ISSN No:-2456-2165

Smart Health Prediction System

Vivek Joshi[1], Shipra Goswami[2], Shalini Goel[3]
Department of Computer Science and Engineering,
Meerut Institute of Engineering and Technology, Meerut, India

Abstract:- With the promises of predictive analysis in machine learning algorithms, prediction of future is no longer a difficult task. Many efforts have already been done to obtain useful knowledge from it. In this paper, we will apply five machine learning algorithms on three different set of medical data. The objective of this research paper is to predict heart disease, diabetes and liver disease by using different machine learning algorithms that are Naïve Bayes Algorithm, Support Vector Machine, Decision Tree, KNN, Logistic Regression and find the most efficient one.

Keywords:- *Machine Learning, Naïve Bayes Algorithm, SVM, KNN, Regression, Decision Tree and Logistic Regression.*

I. INTRODUCTION

Diseases are unusual conditions that give adverse effect on body parts of human being. These are usually the condition of a body that is related with specific symptoms. In human, diseases referred to illness that causes pain, sickness, distress, infection, or death to person afflicted, or similar problem for those who came in contact of infected person. In our research we are mainly focusing on Heart Disease, Diabetes and Liver disease. Heart diseases were found in 71.9% of patients and 28.1% had a previous intervention [8]. Heart attack happens when a part of the heart muscles does not get enough blood to pump therefore it can't get the needed oxygen. The blockage is due to the accumulation of cholesterol, fat and other substances, which build plaque in the arteries. Sometimes, a plaque can rupture and build a clot that block flow of blood and can damage or destroy heart muscle. Its symptoms include chest pain, abnormal heartbeat, light headedness, pain or shoulder discomfort, anxiety, shortness of breath, etc. Diabetes occurs in our body when glucose or blood sugar level is higher than its normal level. A fasting blood sugar level range 72-99 mg/dL is normal and it is good to have blood sugar level less than 180 mg/dL about 2 hours after eating. High blood glucose leads to problem such as kidney disease, stroke, eye problem, nerve damage, foot problem, heart disease, etc. Factors affecting blood sugar level are quantity, time of intake and type of food consumed, age, physical activity, alcohol consumption, stress, dehydration and many more [1]. Liver disease is itself of many types depending upon the origin of its cause. Causes of liver disease are auto immune disease, excessive use of alcohol, viruses, genetics, reactions of medication, street drugs, or toxic chemical. Its symptoms include jaundice (yellowing of the skin and eye colour), abdominal pain, chronic fatigue, vomiting and loss of appetite.

Disease prediction models have the ability to benefit the government and health insurance companies. These can identify patient's disease or critical health conditions. These models can take appropriate action to avoid or minimize the risk of hospital's treatment cost hence improving quality of care. Due to current progress in advancement of tools and techniques of data analytics, disease prediction model can leverage significant amounts of information, such as, clinical diagnosis, demo graphics and clinical measurements, laboratory results, health behaviour, prescriptions and care utilization.

Why need of computer assisted healthcare when there are several doctors present in the world? In India, there is a huge gap in doctor population ratio, there is only 1 doctor on 1456 people so there exist lack of physician to examine people. Many times we need doctor's help urgently but we can't have it due to some reasons. Using our machine learning models users can get to know the status of their health [2].

II. LITERATURE REVIEW

The purpose of our paper is to improve the efficiency of disease prediction models by increasing the accuracy of algorithms. More and more algorithms are used in this so that we can find the best performing algorithm for a particular disease.

Year	Author	Disease	Technique	Accuracy
2021	Nisha Gupta et al. [15]	Heart Disease	SVM D.T.	83% 79%
2017	Ashok Kumar Dwivedi et al. [16]	Heart Disease	SVM N.B. D.T. L.R. KNN	82% 83% 77% 85% 80%
2021	Olta Llaha et al. [17]	Diabetes	SVM N.B. D.T. L.R.	64.70% 68.40% 63.75% 68.02%

ISSN No:-2456-2165

2021	Kishan Patel et al. [18]	Diabetes	D.T. L.R. KNN	Best accuracy is 78% of L.R.
2019	A.K.M Sazzadur Rahman et al. [19]	Liver Disease	SVM N.B. D.T. L.R. KNN	64% 53% 69% 75% 62%
2018	Joel Jacob et al. [20]	Liver Disease	SVM L.R. KNN	75.04% 73.23% 72.05%

Table 1: Accuracy comparison of different research papers

III. SYSTEM DESIGN

This section explains the machine learning process and algorithms with its application in smart health prediction.

A. Machine Learning

Machine learning is a subfield of artificial intelligence. It aims to design and develop algorithms that allow computers to improve their performance over time based on data [3-4]. It is the study of algorithms which improve through their experiences and by using the data. It is the procedure of training machine to learn without explicitly programming. Machine learning algorithm uses data to predict the possible outcome and use train-test-split technique for assessing the performance. Machine learning uses mathematical model of data (algorithms) to find patterns and then using that pattern to carry out its tasks. Machine learning mainly have two objectives:

- Classification of data using existing models.
- Prediction of future outcomes using these models.

Our proposed profound model-based disease prediction includes many stages. All the stages can be summed up in five stages:

- Stage 1: Gathering of datasets from different reliable sources.
- Stage 2: Analyzing datasets using numpy and pandas.
- Stage 3: Performing data cleaning to convert biased data to unbiased data. This step is important because biased data leads to biased prediction which will affect our final prediction accuracy.
- Stage 4: Creating and training models.
- Stage 5: Testing accuracy of each and every model and choosing best model for prediction.

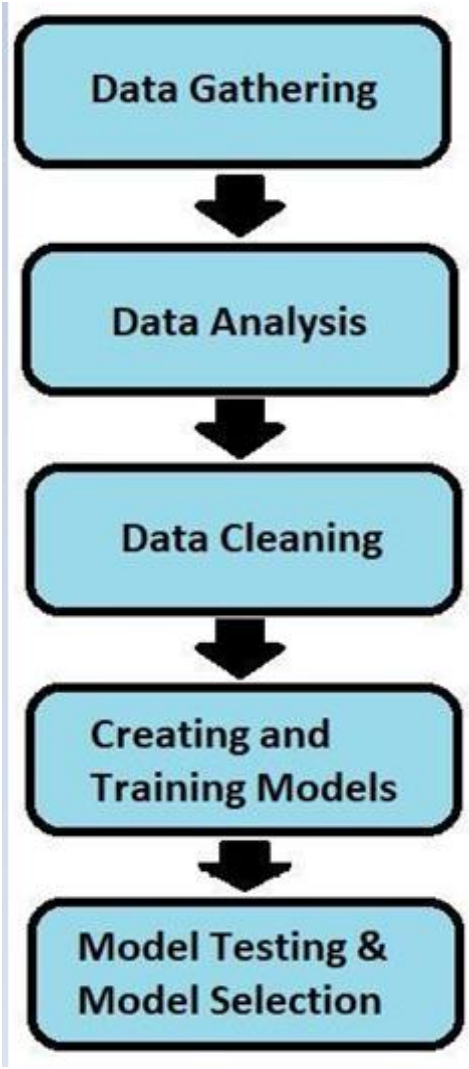

Fig. 1: Flow Diagram

ISSN No:-2456-2165

IV. METHODOLOGY

The Methodology is to implement "Smart Healthcare Prediction System" via Machine Learning techniques, which includes Naïve Bayes Classifier, Support Vector Classifier, Decision Tree, K-Neighbors Classifier, and Logistic Regression. This procedure can be termed as 'Knowledge Discovery Process', this process includes:

- **Data Selection**: Data selection is the process of determining the appropriate source of data and its data type. Data selection process is the actual practice of data collection. This definition differentiate data selection from selectively excluding data that is not useful for research hypothesis. This step includes sorting reliable information into files, gathering information from online databases like Kaggle, Datahub, UCI etc [12].
- **Data Pre-Processing**: This step comprises altering stored data into clean data. Pre-Processed file contains uneven, biased, partial and/or missing information, which might affect models' accuracy or might make our model biased [12]. In order to overcome this obstacle we have used following libraries:
  - Pandas: To read and transform data for further data cleaning.
  - Numpy: To clean and transform data for machine learning models.
  - Sklearn: To split data into testing and training data and to make different machine learning models.
  - Matplotlib: For visual representation of data and comparison of models [15].

- **Algorithms used**: We have applied many algorithms on medical dataset to find best result for a particular type of disease:
  - **Naïve Bayes Algorithm**: Naïve Bayes classification algorithm is a probabilistic classifier. It is based on probability model that include strong independent assumptions. The occurrences of specific features of a class are independent of the absence or presence of other feature according to Naïve Bayesian classifier theorem [13]. This collection of classification algorithms is based on supervised learning algorithm which itself is based on Bayes Theorem [5, 9].

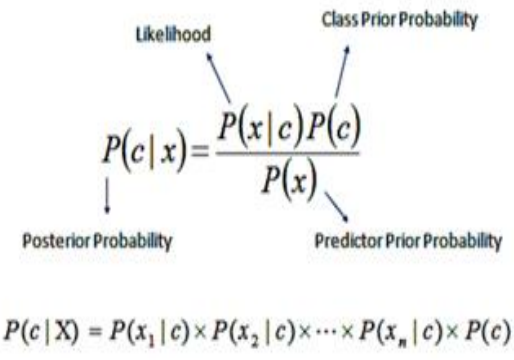

Fig. 2: Bayes' Theorem.

  - **Support Vector Machine:** SVM is used for both regression and classification problem. It is a supervised machine learning algorithm which works to find hyper plane in high dimensional space. The main purpose of the SVM algorithm is to draw a decision boundary that can divide multi dimensional space into different classes so that we can easily classify new data point in correct dimension while classifying data in future. This best decision linear boundary is known as hyper-plane. SVM algorithm can be used for image classification, text categorization, face detection, etc [13-14].
  - **Decision Tree**: Decision Tree is a tree-structured classifier, in which internal node represents dataset features, branches represent the decision rules and each leaf node represents the result. It has two nodes namely, decision node and leaf node. Decision nodes are decisions making node who have multiple branches, whereas leaf nodes are the results of those decisions and do not have any child branches [3, 10].
  - **K - Neighbors Classifier**: KNN is a supervised learning algorithm used for both classification and regression. It tries to predict correct class of a data by calculating the distance between the test dataset and the training dataset. K in KNN represents the k^{th} nearest neighbor, where k is an integer. This algorithm assumes that in the given data similar things exist in close range and then create groups on the basis of the model's learning [10, 14].
  - **Logistic Regression:** Logistic regression predicts the output of a categorical dependent variable using a given set of independent variables. Therefore, the outcome must be a categorical or discrete value. The output can only be a binary (0 or 1, yes or no, true or false, etc.) but in place of 0 and 1, most of the time it gives the probabilistic output which is between 0 and 1 [3].

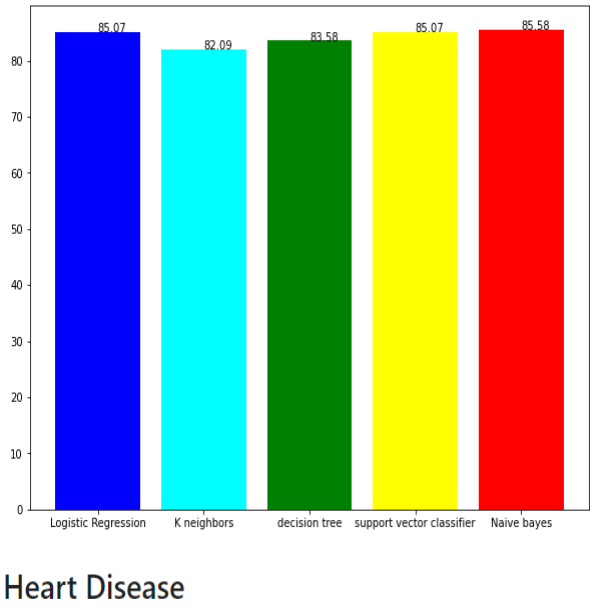

Fig. 3: Algorithm comparison for Heart Disease.

ISSN No:-2456-2165

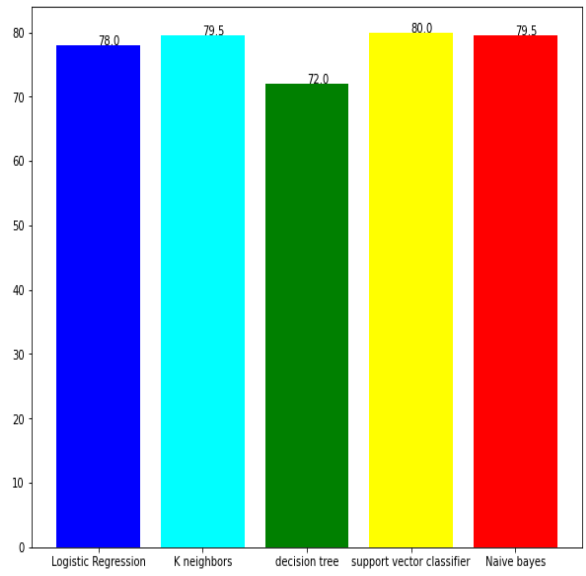

Fig. 4: Algorithm comparison for Diabetes

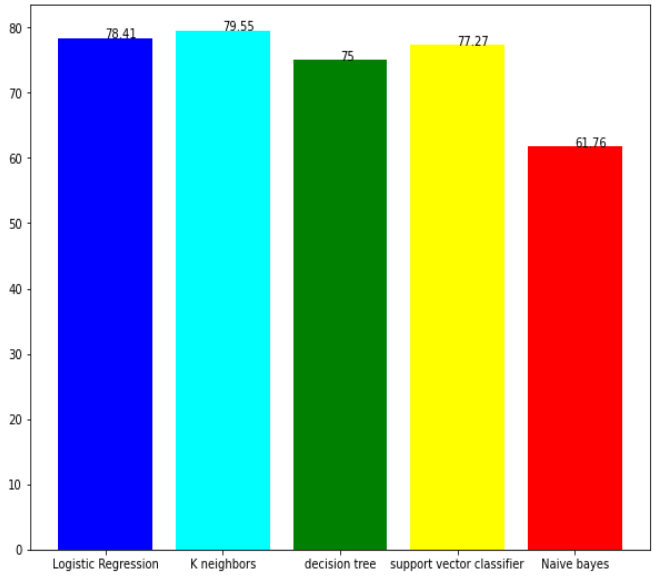

Fig. 5: Algorithm comparison for Liver disease.

V. RESULT

After performing analysis on different machine learning algorithms we concluded that logistic regression is best suited Heart Disease with 85.07% accuracy, SVM is best for Diabetes with 80% accuracy, and k nearest neighbors is best for Liver Disease with 79.55% accuracy.

For detailed comparison refer to accuracy comparison table given below:

	Regression	K - Neighbors	Decision Tree	SVM	Naïve Bayes
Heart Diseases	85.07%	82.09%	83.58%	85.07%	83.58%
Diabetes	78.00%	79.50%	72.00%	80.00%	79.50%
Liver disease	78.41%	79.55%	75%	77.27%	61.76%

Table 2: Accuracy of Algorithms

VI. CONCLUSION

The machine learning plays a vital role in diseases prediction while designing “Smart Health Prediction System”. However, no single machine learning algorithm is best suited to resolve the prediction issues for all healthcare datasets. The combination of several machine learning algorithms or hybrid version of the machine learning algorithm may be the better approach to get best prediction of diseases. Though accuracy is increased by considerable amount, there is further room for improvements in accuracy of models. The future may be in designing a better machine learning model that can address healthcare with real time healthcare dataset. This study does not comprise the complete analysis of all the existing data algorithms and dataset.

REFERENCES

[1.] Shubham Salunke, Shubham Rajiwade, Deepak Yadav, S.K.Sabnis, "smart health prediction system using machine learning", IJRAR – International Journal of Research and Analytical Reviews (IJRAR), EISSN 2348- 1269, P- ISSN 2349-5138, Volume.7, Issue 1, Page No pp.483-488, March 2020.

[2.] Sella Ppan Palaniappan, Rafiah Awang, “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, IJCSNS, Volume 8, August 2018.

[3.] I.-N. Lee, S.-C. Liao and M. Embrechts, “Data mining techniques applied to medical information”. Med.inform. Vol. 25, no. 2, pp81- 102, 2000.

[4.] “Using machine learning algorithms in cardiovascular disease risk evaluation”, Sitar-Taut, V.A., et al., Journal of Applied Computer Science & Mathematics, 2009.

[5.] G. Pooja Reddy, M. Trinath Basu, K.Vasanthi, K.Bala Sita Ramireddy, Ravi Kumar Tenali, “Smart E-Health Prediction System using DataMining”, IJITEE volume 8, issue 6, April 2019.

[6.] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain”, Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March 2016.

[7.] Nikita Kamble, International Journal of Scientific Research in Computer Science Engineering and Information Technology, Vol. 2, Issue 5, 2017, “Smart Health Prediction System Using Data Mining”.

[8.] K. Vembandasamy, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2, Issue 9, September 2015, "Heart Diseases Detection Using Naive Bayes Algorithm".
[9.] Prof. Krishna Kumar Tripathi, International Research Journal of Engineering and Technology (IRJET), Vol.5 Issue: 4, Apr-2018, " A Smart Health Prediction Using Data Mining".
[10.] M. Thiyagaraj, G. Suseendran, International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, September 2019, "Research of Chronic Kidney Disease based on Data Mining Techniques".
[11.] M. Gandhi, "Predictions in Heart Disease Using Techniques of Data Mining," Int. Conf. Futur. trend Comput. Anal. Knowl. Manag., 2015.
[12.] Vidya Zope1, Pooja Ghatge2, Aaron Cherian3, Piyush Mantri4, Kartik Jadhav, IJSRD - International Journal for Scientific Research & Development, Vol. 4, Issue 12, 2017, "Smart Health Prediction using Machine Learning".
[13.] Mr. ChalaBeyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.
[14.] Rakshith D B, Mrigank Srivastava, Ashwani Kumar, Gururaj S P, "Liver Disease Prediction System Using Machine Learning Techniques", IJERT, Vol-10, Issue 6, June 2021.
[15.] Nisha Gupta, Gulbakshee Dharmale, Darshana Parmar, "Heart Disease Prediction using machine learning", 2021 JETIR March 2021, Volume 8, Issue 3.
[16.] Ashok Kumar Dwivedi, "Analysis of computational intelligence techniques for diabetes mellitus prediction", Neural Comput. Appl., vol. 13, no. 3, pp. 1–9, 2017.
[17.] Olta Llaha, Amarildo Rista, "Prediction and detection of diabetes using Machine Learning", 202; Proceedings of RTA-CSIT 2021, May 2021, Tirana, Albania.
[18.] Kishan Patel, Manu Nair, Subham Phansekar, "Diabetes prediction using Machine Learning", 2021 IJSER, Volume 12, issue 3, March 2021.
[19.] A.K.M Sazzadur Rahman, F.M. Javed Mehedi Shamrat, Zarrin Tasnim, Joy Roy, Syed Akhter Hossain, "A Comparative Study on Liver Disease Prediction Using Supervised Machine Learning Algorithms", 2019 IJSTR, Volume 8, issue 11, November 2019.
[20.] Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac, "Diagnosis of Liver Disease Using Machine Learning Techniques", 2018 IRJET, Vol-5, Issue 4, April 2018.