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

Utilizing Data Analytics to Predict Chronic Condition: A Focus on Diabetes

Bharath Srinivasaiah

Solution Architect Sr, Leading Health Insurance Company, Richmond, Virginia, USA

Abstract: ***The United States of America has a crisis of chronic disease.90% of the nation's 4.1 trillion dollars in annual health care expenditure goes towards chronic and mental health condition management [1] [2]. Diabetes has the highest prevalence among chronic conditions in the [3] healthcare industry, impacting 11.6% of the U. S. population [4]. Nearly 38 million Americans have Diabetes, and another 96 million have prediabetes condition, which puts them at risk of developing type 2 diabetes. Diabetes causes severe complications like heart disease, blindness, or kidney failure. In 2022, the estimated cost of diagnosed Diabetes was nearly $400 billion in medical costs and productivity loss [4]. Diabetes is one of the Top 10 conditions resulting in mortality, with an average death of 31.1 per 100, 000 [6]. Early detection and prevention of Diabetes are required to reduce healthcare costs, improve the Quality of life, and prevent early deaths. This white paper explores the opportunity of using predictive data analytics to predict if a person is likely to develop Diabetes. This data helps Healthcare Organizations develop strategies to target interventions, enhance patient outcomes, improve Quality of life, and reduce healthcare costs.***

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

1. Introduction

Chronic conditions are health conditions that last longer and require ongoing medical attention. Management of Chronic Conditions is very challenging due to its complex nature. Many factors like social and economic, health care providers, system - related factors, treatment plans, and related costs impact chronic condition management. Diabetes is one of the top 10 prevalent Chronic diseases in America. It is a hazardous chronic condition that affects how the body converts food into energy and processes blood sugar (Glucose). This is a long - lasting chronic condition with long - term effects including but not limited to heart, kidney, nerves, eyes, and feet. Diabetes is also one of the top 10 conditions leading causes of death in the United States and one of the main contributing factors to the top chronic condition for death, which is heart disease. Diabetes as a condition is rapidly increasing in the United States, exceeding the prior predictions. The projected number of adults with diabetes diagnosis is expected to increase to 39.7 million (13.9%) in 2030 [5]. The number of people with Diabetes aged 65 years and over is expected to increase to 21.0 million in 2030 [5].

Diabetes has become a more significant challenge in the United States Healthcare system as it impacts the Quality of life of Americans and adds to the burden of healthcare costs. Early prediction and prevention of Diabetes is the solution to this problem, for which data analytics is vital. Data Analytics can be broadly classified into four sections: Descriptive, prescriptive, predictive, and diagnostic. In this paper, we explore predictive analytics and use this model to predict if a person is likely to develop Diabetes. This white paper explores the potential of predictive analytics to predict Diabetes in the population. Healthcare Organizations can utilize this predicted data to come up with strategies for targeted interventions.

2. Solution

We will use predictive data model techniques to predict if a person is likely to have Diabetes. Predictive data models are a statistical process that aims to predict the outcome or future based on historical data. All predictive models are made of algorithms and can be classified into two main categories: classification models and regression models [7]. Regression models are primarily used to predict a number, while classification models are used to predict class membership.

Logistic regression is a statistical analysis method to predict a binary outcome for a dependent variable by analyzing the relationship between one or more independent variables [8]. In our case, we will use logistic regression to predict whether a person will likely have Diabetes. By utilizing this method, we can develop a model that healthcare providers can use to identify patients at risk of Diabetes.

To build this regression model, we will use the data from the National Institute of Diabetes and Digestive and Kidney Diseases, which consists of medical information and laboratory analysis. Below are the steps we will be using in the process

- Data Collection: Identify and collect the required data from various sources
- Data Extraction and Transformation: Prepare and clean the collected data for analysis.
- Data Analysis using Visualization: Analyze and explore the data using visualizations to gain insights.
- Model Development: Build predictive models based on the analyzed data.
- Model Validation: Evaluate and validate the model's performance.

a) Data Collection:

This data collection contains Medical and Laboratory information. From the data set in the (. csv) file, we can find several variables; some of them are independent (several

medical predictor variables), and there is only one target dependent variable (Outcome). Below is the information on the dataset attributes

- Pregnancies: Number of pregnancies
- Glucose: Glucose level in blood
- BloodPressure: Blood pressure value
- SkinThickness: Thickness of the skin
- Insulin: Insulin level in blood
- BMI: Body Mass Index
- DiabetesPedigreeFunction: Diabetes Percentage
- Age: Age in number
- Outcome: To express the result, 1 is Yes, and 0 is No

Import the required Python libraries. Below are some of the libraries we would be using in our model.

- Pandas: The library is used for data manipulation and data analysis.
- NumPy: The library is used to work with large multidimensional arrays.
- Matplotlib: Library used for creating visualizations.
- Seaborn: The library is based on Matplotlib and is used to create advanced visualizations.
- Spicy: The library is used for statistical and probabilistic analysis.
- Sklearn: Machine Learning Library featuring various algorithms.

b) Data Extraction and Transformation:

```
In [1]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt
        from scipy.stats import skew
        import sklearn
```

Import the diabetes data set in csv to data frame using the panda's library.

```
In [2]: diabetes_dataset=pd.read_csv(".../diabetes.csv")
```

Data analysis and cleansing are critical steps in predictive model building. This involves using various functions to identify missing values, default values, outliers, errors, inconsistencies, and inaccuracies. By data cleansing, we improve the data quality and the accuracy of the model, reduce bias, save resources and time, improve the model performance, and higher efficiency. We will be using the below functions.

- Describe (): This will generate the data set's descriptive statistics, including counts, mean, std, min, and max values for each attribute.
- Isna () /Isin ({0}): This method verifies if there are any missing values, null values, or 0 values in the data frame attributes.
- Nm. shape (): Returns tuple value

```
In [3]: diabetes_dataset.describe()
```

Out[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

```
In [4]: diabetes_dataset.isna().sum()
Out[4]: Pregnancies                 0
        Glucose                     0
        BloodPressure               0
        SkinThickness               0
        Insulin                     0
        BMI                         0
        DiabetesPedigreeFunction    0
        Age                         0
        Outcome                     0
        dtype: int64

In [5]: diabetes_dataset.shape
Out[5]: (768, 9)

In [7]: diabetes_dataset.isin({0}).sum()
Out[7]: Pregnancies                 111
        Glucose                       5
        BloodPressure                35
        SkinThickness               227
        Insulin                     374
        BMI                          11
        DiabetesPedigreeFunction      0
        Age                           0
        Outcome                     500
        dtype: int64
```

Based on the above analysis, five attributes, Glucose, blood pressure, skin thickness, insulin, and BMI, have 0s in the rows. Missing values, null values, or 0s will adversely impact the model's accuracy. These also create bias in the results. Hence, handling the missing or null or 0 values in the data modeling process is critical.

To replace these values, we would need to verify the data distribution for each impacted attribute and identify if they are skewed. We would use the imputation technique to replace the missing or 0 values with some substitute values. Imputation techniques can be applied to Numerical Variables, categorical variables, or a combination of both. Below are some of the commonly used imputation methods.

- Mean Imputation: If data is Normally distributed
- Median Imputation: If data distribution is skewed
- Mode Imputation: If data is Catageroical

Using the skew () methods, we can find that the data for blood pressure, Insulin, and Age are skewed, and Glucose, Skinthickness, and BMI are close to normally distributed. Using the imputation technique, we will replace the 0's in Blood Pressure, Insure, and Age with their median value, while Glucose, Skinthicknees, and BMI will be replaced with the mean values.

```
In [8]: diabetes_dataset.skew()
```

```
Out[8]: Pregnancies                0.901674
        Glucose                    0.173754
        BloodPressure             -1.843608
        SkinThickness              0.109372
        Insulin                    2.272251
        BMI                       -0.428982
        DiabetesPedigreeFunction   1.919911
        Age                        1.129597
        Outcome                    0.635017
        dtype: float64
```

```
In [9]: attributelist_withskewness=['BloodPressure','Insulin','Age']
        attributelist_withoutskewness=['Glucose','SkinThickness','BMI']
        for clmn in attributelist_withoutskewness:
            diabetes_dataset[clmn]=diabetes_dataset[clmn].replace({0:diabetes_dataset[clmn].mean()})
        for clmn in attributelist_withskewness:
            diabetes_dataset[clmn]=diabetes_dataset[clmn].replace({0:diabetes_dataset[clmn].median()})
```

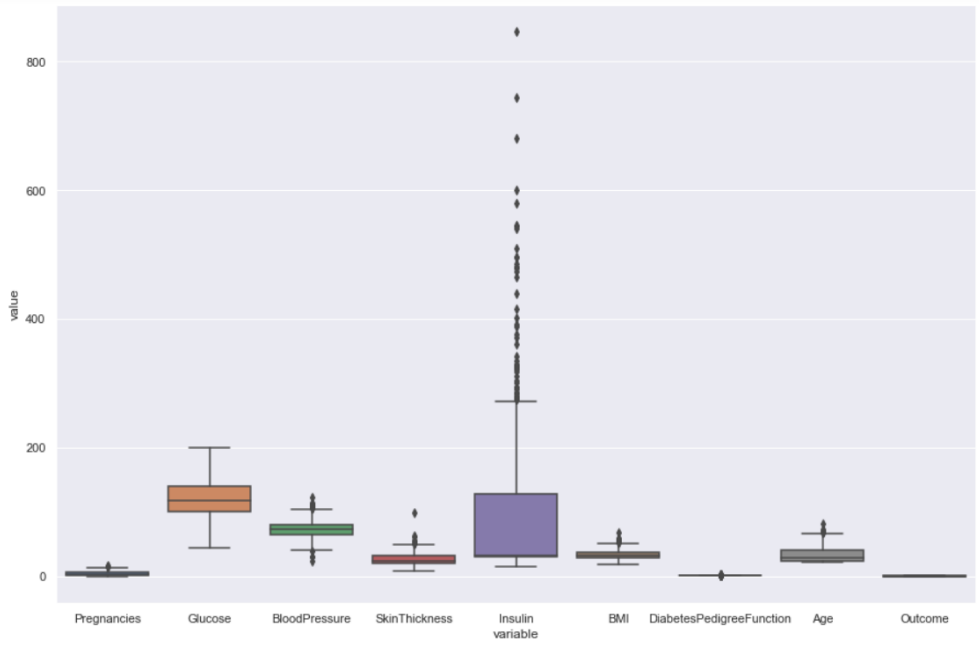
c) Data Analysis using Visualization:

Data Visualization is critical in predictive data analytics, especially when building a model. Visualization helps us unearth data patterns, identify outliers, review data distribution, identify skewness, identify trends, find correlations between variables, and guide the hypothesis. Below are some Visualization methods that can be leveraged to build models.

- Seaborn. boxplot (): It provides a visual summary of the variability of the data values, like mean, upper and lower quartiles, min and max values, and outliers [9].
- Seaborn. Scatterplot (): It is used to identify the relationship between two different variables.
- Seaborn. CatPlot (): It allows users to visualize the relationship between two continuous variables.

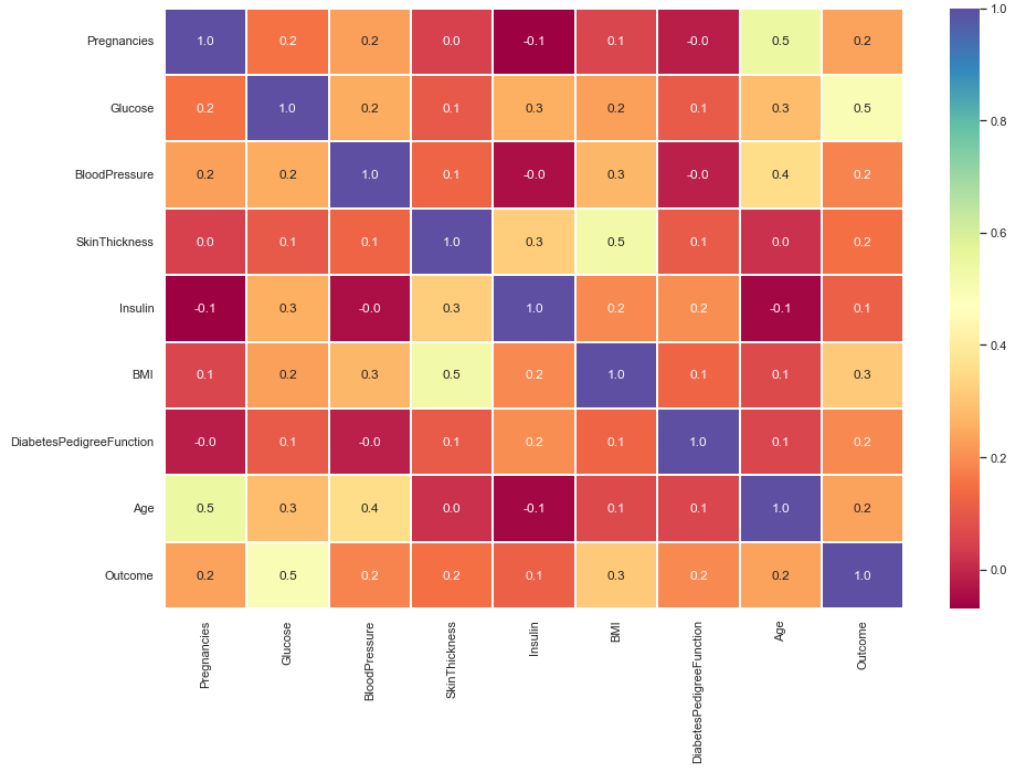
```
In [11]: import seaborn as sns
         sns.set(rc={"figure.figsize":(15, 10)})
         sns.boxplot(x="variable", y="value", data=pd.melt(diabetes_dataset))
```

```
Out[11]: <AxesSubplot:xlabel='variable', ylabel='value'>
```



We will use a correlation matrix to determine the relationship between the variables for feature selection.

```
In [35]: sns.heatmap(diabetes_dataset.corr(), annot=True, linewidths=0.2, fmt='.1f', cmap='Spectral')
Out[35]: <AxesSubplot:>
```



Logistic regression has critical assumptions that must be considered when building the model. Below are the details on some of the essential assumptions.

- The Dependent variable is binary. This is the basic assumption of logistic regression. In this case, the dependent variable's outcome can take only two values, 0 and 1, true and false, pass and fail.
- No multicollinearity exists between the independent variables. This assumption implies that all the predictor variables should be independent.
- There are no outliers in the dataset. This is a critical assumption for logistic regression. Outliers will negatively affect the model and its accuracy.

Based on the above key assumptions and the data visualization, we do not see much multicollinearity between the independent variables in the dataset used to build the model. However, we see outliers in attributes like skin thickness, BMI, and Glucose. For this model to work more accurately, we need to drop the outliers from the dataset before the key attributes can be selected as features to build the model. Identify such rows and remove them from the dataset.

```
In [87]: np.where(diabetes_dataset['Insulin']>500)
Out[87]: (array([  8,  13, 228, 247, 286, 409, 584, 655, 753], dtype=int64),)

In [88]: np.where(diabetes_dataset['BMI']>60)
Out[88]: (array([177], dtype=int64),)

In [89]: np.where(diabetes_dataset['SkinThickness']>60)
Out[89]: (array([445, 579], dtype=int64),)

In [90]: np.where(diabetes_dataset['SkinThickness']>50)
Out[90]: (array([ 57,  86,  99, 120, 211, 275, 445, 532, 579], dtype=int64),)

In [91]: diabetes_dataset.drop(diabetes_dataset[(diabetes_dataset['BMI'] >60)].index, inplace=True)
         diabetes_dataset.drop(diabetes_dataset[(diabetes_dataset['SkinThickness'] >60)].index, inplace=True)
         diabetes_dataset.drop(diabetes_dataset[(diabetes_dataset['Insulin'] >500)].index, inplace=True)

In [92]: diabetes_dataset.shape
Out[92]: (756, 9)
```

d) Model Development:

The critical step in building the model is feature selection. In this step, we will review the exploratory data analysis done in the previous step and select only the essential features that would contribute to the model to predict the outcome more accurately. Based on the analysis, we will use the attributes – Glucose, Blood pressure, Insulin, Skin Thickness, BMI, Age - and dropping pregnancies and diabetes pedigree function as those do not add much value to the model. Y is our output variable, which is the outcome with 0 and 1, not diabetic and diabetic, while X is our input variable or independent variable.

```
In [93]: Y=diabetes_dataset['Outcome']

In [94]: X=diabetes_dataset[['Glucose', 'BloodPressure', 'Insulin','SkinThickness',
                'BMI', 'Age']]
```

The next step in the model - building process is splitting the data. Splitting the data sets into train and test datasets will help to assess the model's performance. While a train data set is used to train the model, test data sets will help evaluate the model. We will use the sklearn linear_model to build the logistic regression model and train them.

```
In [101]: from sklearn.linear_model import LogisticRegression
          model = LogisticRegression()
          model.fit(X_train, Y_train)
          print("Training Score: ", model.score(X_train, Y_train))
          print("Testing Score:  ", model.score(X_test, Y_test))

          Training Score:  0.7649006622516556
          Testing Score:   0.8157894736842105
```

e) Model Validation:

The next step in the process is to test the model. This process evaluates the model's performance on the test data set. Upon testing, we will be creating the confusion matrix. This matrix will help to summarize the outcomes in matrix form, providing details on how many predictions are correct and how many predictions are wrong per class. This plots all the actual values along with the predicted values of a class. Finally, we will test the accuracy score of the logistic model and test it for some random datasets.

```
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
```

This model predicts if a person is likely to have diabetes with an accuracy of 81.57%.

3. Applications of the Solution in Various Organizational Processes

Logistic regression, a predictive data analytics model, has broad applications across various organizations. Below are some of the use cases

A. Predicting Other Chronic Conditions in HealthCare Industry

Chronic conditions are the primary cause of mortality and disability in the United States population.90% of the nation's 4.1 trillion dollars in annual health care expenditure goes towards chronic and mental health condition management [1]. This model can predict if a person is likely to have any specific chronic condition. Healthcare Organizations develop strategies to target interventions, improve quality of life, enhance patient outcomes, and reduce healthcare costs.

B. Financial Institutions to predict if a transaction is fraudulent or not

It's alarming that Credit Card Fraud has significantly risen since the last decade. As per the latest reports in the United States, there have been more than 5 million instances of fraudulent transactions, in the year 2022 related to identify theft, Fraud and other reports [10], causing losses in billions of dollars. We see a very significant increase in the fraudulent transactions. A logistic regression model can predict if the transaction is fraudulent or not based on historical data. This would help save billions of dollars in fraudulent transactions.

C. Predict if an email is Suspicious or not across various organizations.

According to the latest reports, over 90% of cyber - attacks are through suspicious emails [11]. Also referred to as phishing email, it has seen a 1265% increase last year. These emails are one of the primary causes of identity theft, which further results in significant financial losses. Organizations can lose valuable information, resulting in security breaches and financial loss. That's why it's critical to have email security. A logistic regression model can predict if an email is suspicious or not based on certain independent variables in the historical data set and prevent the passing.

4. Benefits of the Solution

This solution offers several benefits to the healthcare industry across the world. Here are the key benefits

a) ***Prevention of Diabetes:***
Diabetes is a chronic disease that, over time, will damage different organs in the body. Possible long - term effects include damage to the heart, nerves, eyes, feet, gums, and kidneys. This can also lead to heart disease and stroke. Healthcare Organizations can use this model to identify the risk population and prevent them from developing type 2 diabetes.

b) ***Enhance Patient Outcomes:***
Using this predictive model, healthcare organizations can develop strategies with a patient - centered approach to improve the outcomes. A person, if identified as likely to develop type 2 diabetes, will also have the hood to develop other chronic conditions like high blood pressure, Cholesterol, Insulin resistance, and cardiovascular issues. The new strategy should include a comprehensive plan to reduce all the associated risks, which will help to enhance patient outcomes.

c) ***Improve Quality of Life***:
With the prediction, in most cases, a person developing diabetes can be prevented. In some instances, there is a hood of people still developing diabetes. In such scenarios, healthcare organizations can develop strategies like targeted interventions in lifestyle, campaigns, and education that can improve the Quality of life.

d) ***Reduce Healthcare Costs:***
Diabetes alone is estimated to cost $400 billion in medical costs and cause productivity loss [3]. We can use this predictive model to slow down the prevalence of diabetes, which will help us reduce healthcare costs.

e) ***Reduce Diabetes Prevalence:***
Diabetes prevalence has rapidly increased in the last decade. By 2030, adult prevalence is expected to increase to 14% of the population. We can use this logistic model to prevent diabetes and help control or slow the prevalence rate.

5. Conclusion

In conclusion, the practical usage of predictive data analytics is required to develop cost - effective healthcare strategies to manage chronic conditions like Diabetes. By using data - driven insights, healthcare organizations can improve the quality of care, improve patient outcomes, reduce prevalence, and reduce healthcare costs [15]. This white paper provides a technical perspective on the vital role of data analytics in addressing challenges posed by chronic conditions like Diabetes. It includes guidance on data - driven solutions to transform healthcare system delivery.

References

[1] Buttorff C, Ruder T, Bauman M. Multiple Chronic Conditions in the United States [PDF - 393kb] Santa Monica, CA: Rand Corp.; 2017.

[2] National Health Expenditure Data: Historical. Center for Medicare & Medicaid Services. December 15, 2021. https: //www.cms. gov/Research - Statistics - Data - and - Systems/Statistics - Trends - and - Reports/NationalHealthExpendData/NationalHealthAccountsHistorical

[3] Hansen, Helen. "Health Outcomes of a Diabetes Supply and Diabetes Self - management Education Program in an At - risk Population. " 2000, https: //core. ac. uk/download/pdf/50516605. pdf.

[4] National Diabetes Statistics Report, *2023* | Diabetes | CDC. https: //www.cdc. gov/diabetes/health - equity/diabetes - by - the - numbers. html#: ~: text=The%20total%20estimated%20cost%20of, indirect%20costs%20attributable%20to%20diabetes.

[5] Lin, J, et al. "Projection of the Future Diabetes Burden in the United States through 2060. " Population Health Metrics, 2018, https: //doi. org/10.1186/s12963 - 018 - 0166 - 4.

[6] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp.261 - - 265). IEEE Computer Society Press.
[7] Meyr, H., et al. "Synthesis of Synchronization Algorithms. " 2001, https: //doi. org/10.1002/0471200573. ch5.
[8] What is Logistic Regression? Jan 2022 - Definition from SearchBusinessAnalytics. https: //www.techtarget. com/searchbusinessanalytics/definition/logistic - regression?amp=1
[9] Hartantyo, R. Y. (Rahadian), et al. "Ameliorative Effect of Infused Watercress on Rat Galactopoiesis Following Maternal Separation. " 2018, https: //media. neliti. com/media/publications/267305 - ameliorative - effect - of - infused - watercres - 62e3bf3d. pdf.
[10] Consumer Sentinel Network - . https: //www.ftc. gov/system/files/ftc_gov/pdf/CSN - Data - Book - 2022. pdf
[11] Top 15 phishing attack statistics (and they might scare you), https: //www.cybertalk. org/2022/03/30/top - 15 - phishing - attack - statistics - and - they - might - scare - you/