WQD7001 PRINCIPLES OF DATA SCIENCE

GA1

Predictive Modeling for Early Intervention: Developing a Diabetes Risk Classification System

Name	Matrix No.
Tan Kai Qi	23051355
Alicia Chua Yang Qin	23060317
Che Jia Min	23053571
Ng Chee Kang	23051336
Sang Yong Han	22087278

Contents

Introduction

Diabetes is one of the most severe health conditions worldwide. It is a long-term disease that is caused by chronic metabolic disorders due to high blood glucose levels in the blood. As per the International Diabetes Federation, there are currently 537 million individuals worldwide experiencing diabetes and the projections indicate that this number will rise to 643 million by 2030 (Ajay Kumar & Kamaldeep Kaur, 2023a).

Diabetes development could be due to factors such as obesity, family history, lack of physical activity, unhealthy diet, lifestyle factors, and so on. According to a WHO report, it indicates that diabetes causes 1.5 million deaths worldwide each year (Jiang et al., 2023). Undiagnosed diabetes can increase the risk of cardiac stroke, diabetic nephropathy, and other disorders (Kaur & Kumari, 2022). Therefore, it is important to enhance awareness of the hazards of diabetes and predict an individual's diabetes risk for early detection so that treatment of diabetes can be in place as early as possible to reduce the risk of severe health problems.

With the increasing availability of electronic health records, it is more effective to identify the signs of diabetes than manual procedures while avoiding human error and complications. There are many factors that can help to predict diabetes properly. These factors such as glucose level and BMI, diabetes, blood pressure, age, pregnancy, and so on. Common symptoms of diabetes are polyuria, polydipsia, polyphagia, sudden weight loss, obesity, etc (Cinar et al., 2023). In this assignment, a predictive model is constructed to study which symptoms tend to happen in individuals with diabetes.

OSEMN framework is used to predict an individual's diabetes risk by involving several steps such as Obtain, Scrub, Explore, Model, and Interpret. The process begins by gathering data from relevant data sources to identify diabetes risk symptoms. The data is then cleaned to ensure accuracy and consistency. Subsequently, a comprehensive analysis of the data to understand the relationships between different variables and their correlation with diabetes risk. By utilizing this enriched dataset, a predictive model is constructed to assess an individual's diabetes risk. Lastly, the outcome of this model is reviewed to enable a tougher understanding of how various symptoms could correlate to an individual's diabetes risk. This study can assist in the early identification of individuals at higher risk, enabling proactive interventions, lifestyle modifications, and targeted healthcare strategies to manage the condition effectively.

Problem Statement

The widespread prevalence of diabetes poses substantial healthcare costs and formidable public health challenges. Diabetes created a significant burden on healthcare systems which required additional resources for both management and treatment. The hindrance caused by late diagnosis creates an obstacle to implementing timely interventions and preventive measures, resulting in exacerbated health outcomes. In its early stages, diabetes might exhibit subtle symptoms or none. This led to individuals ignoring those warning signs or attributing them to other causes, hence delaying diabetes diagnosis and management. Besides, the identification of at-risk individuals based on attributes can be challenging without a systematic approach. This is because risk factors for diabetes can be different in various populations or individuals. Without a structured methodology, identifying risk indicators becomes challenging, leading to potential oversight critical factors. These combined problems highlight the critical need for a more effective and organized approach to recognize, diagnose, and treat diabetes cases as early as possible to reduce the burden on healthcare systems and enhance health outcomes for affected individuals. Through early recognition of at-risk individuals, it can pave the way for a more effective healthcare system, ultimately benefiting individuals, communities, and society as a whole.

Objectives

1. To identify a robust classification model for predicting an individual's risk of developing diabetes based on a set of provided attributes.
2. To enable early intervention by identifying individuals at a high risk of diabetes.
3. To identify prevention efforts by providing actionable insights to individuals, healthcare professionals, and public health authorities

Scope and Domain

This project focuses on utilizing predictive modeling methodologies to categorize individuals into different categories of diabetes risks. The scope includes collecting, processing, and analyzing relevant data while considering various symptoms such as polyuria, polydipsia, polyphagia, sudden weight loss, obesity, and so on associated with diabetes risk. It integrates knowledge and practices from the healthcare domain with advanced data analysis and modeling techniques. This is aimed to develop a robust predictive model that is capable of identifying an individual's diabetes risk which can contribute to healthcare systems by enabling proactive interventions and focused preventive measures to mitigate the risk of diabetes.

Literature Analysis

Diabetes is a long-term disease that causes significant difficulties in a variety of organs, including the heart, eyes, kidneys, and nerves (Sneha & Gangil, 2019). Thus, early diagnosis is important for managing and preventing diabetes from life. To minimize human errors and complications, technological methods will be used for diabetes screening and improving the efficiency of the detection. However, reliable prediction models are desperately needed to identify those at risk before symptoms appear, especially with the increased prevalence of diabetes.

In recent years, artificial intelligence (AI) techniques have played an important role in disease prediction in the health sector. There are various machine learning and deep learning techniques have been excavated in this rising diabetes epidemic for detection and early diagnosis (Ajay Kumar & Kamaldeep Kaur, 2023b). A study on the diabetes prediction model using Boruta feature selection and ensemble learning to achieve better performance on the model (Zhou et al., 2023). With the Boruta feature selection algorithm, suitable attributes able to be identified from the dataset and improve the model performance to have an early diagnosis of diabetes. Ensemble learning is used to improve the robustness and accuracy of classification by combining several single classifiers in different methods. The model can achieve an accuracy rate of 98% in this study.

Based on the research from Madan et al., 2022, 5 various types of models which are Convolutional Neural Network (CNN), Deep Neural Network (DNN), Bi-directional Long-Short Term Memory (Bi-LSTM), Convolutional Neural Network with Long Short-Term Memory (CNN-LSTM) and Convolutional Neural Network with Bi-directional Long Short-Term Memory (CNN-Bi-LSTM), used for diabetes over static PIMA Indian dataset (PIDD) identification. From the analysis results, CNN-Bi-LSTM was able to achieve the highest accuracy, sensitivity, and specificity which are 98.85%, 97%, and 98% respectively. There is a recommendation to have a dashboard for model visualization in real-time situations, to assist doctors in maintaining patients' information and monitoring real-time vital sign statistics.

To identify the relationship between glycated hemoglobin (HbA1C) and random, fasting and postprandial glucose (PPG/PP), Ansari et al., 2023 used Pearson correlation analysis to determine PPG/PP has a better correlation with HbA1C compared to fasting and random glucose. In this study, PPG/PP was found to be more accurate in predicting HbA1C and average glucose levels. However, the sample size from this study is limited. A large sample size is required to increase the precision and reliability of the study.

The analysis study (Rupapara et al., 2023) of the impact of feature selection on the machine learning ensemble classifier has been done for principle component analysis (PCA) and Chi-square (Chi-2). From the analysis results, the Chi-2 feature has higher performance than PCA and original features. However, the researchers used the proposed feature fusion to achieve the highest accuracy which is a 0.85 score for diabetes prediction approaches.

Observations were made after reviewing the literature related to diabetes prediction, this study proposed to use Pearson correlation analysis and Chi-2 to obtain the symptoms that have a better correlation to diabetes.

Methodology

Obtain - Data collection (Description of dataset)

The dataset obtained (Islam et al., n.d.) contains the records of diabetes-related symptoms for 520 individuals. In this dataset, the data is collected from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. The patients were asked to complete a direct questionnaire. In this dataset, there are 520 records and 17 variables. Among these 17 variables, 16 are categorical variables. There's only 1 numerical variable in this dataset (age).

The categorical variables are: gender, polyuria (excessive urination), polydipsia (excessive thirst/drinking), sudden_weight_loss, weakness, polyphagia (extreme hunger), genital_thrush(yeast infection), visual_blurring, itching, irritability, delayed_healing (of the wound), partial_paresis (muscle weakness), muscle_rea stiffness, alopecia (hair loss), obesity, class (presence of diabetes)

Scrub – Data cleaning

The **readr package** from R is used to read the CSV file. Readr is part of the tidyverse ecosystem. It is faster and more memory-efficient.

```
mydf <- read_csv2("diabetes_data.csv")
```

The read_csv2 function is used because the delimiter of the file is (;). In other words, the values or fields are separated by semicolons (;). Next, the **dplyr package** from R is used to clean and manipulate the data.

The glimpse () function is used to view every column in the data frame.

```
glimpse(mydf)
```

Then, the dataset is checked for any missing values with the code below.

```
missing_values <- mydf %>% summarise_all(~ sum(is.na(.)))
```

Furthermore, sum(is.na(.)) is used to check if each element in the column is missing (NA) and sum the number of missing values in the column. Then, the summarize_all function is used to apply it to each column in the data frame. Our analysis shows that there are no missing values in the dataset.

The gender column in the dataset contains the categorical variable (male and female). The categorical variable is then converted to binary values (0=male, 1=female) to ensure compatibility with machine learning models.

```
mydf<- mydf %>% mutate(gender=ifelse(gender=="Male",0,1))
```

The mutate function is used to create a new column and in our case modify the existing column (gender). The ‘if else’ statement is used to convert male and female to binary values (0) and (1).

The format of the data is also consistent as the data for all columns are in binary values (0,1) except for the age column.

Exploratory Data Analysis (EDA)

Data structure

This dataset contained 520 observations, and each described by 17 variables as below:

Age	Gender	Polyuria	Polydipsia	Sudden weight loss
Weakness	Polyphagia	Genital thrush	Visual blurring	Itching
Irritability	Delayed healing	Partial paresis	Muscle stiffness	Alopecia
Obesity	Class			

Age is a continuous numerical variable. The remaining variables are binary categorical, encoded as “0” and “1”. For the gender variable, “0” represents male; “1” represents female, a conventional representation in binary variables for analysis. Those variables include 'polyuria', 'polydipsia', 'sudden weight loss', etc., also follow the binary format where '1' indicates the occurrence of a symptom, and '0' indicates its absence. The 'class' variable is binary data representing the diagnosis, where '1' indicates diagnosed diabetes and '0' does not. Even though stored as integers, each of these binary variables categorizes the presence or absence of a feature, and it is treated as categorical during analysis.

Data Cleaning

Determining the data quality is essential. For the reliability of the results, we took a comprehensive data cleaning process, which is a critical step in exploratory data analysis. We performed several cleaning steps:

- Missing value analysis
- Validity checks
- Consistency confirmation

The first action was to check for missing data, which can cause bias or inaccuracies in our analysis. We conducted a thorough investigation using is.na() function across all variables to check for any missing value. We found that this dataset is complete and has no missing value. Most of the variables are binary, and next, we validated the integrity of the data by ensuring all values fell within the expected range of 0 and 1. We also examined the consistency of data. Our review did not reveal any inconsistencies.

In conclusion, the dataset was clean and well-prepared for the subsequent stages of our analysis.

Data Distribution

Our EDA started with a look into the distribution of age, gender, and diabetes diagnosis within the dataset.

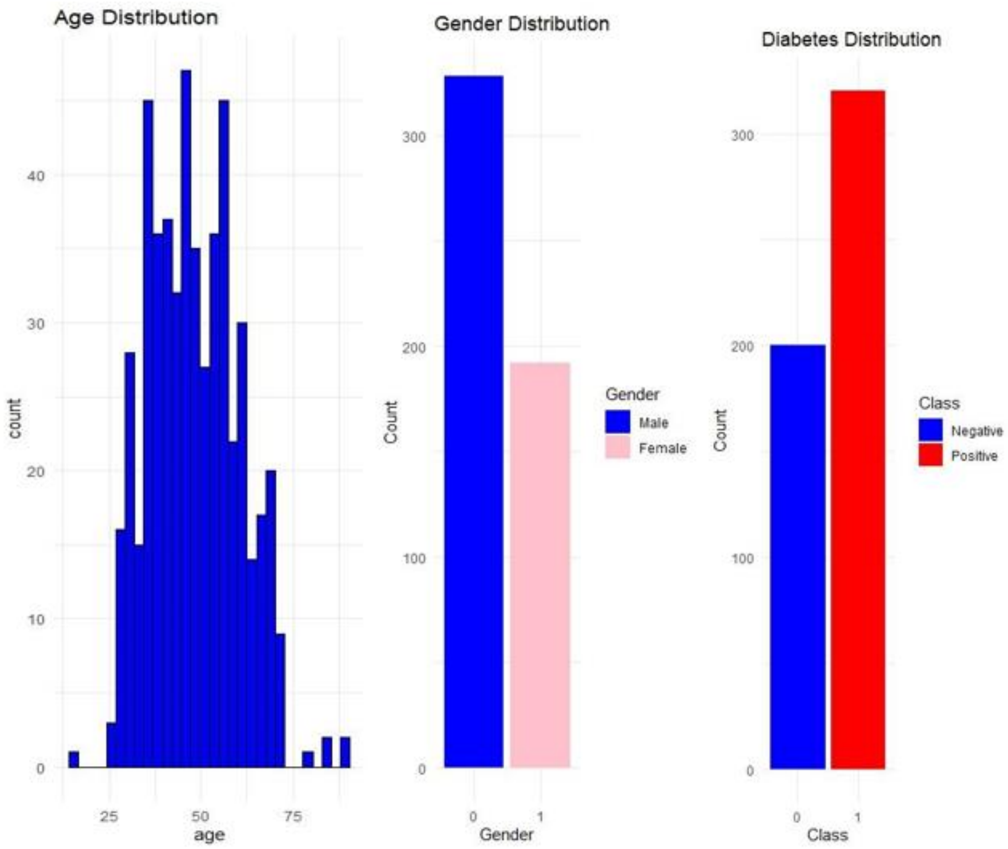


Figure 1: Distribution of age, gender, and diabetes diagnosis

Age Distribution

The dataset showed middle-aged individuals as the leading group, with the 30-59 age group comprising 405 of the 520 participants. Young adults and children (0-29) are minimally represented, with only 20 individuals, while seniors (60-89) comprise 93 participants. Those over 90 are the least represented at just 2 individuals. The histogram between 30 to 59 years visually confirms the dataset's demographic focus on adults in their prime working years, with fewer occurrences of younger and older ages.

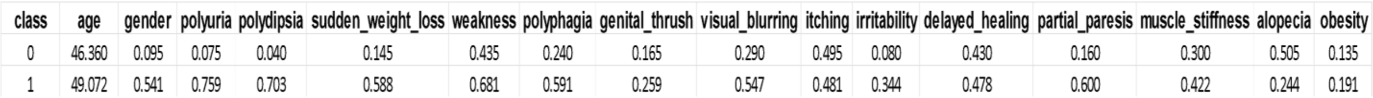
Gender Distribution

There are 328 males and 192 females in this dataset. The numbers show more males than females. The difference is significant. The males account for approximately 63% of the dataset, while females are about 37%. We will consider this imbalance when we process the analysis, especially in research where gender may significantly influence the outcomes or prevalence of the studies.

Diabetes Distribution

The 'class' represents diabetes diagnosis in our dataset, with '1' indicating a positive diagnosis and '0' a negative one. Out of the total, 320 individuals are diagnosed with diabetes, surpassing the 200 individuals without it. We will further analyze this distribution to understand how diabetes relates to other variables in the dataset. This in-depth examination examines the connections between a diabetes diagnosis and different factors.

Relationship and Statistical Analysis



class	age	gender	polyuria	polydipsia	sudden_weight_loss	weakness	polyphagia	genital_thrush	visual_blurring	itching	irritability	delayed_healing	partial_paresis	muscle_stiffness	alopecia	obesity
0	46.360	0.095	0.075	0.040	0.145	0.435	0.240	0.165	0.290	0.495	0.080	0.430	0.160	0.300	0.505	0.135
1	49.072	0.541	0.759	0.703	0.588	0.681	0.591	0.259	0.547	0.481	0.344	0.478	0.600	0.422	0.244	0.191

Figure 2: Sample of the relationship between diabetes and the prevalence of symptoms

Firstly, we examine the differences in the prevalence of symptoms associated with diabetes. There are a few key observations below:

- **Polyuria**: In the diabetic group, **75.9%** report experiencing polyuria, compared to **7.5%** in the non-diabetic group.
- **Polydipsia**: The percentage of individuals reporting polydipsia in the non-diabetic group is **4%**, which significantly increases to **70.3%** in the diabetic group.
- **Sudden Weight Loss**: Only **14.5%** of non-diabetics report sudden weight loss, compared to **58.8%** of people with diabetes.
- **Polyphagia**: **24**% of non-diabetics with polyphagia symptoms compared to **59.1%** among diabetics.
- **Irritability**: Within the group without diabetes, irritability is reported by **8%**, in contrast to the diabetic group, which rises to **34.4%**.
- **Partial Paresis**: Observations of partial paresis show that **16%** of the non-diabetic group experience this symptom, significantly increasing to **60%** in the diabetic group.

The diabetic group also shows increased averages in other variables, but the most pronounced changes are observed in the above mentioned variables. The age variable does not differ significantly between the groups, which means age might be a factor, but it is not the key; the presence of specific symptoms is more indicative of diabetes based on the data.

Class	**Male**	**Female**
0	90.5%	9.5%
1	45.9%	54.1%

Table 1: The distribution of gender variables

The gender variable is slightly complex due to the distribution imbalance. When we performed a comparison between males and females, it showed under the non-diabetic group 90.5% are male and 9.5% are female. In the diabetic group, 45.9% are male and 54.1% are female. The results reflect the high percentage of females with diabetes in this dataset. The following section will review the heatmap and chi-square for these variables.

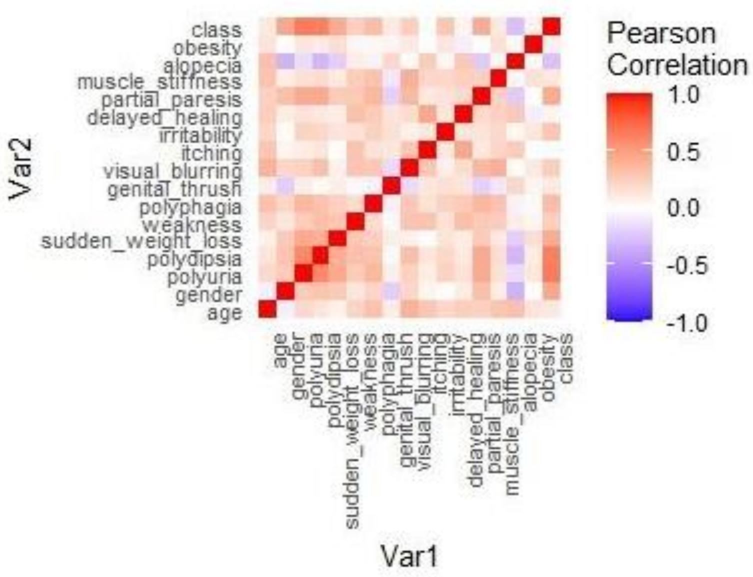


Figure 3: Heatmap for variables

The Pearson correlation coefficient quantifies the relationship between two variables, ranging from -1 for a perfect negative correlation to +1 for a perfect positive correlation.

Analyzing the heatmap of Pearson correlations in our dataset showed a strong positive correlation between polyuria and polydipsia, indicating that these symptoms tend to happen in individuals with diabetes. This observation is consistent with the higher prevalence of both symptoms in the diabetic group we identified earlier. The heatmap also shows a moderate positive correlation between sudden weight loss and polyphagia, which suggests that these symptoms are more frequently reported together in diabetic individuals. Although gender, irritability, and partial paresis positively correlate with diabetes, their interrelatedness appears less significant.

The absence of a strong correlation between age and diabetes status confirms that specific symptoms are more predictive of diabetes than age in our dataset. These insights provide a better understanding of symptom interrelations, setting a firm foundation for the subsequent phase of our project, which involves building a predictive model in part 2.

Symptom	Chi-Squared	Degrees of Freedom	p-Value
Polyuria	227.87	1	< 0.00000000000000022
Polydipsia	216.17	1	< 0.00000000000000022
Sudden Weight Loss	97.296	1	< 0.00000000000000022
Polyphagia	59.595	1	0.00000000001165
Irritability	45.208	1	0.00000000001771
Partial Paresis	95.388	1	< 0.00000000000000022
Gender	103.04	1	< 0.00000000000000022

Table 2: Chi-square test table

This chi-squared test result provided statistical solid support to the observed relationship between various symptoms and the diagnosis of diabetes within the dataset.

The Chi-Squared values for symptoms such as Polyuria and Polydipsia are relatively high, reflecting the strength of the association. The higher the value, the less likely the observed relationship is due to coincidence. The test showed a degree of freedom of 1, typical for a test between two categorical variables.

The associations are statistically significant if the p-value is significantly less than 0.05. Technically, the p-value less than 2.2e-16, reported as '< 0.00000000000000022', indicated a near-zero probability that the observed are random.

The statistical results proved the previous discussion on symptoms, enhancing the hypothesis that these symptoms are linked to diabetes, and it may be considered a strong indicator in the prediction of diabetes.

These associations set a clear direction for part two of the project, where we aim to predict diabetes occurrence based on symptom presentation.

Next: Modeling plan for the Part 2

The EDA provided a clear picture of this dataset to us. Polyuria, polydipsia, sudden weight loss, polyphagia, irritability, partial paresis, and gender will be the variables that we will focus on in the second part of the project. All these variables showed a statistically significant relationship with diabetes.

We mentioned the imbalance of gender distribution in the relationship discussion. We plan to use stratified analysis to examine the relationship between different genders and diabetes. The step will include creating two new subsets, and the logistic regression will be trained based on different subsets of genders to reduce the impact of imbalance distribution.

In the next part of the project, we will develop a logistic regression model. This model uses the variables that show significant correlations with diabetes in our EDA to predict the likelihood of diabetes in individuals. By examining the indicators like the symptoms, we hope the regression model will provide a diabetes risk estimation and enhance the ability to identify early diabetes risk groups that align with our objectives.

Ethical Considerations

Predictive analysis is used in this project with topics: Predictive Modeling for Early Intervention: Developing a Diabetes Risk Classification System. Definition predictive analysis in medical research uses electronic algorithms such as collecting the health record data to improve patients' outcomes and lower health care costs incurred, in conjunction with targeted costs. However, that opportunity raises ethical considerations when doing predictive modeling. Ethical considerations are crucial because they advance the goals of research, such as knowledge, accuracy, and error avoidance. When concluding any research project, it is forbidden by ethical considerations to interpret data incorrectly and present false information. Furthermore, since research typically requires cooperation between researchers and people of diverse dispositions, ethical considerations are essential to foster mutual respect, trust, and collaboration. Below are the ethical considerations that incurred in the research: -

Consent and Privacy (Privacy and confidential)

When developing a predictive analysis, researchers need access the big data to perform the data cleaning and analysis. For example, in this topic is collated using direct questionnaire and diagnosis results from the patients in the Sylhet Diabetes Hospital in Sylhet, Bangladesh. With these steps, consent, and privacy ethical considerations are involved. Because some of the patient information such as age range, habits, and so on that related to the research topics to be used for analyzing as main factors or issues or solutions. Normally, proforma notifications are used to completely protect patient privacy example, notify all patients that the data gathered on them during regular health care may be used in de-identified form in predictive analytics models and added on with their signature as well. Therefore, model developers are allowed to use the patient data that have already been collected without explicit consent with federal regulations which data allow use in any research on human subjects and privacy of health information.

Risk communication (Informed Consent)

With the number of diabetes continuing to increase, a central focus remains where the focus is on how to prevent and control diabetes. Therefore, risk communication is important because giving patients a thorough knowledge of the advantages, disadvantages, trade-offs, and uncertainties associated with any suggested course of therapy is the fundamental goal of risk communication. If misunderstanding on the preventative and actions to control, it may lead to psychosocial harms or familial implications and significantly impact life decisions. People with a family history of diabetes, for instance, did not think they were at risk for the disease since they always felt "different" from their afflicted relatives in significant ways. Not only that, but it also shows that patients who test positive underestimate their risk like assuming risk as an absolute prediction of disease (fatalism). All of that might affect the preventative steps and time for individuals due to personal perspective and controllable on the disease is less. Therefore, with the visible numeric, verbal, and potential effect is needed to maximize patient understanding. Ignoring the reality that civilizations are extremely diverse, with varying experiential effects and attitudes that can change over time, is the idea of a single, static general population. Health providers must thus be aware of these extra variables that could affect how patients interpret and use risk information.

Costly therapy and diabetic complications (Equity)

Numerous studies have demonstrated the increased advantages of aggressive insulin therapy in managing severe diabetes and delaying the onset of complications; however, intensive insulin therapy comes at a high cost. Therefore, the moral conundrum that many physicians face is whether to continue with traditional therapy, which may result in early complications or to begin expensive, intensive therapy with expensive human insulin to prevent future complications. Medical professionals frequently encounter moral conundrums with an economic basis. One such would be the terrible consequence of diabetes, foot gangrene. Restoring the foot is frequently feasible, albeit at a significant cost. For this cutting-edge treatment, the family must take on significant debt. Maybe the only option left to bear this financial burden is amputation. Losing a limb can be extremely debilitating for young diabetics and have an impact on their ability to work. Socioeconomic factors play a big role in the difficult decision to amputate. In a patient with end-stage renal disease who has multisystem failure and for whom a renal transplant is not practical, a similar treatment option may be taken into consideration. The question may well be how long hemodialysis should be continued considering the associated costs and almost certain negative outcomes. These conundrums are likely to arise more frequently as the number of diabetics with complications rises and resources become more limited, which will spark a broader discussion on the moral, social, and financial aspects of managing diabetic complications.

Conclusion

In this project, the Chi-square test has been used to deep learning algorithms for diabetes prediction and provide a potential change from the way “low awareness” to “high awareness” for individuals on diabetes issues. With machine learning and deep learning applications, diabetes data is the most significant component that contributes to medical care systems, especially for those high-rate countries. The paper aims to predict diabetes risk, results can seek to facilitate prevention efforts by providing actionable insights to individuals, healthcare professionals, and public health authorities in the early intervention stage. A total of 520 patients’ data has been used to analyze the diabetes diagnosis. Studies show a significant relationship between various symptoms (polyuria, polydipsia, sudden weight loss, polyphagia, irritability, partial paresis) and the diagnosis of databases. As a result, the government can cooperate with agency/ hospital charity organizations to promote an event for awareness and set up a foundation to help patients who can't afford medical treatment.

Reference

Ajay Kumar, & Kamaldeep Kaur. (2023a). A Novel MCDM-Based Framework to Recommend Machine Learning Techniques for Diabetes Prediction. *International Journal of Engineering and Technology Innovation*. https://doi.org/10.46604/ijeti.2023.11837

Ajay Kumar, & Kamaldeep Kaur. (2023b). A Novel MCDM-Based Framework to Recommend Machine Learning Techniques for Diabetes Prediction. *International Journal of Engineering and Technology Innovation*. https://doi.org/10.46604/ijeti.2023.11837

Ansari, S., Jayeeta Bhadra, Ahirwar, A. K., & Gupta, J. (2023). Correlation analysis of HbA1c versus random, fasting, and postprandial glucose levels as predictors of glycemic control in type 2 diabetes patients. *Asian Journal of Medical Sciences*, *14*(4), 37–43. https://doi.org/10.3126/ajms.v14i4.53072

Cinar, I., Taspinar, Y. S., & Koklu, M. (2023). Development of Early Stage Diabetes Prediction Model Based on Stacking Approach. *Tehnicki Glasnik*, *17*(2), 153–159. https://doi.org/10.31803/tg-20211119133806

Islam, M. M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (n.d.). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. *Volume 992, Pages 113 - 125*, *992*, Manipal. https://doi.org/10.1007/978-981-13-8798-2_12

Jiang, L., Xia, Z., Zhu, R., Gong, H., Wang, J., Li, J., & Wang, L. (2023). Diabetes risk prediction model based on community follow-up data using machine learning. *Preventive Medicine Reports*, *35*. https://doi.org/10.1016/j.pmedr.2023.102358

Kaur, H., & Kumari, V. (2022). Predictive modeling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, *18*(1–2), 90–100. https://doi.org/10.1016/j.aci.2018.12.004

Madan, P., Singh, V., Chaudhari, V., Albagory, Y., Dumka, A., Singh, R., Gehlot, A., Rashid, M., Alshamrani, S. S., & Alghamdi, A. S. (2022). An Optimization-Based Diabetes Prediction Model Using CNN and Bi-Directional LSTM in Real-Time Environment. *Applied Sciences (Switzerland)*, *12*(8). https://doi.org/10.3390/app12083989

Rupapara, V., Rustam, F., Ishaq, A., Lee, E., & Ashraf, I. (2023). Chi-Square and PCA Based Feature Selection for Diabetes Detection with Ensemble Classifier. *Intelligent Automation and Soft Computing*, *36*(2), 1931–1949. https://doi.org/10.32604/iasc.2023.028257

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0175-6

Zhou, H., Xin, Y., & Li, S. (2023). A diabetes prediction model based on Boruta feature selection and ensemble learning. *BMC Bioinformatics*, *24*(1). https://doi.org/10.1186/s12859-023-05300-5