kidney Disease Prediction Using SVM and ANN Algorithms" IJCBR , ISSN (online): 2229- 6166,Volume 6 Issue 2 March 2015.