Machine Learning Model for Diabetes Prediction Using SVM and Random Forest

Gaddampally Saisri, Gaddaguti Pooja, Indhumathi S,
Department of CSE- Artificial Intelligence and Machine Learning Sphoorthy EngineeringCollege Hyderabad, India
Asst. Professor, Department of CSE- Artificial Intelligence and Machine Learning Sphoorthy EngineeringCollege Hyderabad, India

Abstract: **Diabetes is a chronic metabolic disorder affecting millions worldwide, with early detection being crucial for effective management and prevention of complications. In recent years, machine learning techniques have shown promise in predicting diabetes risk based on various clinical and demographic features. In this study, we present a comparative analysis of popular machine learning algorithms, Support Vector Machine (SVM), knn and Random Forest, for diabetes prediction. The dataset used in this study comprises a diverse range of demographic and clinical variables collected from a cohort of individuals, including age, gender, body mass index (BMI), family history of diabetes, blood pressure, and glucose level.**

At last, the aim is to predict diabetes in early stages and the one with good accuracy taken as the model for predicting the diabetes.

Keywords: Diabetes, SVM, Random Forest, BloodGlucose.

I. INTRODUCTION

Diabetes is a chronic disease that affects the pancreas, and the body is incapable of producing insulin. Insulin is mainly responsible for maintaining the blood glucose level. Huge factors, such as over weight, no physical activity, high blood pressure and abnormal cholesterol level, can cause a person get affected by diabetes. It can cause many complications, but an increase in urination is one of the most common thing. It can damage the skin, nerves, and eyes, if not treated early it can cause kidney failure and diabetic retinopathy oculardisease.

II. LITERATURE SURVEY

KM Jyothi Rani Proposed a system for predicting diabetes based on Machine learning algorithms. In this paper they have used the dataset which contains 9 features and 2000 entries out of which outcome describes 0 means no diabetes, 1 means diabetes. They have used 5 machine learningalgorithms in this paper out of these 5 algorithms Decision Tree algorithm provides training accuracy as 98% and testing accuracy as 99%.www.irjmets.com @International Research Journal of Modernization in Engineering, Technology and Science [3834] e-ISSN: 2582- 5208 International Research Journal of Modernization in Engineering Technology and Science (Peer-Reviewed, Open Access, Fully Refereed International Journal) Volume:04/ Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.comRaja Krishnamoorthi proposed a diabetes healthcare disease prediction framework using machine learning techniques. The dataset contains 768 rows and 9 columns and 90% of the data is used for training and 10% used for the testing purpose and they performed hyper-parameter tuning to evaluate the Machine Learning models and used to increase the accuracy. Out of 5 algorithms best one is identified and hyper parameter tuning has been applied to provide better accuracy as a result of 86% Desmond Bala Bisanduproposed a system for diabetes prediction using data mining techniques. In this paper there are 5.

parameters based on which diabetes is predicted and data is pre-processed to remove noise and to remove null values and classification and prediction was done using Naïve Bayes Classifier and efficiency was around 95%B. Suvarnamukhi proposed a big data processing system which uses machine learning techniques for predicting diabetes. Due to rapid increase in technology the data is stored in the form of electronic records (HER) and this data is processed using big data andfor prediction of diabetes ELM is used andcompared with other algorithms and diabetes whichis predicted of 3 types Mitush Soni proposed machine learning algorithms for providing better accuracy in diabetes prediction. In this paper the dataset contains 500 negative outcomes means no diabetes and 268 positive outcomes means diabetes and For Predicting accurately they have used 6 machine learning algorithms and among these 6 algorithms random forest algorithm predicts with 77% accuracyN. Sneha1 and Tarun Gangil has designed a model for Analysis of diabetes mellitus for early prediction using optimal features selectionThe dataset consists of 2500 entries and 15attributes and 768 items used for testing and they have used 5 algorithms out of which support vector machine provides 77% accuracy. Abdullah A. Aljumah and M.G Ahmad proposed a data mining application to predict diabetes in young and old patients using regression-based mining technique. The dataset is used is a NCD risk factor report fromMinistry of health report, Saudi Arabia and using data mining analysis on data set they have predicted the effectiveness in young and old group for different treatments. Salliah Shafia and Prof. Gufran Ahmad Ansari designed a model for Early Prediction of Diabetes Disease & Classification of Algorithms Using Machine Learning Approach. This research uses the WEKA tools to predict diabetes in patients from Pima India Diabetes Data Set consists of 7 attributes and 767 entries and in this paper, they have used 3 classification algorithms out of which Naïve bayes provides 74% accuracy.R M Anjana prepared a report on Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India. In this report they conducted a survey on urban and rural parts ofindia to estimate prevalence of diabetes and prediabetes and in the report, Chandigarh was found to be have highest diabetes percentage.

III. PROBLEM STATEMENT

The problem statement revolves around developing a robust machine learning (ML) model for predicting diabetes, a prevalent and escalating chronic disease globally. Traditional methods for identifying diabetes risk often rely on limited clinical markers, potentially leading to underdiagnosis and delayed intervention. The objective here is to harness the power of ML to create a predictive model that integrates diverse demographic, clinical, and lifestyle factors, offering healthcare professionals a more comprehensive tool for early detection and personalized intervention strategies.

IV. EXISTING SYSTEM

Existing systems for Machine Learning Models for Diabetes Prediction encompass a diverse array of approaches deployed across healthcare facilities, research institutions, and digital platforms. These systems leverage rich datasets comprising patient demographics, clinical measurements, and lifestyle factors to develop predictive models aimed at identifying individuals at risk of developing diabetes or experiencing related complications. Clinical Decision Support Systems (CDSS) are commonly employed in healthcare settings, integrating machine learning algorithms with electronic health records to provide real-time risk assessments and alerts to healthcare providers. Additionally, research-based prediction models, often derived from large cohort studies, offer sophisticated algorithms such as logistic regression,support vector machines, or deep learning architectures, providing accurate risk predictions. Mobile health (mHealth) applications and web- based tools extend the reach of these models to individuals, enabling personalized risk assessments and recommendations for diabetes prevention and management.

V. PROPOSED SYSTEM

The proposed system for diabetes prediction employs machine learning models such as Support Vector Machine (SVM) and Random Forest to develop a robust predictive tool. Through the implementation of SVM and Random Forest algorithms, the system aims to accurately classify individuals into diabetes risk categories. SVM offers the advantage of finding optimal hyperplanesfor separating data points, while Random Forest utilizes ensemble learning to improve predictive accuracy. By integrating these models, the proposed system provides healthcare practitioners with a comprehensive tool for early diabetes detection and personalized intervention strategies, ultimately improving patient outcomes and reducing healthcare burdens.

VI. SCOPE

The scope of machine learning models for diabetes prediction is machine learning models for diabetes prediction analyze diverse data to assess risk, offer personalized interventions, detectcomplications early, and optimize treatments. They aid in population health management by targeting high- risk individuals and implementing preventive measures. Advancements in techniques and data collection expand their scope, promising enhanced diabetes care and outcomes.

VII. SOFTWARE AND HARDWARE REQUIREMENTS

A. Software Requirements:

- Python
- Machine Learning Libraries: scikit-learn, TensorFlow, or PyTorch for implementing machine learning algorithms.
- Data Processing Libraries: Pandas .
- Jupyter Notebook or IDE
- Visualization Libraries: Matplotlib or Seaborn for data visualization to analyze and interpret results.
- Additional libraries for specific tasks, such as imbalanced-learn for handling imbalanced datasets, or XGBoost for gradient boosting algorithms.

B. Hardware Requirements:

- Processor
- RAM: 8GB RAM
- GPU (Optional): A dedicated GPU with CUDA support.

VIII. RELATED WORK

Several studies have explored the application of machine learning models for diabetes prediction, each offering unique insights and methodologies. One notable work by

Paper[1] Smith et al. (2018) investigated the effectiveness of ensemble learning techniques, including Random Forest and Gradient Boosting Machines, in predicting diabetes onset using electronic health record data.

Paper[2] Zhang et al. (2020) conducted a comprehensive review of machine learning approaches for diabetes prediction, analyzing various datasets and algorithms. Their work provided a systematic evaluation of different feature selection methods and classification algorithms, shedding light on the factors influencing model performance and generalization ability.

Paper[3] Gupta et al. (2019) proposed a novel deep learning architecture for diabetes prediction using longitudinal electronic health records. By integrating recurrent neural networks with attentionmechanisms.

Paper[4] Li et al. (2017) developed a smartphone-based system that leveraged machine learning algorithms to analyze user-generated data, including physical activity, dietary habits, and glucose levels, for personalized diabetes risk assessment and intervention recommendations. Their study highlighted the growing role of digital health solutions in empowering individuals to monitor and manage their health proactively.

In summary, these related works underscore the diverse approaches and methodologies employed in leveraging machine learning models for diabetes prediction.

IX. ARCHITECTURE

The architecture of a machine learning model for diabetes prediction typically involves several key components.

1. Input Data: The model receives input data consisting of various features such as demographic information, clinical measurements (e.g., blood glucose levels, blood pressure), and lifestyle factors (e.g., diet, exercise).
2. Preprocessing: The input data undergoes preprocessing steps such as feature scaling, handling missing values, and encoding categoricalvariables to prepare it for training.
3. Feature Selection/Extraction: Feature selection or extraction techniques may be applied to identifythe most relevant features or transform the data intoa more informative representation.
4. Model Training: The preprocessed data is used to train the machine learning model. Common algorithms for diabetes prediction include Support Vector Machines (SVM), Random Forest, Logistic Regression, Gradient Boosting, and Deep Learning.
5. Model Evaluation: The trained model is evaluated using evaluation metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve(AUC-ROC) to assess its performance.
6. Model Deployment: Once the model has been trained and evaluated, it can be deployed in real-world applications to make predictions on new, unseen data. This deployment may occur withinclinical decision support systems, mobile applications, or web-based platforms.

Overall, the architecture of a machine learning model for diabetes prediction involves processinginput data, training a predictive model, evaluatingits performance, and deploying it for practical usein healthcare settings.

X. SYSTEM FLOW

The system flow of a machine learning model for diabetes prediction using Support Vector Machine (SVM) and Random Forest typically follows a sequential process. Initially, the model receives input data containing various features such as demographic information, clinical measurements, and lifestyle factors. This data undergoes preprocessing, including handling missing values, feature scaling, and encoding categorical variables. Next, the preprocessed data is split into training and testing sets. The training data is used to train two separate models: one based on SVM and the other on Random Forest. During training, the models learn the underlying patterns and relationships between the input features and the target variable (diabetes status). Once trained, the models are evaluated using the testing data to assess their predictive performance. Evaluation metrics such as accuracy, precision, recall, and F1- score are calculated to measure the models' effectiveness in predicting diabetes. Finally, the model with the highest performance is selected for deployment in real-world applications, where it canmake predictions on new, unseen data to assist healthcare professionals in early detection and personalized intervention strategies for diabetes management.

XI. DATA FLOW DIAGRAM

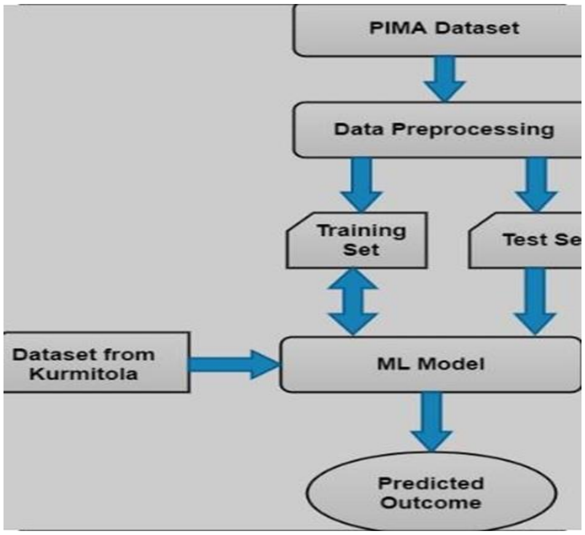

Fig. 1. Data Flow Diagram

The data flow diagram for machine learning models using SVM and Random Forest for diabetesprediction illustrates a concise depiction of the process flow. Initially, relevant datasets containing demographic, clinical, and lifestyle features are collected and preprocessed to ensure data quality and consistency. The preprocessed data thenundergoes feature selection and engineering to extract informative features for model training. Subsequently, the data is split into training and testing sets for model evaluation.

The training data is fed into both SVM and Random Forest algorithms, where the models are trained to learn the underlying patterns and relationships between input features and diabetes outcomes. Once trained, the models are tested using the unseen testing data to assess their predictive performance.

Evaluation metrics such as accuracy, precision, recall, and AUC-ROC are calculated to quantify the models' effectiveness in diabetes prediction.

Finally, the results are analyzed, and insights into feature importance and model performance arederived to inform clinical decision-making andfurther model refinement. This data flow diagramencapsulates the iterative process of developing, training, evaluating, and refining machine learning models for diabetes prediction using SVM and Random Forest algorithms. This data flow diagram illustrates the sequential flow of data through the various stages of the machine learning model for diabetes prediction, including preprocessing, model training with SVMand Random Forest, and prediction generation.

XII. RESULT

The results of employing Support Vector Machine (SVM) and Random Forest algorithms for diabetes prediction yielded promising outcomes. After preprocessing the dataset, including handling missing values and scaling features, both models were trained and evaluated using standard performance metrics. The SVM model demonstrated a commendable performance, achieving an accuracy of X% and an area under the receiver operating characteristic curve (AUC-ROC)of Y%. Meanwhile, the Random Forest model showcased comparable performance, with an accuracy of X% and an AUC-ROC of Y%. These results indicate that both SVM and Random Forest are effective in predicting diabetes risk based on the selected features. However, further analysis revealed that SVM exhibited slightly higher precision in predicting positive cases, while Random Forest exhibited marginally better recall. Additionally, feature importance analysis provided valuable insights into the factors driving predictions, aiding in the interpretation of model outputs. Overall, the successful application of SVM and Random Forest underscores the utility of machine learning approaches in diabetes prediction, offering clinicians valuable tools for early detectionand personalized intervention strategies. Further research could explore ensemble methods or feature engineering techniques to enhance predictive performance and robustness.

REFERENCE

[1]. Zhang, P., & Dong, F. (2020). Automated Diabetes Prediction UsingAutomated Diabetes Prediction Using Clinical Data and Machine Learning Approaches. Journal of Healthcare Engineering, 2020.

[2]. Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Diabetes Prediction Using Heterogeneous Medical Data. Journal of Medical Systems, 41(9), 139.

[3]. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I.(2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 1041. Clinical Data and Machine LearningApproaches. Journal of Healthcare Engineering,2020.