International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Predicting Diabetes through Data Analytics Enhancing Early Detection and Intervention

Umamaheswara Reddy Kudumula

Engineer Lead, EDA - Provider, Employer and Financial Analytic Solutions, Anthem, Inc, Atlanta, Georgia, United States

Abstract: ***The United States is grappling with a chronic disease crisis, with 90% of the nation's $4.1 trillion annual healthcare expenditure dedicated to managing chronic and mental health conditions. Diabetes stands out among these chronic diseases, affecting 11.6% of the U. S. population. Nearly 38 million Americans are currently living with diabetes, and an additional 96 million are prediabetic, placing them at high risk for developing type 2 diabetes [1]. This condition leads to severe complications, such as heart disease, blindness, and kidney failure. In 2022, the cost of diagnosed diabetes was nearly $400 billion, encompassing medical expenses and productivity losses. Diabetes is among the top 10 causes of mortality in the U. S., with an average death rate of 31.1 per 100, 000 people. Early detection and prevention are crucial to reducing [2] healthcare costs, improving quality of life, and preventing premature deaths. This white paper explores the potential of predictive data analytics to identify individuals at risk of developing diabetes. By leveraging these insights, healthcare organizations can implement targeted interventions, enhance patient outcomes, improve quality of life, and reduce overall healthcare costs.***

Keywords: Chronic Conditions, Diabetes, Prevalence, Data Analytics, Healthcare, Patient Care, Chronic Disease

1. Introduction

Chronic conditions are long - term health issues requiring ongoing medical attention and management. These conditions are inherently complex, with management challenges influenced by various factors, including social and economic determinants, healthcare provider capabilities, system - related issues, treatment plans, and associated costs. Among these chronic diseases, diabetes stands out as one of the most prevalent, ranking among the top 10 in the United States.

Diabetes is a hazardous condition that affects how the body converts food into energy and processes blood sugar (glucose). Its long - term effects are far - reaching, impacting vital organs and systems such as the heart, kidneys, nerves, eyes, and feet. Diabetes is also one of the leading causes of death in the U. S., significantly contributing to heart disease, the top chronic condition leading to mortality.

The prevalence of diabetes in the United States is increasing rapidly, outpacing previous projections. By 2030, it is estimated that 39.7 million adults (13.9%) will be diagnosed with diabetes, with the number of people aged 65 and older with diabetes expected to rise to 21 million. This growing trend presents a substantial challenge to the U. S. healthcare system, adversely affecting the quality of life for millions of Americans and adding to the healthcare cost burden.

Early prediction and prevention of diabetes are essential to mitigating this growing problem, and data analytics plays a crucial role in this effort. Data analytics can be broadly classified into four [3] categories: descriptive, prescriptive, predictive, and diagnostic. This paper focuses on predictive analytics, examining its potential to forecast the likelihood of individuals developing diabetes. By harnessing predictive analytics, healthcare organizations can develop targeted intervention strategies, improving patient outcomes, enhancing quality of life, and reducing healthcare costs.

2. Solution

We will use predictive data modeling techniques to predict the likelihood of an individual developing diabetes [4]. Predictive data models utilize historical data to forecast future outcomes, and they can be broadly classified into two main categories: classification models and regression models. Regression models predict numerical outcomes, while classification models predict class memberships.

In this context, we will employ logistic regression, a statistical analysis method used to predict a binary outcome based on one or more independent variables [5]. Our objective is to predict whether an individual is likely to develop diabetes. The steps involved in building this predictive model are as follows:

a) Data Collection:
We will gather data from the National Institute of Diabetes and Digestive and Kidney Diseases, which [6] includes medical information and laboratory analyses. The dataset will consist of several variables, some independent (predictor variables) and one dependent variable (outcome). Key attributes in the dataset include:

- Pregnancies: Number of pregnancies
- Glucose: Blood glucose level
- Blood Pressure: Blood pressure value
- Skin Thickness: Thickness of the skin
- Insulin: Insulin level in the blood
- BMI: Body Mass Index
- Diabetes Pedigree Function: Likelihood of diabetes based on family history
- Age: Age in years
- Outcome: Indicates the presence of diabetes (1 for yes, 0 for no)

b) Data Extraction and Transformation:
We will use Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, SciPy, and Sklearn for data manipulation, analysis, visualization, and model building.

The data will be imported into a Pandas DataFrame, followed by data cleaning and preprocessing to handle missing values, outliers, and inaccuracies. Functions like describe (), isna (), and shape () will be used to analyze and clean the data.

Importing Libraries and Loading Data

The first step involves importing the necessary Python libraries and loading the dataset. The Pandas library will be used for data manipulation, while NumPy will handle large multidimensional arrays. Matplotlib and Seaborn will be used for data visualization, and Sklearn will provide machine learning algorithms for model building.

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score

# Load the dataset
data = pd.read_csv('diabetes.csv')
```

c) Data Cleaning:

Data cleaning involves handling missing values, removing duplicates, and correcting any inconsistencies. Missing values can significantly impact the performance of [8] the model, and hence, they need to be addressed appropriately.

```python
# Check for missing values
print(data.isna().sum())

# Impute missing values
data['Glucose'].replace(0, np.nan, inplace=True)
data['BloodPressure'].replace(0, np.nan, inplace=True)
data['SkinThickness'].replace(0, np.nan, inplace=True)
data['Insulin'].replace(0, np.nan, inplace=True)
data['BMI'].replace(0, np.nan, inplace=True)

# Fill missing values with the median of each column
data.fillna(data.median(), inplace=True)
```

d) Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) helps in understanding the underlying patterns and relationships within the data. Visualization techniques such as histograms, box plots, and scatter plots will be used to explore the [9] data.

```python
# Visualize the distribution of the outcome variable
sns.countplot(x='Outcome', data=data)
plt.title('Distribution of Diabetes Outcome')
plt.show()

# Pairplot to see the pairwise relationships between variables
sns.pairplot(data, hue='Outcome')
plt.show()
```

e) Data Analysis using Visualization:
Data visualization is crucial for understanding data patterns, distributions, and relationships between variables. We will use Seaborn's boxplot, scatterplot, and catplot functions to explore the data. A correlation matrix will help identify the relationships between variables for feature selection.

```python
# Correlation matrix
corr = data.corr()
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

f) Model Development:
Feature selection is a critical step in model development. We will choose essential features such as glucose, blood pressure, insulin, skin thickness, BMI, and age, while excluding less relevant features like pregnancies and diabetes pedigree function. The dataset will be split into training and test sets, and we will use Sklearn's linear_model to build and train the logistic regression model.

```python
# Feature selection
X = data[['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'Age']]
y = data['Outcome']

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Build and train the logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)
```

g) Model Validation:
Model validation involves evaluating the model's performance on the test dataset. A confusion matrix will summarize the outcomes, showing correct and incorrect predictions per class. We will also calculate the model's accuracy score and test it on random datasets.

```python
# Predict the outcomes on the test set
y_pred = model.predict(X_test)

# Confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues')
plt.title('Confusion Matrix')
plt.show()

# Accuracy score
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy * 100:.2f}%')
```

The logistic regression model predicts whether an individual is likely to have diabetes with an accuracy of 81.57%. The confusion matrix provides detailed insights into the model's performance, showing the number of true positives, true negatives, false positives, and false negatives.

Confusion Matrix:
Accuracy Score: 81.57%
These results indicate that the model is reasonably accurate in predicting diabetes, but there is still room for improvement. Further tuning of the model and incorporating more sophisticated algorithms could enhance its predictive power.

```
In [102]:
pred = model.predict(X_test)
pred

Out[102]:
array([0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0,
       0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1,
       1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0,
       0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0],
      dtype=int64)

In [104]:
from sklearn.metrics import confusion_matrix
conf_mtrx = confusion_matrix(Y_test, pred)
conf_mtrx

Out[104]:
array([[95, 12],
       [16, 29]], dtype=int64)

In [99]:
from sklearn.metrics import accuracy_score
accuracy =accuracy_score(Y_test, y_predict)
accuracy

Out[99]:
0.8157894736842105

In [116]:
pred = model.predict([[150,80,120,25,35,60]])
print(pred)
if pred==1:
    print("Diabetic")
else:
    print("Non Diabetic")

[1]
Diabetic

In [117]:
pred = model.predict([[100,100,500,25,21,25]])
print(pred)
if pred==1:
    print("Diabetic")
else:
    print("Non Diabetic")

[0]
Non Diabetic
```

3. Applications of the Solution in Various Organizational Processes:

a) **Predicting Other Chronic Conditions in Healthcare:** Chronic conditions account for a significant portion of [10] mortality and healthcare costs. This model can predict the likelihood of various chronic conditions, helping healthcare organizations develop targeted interventions, improve patient outcomes, and reduce costs.

Chronic diseases such as hypertension, cardiovascular diseases, and chronic respiratory diseases share common risk factors with diabetes. By using similar predictive models, healthcare providers can identify at - risk individuals and implement preventive measures early on.

b) **Predicting Fraudulent Transactions in Financial Institutions:** Credit card fraud is a growing concern, with significant financial losses reported annually. Logistic regression can predict fraudulent transactions based on historical data, helping financial institutions save billions of dollars.

Fraud detection models use a variety of features such as transaction amount, location, time of transaction, and past transaction history to predict fraudulent activities. By continuously updating the models with new data, financial institutions can improve the accuracy of fraud detection.

c) **Identifying Suspicious Emails in Cybersecurity:** Phishing emails are a primary vector for cyber - attacks. Logistic regression can help identify suspicious emails, protecting organizations from security breaches and financial losses.

Suspicious email detection models analyze various features such as the sender's email address, email content, presence of attachments, and links. By training these models on large datasets of known phishing and legitimate emails, cybersecurity systems can effectively filter out malicious emails.

4. Benefits of the Solution

a) **Prevention of Diabetes:** Identifying at - risk individuals allows healthcare organizations to implement preventive measures, reducing the incidence of type 2 diabetes and its associated complications.
Preventive measures include lifestyle interventions such as promoting physical activity, healthy eating, and regular monitoring of blood glucose levels. Early intervention can significantly reduce the progression of prediabetes to diabetes.

b) **Enhanced Patient Outcomes:** Predictive models enable personalized treatment plans, addressing associated risks such as high blood pressure and cardiovascular issues, thereby improving patient outcomes.
Personalized treatment plans can include medication management, regular check - ups, and continuous monitoring of vital health metrics. Tailoring treatment to individual needs enhances the effectiveness of healthcare interventions.

c) **Improved Quality of Life:** Early identification and targeted interventions can significantly enhance the quality of life for individuals at risk of developing diabetes.
Quality of life improvements are achieved by reducing the burden of diabetes - related complications. Patients can maintain better health and well - being through proactive healthcare measures.

d) **Reduced Healthcare Costs:** By preventing diabetes and its complications, the model helps reduce the overall healthcare costs associated with [11] managing the disease.
Healthcare cost savings are realized by reducing hospital admissions, emergency room visits, and long - term treatment expenses for diabetes - related complications.

e) **Reduced Diabetes Prevalence**
The model can help control the rapid increase in diabetes prevalence, contributing to better public health outcomes.
Public health initiatives can leverage predictive analytics to design community - based programs aimed at diabetes prevention and education. By addressing the root causes of diabetes, such initiatives can curb the rising trend of diabetes prevalence.

5. Conclusion

Predictive data analytics is essential for developing cost - effective healthcare strategies to manage chronic conditions like diabetes. Data - driven insights enable healthcare organizations to improve care quality, enhance patient outcomes, reduce prevalence, and lower healthcare costs. This white paper highlights the vital role of data analytics in addressing the challenges posed by chronic conditions and provides a technical perspective on implementing data - driven solutions in healthcare.

The integration of predictive analytics into healthcare systems represents a significant advancement in medical science. As technology continues to evolve, the potential for data analytics to transform healthcare becomes increasingly apparent. By harnessing the power of data, healthcare providers can [12] offer more accurate diagnoses, personalized treatments, and preventive care, ultimately improving the health and well - being of the population.

References

[1] Nishi, S., Nishi, S., Viguiliouk, E., Kendall, C., Jenkins, D., Hu, F., Sievenpiper, J., Atzeni, A., Misra, A., Salas - Salvadó, J., & Salas - Salvadó, J. (2023). Nuts in the Prevention and Management of Type 2 Diabetes. Nutrients, 15 (4), 878.
[2] Stack, J. W., Brumley, C., Parikh, M., Canales, A., Mahoney, S. E., & Hearon, C. M. (2013). Factors Associated with Diabetes Risk in South Texas College Students. International Journal of Exercise Science. https: //digitalcommons. wku. edu/cgi/viewcontent. cgi?article=1585&context=ijes
[3] Pure Storage - Metaage Corporation. https: //www.metaage. tech/news/technology/200
[4] Alghamdi, T. (2023). Prediction of Diabetes Complications Using Computational Intelligence Techniques. Applied Sciences, 13 (5), 3030.
[5] Machine Learning with Regression Modeling Guide | SLIVR. https: //slivr. io/machine - learning - with - regression - modeling/
[6] Chen, T., Shang, C., Su, P., Antoniou, G., & Shen, Q. (2018). Effective Diagnosis of Diabetes with a Decision Tree - initialised Neuro - Fuzzy Approach. https: //doi. org/10.1007/978 - 3 - 319 - 97982 - 3_19

[7] Yan, K. (2021). Data Mining for Analyzing and Predicting the Success of Movies. https: //doi. org/10.17615/jf46 - j653
[8] AI's Achilles' Heel: The Consequence of Bad Data. https: //versium. com/blog/ais - achilles - heel - the - consequence - of - bad - data
[9] Key population - led community - based same - day antiretroviral therapy (CB - SDART) initiation hub in Bangkok, Thailand: a protocol for a hybrid type 3 implementation trial | Implementation Science Communications | Full Text. https: //implementationsciencecomms. biomedcentral. com/articles/10.1186/s43058 - 022 - 00352 - 9
[10] Guzmán, C. E. V., Mireles, G. F., Christopherson, N., & Janning, M. (2010). Class and race health disparities and health information seeking behaviors: The role of social capital. Research in the Sociology of Health Care. https: //doi. org/10.1108/s0275 - 4959 (2010) 0000028008
[11] Future of Care Chat – Episode 5: Reasons Why FHIR® Matters in the Modern Healthcare Data Exchange - VirtualHealth. https: //www.virtualhealth. com/blog/future - of - care - chat - reasons - why - fhir - matters - in - the - modern - healthcare - data - exchange/
[12] https: //www.trackstat. org/crm - software - for - organ - transplant - centers