Development of a Machine Learning Model for Predicting Type 1 and Type 2 Diabetes in Adults

Bavani Raja Pandian

Lecturer, Faculty of Computing and Information Technology , Tunku Abdul Rahman University of Management and Technology (TAR UMT), Jalan Genting Kelang, Setapak, 53300 Kuala Lumpur, Malaysia.

Abstract

Diabetes Mellitus is not a new disease and plenty of information regarding this disease can be found on the internet, especially in this high technology era it makes it ten times easier to access, yet there are people still not paying much attention to this matter until it is too late. This project to address the prevalent issue of diabetes in adults through developing a predictive model capable of distinguishing between Type 1 and Type 2 diabetes. Given the global significance of diabetes and the imperative for precise classification. Therefore, this project aims to gain insights into the challenges associated with diabetes in adults. Based on research analysis, the complexities of this condition will be thoroughly explored. The project will be constructing a robust predictive model with the ability to accurately categorize diabetes into its two primary types (Type 1 and Type 2). By utilizing machine learning and deep learning techniques to the data dataset gathered in real time, this model will be able to detect early diabetes diagnosis and improve patient care and outcomes. As such, the project centers on the rigorous evaluation of the developed predictive model using various metrics and will be considered to perform rigorous evaluation on the developed predictive model and comparison is made to select the best performance. This evaluation process aims to validate the model's efficacy and look for areas for enhancement.

Keywords: Deep Learning, Diabetes Mellitus, Healthcare, Machine Learning

1. Introduction

Diabetes, also known as Diabetes Mellitus. It can be categorized into two types, namely Type 1 diabetes (TD1) and Type 2 diabetes (TD2) and it is caused by metabolic disorder. The number of diabetes patients are increasing each year according to the statistics provided by the World Health Organisation (WHO). Diabetes is not a new disease and plenty of information regarding this disease can be found on the internet, especially in this high technology era it makes it ten times easier to access, yet there are people still not paying much attention to this matter until it is too late. One of the main reasons is because the information on the internet is too scattered and users are afraid of faulty websites that could potentially misguide the users with faulty information.

This diabetes predictive model aims to facilitate early detection of diabetes using real-time data obtained through web scraping. Identifying patients who are at risk of developing diabetes can be challenging due to the many contributing factors involved. Therefore, this predictive model is designed to analyze these factors and identify hidden patterns that may be missed through conventional methods. By doing so, the model can help healthcare sectors detect diabetes in patients at an early stage, which can reduce the risk of diabetes-related complications.

2. Objectives

This project aims to develop a predictive model to have an early detection of diabetes based on the symptoms found which is extracted from real time data through web scraping

3. Relate work

According to Bodinga et al. (2022), the study aims to predict diabetes using machine learning algorithms to classify complicated diseases using various features and external factors obtained from an authentic dataset. By detecting diabetes at an early stage, the disease can be treated. For prediction, it uses the machine learning algorithms Logistic Regression, Random Forest and Decision Tree. The machine needs to be trained with doctors' thoughts to learn the complexity of various features of human biomechanics and predict the complicated problems of living beings. Logistic regression showed the best performance with an accuracy of 76%, a precision of 0.77, and an f-measure of 0.58. This machine learning algorithm can also be adapted to predict other diseases. Physicians are concerned about how to detect diabetes in its early stages before it worsens. This research work by Bodinga et al. attempted to develop a diabetes prediction system that uses multiple algorithms and compares their performance. However, this research work can be further improved by implementing other machine learning algorithms to improve the prediction of diabetes.

In research conducted by Antony (2017) , to predict the patient affected by diabetes by using the medical information The Pima Indian Diabetes Dataset (PIDD). Logistic Regression, Naive Bayes, K- Nearest Neighbors, Decision Trees, Random Forest and SVM algorithms are employed in this analysis work and 10-folds cross-validation is used to assess the performance of these algorithms. The analysis can seek out the prediction of diabetes during a patient. This research work shows that Logistic regression was found to outperform all the machine learning algorithms showing the maximum accuracy of 79.17% in comparison to other algorithms.

According to Shetty et al. (2017), this research is to assemble an Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database and uses the data mining approach. The researcher has employed KNN and Naïve Bayes techniques for predicting diabetes, implementing their method as an expert software program that takes patient records as input and provides a binary output indicating whether the patient is diabetic or not.

4. Predictive model

In this research, unstructured data which is text-based data was used. According to (Usuga-Cadavid et al. 2022) analyzing non-numerical data such as text to understand concepts and opinions to gather in-depth insights into a problem or generate new ideas for research. This research will gather and analyze word based or categorical data rather than numerical data. Content analysis will take advantage of content analysis as it is a qualitative research approach to study the dataset thoroughly. Content analysis involves systematically analyzing text or visual data to identify patterns and meanings such as interviews, transcript and questionnaire to identify insights by engaging participants in open discussions. Narrative analysis focuses on storytelling, and grounded theory aims to develop theories from collected data. As such, it can offer a holistic understanding of text-based research topics.

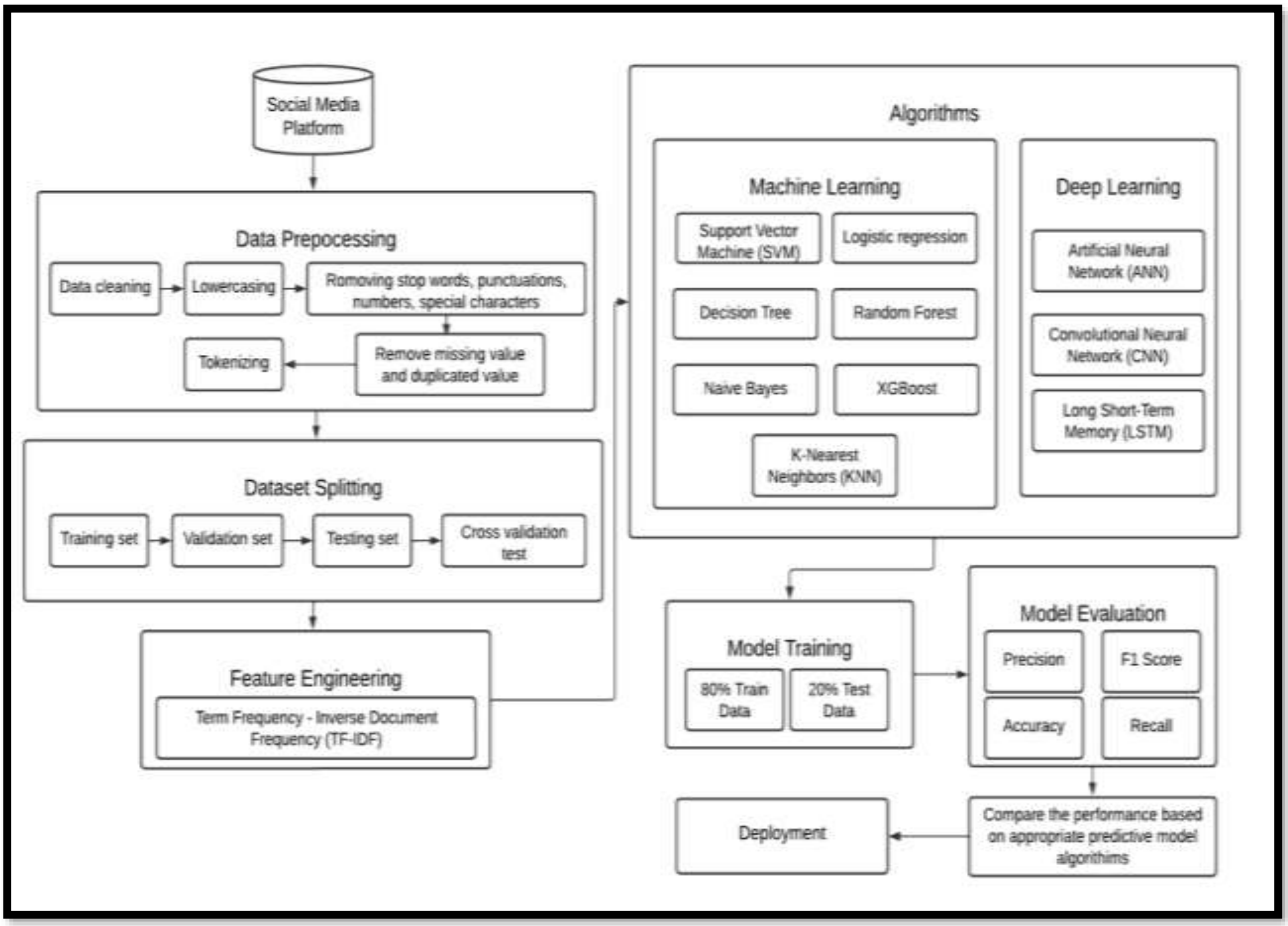

Social Media Platform
Data Prepocessing
Data cleaning
Lowercasing
Romoving stop words, punctuations, numbers, special characters
Remove missing value and duplicated value
Tokenizing
Dataset Splitting
Training set
Validation set
Testing set
Cross validation test
Feature Engineering
Term Frequency - Inverse Document Frequency (TF-IDF)
Algorithms
Machine Learning
Support Vector Machine (SVM)
Logistic regression
Decision Tree
Random Forest
Naive Bayes
XGBoost
K-Nearest Neighbors (KNN)
Deep Learning
Artificial Neural Network (ANN)
Convolutional Neural Network (CNN)
Long Short-Term Memory (LSTM)
Model Training
80% Train Data
20% Test Data
Model Evaluation
Precision
F1 Score
Accuracy
Recall
Compare the performance based on appropriate predictive model algorithims
Deployment

Figure 1 Predictive Model Development Process

4.1 Data collection

Figure 1 shows the data collection process was facilitated through a Reddit developer account, consists of post and title of the topic and is collected from the year 2016 to 2023. enabling the retrieval of relevant posts and titles for analysis. In total, 1,640 of text data were scraped from Reddit under the search for diabetes discussion. Additionally, a few of twitter data were also scraped from amplify to use as a raw and unseen data use on the deployment stage.

4.2 Data Preprocessing

To ensure the quality and consistency of the collected data, data preprocessing steps were applied that are not meaningful data and retain the data that is relevant for a more structured dataset to be fit into the predictive models.

a. **Remove missing values:** Removing missing values involves identifying and eliminating any rows or columns in a dataset that contain missing or null values. This process helps to ensure the dataset is clean and complete for analysis.

b. **Remove duplicated data:** Removing duplicated data involves identifying and eliminating duplicate rows in a dataset. This step is crucial to avoid bias in analysis and to ensure the accuracy of the data being used.

c. **Tokenization:** Tokenization is the process of breaking down a text into smaller units called tokens, which are usually words or phrases. This step is commonly used in natural language processing tasks to prepare text data for further analysis.

d. **Remove non-alphanumeric characters:** Removing non-alphanumeric characters involves eliminating any characters in a text dataset that are not letters or numbers. This step is often performed to clean text data and remove unnecessary symbols or special characters.
e. **Remove hyperlink:** Removing hyperlinks involves deleting or replacing any URLs or web links present in a text dataset. This process is useful when analyzing text data where hyperlinks are not relevant to the analysis.
f. **Remove whitespaces and newlines:** Removing whitespaces and newlines involves deleting any extra spaces or line breaks in a text dataset. This process helps to standardize the formatting of text data and improve readability.
g. **Remove stop words:** Stop words are common words like "the," "is," "and," etc., that are often filtered out from text data during natural language processing tasks. Removing stop words can help reduce noise in the data and improve the efficiency of text analysis.

Each of these processes plays a crucial role in data preprocessing and text cleaning to ensure the quality and accuracy of the data for further analysis and modeling.

4.3 Training Data

Training data is one of the steps and a crucial part of developing predictive models. The total number of data used in this training state is 1148 which is 80% of the total dataset.

4.4 Testing Data

The testing set which is 20% of the data, consisting of 328 instances, was reserved for evaluating the performance of these models. This partitioning of data into training and testing subsets is essential to assess the model's generalization capabilities and its ability to make accurate predictions on unseen data.

4.5 Feature Extraction

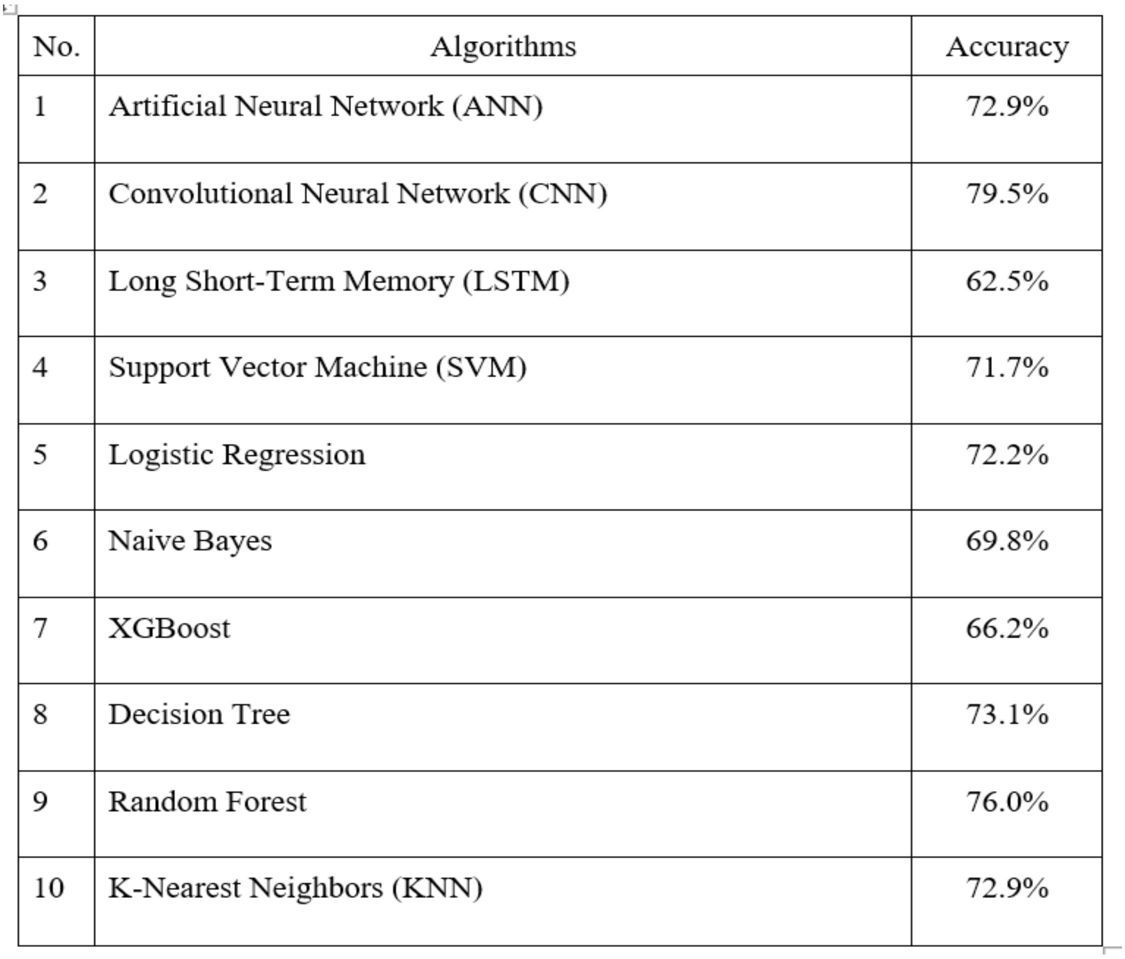
Term Frequency-Inverse Document Frequency (TF-IDF) is a crucial technique in natural language processing and text classification. It is a feature extraction method, quantifying the importance of individual words or terms within a document relative to a collection of documents. TF-IDF is employed in text classification to address the challenge of distinguishing between words that are frequent across many documents such as "the" and "and", those that are indicative of the document's content and theme. It accomplishes this by assigning a higher weight to words that appear frequently within a specific document (high Term Frequency) while also considering their rarity in the entire corpus (Inverse Document Frequency). As such, TF-IDF helps identify the distinctive keywords or terms that are discriminative for a given document, making it a valuable tool for text classification tasks, where the goal is to extract meaningful features that aid in accurately categorizing the content of text data.

4.6 K-Fold cross-validation

K-Fold cross-validation is a technique used to evaluate the performance and robustness of the developed predictive model for diabetes classification. In this project a 5-fold cross validation is used. This method divides the dataset into five subsets, allowing the model to be trained and tested multiple times, ensuring that its accuracy is not a result of random data partitioning. During each iteration, one subset serves as the test data, while the others collectively form the training data. Performance metrics, including accuracy, precision, recall, and F1-score, are computed for each fold, culminating in the mean accuracy across all iterations. K-fold cross-validation incorporates confidence in the model's performance by avoiding overfitting and offering a more comprehensive ability to accurately classify diabetes data into Type 1 and Type 2. This rigorous evaluation approach is vital in ensuring the reliability and effectiveness of the predictive model.

4.7 Model Evaluation

he decision to deploy the Convolutional Neural Network (CNN) model, which achieved the highest accuracy of 79.5% according to Table 2, for predicting raw and unseen data is grounded in its exceptional performance on the test dataset. With its robust accuracy, the CNN model showcases its ability to generalize effectively to new data points, crucial for real-world applications. Leveraging CNNs' prowess in feature extraction, particularly in image recognition tasks, ensures that the model captures pertinent patterns within the input data. While CNNs are known for their complexity, techniques for interpreting their decisions can provide insights into the model's predictions, vital for sensitive applications. Considering scalability, infrastructure needs, consistent data preprocessing, ongoing monitoring, and potential updates, the deployment of the CNN model promises reliable and accurate predictions for unseen data, paving the way for impactful real-world applications.



No.	Algorithms	Accuracy
1	Artificial Neural Network (ANN)	72.9%
2	Convolutional Neural Network (CNN)	79.5%
3	Long Short-Term Memory (LSTM)	62.5%
4	Support Vector Machine (SVM)	71.7%
5	Logistic Regression	72.2%
6	Naive Bayes	69.8%
7	XGBoost	66.2%
8	Decision Tree	73.1%
9	Random Forest	76.0%
10	K-Nearest Neighbors (KNN)	72.9%

Table 2 Results of classifiers

4.8 Conclusion

This predictive model was developed to cater the growing global prevalence of diabetes, highlighting the importance of precise classification and early diagnosis for effective patient care. Additionally, the model contributes significantly to various aspects of healthcare. Firstly, it plays a crucial role in helping pre-diabetic patients avoid further implications of diabetes by enabling early detection and intervention. This proactive approach not only enhances patients' quality of life but also reduces the long-term costs associated with diabetes disease treatment. Moreover, the model facilitates healthcare resource allocation efficiently through early detection. By accurately identifying diabetes types at an early stage, healthcare providers can allocate resources more effectively, ensuring that patients receive the right care at the right time.

CONFLICT OF INTEREST

The authors declare no potential conflicts of interest in connection with the research, authorship, and/or publication of this article

References

1. American Diabetes Association (2022). Type 2 Diabetes - Symptoms, Causes, Treatment | ADA. [online] diabetes.org. Available at: https://diabetes.org/diabetes/type-2.
2. Danasingh, Asir Antony. (2017). Diabetes Prediction Using Medical Data. Journal of Computational Intelligence in Bioinformatics. 10. 1-8.
3. Usuga Cadavid, Juan Pablo & Lamouri, Samir & Grabot, Bernard & Fortin, Arnaud. (2021). Using deep learning to value free-form text data for predictive maintenance. International Journal of Production Research. 60. 1-28. 10.1080/00207543.2021.1951868.
4. Bodinga, Bello & Abdulsalam, Mukhtar & Buhari, Bello & Mansur, Muzzammil. (2022). On The Analysis of Some Machine Learning Algorithms for the Prediction of Diabetes. International Journal of Advanced Networking and Applications. 14. 5294-5299. 10.35444/IJANA.2022.14109.
5. D. Shetty, K. Rit, S. Shaikh and N. Patil, "Diabetes disease prediction using data mining," *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India, 2017, pp. 1-5, doi: 10.1109/ICIIECS.2017.8276012.