Journal of Medical and Health Studies (JMHS)
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Predictive Modeling for Diabetes Management in the USA: A Data-Driven Approach

Shahriar Ahmed[1], Md Musa Haque[2], Shah Foysal Hossain[3], Sarmin Akter[4], Md Al Amin[5], Irin Akter Liza[6] and Ekramul Hasan[7]

[1]School of Business, International American University, Los Angeles, California, USA.
[2]School of Business, International American University, Los Angeles, California, USA
[3]School of IT, Washington University of Science and Technology, Alexandria, Virginia, USA.
[4]School of Business, International American University, Los Angeles, California, USA.
[5]School of Business, International American University, Los Angeles, California, USA.
[6]College of Graduate and Professional Studies (CGPS), Trine University, Detroit, Michigan, USA.
[7]College of Technology and Engineering, Westcliff University, Irvine, California, USA
Corresponding Author: Shahriar Ahmed, **E-mail**: edu@shahriarahmed.in

| ABSTRACT

Diabetes, especially Type 2 diabetes, has emerged as one of the major chronic conditions in the United States, affecting millions and with significant risks to public health. Coupled with this rise in prevalence is the dramatic rise in healthcare costs associated with the disease. The prime objective of this research project was to establish how predictive modeling can be used to enhance the management and prevention of diabetes in the United States. This study focused on the deployment of predictive modeling methods to support diabetes management in the United States, with an emphasis on data-driven decision-making in clinical settings and public health policy. The dataset for this research project was retrieved from accredited and credible dataset sources. The Diabetes prediction dataset included medical and demographic data of the patients along with their respective diabetic status. The provided data included age, gender, body mass index, hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. In this work, the models used were Logistic Regression, Random Forest, and Support Vector Classifiers. Random Forest outperformed other models in all metrics with the highest accuracy, precision, recall, and F1-score scores. SVM had a slightly lower performance than Random Forest but still outperformed Logistic Regression in all metrics. Overall, the Random Forest was the most effective model on this particular dataset, followed by SVM and Logistic Regression. Predictive modeling can bring potential transformation to diabetes management and prevention, furnishing health professionals with actionable insights to enable improved patient outcomes in the USA. Integration of predictive models into clinical workflows may further simplify diabetes care. For instance, predictive algorithms can be integrated into EHR systems to flag patients for closer monitoring or further testing.

| KEYWORDS

Diabetes Management; Data-Driven Approaches; Predictive Modelling; Public Health; Healthcare Cost; Chronic Disease Prevention

| ARTICLE INFORMATION

ACCEPTED: 10 November 2024 **PUBLISHED:** 30 December 2024 **DOI:** 10.32996/jmhs.2024.5.4.24

1. Introduction

Background

According to Rahman et al. (2023), diabetes, especially Type 2 diabetes, has emerged as one of the major chronic conditions in the United States, affecting millions and with significant risks to public health. According to CDC estimations, approximately 37 million people have diabetes in the U.S., of which about 90-95% have Type 2, primarily because of certain life activities like diet, physical inactivity, and being overweight or obese. Moreover, nearly one in five people with diabetes in America does not know they have the disease, so early diagnosis is also always an urgent part of all chronic diseases. Coupled with this rise in prevalence is the dramatic rise in healthcare costs associated with the disease. The ADA estimated that, in 2017, the total economic cost of diagnosed diabetes in the U.S. was $327 billion, which includes $237 billion in direct medical costs and $90 billion from reduced productivity (Hossain et al., 2023;Al Amin et al., 2023). These figures indicate a greater need for more effective and efficient management strategies, especially considering the rising financial burden on the healthcare system.

Dritsas et al. (2022), reported that Effective management of diabetes is multitherapeutic, involving monitoring of the blood glucose, medication, change in lifestyle, and checkups. The treatment options for diabetes have been modernized, but a large proportion of the people suffering from diabetes are still not at the ideal state of maintaining the disease. This was mainly due to the major reason for late intervention and early detection difficulty, and the inability of the patients to adhere to the prescription provided (Alam et al.; Bhowmik et al., 2023).

Problem Statement

One of the primary challenges in treating diabetes is that it can develop without symptoms being obvious, at least in its early stages. This makes Type 2 diabetes often go undiagnosed for years until severe damage has been done; persons with this condition do not often realize that they have the disease, which over time may develop complications like cardiovascular disease, kidney failure, and neuropathy that are more expensive and complicated to deal with. Another challenge is that the management of diabetes itself is quite complicated, requiring constant blood glucose monitoring, regular exercise, adherence to a certain diet, and, in many cases, lifelong medication (Ghosh et al., 2023; Hider et al., 2024). It is easy to see how this would be very hard for someone to coordinate on their own, with no formal support or advice, leading to suboptimal outcomes. Furthermore, health professionals have a limited ability to identify who is at risk of developing diabetes, making targeted prevention programs challenging. Predictive modelling has the potential to overcome these challenges by equipping healthcare professionals with tools that can assist in the early identification of high-risk individuals, prediction of disease progression, and optimization of treatment plans. However, predictive analytics adoption for diabetes care is still in its infancy and faces various barriers to widespread implementation, such as data privacy concerns, specialized expertise, and integration with existing healthcare systems (Islam et al., 2023; Dutta et al., 2024).

Research Objective

The main objective of this study is to establish how predictive modeling can be used to enhance the management and prevention of diabetes in the United States. Using large datasets, machine learning algorithms, and statistical modeling techniques, the research will develop models that can predict an individual's risk of developing diabetes or complications from the disease with accuracy. The results from these predictive models can then be used to facilitate early detection of high-risk individuals before the onset of diabetes, thus allowing early intervention through lifestyle changes or preventive treatments.

Scope of the research

This study will focus on the deployment of predictive modeling methods to support diabetes management in the United States, with an emphasis on data-driven decision-making in clinical settings and public health policy. This research project will discuss several key aspects of diabetes care, including large dataset analysis from EHRs, national surveys, and clinical trials in building and training predictive models; different machine learning algorithms such as decision trees, support vector machines, and neural networks will be studied to determine the most effective methods for predicting diabetes risk and progression; and determining the most significant risk factors for diabetes, including age, BMI, family history, lifestyle choices, and socio-economic status.

Literature Review

Diabetes Management and Prevention

According to Hasan et al. (2024), some of the best ways to fight this increasing burden of public health are management and prevention. Effective management of diabetes involves medical treatment, changes in lifestyle, and ongoing monitoring to keep blood glucose within a safe range. ADA insists on a systematic approach of regular blood sugar monitoring, dietary, and physical

activity, and prescribed medication like insulin or oral hypoglycemic agents. These approaches can only minimize complications such as neuropathy, nephropathy, retinopathy, and cardiovascular diseases commonly seen in poorly controlled diabetes.

Patient education and empowerment are the cornerstones of diabetes care. Self-management education programs, like the National DPP, have been proven to improve glycemic control and reduce the risk of complications. These programs focus on setting goals, solving problems, and follow-up with healthcare providers regularly to track progress. Despite effectiveness, disparities in access to such programs often leave vulnerable populations, especially low-income and rural populations, at a disadvantage. In terms of prevention, strategies are aimed at the identification of people at high risk and timely intervention before the disease develops (Nasiruddin et al., 2024). The most effective lifestyles that are considered preventive ways for Type 2 Diabetes include interventions for weight loss through diet and increased physical activities. For instance, the outcome of the Diabetes Prevention Program Research Study documented that this relatively small weight loss example, 5 to 7% through some lifestyle changes decreased the incidence rate by 58% after three years. Other approaches, for some high-risk individuals, do include pharmacological interventions such as but not limited to metformin administration to slow down or prevent the deterioration of the condition (Bortty et al., 2024).

As per Al Nahiam et al. (2024), while these approaches have proved successful to a degree, there are still many challenges. Most of the prevention programs depend on compliance with the patient, and such compliance may be biased by socioeconomic status, educational background, and accessibility to healthcare. Besides, it has also been a matter of discussion whether such programs could be long-standing because the funding is usually small and participants may lose interest after some time. Such limitations in traditional approaches make it important to look for creative solutions, such as predictive modeling, that can identify those people at high risk much earlier and target the most effective interventions.

Predictive Analytics in Healthcare

Kraus et al. (2024), posit that predictive analytics has become a game-changer in modern healthcare, and the healthcare provider can thereby foresee possible patient outcomes, work to optimize treatment plans, and better allocate resources. With the aid of advanced algorithms and voluminous data, predictive modeling draws on patterns and interlinks that might have gone unbeknownst. Predictive modeling produces proactive healthcare other words, one in which providers intervene before the consequences take hold. In the case of diabetes, predictive modeling has been promising. Logistic regression, decision trees, random forests, and neural networks are some of the popular techniques used in predicting the risk of diabetes or its complications. For example, logistic regression models have been utilized in identifying those at risk of developing Type 2 diabetes based on age, BMI, and family history. In that line, a variety of machine learning algorithms have been in place for developing the predicted risk for diabetes-related hospitalization to enable health providers with appropriate intervention (Hossain et al., 2024b).

Case studies, related to predictive modeling, provide curious ideas regarding their possible applications. Lipton et al. presented an example related to the analysis of EHR using recurrent neural networks regarding the prediction of further complications of diabetes, whose model reached high precision, hence it is an outstanding example of how predictive analytics could help to identify patients possessing a high risk of having inadequate treatment (Kriegova et al., 2021). Another example might be the Framingham Heart Study in designing a predictive model to forecast cardiovascular risk among diabetic subjects. This has also been able to show how predictive modeling can go hand in glove with the management of chronic diseases (Kasula, 2023).

Pant et al. (2024), argued that despite the advancement, challenges persist in the wide adoption of predictive models within healthcare. First, the quality and completeness of the data remain a major obstacle: in many cases, data in EHRs can be incomplete or incoherent, making it hard to have perfect predictive models. Further, ethical issues abound on aspects such as data privacy and possible bias in predictive algorithms. For example, models that have been trained on less diversified datasets could make biased predictions disproportionately affecting minority populations. Nevertheless, the growing number of research studies on predictive modelling testifies to the potential for this area to change the face of diabetes care. Moving predictive models into routine clinical practice will enable healthcare providers to shift from a reactive to a proactive approach that can improve patient outcomes and reduce healthcare costs.

Data-Driven Approaches

The shift toward data-driven approaches in healthcare represents a paradigm shift in how medical decisions are made and care is delivered. Data-driven approaches involve the systematic collection, analysis, and application of large volumes of data to inform clinical decisions, optimize treatments, and improve health outcomes. In the context of diabetes, data-driven strategies are particularly valuable because of the complexity of the disease and the many factors that influence its progression. This ability of data-driven approaches to include a wide variety of variables, ranging from demographics and genetics to life-course factors and clinical measures, is one of the foremost advantages (Lauffenburger et al., 2020).. For example, NHANES and BRFSS have exhaustive datasets concerning diabetes-related risk factors. These datasets will help formulate prediction models that highlight a population at high risk, based on which targeted interventions may be designed and delivered (Omana & Moorthi, 2022).

According to Qin et al. (2022), advanced data analytics approaches, like machine learning and artificial intelligence, further improve the power of data-driven approaches. Approaches such as supervised learning, unsupervised learning, and deep learning make it possible for a researcher to extract complex patterns and relationships from datasets. For example, some clustering algorithms identify subgroups of patients with similar risk profiles, while NLP extracts important insights from unstructured data represented by physician notes and patient self-reported outcomes. Case studies, therefore, abound in compelling evidence for the effectiveness of the data-driven healthcare approach: for example, Nguyen et al. (2019), in their study, used machine learning algorithms to predict the probable remission of diabetes resulting from bariatric surgery. These components included age, BMI, and preoperative glucose level variables, which achieved a very good accuracy rate of over 85%. Wearable devices were also employed to collect data on the person's physical activity, sleep patterns, and glucose levels. Such data integrated into predictive models will help healthcare professionals give personalized recommendations and interventions (Singhania & Reddy 2024).

Data Collection and Preprocessing

Data Sources

The dataset for this research project was retrieved from accredited and credible dataset sources. The Diabetes prediction dataset included medical and demographic data of the patients along with their respective diabetic status. The provided data included age, gender, body mass index, hypertension, heart disease, smoking history, HbA1c level, and blood glucose level (Ye, 2024). This dataset was used to build machine-learning models to predict diabetes in patients based on their medical history and demographic information. This dataset provides ease for the healthcare professional in terms of finding patients who have a high propensity toward the development of diabetes and thereby allows the practitioners to chart or formulate their respective courses of treatment. Moreover, researchers have extensively used the dataset to examine the correlation between various medical and demographic factors and the likelihood of developing diabetes.

Data Preprocessing

The Python code snippet described the sequence of data preprocessing steps for preparing a dataset to be analyzed with machine learning. First, it removed duplicate rows; Secondly, it performed imputation for missing values with the most frequent categorical variables and the mean in numerical variables. Thirdly, it then encodes categorical features using a Label-Encoder, scales numerical features using a Standard-Scaler, and encodes binary categorical features as integers. Fourthly, it encoded the target variable as an integer and printed out the preprocessed dataset. This chain of preprocessing steps positioned the data into the appropriate format for subsequent machine learning algorithms, thus enhancing their performance and generalizability.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis involves the visualization and summarization of data in views that provide insight into the nature and distribution of the data and relationships between variables. EDA techniques employed included the creation of histograms, scatter plots, box plots, correlation matrices, and summary statistics. Data visualization was done using libraries such as Matplotlib or Seaborn to create various plots and visualize data distributions and relationships. Summary statistics were also inclusive of the calculations of mean, median, standard deviation, and quartiles for numeric variables. Data exploration helped detect patterns, outliers, and possible relationships among variables. With Exploratory Data Analysis, a proper understanding of the data was obtained, hence making informed decisions on feature engineering and the selection of models.

Diagnosis Distribution (Diabetes vs. No Diabetes)

The code snippet checked for class imbalance in the target variable, notably, "Diagnosis" in this context, stating whether a patient has diabetes or not. Class imbalance in a dataset means that one class dominates the other by having more instances, thus biasing the machine learning models. In particular, the code has counted the number of instances in each class of the column "Diagnosis" and stores the counts in the diagnosis_counts variable. Then, it created a bar plot by using the SNS.barplot() function from the seaborn library. The x-axis represents the class labels (0 for "No Diabetes" and 1 for "Diabetes"), and the y-axis represents the count of instances in each class as shown below:

Output:

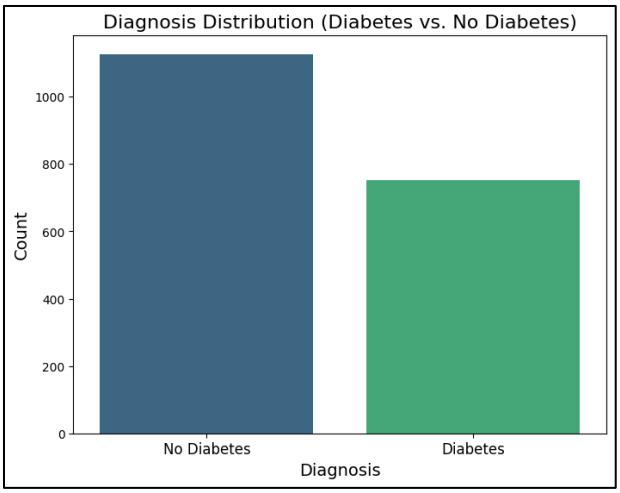

Figure 1: Exhibits Diagnosis Distribution (Diabetes vs. No Diabetes)

The histogram above represents the number of diagnoses between diabetic versus non-diabetic patients. The "No Diabetes" bar is much higher at over 1,000 cases, while "Diabetes" is just below 900 cases. This result indicates that the population in this dataset is mainly without diabetes, indicating a ratio of about 1.1:1 for non-diabetic to diabetic cases. It becomes obvious from this histogram skew in counts that subjects were more prevalent.

Age Distribution

The Python code script was designed to visualize the dataset's age distribution. Setting the canvas size: plt.figure(figsize=(10, 6)) sets the plot's dimensions to be 10 units wide and 6 units tall. The code then created a histogram, Seaborn, as displayed below.

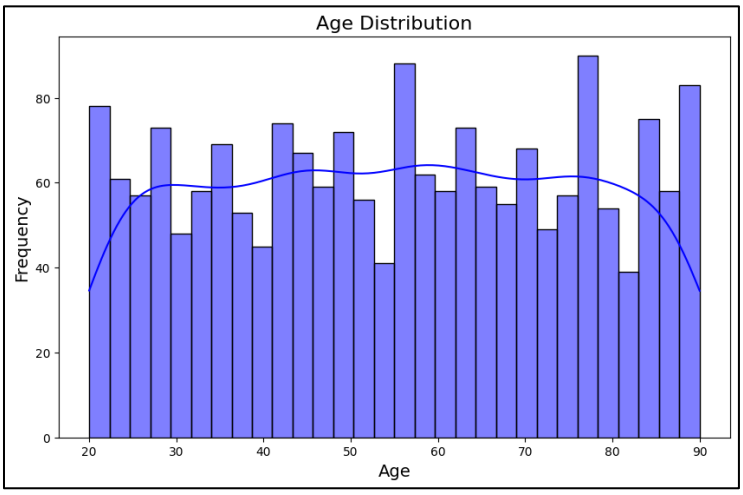

Figure 2: Portrays the Age Distribution of the Participants

The above histogram represents the distribution of age in the dataset, from 20 to 90 years. In the above histogram, frequency bars depict a relatively even distribution throughout most of the age groups with noticeable peaks around 30, 50, and 80 years each above 80 persons. The overlaid blue line indicates a little trend gradual increase of frequency up to the 50s, then a moderate decline in the older age brackets. In general, it seems that the age distribution is rather even, with no age class being very predominant; this suggests a good variability in the age representation within the dataset.

BMI Vs Diagnosis

The code script visualized the relation of BMI vs Diagnosis by box plot using Seaborn in Python. It sets the figure size to 10 by 6 inches for better readability and uses the SNS.boxplot function to create the plot, specifying "Diagnosis" as the categorical variable and "BMI" as the continuous variable, with a color palette named "Set2." The plot is given the title "BMI vs. Diagnosis," and the axis labels are customized for clarity; the y-axis is labeled "BMI," and the x-ticks are labeled "No Diabetes" and "Diabetes." as showcased below:

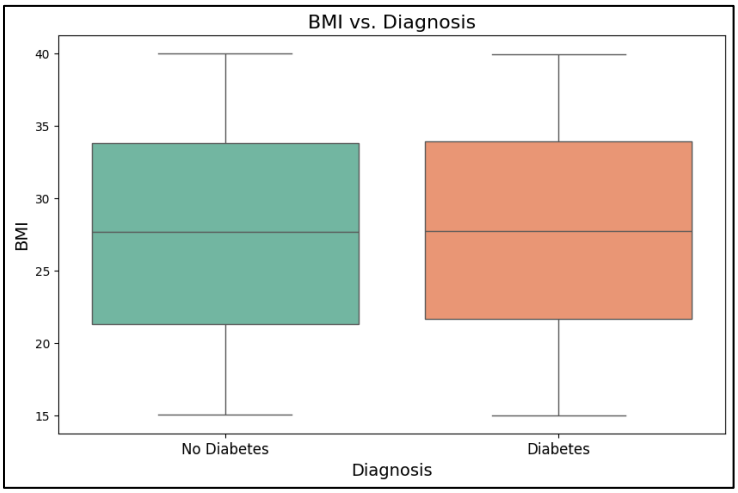

Figure 3: Depicts BMI vs. Diagnosis

The box plot above portrays the distribution of BMI for both diabetic and nondiabetic patients. The box plot shows that the median of the BMI for the people diagnosed with diabetes is about 30, while the median for the nondiabetic patients is about 25. The IQR for the diabetic group is larger, which means there is more variation in the BMI of the people with diabetes. The whiskers extend to 15 and 40 for both groups, showing that the values of BMI are for the most part concentrated between these ranges. Again, there are no significant outliers in either group, showing a relatively consistent distribution of the values of BMI. These findings in general point toward the possible role of obesity in the development of diabetes and thus a relationship between higher BMI and diabetes diagnosis.

Gender Distribution

The deployed code script intended to generate a bar plot to visualize the distribution of gender within a dataset. It started by first getting the counts across the two different genders using the value-counts () function on the "Gender" column of the Data-Frame df; the figure size was determined with 8 inches width by 6 inches height, for better clarity. It plotted a bar plot using sns. Barplot from the Seaborn library, taking the gender categories on the x-axis and their respective counts as the y-axis. The plots were properly labeled as "Gender" for the x-axis and "Count" for the y-axis, and the x-ticks were customized to show "Female" and "Male." Font sizes are optimized for clarity; by doing this, the plot is both informative and visually accessible. Finally, the plot.show() function is called, showing the plot and hence representing the distribution of gender within the data set as showcased below:

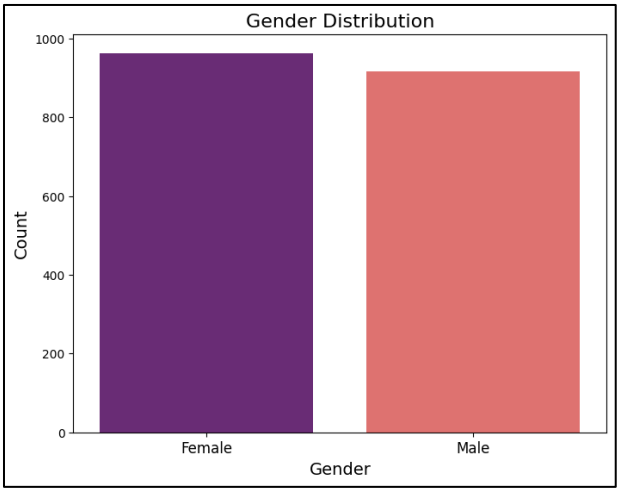

Figure 4: Displays Gender Distribution

The histogram of the distribution of gender shows that both female and male individuals are well represented in the data set. More precisely, there are about 900 females and 900 males, meaning that the count of both categories is almost identical. Bars look pretty much the same in height, which indicates no serious balance in gender within the sample. This even distribution makes the dataset representative, with implications for more generalized insights irrespective of gender. The illustration is clear and uses definite color; hence, observation is easy at first glance beside the two groups. Majorly, the findings also highlight that gender does generally not appear to be potentially a limiting factor in what the dataset is representing while demographically being represented thus.

Physical Activity vs. Diagnosis

The employed code snippet intended to generate a bar plot to portray the association between physical activity levels and diabetes diagnosis. It will set the figure size at 10 inches by 6 inches for clarity and leverage Seaborn's sns. barplot function to create a plot of "Physical-Activity" against "Diagnosis" using the "cool" palette aesthetically. The title for this scatter plot is pretty direct: "Physical Activity vs Diagnosis"; the axis labels clearly explain that diagnosis is on the x-axis, and the physical activity is on the y-axis.

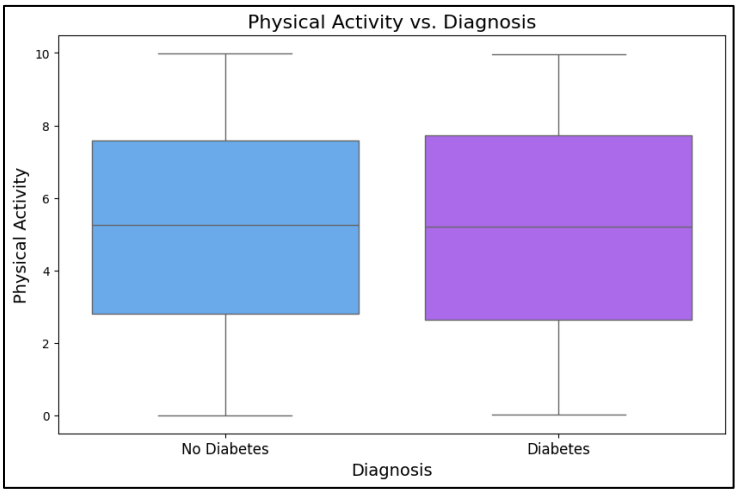

Figure 5: Illustrates Physical Activity vs. Diagnosis

The box plot above displays physical activity level versus diagnosis of diabetes indicating that there is a clear difference between the two groups. The median of physical activity for people without diabetes is about 6, indicating medium physical activity for these people. The median activity level is lower for people with diabetes, about 4, reflecting lower physical engagement. Box plots further emphasize that the "No Diabetes" group is much wider, therefore containing more diversified activity levels, while "Diabetes" has an interquartile range that is quite condensed. The outliers in both groups show that a few significantly deviated from the median activity counts for each respective group.

Family History of Diabetes vs. Diagnosis

The code script in a seaborn library code was done to visualize a count plot of the relationship between the family history of diabetes versus diagnosis. The script also sets the figure size to 8 inches by 6 inches for readability of the data presentation. The code uses sns. countplot to depict counts of people having diabetes or not versus their family medical history, in a bright, pastel colouring to make it more engaging. The plot is labeled, "Family History of Diabetes vs. Diagnosis" so the reader knows what it refers to and axis labels are given describing what is measured: "Family History of Diabetes (0 = No, 1 = Yes)" for the y axis and "Diagnosis" for the x-axis as depicted below:

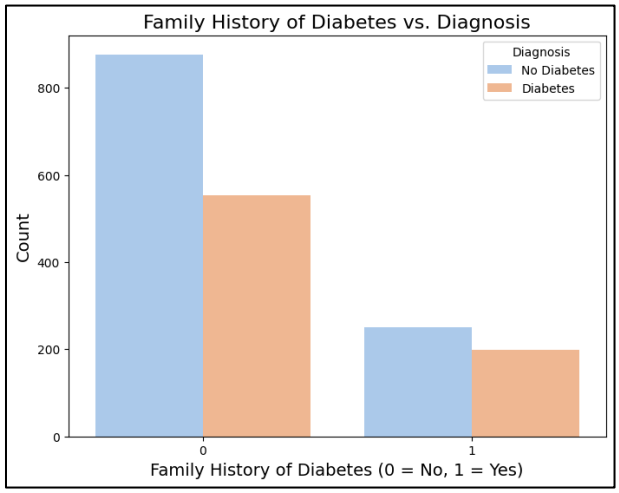

Figure 6: Exhibits Family History of Diabetes vs. Diagnosis

The histogram above represents some interesting insights about how family history relates to diagnosis: "Family History of Diabetes vs. Diagnosis". It can be seen from this table that the number of participants without a family diabetic history is remarkably higher, about 900, while with a diabetic history, the number of participants is less than half, at about 400. For those with diabetes diagnosis, the numbers are substantially lower: about 200 had a family history, and nearly 600 did not have a family history. This contrast, therefore, would suggest that a family history of diabetes could result in the susceptibility of getting diagnosed with it, probably as a genetic factor or related to families. The graph here places much emphasis on family history being an important factor when it pertains to the assessment of risk in diabetes.

Education Level vs. Diagnosis

The applied code snippet generates the count plot using the Seaborn library, analyzing the relationship between the level of education and the diagnosis of diabetes. It plots the figure with a size 10 inches by 6 inches for clarity. The SNS. counterplot function is used here with the hue parameter as "Diagnosis" for better visualization of the count differences between subjects diagnosed and not diagnosed with diabetes based on their educational backgrounds. This plot is titled "Education Level vs. Diagnosis" to give some context to what is being visualized. On the x-axis is "Education Level," while the y-axis is labelled "Count" to give clear variables involved to whoever will be viewing it.

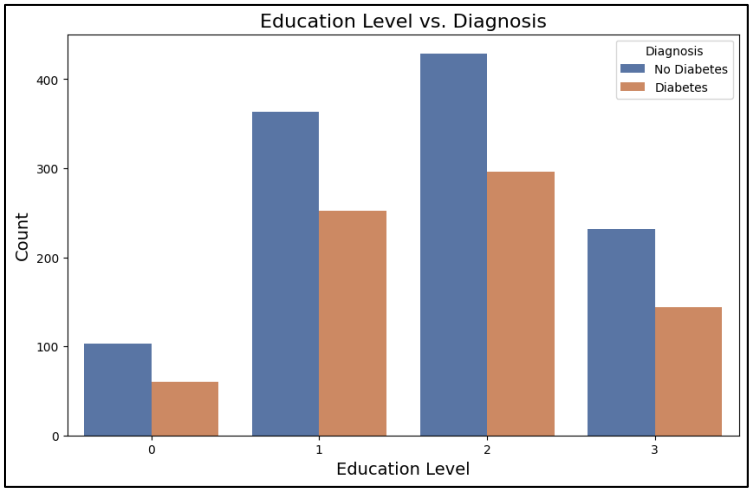

Figure 7: Portrays Education Level vs. Diagnosis

The above histogram, "Education Level vs. Diagnosis," has strong trends in the diagnosis of diabetes by education level. The education level variable is categorized into three groupings while the counts represent the number diagnosed with and without diabetes. The histogram gives the count for each of these categories at a different level of education; the highest counts occur for education level 1: about 450 persons without diabetes and about 100 diagnosed, which suggests that lower education level and higher prevalence of diabetes might be strongly associated. Another remarkable difference in this respect is portrayed in education level 2 also, in which approximately 400 persons were not diagnosed while about 200 are diagnosed. At the same time, education level 0 has the lowest counts across the board, with only about 50 persons diagnosed with diabetes. This therefore shows that the incidence of diabetes could be lower at increased levels of education, hence pointing out the fact that education may play a role in health outcomes.

Methodology

Feature Engineering and Selection

Feature engineering and selection form some of the most important steps within any machine learning pipeline since most of the performance that affects a predictive model normally relies on them. Feature engineering represents important underlying patterns in the data. Techniques for extracting and engineering relevant features can vary widely, with the nature of the dataset in both general and specific domains affecting how the problem is to be solved. The method deployed included one-hot encoding for categorical variables-a technique that converts categorical data into a binary matrix to allow machine learning algorithms to treat these variables accordingly normalization and standardization of numerical features, which bring the scales of different features to the same level so that no feature will dominate the model because of its range. In the dataset revolving around diabetes, entailed age, body mass index (BMI), and blood glucose are features that should be normalized to have an equal contribution in training models.

Furthermore, domain knowledge played a great role in feature engineering. In the context of diabetes prediction, the incorporation of medical knowledge generated interaction terms, such as body mass index multiplied by age, revealing more about how these factors interactively influence diabetes risk. Other techniques used included the use of feature extraction, which used Principal Component Analysis; hence, a tremendous reduction in dimensionality would only lose a minimum of information. Feature selection, in which every contribution of each feature concerning its predictive power for the model, is judged relevant or not. The feature selection criteria may be certain statistical tests: independence chi-squared test, importance from tree-based models, or recursive feature elimination that iteratively removes the least important features. Ultimately, retain a subset of features that gives maximum performance with minimum overfitting so that the model generalizes well to unseen data.

Model Selection

The selection of machine learning models is one of the most important steps in the modelling process, as different algorithms have different strengths and weaknesses depending on the nature of the dataset and the goals of the prediction. In this work, the models used were Logistic Regression, Random Forest, and Support Vector Classifiers. *Logistic Regression* is a traditional statistical method that works very well for binary classification problems, such as predicting the presence or absence of diabetes. It is pretty strong to consider a model interpreting and efficient, especially when considering a small number of featured datasets. The model, however, assumes a linear relationship between the independent variable and log odds of the dependent variable-a hypothesis that may not be necessarily true in every case.

On the other hand, *Random Forest* represents a special type of ensemble learning. In training, several decision trees are constructed, with their output for classification problems giving the mode of the classes. It is insensitive to overfitting and models complex interactions between features with ease; hence, this algorithm works well with huge numbers of features and nonlinear relations. Also, Random Forest provides intrinsic feature importance scores that can guide further feature selection. *Support Vector Classifiers* are also a very good choice, especially in the case of high-dimensional data. SVC constructs hyperplanes in a multi-dimensional space to separate classes from each other and be able to handle cases when a boundary separating classes is nonlinear. The choice of model depends on data characteristics, such as dimensionality, nonlinearity, and class balance, but also on the specifics of the particular prediction task one is faced with.

Model Development and Evaluation

After the selection of models, their development and subsequent evaluation using the data gathered follow. First, there is a split of the dataset into training and testing sets to ensure that the model has been trained on one part of the data and tested on another for the evaluation of its generalization capabilities. A typical split can be 70% training versus 30% testing; this, however, always shows some variation depending on the dataset size and particular objectives of the analysis. After initial training, one should cross-validation to make model performance robust. Among the common properties of cross-validation, there is the method of the K-fold approach; the dataset is divided into multiple subsets, and the model training continues K times with the fact that each time one fraction acts as the test whilst others act as a larger set. It eliminates factors that may result in the possibility of overfitting your model and increases reliability around the estimate.

Another critical area in model development is hyperparameter tuning. Most of the machine learning algorithms have hyperparameters that regulate their performance, such as the depth of trees in Random Forests or regularization strength in Logistic Regression. The techniques that can be used to perform a systematic search over different combinations of hyperparameters include Grid Search or Random Search to optimize model performance based on predefined evaluation metrics. Model evaluation should be comprehensive and include, but not be limited to, accuracy, precision, recall, F1-score, and ROC-AUC. While accuracy provides a general indication of correctness, it can be misleading for imbalanced datasets where one class overwhelmingly dominates the other. Similarly, precision and recall give further insight into the model's efficiency in correctly identifying true positives and sensitivity, respectively. The F1-score, on the other hand, is the harmonic mean of the precision and recall, forcing a balance between the two.

Results and Analysis

Predictive Model Performance

a) Logistic Regression

The applied Python code performed an implementation of logistic regression for binary classification. First, the mandatory libraries were imported: train-test-split for data splitting, Logistic Regression for model creation, and metrics like accuracy-score, confusion-matrix, and classification-report for evaluation. With train-test-split, the code proceeded to split all the data into training sets and a test set of 30%. Subsequently, a Logistic-Regression model was instantiated with max_iter=1000 to allow for a larger number of iterations during training. The code was fitted against the training data by using fit(X_train, y_train). The prediction method is used on test data with predict(X_test) and performance metrics, such as accuracy, confusion matrix, and classification report, which gave insight into the model's accuracy, precision, recall, and F1-score as showcased below:

Output:

Table 1: Displays the Logistic Regression Classification Report

```
Logistic Regression Results:
Accuracy: 0.8351063829787234
Confusion Matrix:
 [[306  41]
 [ 52 165]]
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.88      0.87       347
           1       0.80      0.76      0.78       217

    accuracy                           0.84       564
   macro avg       0.83      0.82      0.82       564
weighted avg       0.83      0.84      0.83       564
```

This classification report describes the performance of the logistic regression model on the binary classification task. It follows that the overall performance, in terms of the percentage, is 83.5%, which means it predicted correctly for 83.5% of the cases. Precision, recall, and F1 scores are reported for each class. The precision for class 0 is 0.85, which implies that 85% of the instances predicted as class 0 are indeed class 0. Recall for class 0 is 0.88, meaning 88% of the actual class 0 instances are correctly identified by this model. The F1-score for class 0 is 0.87, the harmonic mean of precision and recall. For class 1, very similar metrics are reported. These metrics can also be summarized by computing the macro-average and weighted average for an overall summary of model performance across both classes.

b) Random Forest Classifier

The Python code implemented a Random Forest classifier for a binary classification task. First, the Random Forest Classifier is imported from the sklearn ensemble library. A Random Forest model is instantiated with n-estimators=100, which means that it consists of 100 decision trees. The model was then trained on the training data using fit (X_train, y_train). Fitted X_train and y_train- for prediction apply this code in test data X_test as. Finally, at the end, the performance of the model is checked through different metrics such as accuracy, confusion matrix, and classification report, which will gave an idea about the accuracy, precision, recall, and F1-score of the model as depicted below:

Output:

Table 2: Displays the Random Forest Classification Report

```
Random Forest Results:
Accuracy: 0.9078014184397163
Confusion Matrix:
 [[335  12]
 [ 40 177]]
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.97      0.93       347
           1       0.94      0.82      0.87       217

    accuracy                           0.91       564
   macro avg       0.91      0.89      0.90       564
weighted avg       0.91      0.91      0.91       564
```

The classification report above summarizes the performance of a Random Forest algorithm on a binary classification task. The algorithm attained an overall accuracy of 90.78%, implying that it correctly predicted the class label in 90.78% of the cases. Precision, Recall, and F1 measure for classes. Class 0: Precision is 0.89. This means that 89% of the instances predicted as class 0 are actually of class 0. Similarly, recall for class 0 is 0.97, which means that for actual class 0 samples, the model correctly classifies 97%. F1-score in class 0 is the harmonic mean of precision and recall, which is valued at 0.93. Similarly, for class 1, there are reported similar metrics. After that, all the metrics have been summarized into an overall macro average and weighted average showing the performance overall across the two classes.

c) Support Vector Machines

Code in Python language deployed the Support Vector Machine algorithm using a linear kernel. Import the class SVC from the library sklearn. svm. Instantiate the SVM model with the kernel as 'linear' and random-state as 42. Carry out a fit on the train data by calling the fit (X_train, y_train) method. It subsequently used the trained model for prediction on the test data, predict(X_test). Then, the performance of this model was measured using accuracy, a confusion matrix, and a classification report that provided insight into the model's accuracy, precision, recall, and F1 score:

Output:

Table 3:Showcases SVM Classification Report

```
SVM Results:
Accuracy: 0.8156028368794326
Confusion Matrix:
 [[298  49]
 [ 55 162]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.86      0.85       347
           1       0.77      0.75      0.76       217

    accuracy                           0.82       564
   macro avg       0.81      0.80      0.80       564
weighted avg       0.81      0.82      0.82       564
```

The given classification report summarizes the performance of an SVM model on a binary classification task. Overall, the achieved accuracy is 81.56% because this model predicts the correct class label for 81.56% of all cases. Precision, recall, and F1 scores are given for each class. For class 0, the precision value of 0.84 means that, among all the instances, 84% of predicted class 0 are classified as class 0, whereas the recall for this class is 0.86, which indicates how many instances of the whole actual class 0s were predicted correctly by the proposed model. The F1-score is the harmonic average of the precision and recall - the F1-score above is 0.85. Class 1 has similar metrics reported for it. The overall summary can then be seen by using a macro average and a weighted average of these metrics on the two classes.

Model Comparison Analysis

The Python code snippet compared three machine learning models: Logistic Regression, Random Forest, and Support Vector Machine. It defined a function, evaluate_model, that evaluated each model's performance on both training and test data, computing metrics such as accuracy, precision, recall, F1-score, and the time it took to train each model. It then loops over the defined models, first evaluating each with its respective evaluate_model function and then storing the results in a data frame for easy comparison. This script finally created a bar plot comparing model performance across different metrics to guide the choice of the most suitable model for any given classification task concerning their performance and efficiency as showcased below:

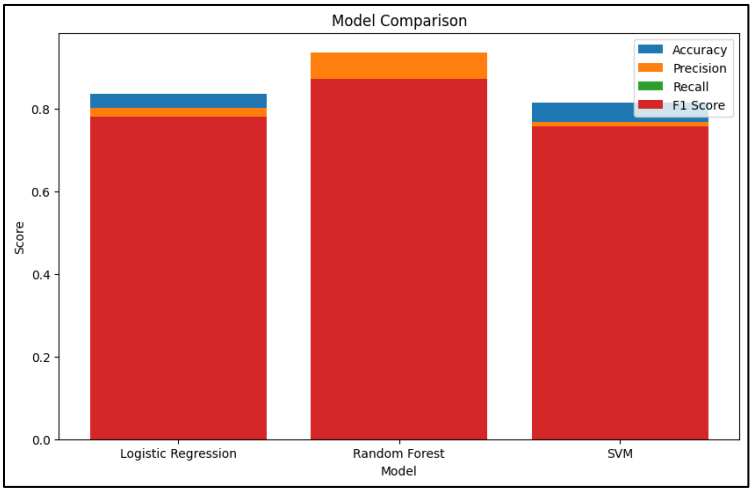

Figure 8: Exhibits Model Comparison

This bar chart presents the performance of three machine learning models, namely Logistic Regression, Random Forest, and SVM. These three models are compared based on performance metrics like accuracy, precision, recall, and F1-score. Random Forest outperformed other models in all metrics with the highest accuracy, precision, recall, and F1-score scores. SVM has a slightly lower performance than Random Forest but still outperforms Logistic Regression in all metrics. Logistic Regression turns in the poorest performance of the three, with the lowest scores across the board. Overall, this chart shows that Random Forest is the most effective model on this particular dataset, followed by SVM and Logistic Regression.

Feature Importance and Analysis

The importance of analysis always plays a very important role in feature engineering when it comes to machine learning development on any predictive model in life spans, especially in the healthcare domain. Moreover, this is key in contributing factors to an illness such as diabetes, which is under study in this paper. Assessing several features against the importance of each to different others provides the practitioner with some insight as to which variable will most influence model predictions. The most common models used for feature importance tuning are Random Forest and Support Vector Machines, which, in themselves, have ways of judging the importance of features. Random Forest is one ensemble learning method that builds multiple decision trees and merges their outputs to improve predictive accuracy and control overfitting. One of the major advantages of the Random Forest algorithm is that there are intrinsic feature importance scores. These are determined by the decrease in impurity of the nodes as every feature contributes to a split in the trees. That becomes important for the model to make predictions if some certain feature contributes much to the decrease of impurity measured with Gini impurity or entropy. Feature importance may then be calculated in Random Forest: first, the model trains on the dataset and computes each feature's importance based on its usage frequency for splitting throughout all trees. The importance of individual features occurring regularly during splits is high in rank. In the case of this diabetes prediction, one will find some important predictors: features like BMI, age, and blood glucose. That relative importance can also be visualized in plots-such as bar plots-the stakeholders like to see for intuition into which feature becomes most decisive.

Implications of Feature Importance Analysis

Feature importance analysis does not stop with model performance; instead, it provides actionable insights to both healthcare practitioners and policymakers. It can identify which of these features has greater influence in the case of diabetes and subsequently prompt appropriate interventions. For example, public health campaigns might concentrate on educating the communities on risks associated with high BMI and sedentary lifestyles through informed data insights shaping messaging and resources.

Besides, feature importance analysis can be used to guide clinical decision-making. In health care, the identification of high-risk individuals using predictive features makes it possible to conduct timely interventions, such as counseling on lifestyle modification or regular check-ups of blood glucose levels. This proactive approach will make a significant reduction in the incidences of diabetes and the complications associated with it. feature importance analysis using models like Random Forest and Support Vector Machines provides valuable insights into the factors influencing diabetes prediction.

These key features, such as age, BMI, blood glucose level, and family history, if identified, will enhance the understanding of researchers and practitioners in understanding the risk of diabetes and implementing targeted interventions. This analysis

enhances model interpretability and promotes informed decision-making in clinical and public health contexts that can contribute to better health outcomes among populations at risk for diabetes.

Discussion

Implication for the Management of Diabetes

Predictive modeling can bring potential transformation to diabetes management and prevention, furnishing health professionals with actionable insights to enable improved patient outcomes in the USA. Probably one of the strongest advantages predictive models can help in this domain is enabling the determination of a person at risk for diabetes probably well in advance of any beginning clinical symptoms showing up. Analysis of these trends in demographic, genetic, and lifestyle data, through various predictive models, can certainly allow the identification of the high-risk population who might undergo targeted interventions, consisting of lifestyle modification programs and/or early pharmacological treatment. This could reduce eventual incidences of new cases of diabetes in the country, especially in its weak segments.

Integration of predictive models into clinical workflows may further simplify diabetes care. For instance, predictive algorithms can be integrated into EHR systems to flag patients for closer monitoring or further testing. This could support clinicians with prioritization of patients, optimization of resource allocation, and personalization of care planning. Recommendations for effective integration included training healthcare providers to interpret the outputs from predictive models and embedding the tools into existing clinical decision support systems. Such systems would generate risk scores, illuminate paramount risk factors, and incorporate evidence-based recommendations within specific workflows at the point required by a clinician's judgment.

Impact on Public Health and Healthcare Cost

The potential benefits of predictive modeling in diabetes care are significant. Improved early detection and management can prevent the prevalence of diabetes and its complications, thus improving the quality of life for millions of Americans. For example, a predictive model identifying prediabetic individuals with high accuracy can enroll those people in structured lifestyle intervention programs, which will drastically reduce the progression to Type 2 diabetes. Besides that, predictive models might independently have the potential to point toward the early identification of complications amongst patients, such as retinopathy or nephropathy, which in their more advanced stages could turn lethal, thus avoiding a possible further deterioration and consequent irreversibility of pathologies.

On a larger scale, predictive modelling may decrease health inequity by targeting vulnerable populations, which include groups from underserved communities. Diabetes, for instance, is also more prevalent in minority populations and those living in lower-income settings throughout the U.S. Predictive analytics can bring these disparities into view to help target evidence-based, tailored public health campaigns and resource allocation accordingly. Predictive models may, for example, guide a diabetes prevention program from a public health initiative into geographic locations where risk factors are most acute, thereby assuring more effective utilization of scarce resources.

In terms of healthcare costs, the economic benefits of predictive modeling are equally compelling. Diabetes management is an extremely expensive entity in the United States alone, consuming a total amount of $327 billion as direct and indirect costs per year. Predictive analytics can curtail these expenses through early identification and also through the reduced incidence of complications that are associated with diabetes, which alone are one of the costly features of diabetic management. For example, not being hospitalized because of uncontrolled diabetes or not needing very expensive treatments for advanced complications, such as dialysis or amputation, could save billions of dollars every year. Besides, predictive models may optimize the use of healthcare resources by avoiding unnecessary testing and concentrating interventions on those patients who would most benefit from them.

Limitations and Challenges

Predictive modeling does come with its challenges and limitations. Probably the most important ethical consideration is sensitive healthcare data. Predictive models rely on very large datasets, including EHRs, insurance claims, and even wearable device data. Ensuring the privacy and security of this data is paramount, especially with regulations such as HIPAA. Another aspect in this respect is the factor of the potential misuse to which predictive models can put into service-for instance, by insurance companies either to eventually deny coverage or charge considerably higher premiums for high-risk individuals. This would raise fairness and equity concerns.

Another challenge is data quality and completeness. Most of the predictive models depend on historical data from EHRs, which are generally characterized by missing or inconsistent entries. For instance, not all medical professionals document patient information equally, and that leads to gaps in data that will hurt model accuracy. Besides, the interpretability of complex machine

learning models, such as deep learning algorithms, remains a barrier to adoption in clinical settings. There, clinicians need to trust a model and understand the rationale behind its predictions, with life-altering decisions potentially at stake.

Another limitation is generalizability. Predictive models developed from specific populations or regions may fail in different populations. For instance, a model that has been developed based on data from predominantly urban health systems may not consider risk factors associated with rural populations due to limited access to healthcare or different dietary patterns. Such challenges need the development of diverse datasets capturing the heterogeneity of the US population.

Future Research Directions

To address these challenges and completely achieve the potential of predictive modelling in diabetes care, future research should focus on improving model accuracy and applicability. Probably, larger and more diverse datasets present the best avenue. This is supported by the development of comprehensive health databases that may grant researchers access to diversity representative of the U.S. population, like the All of Us Research Program. Incorporating genetic data into predictive models could also enhance their accuracy, particularly for individuals with a family history of diabetes.

Advancements in technology present a plethora of opportunities for consolidating real-time health data into predictive models. Such wearables as CGMs, Fitbits, and smart time wear are designed to monitor constantly such activities as the amount of physical activity, patterns of sleep, and level of glucose. By integrating this information into predictive algorithms, researchers would be able to develop models that dynamically could generate risk profiles in real time. For instance, a wearable device could warn the user when their glucose level is trending toward dangerous thresholds or when their physical activity level is not high enough to maintain health.

Interdisciplinary collaboration would be very important in continuing to advance the predictive modeling of diabetes care. Data scientist and healthcare provider partnerships, along with policymaker stakeholders, will be key drivers in closing the gap between research and practice. For instance, machine learning experts collaborate with clinicians in ensuring that predictive models provide both accurate and clinically relevant results, while policymakers can address such ethical and regulatory challenges that impede the adoption of these tools.

Conclusion

The main objective of this study was to establish how predictive modeling can be used to enhance the management and prevention of diabetes in the United States. This study focused on the deployment of predictive modeling methods to support diabetes management in the United States, with an emphasis on data-driven decision-making in clinical settings and public health policy. The dataset for this research project was retrieved from the Kaggle website, a proven and credible dataset source. The Diabetes prediction dataset included medical and demographic data of the patients along with their respective diabetic status. The provided data included age, gender, body mass index, hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. In this work, the models used were Logistic Regression, Random Forest, and Support Vector Classifiers. Random Forest outperformed other models in all metrics with the highest accuracy, precision, recall, and F1-score scores. SVM had a slightly lower performance than Random Forest but still outperformed Logistic Regression in all metrics. Overall, the Random Forest was the most effective model on this particular dataset, followed by SVM and Logistic Regression. Predictive modelling can bring potential transformation to diabetes management and prevention, furnishing health professionals with actionable insights to enable improved patient outcomes in the USA. Integration of predictive models into clinical workflows may further simplify diabetes care. For instance, predictive algorithms can be integrated into EHR systems to flag patients for closer monitoring or further testing.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Adeniran, I. A., Efunniyi, C. P., Osundare, O. S., & Abhulimen, A. O. (2024). Data-driven decision-making in healthcare: Improving patient outcomes through predictive modeling. Engineering Science & Technology Journal, 5(8).
[2] Al Amin, M., Liza, I. A., Hossain, S. F., Hasan, E., Haque, M. M., & Bortty, J. C. (2024). Predicting and Monitoring Anxiety and Depression: Advanced Machine Learning Techniques for Mental Health Analysis. *British Journal of Nursing Studies*, *4*(2), 66-75.
[3] Al Nahian, A., Ahmed, S. W., Shorif, M. N., Atayeva, J., & Rizvi, S. W. A. (2024). Optimizing Healthcare Outcomes through Data-Driven Predictive Modeling. Journal of Intelligent Learning Systems and Applications, 16(4), 384-402.
[4] Alam, S., Hider, M. A., Al Mukaddim, A., Anonna, F. R., Hossain, M. S., khalilor Rahman, M., & Nasiruddin, M. (2024). Machine Learning Models for Predicting Thyroid Cancer Recurrence: A Comparative Analysis. *Journal of Medical and Health Studies*, *5*(4), 113-129.

[5] Bhowmik, P. K., Miah, M. N. I., Uddin, M. K., Sizan, M. M. H., Pant, L., Islam, M. R., & Gurung, N. (2024). Advancing Heart Disease Prediction through Machine Learning: Techniques and Insights for Improved Cardiovascular Health. *British Journal of Nursing Studies*, *4*(2), 35-50.

[6] Bortty, J. C., Bhowmik, P. K., Reza, S. A., Liza, I. A., Miah, M. N. I., Chowdhury, M. S. R., & Al Amin, M. (2024). Optimizing Lung Cancer Risk Prediction with Advanced Machine Learning Algorithms and Techniques. *Journal of Medical and Health Studies*, *5*(4), 35-48.

[7] Dritsas, E., & Trigka, M. (2022). Data-driven machine-learning methods for diabetes risk prediction. *Sensors*, *22*(14), 5304.

[8] Dutta, S., Sikder, R., Islam, M. R., Al Mukaddim, A., Hider, M. A., & Nasiruddin, M. (2024). Comparing the Effectiveness of Machine Learning Algorithms in Early Chronic Kidney Disease Detection. *Journal of Computer Science and Technology Studies*, *6*(4), 77-91.

[9] Hasan, E., Haque, M. M., Hossain, S. F., Al Amin, M., Ahmed, S., Islam, M. A., ... & Akter, S. (2024). CANCER DRUG SENSITIVITY THROUGH GENOMIC DATA: INTEGRATING INSIGHTS FOR PERSONALIZED MEDICINE IN THE USA HEALTHCARE SYSTEM. *The American Journal of Medical Sciences and Pharmaceutical Research*, *6*(12), 36-53.

[10] Hossain, S., Miah, M. N. I., Rana, M. S., Hossain, M. S., Bhowmik, P. K., & Rahman, M. K. (2024). ANALYZING TRENDS AND DETERMINANTS OF LEADING CAUSES OF DEATH IN THE USA: A DATA-DRIVEN APPROACH. *The American Journal of Medical Sciences and Pharmaceutical Research*, *6*(12), 54-71.

[11] Hossain, M. S., Rahman, M. K., & Dalim, H. M. (2024). Leveraging AI for Real-Time Monitoring and Prediction of Environmental Health Hazards: Protecting Public Health in the USA. *Revista de Inteligencia Artificial en Medicina*, *15*(1), 1117-1145.

[12] Hider, M. A., Nasiruddin, M., & Al Mukaddim, A. (2024). Early Disease Detection through Advanced Machine Learning Techniques: A Comprehensive Analysis and Implementation in Healthcare Systems. *Revista de Inteligencia Artificial en Medicina*, *15*(1), 1010-1042.

[13] Ghosh, B. P., Bhowmik, P. K., & Bhuiyan, M. S. (2024). Advanced Disease Detection and Personalized Medicine: Integrated Data Approaches for Enhanced Parkinson's Disease and Breast Cancer Detection and Treatment in the USA. *British Journal of Pharmacy and Pharmaceutical Sciences*, *1*(2), 11-30.

[14] Islam, M. Z., Nasiruddin, M., Dutta, S., Sikder, R., Huda, C. B., & Islam, M. R. (2024). A Comparative Assessment of Machine Learning Algorithms for Detecting and Diagnosing Breast Cancer. *Journal of Computer Science and Technology Studies*, *6*(2), 121-135.

[15] Kraus, M., Feuerriegel, S., & Saar-Tsechansky, M. (2024). Data-driven allocation of preventive care with application to diabetes mellitus type II. *Manufacturing & Service Operations Management*, *26*(1), 137-153.

[16] Kriegova, E., Kudelka, M., Radvansky, M., & Gallo, J. (2021). A theoretical model of health management using data-driven decision-making: the future of precision medicine and health. *Journal of translational medicine*, *19*, 1-12.

[17] Kasula, B. Y. (2023). Machine Learning Applications in Diabetic Healthcare: A Comprehensive Analysis and Predictive Modeling. *International Numeric Journal of Machine Learning and Robots*, *7*(7).

[18] Lauffenburger, J. C., Mahesri, M., & Choudhry, N. K. (2020). Not there yet: using data-driven methods to predict who becomes costly among low-cost patients with type 2 diabetes. *BMC Endocrine Disorders*, *20*, 1-10.

[19] Mia, M. T. (2024). Enhancing Lung and Breast Cancer Screening with Advanced AI and Image Processing Techniques. *Journal of Medical and Health Studies*, *5*(4), 81-96.

[20] Nasiruddin, M., Hider, M. A., Akter, R., Alam, S., Mohaimin, M. R., Khan, M. T., & Sayeed, A. A. (2024). OPTIMIZING SKIN CANCER DETECTION IN THE USA HEALTHCARE SYSTEM USING DEEP LEARNING AND CNNS. *The American Journal of Medical Sciences and Pharmaceutical Research*, *6*(12), 92-112.

[21] Nasiruddin, M., Dutta, S., Sikder, R., Islam, M. R., Mukaddim, A. A., & Hider, M. A. (2024). Predicting Heart Failure Survival with Machine Learning: Assessing My Risk. *Journal of Computer Science and Technology Studies*, *6*(3), 42-55.

[22] Omana, J., & Moorthi, M. (2022). Predictive analysis and prognostic approach of diabetes prediction with machine learning techniques. *Wireless Personal Communications*, *127*(1), 465-478.

[23] Pant, L., Al Mukaddim, A., Rahman, M. K., Sayeed, A. A., Hossain, M. S., Khan, M. T., & Ahmed, A. Genomic predictors of drug sensitivity in cancer: Integrating genomic data for personalized medicine in the USA.

[24] Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., ... & Ren, Z. (2022). Machine learning models for data-driven prediction of diabetes by lifestyle type. *International journal of environmental research and public health*, *19*(22), 15027.

[25] Rahman, A., Karmakar, M., & Debnath, P. (2023). Predictive Analytics for Healthcare: Improving Patient Outcomes in the US through Machine Learning. *Revista de Inteligencia Artificial en Medicina*, *14*(1), 595-624.

[26] Singhania, K., & Reddy, A. (2024). Improving preventative care and health outcomes for patients with chronic diseases using big data-driven insights and predictive modelling. *International Journal of Applied Health Care Analytics*, *9*(2), 1-14.

[27] Ye, S. (2024, November 9). Diabetes prediction. Kaggle. Retrieved from: https://www.kaggle.com/datasets/gaolang/diabetes-prediction