Disease Prediction Using Patient History

[*]Nikhil Butle, [#]Dr. Kirti Wanjale

[*,#]Department of Computer Engineering, Vishwakarma Institute Of Information Technology, Pune, India.

Abstract:- There is an increasing number of chronic disease patients day by day and it is an important topic to be addressed. Most of the time the patient has to take long-term medicines because the treatment doesn't start at the early stage. To solve this problem this research has built a machine learning model that helps in detecting chronic disease so its treatment can start as soon as possible. Model is, particularly for diabetes, heart, and Parkinson's disease using a machine learning algorithm logistic regression, and Support Vector Machine all with an accuracy greater than 75 percent, helps to detect whether the patient is suffering from a particular disease or not. Further, this model also has a great future scope and the most famous one is to use this model in edge devices like mobile and smartwatches.

Keywords - Disease, Prediction, patient.

I. INTRODUCTION

Human Disease Prediction is a very important aspect of human life. Beforehand disease prediction in humans is a vital process in disease treatment. Right from the start, a doctor has controlled almost everything by himself which takes a lot of time. The healthcare and medical industry depends on innovation to improve the efficiency of logistics. It is the impetus behind novel remedies, cures, and treatments. Innovation is also what keeps the medical sector updated and applicable. There has been a great deal of development in the medical field. Innovation is required to advance in several domains. Among these are creating novel therapies for diseases, figuring out how to enhance patient care, and speeding up and streamlining medical procedures. To solve this problem machine learning comes into the picture. The field of machine learning makes predictions using historical data. The comprehension of a computer system that allows the machine learning model to gain knowledge from the data and the experience is known as machine learning. There are two stages to the machine learning algorithm: training and testing.

For many years, technology like machine learning has had a hard time figuring out what's wrong with a person by looking at their symptoms and medical background. This study uses comprehensive machine learning ideas to monitor patient health. With the use of machine learning models, we can swiftly clean and process data and produce results more quickly. Physicians will diagnose patients correctly by utilizing this approach. That means the patient will receive good care which boosts the development of patient healthcare services. Healthcare is an ideal instance of how to introduce machine learning in the medical profession.

II. LITERATURE REVIEW

In the early 2000s, machine learning algorithms began to be used for disease prediction in a wider range of medical fields. For example, researchers developed an algorithm to estimate the danger of heart disease, stroke, diabetes, and other chronic disease. Machine Learning algorithms were also used to develop algorithms to detect illnesses early such as detecting breast cancer or mammograms or lung cancer on a lung scan. Various algorithms can be used for predicting disease. Some of the most commonly used algorithms are Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Deep Learning. Logistic regression is a statistical method used for binary classification. Logistic regression aims to forecast the likelihood that a specific instance falls into a specific category. A decision tree is a popular machine-learning algorithm used for both classification and regression tasks. Recursively dividing the data into subsets according to the values of the input features is how it operates. Making judgments based on the attributes in order to build a model that predicts the target variable is the main objective. Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes or the mean prediction of the individual trees. It's a type of bagging algorithm, where the dataset is divided into multiple subsets, and a decision tree is trained on each subset. A Support Vector Machine is a supervised machine-learning algorithm used for classification and regression tasks. SVM works exceptionally effectively in high-dimensional environments and is especially well-suited for classification issues. The primary goal of support vector machines is to identify the hyperplane in feature spaces that divide various classes. The accuracy of every algorithm varies and depends on which test case it is used And the quality of data used to train the model. Some of the early research has shown that for particular disease prediction, some techniques show the highest accuracy score among all [2][3]. For example, logistic regression provides a precision of 85-90 percent for heart disease, the decision tree provides

 DOI : 10.35291/2454-9150.2024.0092

a precision of 85-90 percent for breast cancer [1], the random forest provides a precision of 90-95% for Stroke, support vector machine provides the precision of 85-90 percent for diabetes and deep learning provides a precision of 90-95 percent for Alzheimer's disease. In the 1960s Dr.Larry Weed was a physician and medical informaticist who developed one of the first computer-based medical record systems. In the 1970s Dr. Jack Smith was a computer Scientist and medical informaticist who developed the first successful machine learning algorithm for predicting the risk of heart disease. Dr. John Tukey was a statistician and mathematician who developed many statistical methods that are used in machine learning algorithms [4]. A recent study developed a support vector machine model to estimate the danger of heart disease according to the patient's medical background and lifestyle factors. The model was trained and tested on a dataset of over 10,000 patients [6]. The results showed that the SVM model was able to estimate the risk of heart disease with an accuracy of over 90%.In the same year, a study developed an SVM model to predict the risk of breast cancer based on a patient's mammogram results. The models were trained and tested on a dataset of

over 20,000 patients. The results showed that the SVM model was able to forecast the danger of breast cancer with an accuracy of over 95%.In the same year, the SVM model was developed to estimate the risk of diabetes based on the patient's medical background and lifestyle factors [5]. The model was trained and tested over a dataset of over 5,000 patients. The results showed that the SVM model was able to estimate the risk of diabetes with an accuracy of 85% [7]. All these factors lead to the conclusion that SVM models can be used to develop accurate and reliable models for predicting a variety of diseases[10]. Because of their capacity to manage high-dimensional data using fewer training set examples, SVM models are particularly well-suited for disease prediction tasks. Many other recent studies have shown the effectiveness of SVM models for disease prediction [9]. SVM models are now being used to predict a broad range of diseases, including Alzheimer's disease, Parkinson's disease, and various types of cancer [8]. A highly reliable and accurate SVM model can identify individuals at high risk for developing certain conditions, so they can be monitored closely and receive medication as soon as possible which would lead to a reduction in the burden of chronic diseases, improved patient outcomes, and lower healthcare cost. All three datasets are available on Kaggle.

III. METHODOLOGY

In this project, this study has used a prediction algorithm. A prediction Algorithm, sometimes referred to as a predictive model or algorithm, is a computational technique or collection of guidelines used to anticipate or predict future results or events using past data and trends. These algorithms are an essential part of data science and machine learning, and they are used for tasks including time-series forecasting, regression, and classification across many industries. Three different types of disease prediction:- heart disease prediction, diabetes disease prediction, and Parkinson's disease prediction are the chronic disease predict in this study. The heart disease prediction study used a logistic regression model and in the other two, this study used a Support Vector Machine. Logistic regression is a statistical model that is often used in situations where the output is true or false (Binary Situations). This method has a lot of advantages it is a good fit for binary classification tracks, easy to interpret, is robust to outliers and it is computationally efficient. Since our main goal is to predict whether the person has a heart disease or not the situation becomes a true or false situation. First, we load the CSV dataset, the dataset is taken from Kaggle and then clean the dataset. This is a very necessary step because there can be a null value or

bad value in the dataset and in that case, we either take the mean or step deviation of the dataset. After that find out the metadata about the data i.e. mean, count of each value, standard deviation, minimum, and maximum. Divide the dataset into two parts, X has features and Y has outcomes. Split the dataset into training and testing data using train_test_split function imported from the Sklearn module. Research has divided 80 percent of the data into training and 20 percent to test. Call the logistic regression model and fit the training dataset. Now that your model is ready test this model with your testing dataset and then check the accuracy of the training and testing dataset. If Both model gives an accuracy greater than 75 percent then the model is suitable for predicting the disease. The only difference between the heart disease model and the other two models is that in the other model, research used SVM, and in the heart disease model. Support Vector Machine(SVM) is particularly well-suited for classification tasks with high-dimensional data. The model is not robust to noise and can learn complex relationships between features. Moreover, they are interpretable. SVM uses some factor on which it makes the decision whether the given statement is true or false in other words whether the output is 1 or 0. The study used Google Collab to test and calculate the accuracy of the three models. Google Colab was the best option as all the code remained secured in the cloud and their in-build Python environment. It is very easy to import the library and the code and e divide it into chunks which makes it easy to debug and solve errors all the datasets have been imported from Kaggle.

In addition to loading and cleaning the dataset, it's crucial to perform exploratory data analysis (EDA) to gain insights into the characteristics and distributions of the features. EDA can involve visualizations such as histograms, box plots, and scatter plots to understand the relationships between variables and identify any potential outliers or anomalies. Furthermore, feature engineering might be employed to extract meaningful information from the dataset and enhance the predictive power of the models. This could include

 DOI : 10.35291/2454-9150.2024.0092

techniques such as one-hot encoding categorical variables, scaling numerical features, and creating new features through transformations or combinations of existing ones. Once the dataset is prepared, hyperparameter tuning can be conducted to optimize the performance of the models. This involves selecting the best parameters for the algorithms, such as the regularization parameter in logistic regression or the kernel function in SVM, through techniques like grid search or random search. Cross-validation is another essential step to evaluate the generalization performance of the models. By splitting the dataset into multiple subsets and training/testing the models on different combinations of these subsets, cross-validation provides more reliable estimates of the models' performance and helps prevent overfitting.

Additionally, it's important to assess the models' performance using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC), depending on the specific requirements of the problem. This allows for a comprehensive understanding of how well the models are predicting the outcomes of interest. Finally, interpreting the results and communicating the findings effectively are essential aspects of any predictive modeling project. This involves analyzing the feature importance, understanding the decision boundaries of the models, and potentially providing actionable insights or recommendations based on the predictions. By following a systematic approach encompassing these steps, the study can ensure robustness, reliability, and interpretability in its predictive modeling efforts across the three chronic diseases of interest.

IV. RESULT AND DISCUSSION

This research has prepared a model that predicts chronic diseases(Diabetes, Heart, and Parkinson's) using a machine learning algorithm. The implementation of the model was done in Google Colab as it is fast and efficient also it provides section-wise running of the code which makes it flexible. The model's motive is to predict the chronic disease so the treatment can be done as early as possible. The table below shows the accuracy rate of each model.

SR.NO	DISEASE	ALGORITHM USED	TRAINING ACCURACY	TESTING ACCURACY
1	Heart	Logistic Regression	0.8512396694214877	0.8196721311475410
2	Diabetes	Support Vector Machine	0.7866449511400652	0.7727272727272727
3	Parkinson's	Support Vector Machine	0.8846153846153846	0.8717948717948718

Table 1.1 The accuracy of all the model

From the above table, the model satisfies the condition of scoring accuracy of more than 75 percent and thus we can say that the algorithm used is acceptable for the model

The above graphs show the accuracy rate in each phase of training and testing the model. As can see the training dropped to 85 percent and testing rose to 72 percent it is because the quality of data affects the accuracy

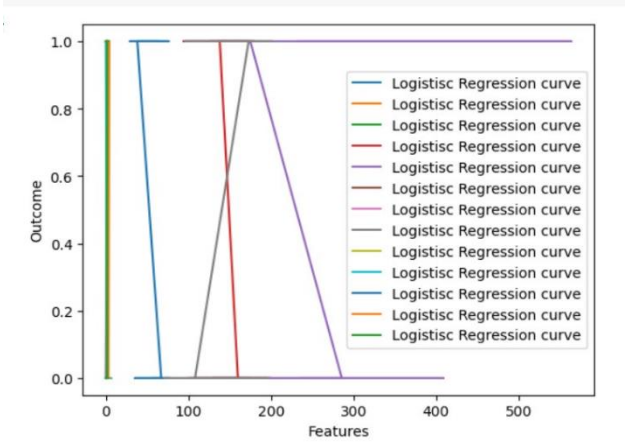

Fig 1.2 Accuracy graph of the Diabetes Model

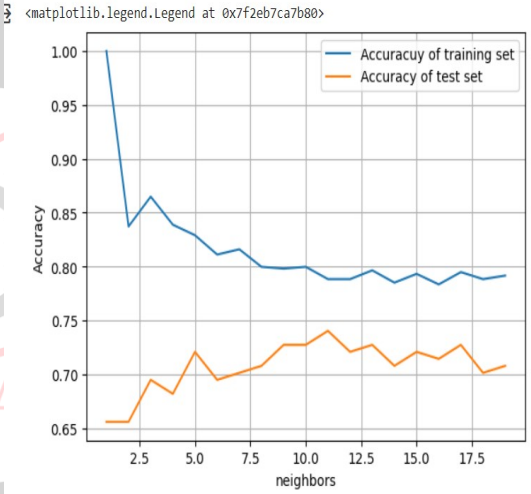

Fig 1.3 Logistic Regression Curve

From the above Logistic Regression curve, we can say that 4 factors affect the outcome of the heart disease prediction.

This research project focused on developing machine learning models for predicting three chronic diseases: Diabetes, Heart disease, and Parkinson's disease. The primary objective was to create accurate predictive models that could assist in early detection and initiation of treatment, thereby improving patient outcomes. To implement these models, Google Colab was chosen as the computing platform. Google Colab offers several advantages, including speed, efficiency, and flexibility. Its ability to execute code section-wise allows for better organization and debugging, facilitating smoother development and testing of the models. The models were trained and evaluated using various machine learning algorithms, with each disease prediction task employing a specific approach tailored to its characteristics and requirements. For instance, logistic

 DOI : 10.35291/2454-9150.2024.0092

regression was utilized for heart disease prediction, while Support Vector Machines (SVM) were employed for both diabetes and Parkinson's disease prediction.

A critical aspect of model development was the preprocessing of the dataset. This involved loading and cleaning the data to handle any missing or erroneous values. Additionally, exploratory data analysis (EDA) was performed to gain insights into the dataset's characteristics and distributions, informing subsequent feature engineering steps. Feature engineering aimed to extract relevant information from the data to enhance the models' predictive performance. Techniques such as one-hot encoding for categorical variables and scaling for numerical features were employed to ensure compatibility with the algorithms used. Hyperparameter tuning was conducted to optimize the models' performance. This involved selecting the best parameters for each algorithm through techniques like grid search or random search, thereby improving the models' ability to generalize to unseen data. Cross-validation was employed to assess the models' generalization performance and mitigate overfitting. By splitting the dataset into multiple subsets and training/testing the models on different combinations of these subsets, more reliable estimates of performance were obtained. Evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) were used to assess the models' performance comprehensively. Interpretation of these metrics provided insights into the models' strengths and weaknesses, guiding further refinement and optimization efforts. The accuracy rates of the developed models were documented and analyzed to gauge their effectiveness in predicting each chronic disease. The findings were presented in a tabular format, showcasing the performance of each model across the different disease prediction tasks. In conclusion, this research project exemplifies the application of machine learning techniques in healthcare for predictive modeling of chronic diseases. Through careful model development and evaluation, the aim is to contribute towards early detection and intervention, ultimately improving patient outcomes and quality of life.

V. CONCLUSION

A disease Prediction System can predict all chronic diseases and this is more budget-friendly more importantly it detects whether you have that particular chronic disease or not so that its treatment can start as soon as possible. This confirms that the Support Vector Machine algorithm is reliable for complex and high-dimensional data. The study shows that logistic regression gives the factor that is dependent on the result in our case the number is four

Overall we can conclude that the model is fit and can be used for commercial purposes according to the accuracy of the model. This model can also help in predicting the risk of the disease. Although there are some restrictions to this model. This model has not been trained to the very huge dataset where numbers go in thousand which affect the accuracy of the data

VI. FUTURE ASPECTS

Future aspects of this model are more promising as more datasets would be available for training and testing the model which will make it more accurate and reliable. We can use Artificial Intelligence (AI) to make it near to perfect model. A disease prediction system can be integrated with electronic health records (ERH) to give real-time insights into patients' high risk of chronic disease. The model can also be integrated with edge devices using Edge AI Like fit-bands and smartwatches and can be integrated with Web and Mobile applications so that patients can detect the risk of disease anywhere in the world.

REFERENCES

[1] Reddy, Palle Pramod, et al. "Disease Prediction using Machine Learning." INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT) 9.5 (2021): c205-c208.

[2] Gaurav, K., et al. "Human Disease Prediction using Machine Learning Techniques and Real-life Parameters." International Journal of Engineering 36.6 (2023): 1092-1098.

[3] Gomathy, C. K., and Mr A. Rohith Naidu "The prediction of disease using machine learning." International Journal of Scientific Research in Engineering and Management (IJSREM) 5.10 (2021).

[4] Keniya, Rinkal, et al. "Disease prediction from various symptoms using machine learning." Available at SSRN 3661426 (2020).

[5] Jindal, Harshit, et al. "Heart disease prediction using machine learning algorithms." IOP conference series: materials science and engineering. Vol. 1022. No. 1. IOP Publishing, 2021.

[6] Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram. "Designing disease prediction model using machine learning approach." 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019.

[7] Ben-Hur, Asa; Horn, David; Siegelmann, Hava; Vapnik, Vladimir N. "Support vector clustering" (2001);". Journal of Machine Learning Research. 2: 125–137.

[8] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine Learning 20 (1995): 273-297

[9] Allison, Paul D. "Measures of fit for logistic regression." Proceedings of the SAS global forum 2014 conference. Cary, NC, USA: SAS Institute Inc., 2014.

[10] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol.2. Oxford: Clarendon, 1892, pp.68–73.

[11] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[12] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[13] K. Elissa, "Title of paper if known," unpublished.

[14] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[15] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[16] M. Young, The Technical Writer's Handbook. Mill Valley, CA: Univer- sity Science, 1989.

 DOI : 10.35291/2454-9150.2024.0092