A Machine Learning Model for Diabetes

Josita Sengupta[1], Koushik Pal[2], Mallika Roy[3], Jishnu Nath Paul[4]

Department of Electronics& Communication Engineering, Guru Nanak Institute of Technology, Kolkata, India

---***---

Abstract - *The prediction of diseases benefits greatly from machine learning methods. Based on the numerous symptoms that users enter as input to the system, the Ailment Prediction system uses predictive modeling to anticipate the user's disease. The algorithm analyses the user's vivid symptoms and returns the likelihood that the disease will occur as an output. Different Supervised Machine Learning methods are used for disease prediction. Accurate medical data analysis enhances patient care and early illness identification as big data usage grows in the biomedical and healthcare industries. This project's overarching goal is to help academics and professionals choose an appropriate machine learning algorithm for the healthcare industry. This study aims to give essential knowledge about the supervised machine learning algorithms utilized in the healthcare industry. As a result of compiling a data table on the efficacy of learning algorithms for various diseases, we were able to compare which algorithm worked best for which specific type of condition. These publications will aid researchers and practitioners in understanding the role supervised machine learning algorithms have in the healthcare industry and the accuracy of various supervised machine learning algorithms.*

Keywords: Machine Learning, Diabetes, Supervised learning, Unsupervised Learning, Semi-supervised Learning, Reinforcement Learning

1. INTRODUCTION

Computers may improve their performance utilizing minimal data thanks to the science of machine learning. This field concentrates on learning and understanding through the study of computer systems. Machine learning works on two methods: training and testing. Machine learning technology has improved extraordinarily in a few decades to predict the disease of patients by looking at their symptoms. Machine learning technology serves as a huge blessing to the healthcare sector in identifying health issues easily. Machine learning uses the various symptoms of the patients as an input and gives an output based on that. Machine learning helps doctors to detect a disease at an early stage. For example, Heart-related problems, Cancer, etc. Healthcare cases depend a lot on machine learning for diagnosis and analysis. As a result, it can be said that healthcare is one of the most important uses of machine learning in the medical field.

2. EXISTING SYSTEMS

Despite the existence of advanced computer systems, medicos still need proficiency in X-rays and surgical operations, however, technology optically discerns it back after understanding. This process still depends on their knowledge and experience in the medical field to understand the factors starting from past medical history, atmosphere, sugar, blood pressure, weather, and some different aspects. A significant number of variables are offered as it takes a lot to comprehend the full process, but no model is prosperously analyzed. We can manufacture models by using machine learning medical records to give outputs efficiently by analyzing the data. Machine learning helps doctors to give many precise decisions for treatment choices, which results in the improvement of patient's health.

3. PROPOSED SYSTEM

The proposed system allows us to detect the disease by analyzing the symptoms. This system utilizes many supervised learning methods for the assessment of the model. Diseases are identified by this system based on the said symptoms. This system is operated with the help of machine learning technology. Diseases like breast cancer, diabetes, kidney diseases, liver diseases, and heart diseases can be predicted using this system.

4. DIABETES DISEASES

Diabetes is a metabolic disease that causes a rise in blood sugar. People having diabetes suffer from dangerous diseases related to the heart, kidneys, and sight; that's why it's a lot more dangerous for them. Lack of energy, feeling thirsty, blurry eyesight, lightweight, urinating more often, etc. are some of the symptoms of diabetes. The health centres accumulate the obligatory data for the diagnosis of the disease through tests and the treatments that are required based on this. The healthcare field depends on the information they gather by analyzing the disease and later operating based on that. This sector is predicated on an immense amplitude of data. Utilizing sizably voluminous data analytics that works with immense data sets and recovers rear data and the connections to process those received data to give an accurate output. Some of the effects and symptoms of diabetes are given below:

- Women suffering from any type of diabetes have high blood sugar during their pregnancy which risks the birth of the baby. The health of the mother can also deteriorate a lot because of this.

- Diabetes causes a huge rise in blood glucose. A huge amount of glucose in the blood can give rise to various health problems after some time.

- Diabetic patients suffer much more from high blood pressure than those without it. Without any treatments, this high blood pressure can cause heart disease and strokes.

- Skin sickness is one of the evident symptoms of diabetes. A high rise in blood glucose leads to inadequate blood flow which causes severe skin damage. Destruction of skin cells leads to high sensitivity towards temperature and pressure.

- Type 1 diabetes happens due to the lack of production of insulin by the cells of the pancreas. As a result, glucose can't provide energy in the body. Thus, type 1 diabetes can take lives if insulin is not injected for its lifetime.

- A diabetic patient needs to have a Body Mass Index (BMI) of 25 or less and maintain a healthy diet to prevent the disease.

- Middle-aged persons have a higher prevalence of type 2 diabetes. However, today's youth are also growing greatly.

- Diabetes Pedigree Function is used to indicate whether the disease has been passed down to the children through their parents or not.

Fig1: Diabetes

5. DIFFERENT TYPES OF LEARNING

5.1 Supervised learning

It is a method where a machine is taught using well-labeled data. This well-labeled data is targeted to train the algorithms into classifying data or predicting the accurate result. The processes under the supervised learning algorithm are

Classification – It is a type of algorithm that classifies a dataset.

Regression –It is a particular kind of supervised learning technique used to forecast continuous outcomes.

5.2 Unsupervised Learning

It is a method in which the machine works without supervision. Since the data, in this case, is unlabeled, the machine must explore the dataset and look for hidden patterns to anticipate the output without human involvement. The types under unsupervised learning are:

Clustering: It is a type of data mining technique for grouping unlabelled data predicated on their homogeneous attributes or differences.

Association: It is another type of unsupervised learning that uses different rules and discovers patterns in data.

5.3 Semi-supervised Learning

Semi-supervised learning refers to learning that combines supervised and unsupervised methods. In semi-supervised learning, the data is fused with a little amount of labeled data with a myriad amount of unlabelled data.

5.4 Reinforcement Learning

Reinforcement learning is a type of machine learning where an agent learns to interact with its circumventing environment by invoking actions and discovering errors or rewards. It's all about taking appropriate action and maximizing as many rewards as possible in a particular situation.

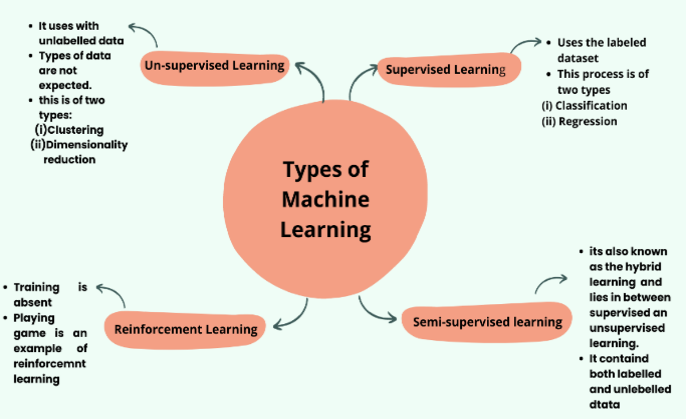
• It uses with unlabelled data
• Types of data are not expected.
• this is of two types:
(i)Clustering
(ii)Dimensionality reduction
Un-supervised Learning
Supervised Learning
• Uses the labeled dataset
• This process is of two types
(i) Classification
(ii) Regression
Types of Machine Learning
• its also known as the hybrid learning and lies in between supervised an unsupervised learning.
• It containd both labelled and unlebelled dtata
• Training is absent
• Playing game is an example of reinforcemnt learning
Reinforcement Learning
Semi-supervised learning

Fig 2: Types of Machine Learning

6. MACHINE LEARNING MODEL DEVELOPMENT TECHNIQUES

The procedures required in creating a machine learning algorithm are as follows:

6.1 Data access and collection

The first step in building a model is to gather data from numerous sources and sort out features that have the most impact on the study. This is a crucial step because the quality and quantity of the information we collect will directly determine how effective the predictive model can be.

6.2 Data preparation

It's time to assess the data now that it has been collected. This aids in avoiding significant problems and ensuring better outcomes. We may also pre-process at this level by doing the following actions:

Data formatting: Data must be recorded in a standard format that our machine learning can comprehend, such as XML, CSV, etc., to use the gathered data in machine learning.

Data Cleaning: Data cleaning is mostly utilized in machine learning algorithms to remove noise, inaccurate data, and duplicate data from our dataset.

Data Reduction: Minimizing the sample data for obtaining a better predictive model in machine learning.

6.3 Data Transformation

After data cleaning, we convert the clean data into another format that will allow us to recover information. This is a crucial step since it improves the prediction model's precision.

6.4 Model Training

In model training, the collection of clean and processed data is divided into training data and testing data. The training dataset is used to train the model, while the testing dataset is used to assess the model's performance on the data. Due to the usage of labeled data here, this characteristic is not present in unsupervised learning.

6.5 Evaluation

The model that was created after the algorithm was tested might now be utilized to make predictions in real-time.

7. TYPES OF SUPERVISED MACHINE LEARNING ALGORITHMS

7.1 LOGISTIC REGRESSION

The statistical method utilized in machine learning is logistic regression. It is more frequently employed to address categorization issues. Logistic regression is used to forecast a dependent variable's categorical output. The result must thus be a discrete or categorical value. It can be either True or False, Yes or No, 0 or 1, etc., but rather than providing the precise values of 0 and 1, it provides the probability values that fall between 0 and 1.

7.2 RANDOM FOREST

Random Forest or random decision forest is constructed by utilizing multiple decision trees and the final decision is obtained by majority votes of the decision trees. It may be utilized to address issues with regression or classification. . The key benefits provided by random forest when used in regression or classification problems are Minimizing overfitting: Decision trees that are extremely massive run the risk of overfitting within training data, thus causing the classification output to drastically differ for a small change in the input value. They are hypersensitive to their training data, which may result in more errors in the test dataset (value).

7.3 NAIVE BAYES

Based on the Bayes theorem, the probabilistic supervised machine learning technique known as Naive Bayes is utilized to resolve classification issues. The Bayes Theorem calculates the probability that an event will occur given the probability that an earlier event will occur. The following equation provides the mathematical version of Bayes' theorem: $P(A|B) =P(B|A) P(A) /P (B)$

According to the equation above, we may say that:

$P(B|A)$ is the posterior probability of the given class where B is the target and A is the attributes. $P (B)$ denotes the class prior probability.

$P(A|B)$ is the chance of predictor given class.

7.4 SUPPORT VECTOR MACHINE

Both classification and regression are performed using a supervised machine learning technique known as the Support Vector Machine (SVM). In the beginning, each data point is mapped into an n-dimensional attribute space (n being the total number of attributes). The process then determines the hyperplane that splits the data points into two classes (divisions), maximizing the margin distance between the classes (divisions) and minimizing classification mistakes. The value of each attribute is then calculated to be equal to the value of that specific coordinate, transforming each data point into a point in an n-dimensional space. The next step is to look for the hyperplane that separates the two classes by the greatest margin.

7.5 K-NEAREST NEIGHBOUR

K-Nearest Neighbour is one of the most straightforward machine learning algorithms that can do both classification and regression tasks. Its foundation is the supervised learning approach. . K-NN makes no assumptions about the underlying data because it is a non-parametric approach. As a result of saving the training dataset rather than instantly learning from it, the method is sometimes referred to as a lazy learner. Instead, it acts while categorizing data by using the dataset. KNN categorizes fresh data into a category that is relatively close to the training data by merely saving the information during the training phase.

7.6 XG BOOST

XG Boost which stands for Extreme Gradient Boosting is a distributed, scalable gradient-boosted decision tree (GBDT) machine learning framework. Artificial neural networks frequently outperform all other algorithms or frameworks in prediction issues involving non-structured data. However, decision tree-based algorithms are currently thought to be best-in-class for small- to medium-sized structured/tabular data

8. Evaluating the model and results

Summarizing the results obtained from our model in the following table:

Disease / Algorithms	Diabetes
Random Forest Classifiers	77
Logistic Regression	80
KNN	74
Naive Bayes	75.97
Support Vector Machine	76
XG Boost	79.87

9. LITERATURE SURVEY

In Paper 1: Disease Prediction using Machine Learning: In the year 2019, this paper was published by Kedar Pingale, Sushant Surwase, Vaibhav Kulkarni, Saurabh Storage, Prof. Abhijeet Karve. In their paper, they used the data provided by the patients as input and gives the probability of disease as output. They have used the Naive Bayes Classifier to predict the disease. With the improvement of technology in the medical field, doctors can now detect a disease at an early stage.

In Paper 2: Diabetes diagnosis using machine learning: In the year 2021, this paper was published by Boshra Farajollahi, Maysam Mehmannavaz, Hafez Mehrjoo, Fateme Moghbeli, Mohammad Javad Sayadi. In their paper, they have discussed the chronic disease diabetes which is dangerous for our health. They wanted to evaluate the performances of logistic regression (LR), decision tree (DT), random forest (RF), and other models for diabetes classification. Through this, they wanted to differentiate between various methods of detecting disease.

Paper 3: Multiple Disease Prognostication Based on Symptoms Using Machine Learning Techniques: This paper was published by Kajal Patil, Sakshee Pawar, Pramita Sandhyan, and Jyoti Kundale in the year 2022. In this paper, they have discussed the importance of health care for every human being. They also made a model which can process output by using a vast source of information. Since the output will be based on the collected data, and that's why it will be different for every person.

In Paper 4: Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence: This paper was published by Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, and Razavi AR in the year 2019. In this paper, they have discussed the models they made to predict the presence of breast cancer by analyzing the information they collected. To evaluate the performance of the models, their accuracy, sensitivity, and other features were compared. The models are used for the early detection of the disease.

10. CONCLUSION

In this study, we investigated the use of ML-based techniques in healthcare. To do this, we first provided an overview of machine learning and its use in healthcare. We categorized machine learning (ML)-based approaches in medicine based on learning techniques (unsupervised learning, supervised learning, semi-supervised learning, and reinforcement learning), data pre-processing approaches (data cleaning approaches, data reduction approaches, and data formatting approaches), and assessment approaches. Because medical data is developing at an increasing rate, current data must be processed to forecast precise illnesses based on symptoms. To categorize patient data, several generic illness prediction systems based on machine learning algorithms, including Logistic Regression, Random Forest, Naive Bayes, Support Vector Machine, K-Nearest Neighbour, and XG Boost have been presented.

11. REFERENCES

1) Pingale, K., Surwase, S., Kulkarni, V., Sarage, S., & Karve, A. (2019). Disease Prediction using Machine Learning.

2) Balasubramanian, Satyabhama, and Balaji Subramanian. "Symptom-based disease prediction in the medical system by using Kmeans algorithm." International Journal of Advances in Computer Science and Technology 3

3) Abu-Jamous, B., Fa, R., Nandi, A.K., 2015a. Feature Selection. Integrative Cluster Analysis in Bioinformatics. John Wiley & Sons, Ltd.

4) A. K. Jain and R. C. Dubes. Algorithms for clustering data. Prentice-Hall, Inc., 1988

5) Tapak L, Mahjub H, Hamidi O, Poorolajal J. Real-data comparison of data mining methods in prediction of diabetes in Iran. Healthc Inform Res. 2013; 19(3):177–85.

6) Ahmad LG, Eshlaghy A, Poorebrahimi A, Ebrahimi M, Razavi A. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013;4(124)

7) Toshniwal D, Goel B, Sharma H. Multistage Classification for Cardiovascular Disease Risk Prediction. In: International Conference on Big Data Analytics; 2015. p. 258–66. Springer.

8) Mustaqeem A, Anwar SM, Majid M, Khan AR. Wrapper method for feature selection to classify cardiac arrhythmia. In: Engineering in Medicine and Biology Society (EMBC), 39th Annual International Conference of the IEEE; 2017. p. 3656–9. IEEE.