"A NEW MEDICAL DECISION SUPPORT SYSTEM FOR DIABETES AND REVERSE DIABETES PREDICTION USING MACHINE LEARNING TECHNIQUES"

Harshitha M[1], Priya S K[2], Yashaswini H S[3], Prajwal J[4], Bhrigu S [5]

[1]*Assistant Professor, Dept. of Information Science & Engineering, VVIET, Karnataka, India*

[2]*Student, Dept. of Information Science & Engineering, VVIET, Karnataka, India*

[3]*Student, Dept. of Information Science & Engineering, VVIET, Karnataka, India*

[4]*Student, Dept. of Information Science & Engineering, VVIET, Karnataka, India*

[5]*Student, Dept. of Information Science & Engineering, VVIET, Karnataka, India*

ABSTRACT

Diabetes, a chronic metabolic disorder, leads to sustained elevation in internal blood glucose levels due to insufficient insulin production or cells' inability to respond to it. Its prevalence and impact on health make handling confidential healthcare data crucial. While diabetes is a one of the global health concern disease type, there's a gap in real-time applications addressing it, especially in early prediction and dietary recommendations. Our project aims to bridge this gap by developing an application that predicts diabetes early on, identifies potential reversals using trending machine learning algorithms like Random Forest, KNN, or Decision Tree, and offers tailored dietary plans. This real-time medical system, developed using Microsoft tools such as Visual Studio and SQL Server, promises to be a valuable asset for hospitals and doctors.

Keyword: ***Model, Random Forest, K-Nearest Neighbors, Diabetes Mellitus, Machine Learning***

1. INTRODUCTION

One of the major reason for patients death is diabetes mellitus, includes a group of metabolic disorders impacting millions globally. Early detection is vital because of severe complications associated with the condition. Numerous studies, often utilizing the Pima Indian diabetes dataset, have investigated diabetes identification, typically focusing on complex techniques without thoroughly comparing common methods. Diabetes is characterized by increased blood sugar levels, leading to symptoms such as more urination, more thirst, hunger, and weight loss. Without ongoing treatment, diabetes can result in life-threatening complications. Diagnosis generally involves measuring a 2-hour post-load plasma glucose level of at least 200 mg/dL, underscoring the importance of timely identification to prevent serious health outcomes. An automation for diabetes prediction using efficient machine learning algorithms is crucial. This real-time application would greatly assist hospitals and healthcare providers in managing patients more effectively. Our proposed system aims to enhance disease prediction processes, enabling healthcare professionals to deliver superior patient care.

2. LITERATURE SURVEY

2.1 IEEE Paper Title: Association Rule Extraction from Medical reports of Diabetic Patients

Year of Publications: 2020

Description: Medical databases are invaluable for enhancing medical diagnosis. With the growth of electronic medical record systems and advancements in medical technology, healthcare institutions generate an increasing amount of medical text data. Unfortunately, much of this data is underutilized, stored without fully leveraging its potential. Proper utilization of this medical information could revolutionize the field of medicine.

This work presents a novel approach to extracting meaningful association rules from medical transcripts. These rules reveal relationships between diseases, symptoms, medications, and patient demographics, such as age groups more susceptible to certain diseases. The methodology combines Natural Language Processing (NLP) tools with data mining algorithms, specifically the Apriori algorithm and FP-Growth algorithm, to uncover these associations. Additionally, correlation measures like lift are used to identify interesting rules, facilitating the extraction of relevant insights from the data.

Drawbacks

- ✧ Used to find the relationships among various diabetes parameters.
- ✧ Provides less accurate results.
- ✧ Not suitable for predicting diabetes.

2.2 IEEE Paper Title: Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers

Year of Publications: 2021

Description: Diabetes, a chronic condition characterized by increased high blood sugar levels, poses significant health risks that can be mitigated with early and accurate prediction. However, predicting diabetes accurately is challenging due to limited labeled data and the presence of outliers or missing values in datasets. This study proposes a robust framework for diabetes prediction, which includes outlier rejection, missing value imputation, data standardization, feature selection, K-fold cross-validation, and the use of efficient machine learning algorithms such as k-nearest neighbor algorithm, decision tree, random forest algorithm, AdaBoost, Naive Bayes, XGBoost, and multilayer perceptron (MLP).

Additionally, we introduce a weighted ensembling technique to enhance prediction accuracy, with weights determined by the Area Under the ROC Curve (AUC) of each machine learning model. The AUC is used as the performance metric and is optimized through hyperparameter tuning using grid search. This comprehensive approach aims to enhance the robustness and accuracy of diabetes prediction, providing valuable insights for early intervention and management of the condition.

Drawbacks

- ✧ The model is constructed using ML algorithms.
- ✧ It cannot be used in real-time applications.
- ✧ It does not predicts reverse diabetes prediction.

2.3 IEEE Paper Title: A Novel Approach to Predict Diabetes by Using Naive Bayes Classifiers

Year of Publications: 2020

Description: Diabetes is a potentially life-threatening chronic condition marked by elevated blood glucose levels. This study aims to analyze a diabetic patient database to facilitate early disease prediction. The experiment uses the Naïve Bayes algorithm method for finding diabetes.Data mining involves discovering useful information from datasets and transforming it into an analyzable format. The training datasets used in this study contains 1865 instances of diabetic patients, each with various attributes, collected from hospital records. The results demonstrate that the proposed method achieves a higher accuracy (0.96) in predicting diabetes compared to traditional methods.

The system employs the Naïve Bayes Classifier to provide output through a web interface, indicating whether an individual has diabetes based on input parameters such as insulin levels and age. This approach significantly enhances the accuracy and usability of the prediction system.

Drawbacks

- ✧ The model is developed using machine learning algorithms.
- ✧ It is not suitable for real-time use.
- ✧ It does not support reverse diabetes prediction.

2.4 IEEE Paper Title: Multi-Agent System Based on ML for Early Diagnosis of Diabetes

Year of Publication: 2020

Description: Diabetes is a growing global concern, with Morocco alone witnessing a diagnosis of over 2 million individuals aged 18 and older. Detecting diabetes early is paramount for effective diagnosis and treatment. In response, this qualitative investigation aims to refine diagnostic procedures by empowering specialized software modules for medical diagnosis with greater autonomy and initiative. A Multi-Agent System (MAS) is proposed as a robust and dependable tool for distributed diagnostics. This study endeavors to develop a novel MAS that assesses the performance of three prominent machine learning algorithms: artificial neural networks (ANN), support vector machines (SVM), and logistic regression (LR), utilizing a diabetes database. The system will utilize a controller agent to aggregate classifications from these algorithms via a majority voting mechanism, thereby enhancing classification accuracy. Moreover, the research explores the present challenges and gaps in integrating ML algorithms within multi-agent systems.

Drawbacks

- ✧ SVM and regression ML algorithms are employed to generate graphical results.
- ✧ Not suitable for real-time applications.
- ✧ The developed model lacks real-time applications useful for hospitals.

3. LITERATURE SURVEY/GAP ANALYSIS

Many research works done on predicting diabetes disease using efficient machine learning algorithms and methods, few works just presented idea and few works done implementation. Several papers have utilized tools like Python, R, and Weka, focusing on static datasets lacking real-time applications or the ability for reverse diabetes prediction. In real-world scenarios, patients typically undergo manual diagnoses by doctors based on various tests, a process demanding substantial medical expertise, time, equipment, and expense.

The proposed system introduces a novel approach by aiming to predict and reverse diabetes using machine learning algorithms. Unlike prior models dependent on static data, it operates with dynamic datasets. Designed as a GUI-based application for hospitals, it streamlines usage for both doctors and patients, a unique development. This real-time application harnesses machine learning models for immediate predictions, a feature not previously realized. With a simple button click, the system predicts both diabetes and reverse diabetes. It employs a broader range of parameters and larger datasets for prediction, thereby enhancing accuracy and reliability.

Drawbacks

- Reverse diabetes prediction is not included.
- Previous works are confined to models.
- Lack of real-time implementations.
- Utilization of static datasets.
- Demands more time.
- Requires medical equipment.
- Higher cost implications.

4. MAJOR OBJECTIVES OF PROPOSED WORK

1. **To build an automated medical system which will automatically predicts diabetes disease using efficient supervised learning techniques.**

   Methodology: We use Visual Studio for front-end design in developing an automated system. Being a Microsoft technology, Visual Studio is well-suited for constructing real-time projects due to its comprehensive library support and the availability of essential packages.

2. **To program data preprocessing algorithms to process data and removing irrelevant data using "binning method"**

   Methodology: Data preprocessing involves the application of the binning method to filter out irrelevant data and extract pertinent information for subsequent analysis. This technique is highly beneficial for handling medical datasets and can also be utilized to address missing values.

3. **To develop a real time application with machine learning model to predict diabetes patients using machine learning algorithms.**

   Methodology: Following data preprocessing, the refined data undergoes processing through the KNN (K-Nearest Neighbors) algorithm, an efficient supervised learning algorithm. Specifically engineered to manage numerical data, the KNN algorithm is distinguished for its rapid processing capabilities. It is well-suited for both small and large datasets and is employed to forecast diabetes occurrences in patients.

4. **To develop a real time application with machine learning model to predict reverse diabetes patients using machine learning algorithms.**

   Methodology: After diabetes prediction, the data undergoes further processing using the Naïve Bayes algorithm, another efficient supervised learning method. Renowned for its versatility across different data types, the Naïve Bayes algorithm boasts rapid processing capabilities. It accommodates a wide array of parameters and is leveraged to forecast the potential reversal of diabetes in patients.

5. PROPOSED METHODOLOGY

Medical datasets containing disease information undergo processing through machine learning algorithms. Particularly, supervised learning algorithms are utilized to scrutinize these datasets, with the goal of attaining high accuracy and efficiency. Through the application of these sophisticated algorithms, enhanced outcomes are achieved.

Diabetes and Reverse Diabetes Prediction Process

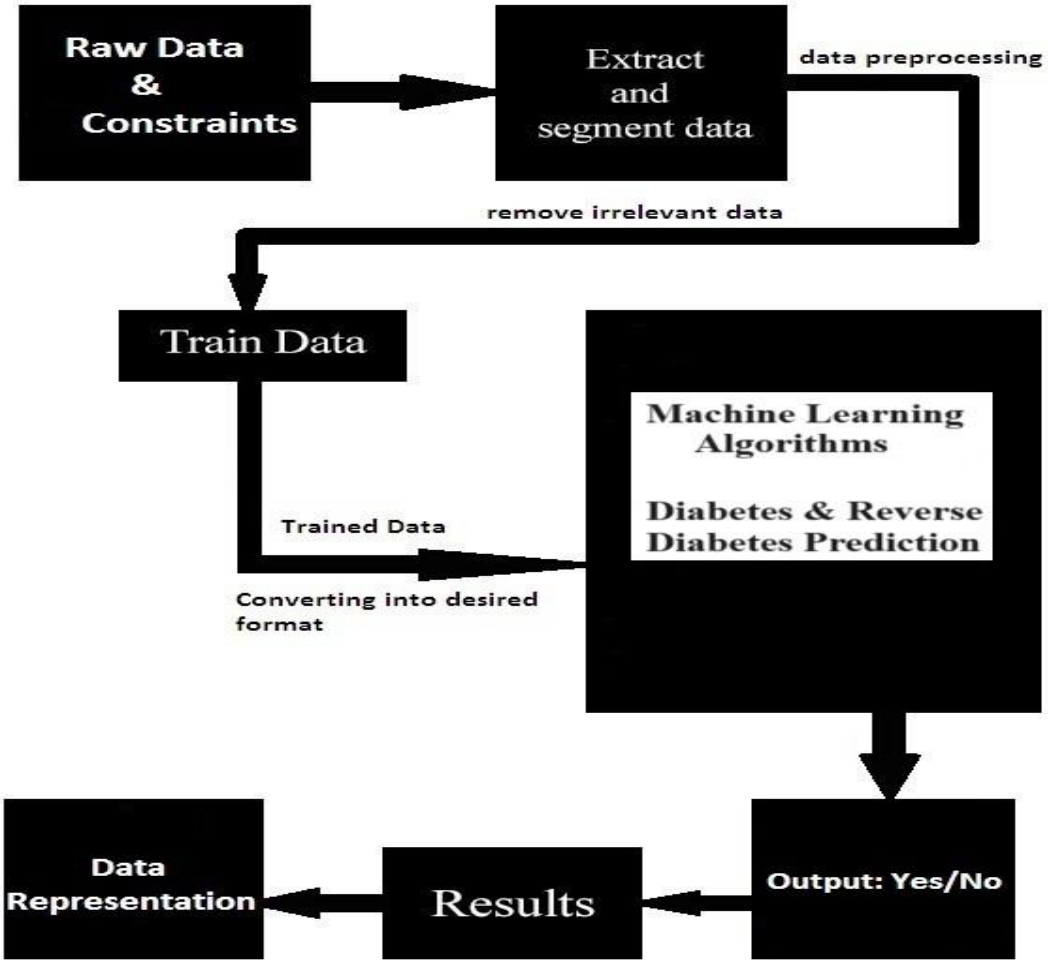

Fig -1: Methodology

Step 1: Training datasets essential for processing and predicting results are initially sourced from diverse online platforms such as Kaggle, Dataworld, Data.gov.in, GitHub, and others. These platforms offer raw data directly, facilitating the collection of pertinent information for analysis and modeling.

Step 2: Following dataset collection, it is imperative to comprehend and preprocess the raw data. During this phase, we tackle missing values, eradicate irrelevant information, and extract relevant data. Extraneous details like serial numbers, patient IDs, mobile numbers, and addresses are eliminated, retaining only the essential data for subsequent processing.

Step 3: After completing data preprocessing, the refined dataset is inputted into machine learning models. Supervised learning algorithms are utilized to construct models with the objective of predicting and potentially reversing diabetes.

Step 4: The ML model is trained using algorithms such as the KNN algorithm, Decision Tree algorithm, and Naïve Bayes algorithm. Both diabetes prediction and reverse diabetes prediction are carried out using these algorithms.
Step 5: Initially, the ML model predicts whether a patient has diabetes (YES or NO). If the patient is classified as YES, then another machine learning model is utilized to predict the potential reversal of diabetes.

Step 6: To assess the machine learning model, we partition the datasets into training and testing sets using a 90:10 ratio. Subsequently, we evaluate the model's performance and compute its accuracy.

Step 7: The final results are presented on a graphical user interface (GUI), and data visualization techniques are utilized.

Step 8: Patients will receive appropriate treatments to cure the disease at the earliest possible stage.

6. ALGORITHMS USED

6.1 KNN Algorithm Steps

How KNN works ?

1. Determine K (no of nearest neighbors)

2. Calculate distance(Euclidean, Manhattan)

3. Detemine K-minimum distance neighbors

4. Gather category Y values of nearest neighbors

5. Use simple majority of nearest neighbors to predict value of qurey instance

6.2 Random Forest Algorithm Steps

Operation of Random Forest

The working of random forest algorithm is as follows.

1. A random seed is chosen to facilitate the random sampling of a subset of data points from the training dataset. This ensures that the original distribution of classes within the dataset is maintained during the sampling process. From this selected dataset, a random subset of attributes is chosen based on user-defined criteria, not all input variables are considered to avoid excessive computation and reduce the risk of overfitting.

2. In a dataset with M input attributes, a random selection process is employed to choose R attributes for each decision tree, where R is a value less than M.

3. Utilizing this subset of attributes, the decision tree model is constructed by iteratively identifying the best split at each and every node based on the Gini index value. This method or recursive process continues until a stopping criterion is reached, indicating that the resulting leaf nodes are too small to further subdivide.

7. EXPERIMENTAL RESULTS

7.1 KNN Algorithm Results

Discussion

We have developed a real-time application aimed at benefiting society, leveraging Microsoft technologies. Our project focuses on utilizing diabetes datasets trained using the K-nearest neighbors (KNN) algorithm, yielding highly promising results. Our KNN algorithm implementation is designed to accommodate dynamic datasets efficiently. With our proprietary KNN library, we have achieved an impressive accuracy rate of approximately 93.2%. Moreover, our prediction process operates swiftly, typically completing within 1500 milliseconds.

Constraint	KNN Algorithm
Accuracy	97.18 %
Time (milli secs)	1606
Correctly Classified (precision)	97.18 %
InCorrectly Classified (Recall)	2.82 %

7.2 RF Algorithm Results

Constraint	RF Algorithm
Accuracy	91.18 %
Time (milli secs)	2606
Correctly Classified (precision)	91.18 %
InCorrectly Classified (Recall)	8.82 %

8. CONCLUSION

Diabetes stands as a major contributor to global mortality rates. Detecting this condition early can significantly impact outcomes, prompting the building of machine learning models. Our system is dedicated to pinpointing and potentially reversing diabetes by analyzing specific parameters. It aids healthcare professionals in forecasting diabetes during its nascent stages, facilitating prompt and tailored interventions. Leveraging a variety of ML methods enhances the accuracy of predictions. Through automation, our system employs efficient data science or machine learning algorithms, notably utilizing the effective supervised learning technique Naive Bayes. This approach efficiently processes medical data to produce predictive insights.

9. REFERENCES

1. "Machine Learning Techniques for detection and classification of Diabetes and Cardiovascular Diseases", Year of Publications: 2020.

2. "Diabetes Prediction Using Assembling of Different Machine Learning Classifiers", Year of Publications: 2021

3. "A Novel Approach to Predict Diabetes by Using Naive Bayes Classifiers", Year of Publications: 2020

4. "Multi-Agent System Based on Machine Learning for Early Diagnosis of Diabetes", Year of Publications: 2020