SYMPTOMS ANALYSIS AND DISEASE PREDICTION USING MACHINE LEARNING

Divya Patel
UG scholar
Dept. of CSE
MIET, Meerut, India
divya.patel.cs.2019@miet.ac.in

Divya Tyagi
UG Scholar
Dept. of CSE
MIET, Meerut, India
divya.tyagi.cs.2019@miet.ac.in

Vivek Kumar
Assistant Professor
Dept. of CSE
MIET, Meerut, India

ABSTRACT

Data science and machine learning have evolved along with modern technology, paving the way for healthcare organisations and medical facilities to better serve their patients by assisting in the early detection of diseases. Massive volumes of data, some of which are concealed, are gathered by the health care sectors and are used for making informed judgments.

We are putting forth a system for diagnosing diseases including chronic kidney disease, liver disease, heart disease, breast cancer and diabetes so they may be treated early on and the user can learn more about them.

Medical diagnoses necessitate visiting a doctor, making an appointment for a consultation, and waiting for blood results to seek a doctor's consultation in order to obtain correct disease indicators. When we don't feel well, the first thing we do is examine our temperature to get a rough estimate or baseline idea of how fevered we are. If the temperature is high enough, we then consult a doctor. In a similar way, this disease prediction system can be used to determine the disease's approximate severity and can advise us as to whether we should seek emergency medical assistance or not, or at the very least begin some home cures for the condition to provide temporary comfort.

The Disease Prediction approach, which focuses on predictive modeling, makes disease predictions for users based on the symptoms they supply as input. As an output, the method returns the likelihood of the condition after analysing the user's symptoms as input.

Keywords: Machine Learning, SVM, Random Forest, Decision Tree, K-NN.

1. INTRODUCTION

The current study aimed to forecast chronic kidney disease, liver disease, breast cancer and diabetes according to the symptoms shown by the patient. Nowadays, machine learning techniques are crucial for disease prediction from medical databases in the healthcare industry. Machine learning is being used by numerous researchers and businesses to enhance medical diagnosis [6].

Globally, chronic kidney disease is an important public health concern because it can result in renal failure, cardiovascular ailment, and early mortality. CKD prevalence varies by age group, gender, socioeconomic level, and geographic region; claims one study. Liver, which is the largest internal organ of human body, performs various crucial functions relating to immunity, digestion, metabolism, and nutrition storage [7]. When predicting heart diseases, data science and machine learning (ML) can be very helpful by taking into account many risk factors such as high blood pressure, high cholesterol, an abnormal pulse rate, diabetes, etc. Breast cancer, which accounts for 30% of all female cancers worldwide and is regarded as a multifactorial disease, is the most prevalent cancer in women. It takes a variety of criteria, including demographic, laboratory, and mammographic risk markers, to accurately estimate the risk of breast cancer. As a result, multifactorial models that analyse a variety of risk factors can be efficient in determining the risk of breast cancer through a more precise study [8]. Diabetes occurs when the body cannot adequately utilize the manufactured insulin or when the pancreas fails to make enough of it. Blood glucose levels in diabetics are higher [9].

The prediction is made using various machine learning algorithms including K-Nearest Neighbors (KNN), Naive bayes, and Support vector machine(SVM). The model predicts that whether the patient has high or low risk of developing the disease based on the symptoms [10].

2. LITERATURE REVIEW

"Disease Prediction by using Machine Learning" is an info mining concept that was introduced by the author of paper [1] [11]. Data analysis is a crucial component of every field, thus the stage's best development is incorporating that method into healthcare foundations. Data mining makes predictions about the information needed for the rapidly expanding medical sector. It is intended to explain the total healthcare systems and is used for (i) healthcare data analysis, (ii) management, and (iii) prediction. Data analysis is used in these sorts of processes to create the treatment procedures, which rely on the use of machine learning to retrieve information on diseases. Due to its effectiveness, decision trees are used to anticipate disease outbreaks [12] [13]. The results of this experiment demonstrate how the outcome is connected to the condition's symptoms. The idea picks a training set, such as the symptoms of a medical patient, uses a decision tree to make predictions, then asks the patient for their symptoms to achieve an accurate diagnosis. Simply the patient-related information is predicted by this method, which only requires minimal effort and expense.

K. Vijiya Kumar [2] [14] suggested the Random Forest algorithm to build a system which can predict a patient's early onset of diabetes more accurately. The findings showed that the proposed model yields the best predictions and that the model is capable of reliably and effectively forecasting the diabetes condition.

Predicting diabetes onset was described by Nonso Nnamoko et al. [3] [15], it uses a meta-classifier to combine the results of five popular classifiers. The outcomes are contrasted with similar research from the literature that made use of identical dataset. It has been demonstrated that the suggested approach can predict diabetes more accurately.

P.Kuppan [4] [16] in their research employed J48, Decision Table, Naive Bayes, and Data Mining algorithms to analyse the data pertaining to Liver disease. Characteristics like patient's medical history, obesity, diabetes, smoking, drinking, were used. It has been determined based on the provided database [17].

S. Ramya and Dr. N. Radha [5] [18] investigated the duration of diagnosis and increased the accuracy of diagnosis using several machine learning classification methods. The proposed work focuses on categorising the various stages of CKD based on their severity. The analysis findings show that RBF method generates 85.3% accuracy, which is superior than the other classifiers when compared to other algorithms like Basic Propagation Neural Network.

3. METHODOLOGY

The proposed system is used to predict diseases on the basis of symptoms analysis and compare the classification accuracy of ML algorithms - Support Vector Machine (SVM), K-nearest neighbor (KNN) and Naive Bayes. The first step is to collect and clean the data by removing inconsistent and noisy data from the dataset, filling out the null and missing values then converting to binary attribute from nominal attribute. The second step is feature selection, it is used to choose the significant attributes by removing insignificant attributes from the dataset. On the basis of findings, attributes have been selected. In the next step, data transformation is done. The fourth step is training the model. In this step, the model is trained by machine learning algorithms in order to predict outcomes from new data. In the fifth or last step, the model with the highest accuracy is opted for disease prediction based on the accuracy of the various classification models.

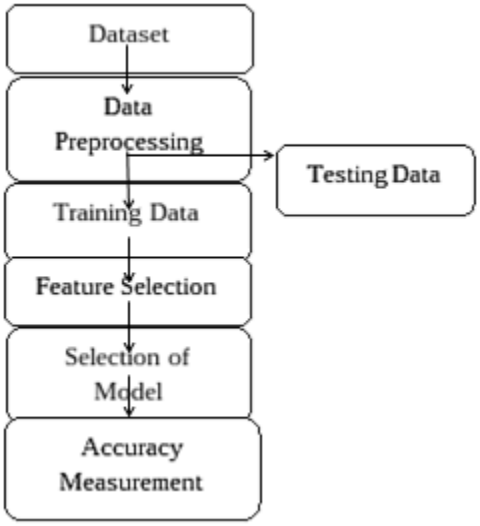


Fig.3. Flow Chart of proposed system

3.1. Dataset

The model is designed using the disease dataset. In the dataset, there are 570 rows and 32 columns. The columns indicate different types of symptoms.

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	commpactness_mean	concavity_mean	concavepoints_mean	symmetry_mean	fractal_mean	radius_se
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095
842517	M	20.57	17.77	132.9	1326	0.8474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09553	0.235	0.07389	0.3063	1.002
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063
84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.082443	0.2976
845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05667	0.3795

Fig.3.1. Sample dataset

3.2 Data Preprocessing

Before starting model building, data preprocessing is required for the removal of unwanted noise and elimination of superfluous attributes from the dataset to make data consistent which improves the accuracy and quality of data. There are three steps in data preprocessing to transform raw data into clean and consistent data set. The first step is Data cleaning in which elimination of missing and noisy data occurs, the second step is Data transformation in which data is transformed into appropriate form which is suitable for mining process, the third step is Data reduction where the amount of data that is required to store on the system is reduced.

3.3. Feature Selection

Feature selection is a data preprocessing technique in data mining which is used for reduction of data by removing insignificant attributes from the dataset and by the removal of unwanted noise. This technique is used to make the model more accurate. It enhances the comprehensibility of data, helps bring out better visualization of data, reduces time required for model training and improves the accuracy and performance of prediction. Before any model is applied to the data, it will be a good choice to remove unwanted and inconsistent data to get more accurate results in less time. It increases the prediction power of the algorithms by selecting the most crucial variables and eliminating the redundant or noisy ones.

3.4 Algorithms

The following machine learning algorithms have been used for prediction.

- K - Nearest Neighbor (KNN)

- Naive Bayes Algorithm
- Support Vector Machine (SVM)

3.4.1. K - Nearest Neighbors (KNN)

K-Nearest Neighbors is considered as one of the simplest Machine Learning algorithms. It uses the supervised learning approach. In this algorithm a new case is assigned to a category based on how closely it resembles to the other categories. With the KNN method, you can classify new data on the basis of its similarity with the old one. KNN is an example of lazy learner because it does no training at all when you supply the training data but memorizes the training dataset. The KNN classifies the new input data into a group or category that is quite close to the old data that was stored at the time of training.

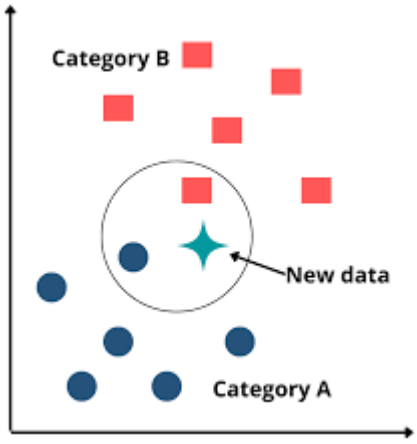


Fig.3.4.1. K - Nearest Neighbor

3.4.2. Naive Bayes

The Naive bayes algorithm of ML is a form of supervised learning and it is a classification algorithm that is based on the Bayes theorem. It is a probabilistic classifier, since it predicts the result based on the probability of an object. The classifier considers that the existence of one feature is independent to the existence of any other feature in that class. This algorithm is mainly used in the creation of classifiers. Classifiers are used to determine that the given inputs belong to which class. This technique is used for building models that are used to assign class labels to problem instances and those labels are drawn from the dataset. Baye's theorem can be mathematically stated as:

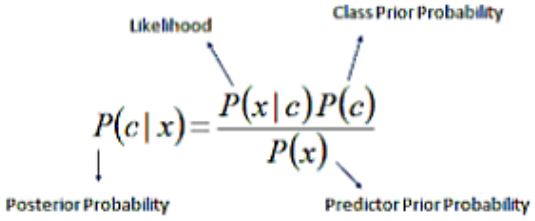


Fig.3.4.2. Naïve Bayes

3.4.3 Support Vector Machine (SVM)

Support Vector Machine is a Supervised Learning algorithm, which can be employed for both Regression and Classification problems. The goal of the algorithm is to create a boundary that is known as a decision boundary which can separate n-dimensional space into different groups or classes and that decision boundary is known as hyperplane. The hyperplane is created with the help of extreme points or vectors. These extreme points are termed as support vectors, that is why the algorithm is termed as Support Vector Machine. Hyperplanes are decision boundaries which is used to classify the data points. The points falling on different side of the hyperplane can be assigned to different classes.

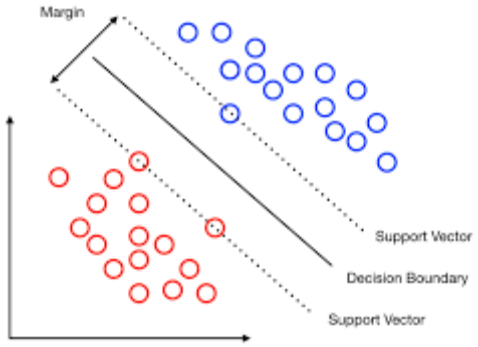


Fig.3.4.3. Support Vector Machine

3.5 Result

The performance of the model can be evaluated on the basis of its accuracy.

Accuracy- It is a deciding factor used in those models in which classification algorithms have been used. It signifies the ability of algorithm to predict the result accurately.

True positive (TP): It is the case of correct prediction of the model as high risk.

True negative (TN): It is the case of correct prediction of the model as low risk.

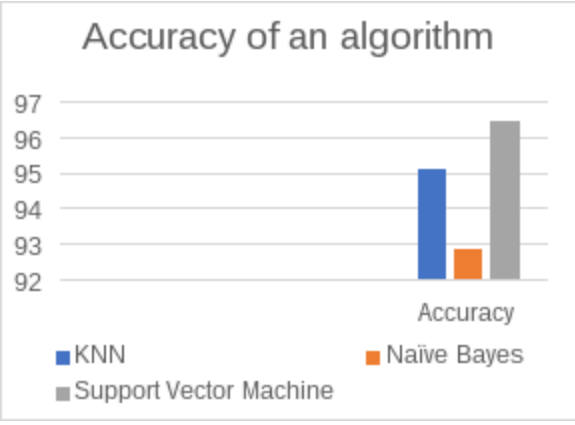
False positive (FP): It is the case of incorrect prediction of low risk as high risk.

False negative (FN): It is the case of incorrect prediction of high risk as low risk.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Algorithm	**Accuracy%**
KNN	95.12%
Naïve Bayes	92.9%
Support Vector Machine	96.49%

This can also be shown in the form of a graph.



3.6 Output Screens

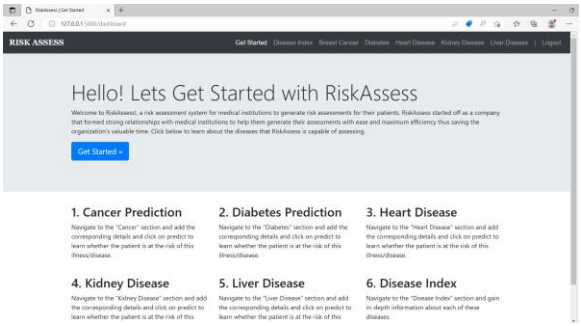


Fig.3.6.1. User Interface

This screen shows the user interface after running the project in local host. Here user gets various options and can choose whatever operation he wants to do.

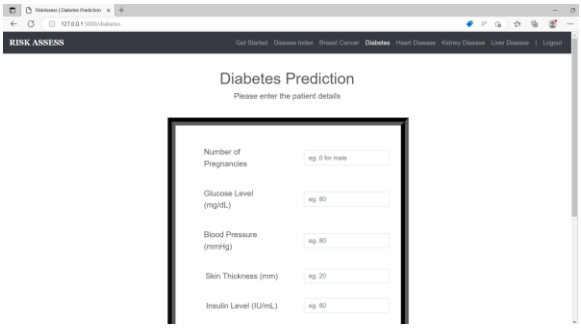


Fig.3.6.2. Diabetes Prediction entry form

This screen shows the diabetes prediction system where the user inputs the required information to predict the result.

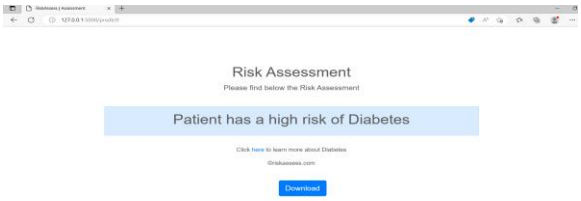


Fig.3.6.3. Predicted Result of diabetes prediction

This screen shows the result predicted by the system as per the details entered by the user.

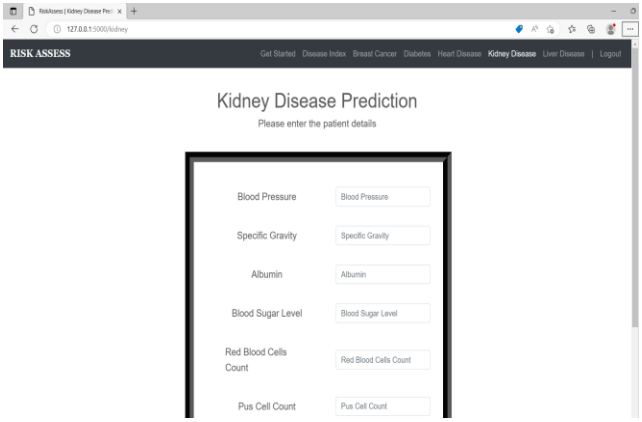


Fig.3.6.4. Kidney Disease Prediction entry form

4.CONCLUSION AND FUTURE WORK

The goal is to predict different diseases based on symptoms. The proposed system can predict diseases on the basis of symptoms entered by the user so it can be cured at an early stage and the user can search about these diseases.
In future the efficiency of the system can be increased by using other machine learning algorithms. In this project we are taking textual data input but in future we can build a system that can take images as well.

5. REFERENCES

[1]. Sayali Ambekar and Dr.Rashmi Phalnikar. "Disease Prediction system by using Machine Learning". International journal of computer engineering and applications, Volume XII, special issue, May 18. ISSN: 2321-3469.

[2]. K. Vijiya Kumar, B. Lavanya, S. Sofia Caroline, "Random Forest Algorithm for Diabetes prediction". International Conference on Networking and Systems Computation Automation, 2019.

[3]. Nonso Nnamoko, David England, Abir Hussain, "Predicting Diabetes: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[4]. P. Kuppan, N. Manoharan. "Analysis of Liver Disease using Naïve Bayes, Data Mining Algorithms and Decision Table", International Journal of Computing Algorithm, vol. 6, no. 1, pp. 2278-239, 2017.

[5]. S. Ramya, Dr. N. Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative
Research in Computer and Communication Engineering,Vol. 4, Issue 1, January 2016.

[6] Narayan, V., & Daniel, A. K. (2022). FBCHS: Fuzzy Based Cluster Head Selection Protocol to Enhance Network Lifetime of WSN. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 11(3), 285-307.

[7] Narayan, V., & Daniel, A. K. (2019). Novel protocol for detection and optimization of overlapping coverage in wireless sensor networks. Int. J. Eng. Adv. Technol, 8.
[8] Awasthi, S., Srivastava, A. P., Srivastava, S., & Narayan, V. (2019, April). A Comparative Study of Various CAPTCHA Methods for Securing Web Pages. In 2019 International Conference on Automation, Computational and Technology Management (ICACTM) (pp. 217-223). IEEE.

[9] Narayan, V., & Daniel, A. K. (2022). Energy Efficient Protocol for Lifetime Prediction of Wireless Sensor Network using Multivariate Polynomial Regression Model. Journal of Scientific & Industrial Research, 81(12), 1297-1309.

[10] Choudhary, S., Narayan, V., Faiz, M., & Pramanik, S. (2022). Fuzzy approach-based stable energy-efficient AODV routing protocol in mobile ad hoc networks. In Software Defined Networking for Ad Hoc Networks (pp. 125-139). Cham: Springer International Publishing.

[11] Srivastava, S., & Sharma, S. (2019, January). Analysis of cyber related issues by implementing data mining Algorithm. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 606-610). IEEE.

[12] Srivastava, S., Yadav, R. K., Narayan, V., & Mall, P. K. (2022). An Ensemble Learning Approach For Chronic Kidney Disease Classification. Journal of Pharmaceutical Negative Results, 2401-2409.

[13] Srivastava, S., & Singh, P. K. (2022). Proof of Optimality based on Greedy Algorithm for Offline Cache Replacement Algorithm. International Journal of Next-Generation Computing, 13(3).

[14] Salagrama, S., Kumar, H. H., Nikitha, R., Prasanna, G., Sharma, K., & Awasthi, S. (2022, May). Real time social distance detection using Deep Learning. In 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) (pp. 541-544). IEEE.

[15] Mahadani, A. K., Awasthi, S., Sanyal, G., Bhattacharjee, P., & Pippal, S. (2022). Indel-K2P: a modified Kimura 2 Parameters (K2P) model to incorporate insertion and deletion (Indel) information in phylogenetic analysis. Cyber-Physical Systems, 8(1), 32-44.

[16] Tyagi, N., Rana, A., Awasthi, S., & Tyagi, L. K. (2022). Data Science: Concern for Credit Card Scam with Artificial Intelligence. In Cyber Security in Intelligent Computing and Communications (pp. 115-128). Singapore: Springer Singapore.

[17] Awasthi, S., Kumar, N., & Srivastava, P. K. (2021). An epidemic model to analyze the dynamics of malware propagation in rechargeable wireless sensor network. Journal of Discrete Mathematical Sciences and Cryptography, 24(5), 1529-1543.

[18] Singh, M. K., Rishi, O. P., Awasthi, S., Srivastava, A. P., & Wadhwa, S. (2020, January). Classification and Comparison of Web Recommendation Systems used in Online Business. In 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM) (pp. 471-480). IEEE.