**HAP 780 Final Project Report**

**ANALYSIS OF LIFESTYLE AS A HEALTH INDICATOR FOR DIABETES**

**Team:** Mukesh Rajmohan

Priyanka Avagadda

Pranavi Rao Lingala

**ABSTRACT**

**Objective:** In this study, we evaluate the top 10 lifestyle factors which are critical for the prevalence or development of Diabetes in American population. We analyse which data mining model pulls out the best results for the study.

**Data:** We use CDC (Centre for Disease Control and Prevention) dataset for data analysis. Our dataset, ‘Diabetic Health Indicators’ includes healthcare statistics and lifestyle survey information. The sample size is 253,680 and includes 23 variables.

**Method:** We use SQL for data preprocessing and Weka 3.8 for predictive model analysis.

After randomly splitting the data into 80 and 20 percent, then we use the 80 percent of the data for feature selection as well as training the prediction models. As a result, we construct six prediction models using algorithms such as Logistic Regression, Naïve Bayes, Decision Tree, J48, Random Forest and Random Tree. We use ROC area, precision and recall values to compare the accuracy of these models with each other.

**Result:** Logistic Regression was the most reliable model for predicting Diabetes based on lifestyle with the best ROC curve area (0.797) and a high recall value (0.975).

**Keywords**: Lifestyle factors, Diabetes, CDC data, predictive model analysis.

1. **INTRODUCTION**

**Diabetes** is a chronic medical condition that occurs when the body is unable to effectively regulate blood sugar (glucose) levels. Glucose is the primary energy source for the body's cells, and insulin, a hormone produced by the pancreas, facilitates the uptake of glucose into cells.

* 1. **Types of Diabetes**
* Type 1 Diabetes:

An autoimmune condition where the immune system attacks and destroys insulin-producing beta cells in the pancreas. Commonly diagnosed in children and young adults but can occur at any age.

* Type 2 Diabetes:

A condition where the body becomes resistant to insulin or the pancreas doesn't produce enough insulin. It is more common in adults but increasingly seen in children and adolescents.

* 1. **Relationship between lifestyle factors and Diabetes**
* Type 1 Diabetes:
  + Lifestyle Factors: There is no direct correlation between lifestyle factors and the onset of Type 1 diabetes.
* Type 2 Diabetes:
  + Lifestyle Factors: A strong correlation exists between lifestyle factors and the risk of developing Type 2 diabetes.
  + Obesity: Excess body weight, particularly visceral fat, is a major risk factor as it leads to insulin resistance.
  + Physical Inactivity: Sedentary behaviour increases the likelihood of insulin resistance and weight gain.
  + Unhealthy Diet: Diets high in processed foods, sugars, and unhealthy fats contribute to weight gain and metabolic dysregulation.
  + Smoking and Alcohol: These habits can worsen insulin resistance and increase inflammation, elevating the risk.
  1. **Significance of Diabetes as a Global Problem**

Diabetes is a critical public health challenge with widespread implications for individuals, healthcare systems, and societies. Its significance lies in its global prevalence and potential for severe complications if left unmanaged.

* Rising Global Prevalence
  + Epidemic Scale: According to the International Diabetes Federation (IDF), over 500 million adults worldwide live with diabetes, a figure projected to rise to 783 million by 2045.
  + Undiagnosed Cases: Many individuals remain unaware of their condition, increasing the risk of complications and strain on healthcare systems.
  + Affecting All Age Groups: Type 2 diabetes, once predominantly seen in adults, is now increasingly diagnosed in children and adolescents due to rising obesity rates.
* Impact on Health

Diabetes significantly increases the risk of:

* + Cardiovascular Diseases: People with diabetes are at higher risk of heart attacks and strokes.
  + Kidney Failure: Diabetes is a leading cause of chronic kidney disease and end-stage renal disease.
  + Blindness: Diabetic retinopathy, caused by high blood sugar damaging blood vessels in the retina, is a major cause of vision loss.
  + Amputations: Poor circulation and neuropathy due to diabetes can lead to non-healing ulcers, often resulting in limb amputations.

1. **METHOD**
   1. **Dataset**

The dataset for this study is from CDC (The Centre for Disease Control and Prevention), a [national public health agency](https://en.wikipedia.org/wiki/National_public_health_institutes) of the United States. It is a [United States federal agency](https://en.wikipedia.org/wiki/Federal_agencies_of_the_United_States) under the [Department of Health and Human Services](https://en.wikipedia.org/wiki/United_States_Department_of_Health_and_Human_Services). Our dataset, “Diabetes Health Indicators” contains healthcare statistics and lifestyle survey information about people in the US along with their diagnosis of diabetes. The sample size is 253,680 and includes 23 variables. The target variable for classification is whether a patient has pre-diabetic, diabetic, or healthy.

**A screenshot of a graph

Description automatically generated**

* 1. **Variables**

In this study, the sample consists of 253,680 patients. The independent variables include

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Patient\_ID | Unique identifier for each patient |
| Age | Age of the patient |
| Sex | Sex of the patient |

The dependent variables include

|  |  |
| --- | --- |
| **Variable** | **Description** |
| HighBP | High blood pressure |
| HighChol | High cholesterol |
| CholCheck | Cholesterol check |
| BMI | Body Mass Index |
| Smoker | Smoking status |
| Stroke | History of stroke |
| HeartDiseaseorAttack | History of heart disease or attack |
| PhysActivity | Physical activity |
| Fruits | Fruit consumption |
| Veggies | Vegetable consumption |
| HvyAlcoholConsump | Heavy alcohol consumption |
| AnyHealthcare | Any healthcare insurance |
| GenHlth | General health |
| MentHlth | Mental health |
| PhysHlth | Physical health |
| DiffWalk | Difficulty in walking |
| NoDocbcCost | No doctor’s visit due to cost issues |

* 1. **Data Preprocessing**
* **Importing data into database:**

The dataset was imported into the Postgres database for structured data analysis. The import process ensured all rows and columns were correctly mapped to the corresponding database table schema, with appropriate data types assigned to each column.

* **Checking for missing and duplicate values:**

Extensive SQL queries were executed to identify missing and duplicate values in the dataset. It was determined that:

* Missing values: None were found in any column of the dataset.
* Duplicate rows: No duplicate entries were present, confirming data integrity.
* **Validation of Binary Values:**

Binary variables within the dataset were validated to ensure they contained only acceptable values (e.g., 0 and 1 or Yes and No).

The following SQL query was executed to identify and delete any records containing invalid binary values:

-- Validate Binary Values

**delete** **FROM** diabetes

**WHERE** "HighBP" **NOT** **IN** (0, 1)

**OR** "HighChol" **NOT** **IN** (0, 1)

**OR** "Smoker" **NOT** **IN** (0, 1)

**OR** "Stroke" **NOT** **IN** (0, 1)

**OR** "HeartDiseaseorAttack" **NOT** **IN** (0, 1)

**OR** "PhysActivity" **NOT** **IN** (0, 1)

**OR** "Fruits" **NOT** **IN** (0, 1)

**OR** "Veggies" **NOT** **IN** (0, 1)

**OR** "HvyAlcoholConsump" **NOT** **IN** (0, 1)

**OR** "AnyHealthcare" **NOT** **IN** (0, 1)

**OR** "NoDocbcCost" **NOT** **IN** (0, 1)

**OR** "DiffWalk" **NOT** **IN** (0, 1)

**OR** "Sex" **NOT** **IN** (0, 1);

* **Aggregation of Age Categories:**

Age values in the dataset were initially grouped using a 13-level age categorization method. These were further aggregated into three broader categories for simplified analysis. The following SQL query was executed:

-- Aggregate Age Categories

**ALTER** **TABLE** public.diabetes **ADD** **COLUMN** "age\_category" **integer**;

**ALTER** **TABLE** public.diabetes **ADD** **COLUMN** "age\_category\_desc" **varchar**;

**UPDATE** diabetes

**SET** "age\_category" = **CASE**

**WHEN** "Age" **BETWEEN** 1 **AND** 3 **THEN** 1 -- Young

**WHEN** "Age" **BETWEEN** 4 **AND** 7 **THEN** 2 -- Middle-aged

**WHEN** "Age" **BETWEEN** 8 **AND** 13 **THEN** 3 – Older

**ELSE** "Age"

**END**;

**UPDATE** diabetes

**SET** "age\_category\_desc" = **CASE**

**WHEN** "age\_category" = 1 **then** **'Young'**

**WHEN** "age\_category" = 2 **then** **'Middle-aged'**

**WHEN** "age\_category" = 3 **THEN** **'Older'**

**ELSE** **'Unlisted'**

**END**;

This aggregation simplified the analysis into the following categories:

* Young: Ages corresponding to levels 1 to 3.
* Middle-aged: Ages corresponding to levels 4 to 7.
* Older: Ages corresponding to levels 8 to 13.
* **Descriptive Variables for Specific Factors:**

Descriptive labels were assigned to certain variables to improve interpretability. For example:

1. Mapping Education Levels to Descriptive Categories

Education levels were mapped into descriptive categories using the following SQL query:

-- Map Education Levels to Descriptive Categories

**ALTER** **TABLE** public.diabetes **ADD** **COLUMN** "education\_desc" **varchar**;

**UPDATE** diabetes

**SET** "education\_desc" = **CASE**

**WHEN** "Education" = 1 **THEN** **'No School'**

**WHEN** "Education" = 2 **THEN** **'Elementary'**

**WHEN** "Education" = 3 **THEN** **'Some High School'**

**WHEN** "Education" = 4 **THEN** **'High School Graduate'**

**WHEN** "Education" = 5 **THEN** **'Some College'**

**WHEN** "Education" = 6 **THEN** **'College Graduate'**

**END**;

1. Converting Income Levels into Descriptive Categories

Income levels were converted into descriptive categories using the following SQL query:

-- Convert Income Levels into Descriptive Categories

**ALTER** **TABLE** public.diabetes **ADD** **COLUMN** "income\_desc" **varchar**;

**UPDATE** diabetes

**SET** "income\_desc" = **CASE**

**WHEN** "Income" = 1 **THEN** **'Less than $10,000'**

**WHEN** "Income" = 2 **THEN** **'$10,000 - $15,000'**

**WHEN** "Income" = 3 **THEN** **'$15,000 - $20,000'**

**WHEN** "Income" = 4 **THEN** **'$20,000 - $25,000'**

**WHEN** "Income" = 5 **THEN** **'$25,000 - $35,000'**

**WHEN** "Income" = 6 **THEN** **'$35,000 - $50,000'**

**WHEN** "Income" = 7 **THEN** **'$50,000 - $75,000'**

**WHEN** "Income" = 8 **THEN** **'$75,000 or more'**

**END**;

These descriptive mappings improved the readability and interpretability of the data.

* **Addition of New Lifestyle Factors for Analysis:**

Several new derived variables were created to enhance analytical insights:

1. Lifestyle Score

A composite lifestyle score was calculated based on multiple factors, including physical activity, fruits, and vegetable consumption status. This score provided a single metric to evaluate overall lifestyle health.

-- Create Lifestyle Score (Fruits, Veggies, PhysActivity)

**ALTER** **TABLE** diabetes **ADD** **COLUMN** "LifestyleScore" **INTEGER**;

**UPDATE** diabetes

**SET** "LifestyleScore" = "Fruits" + "Veggies" + "PhysActivity";

1. BMI\_PhysActivity

A combined variable representing the interaction between Body Mass Index (BMI) and physical activity levels was generated to assess their joint effect on health outcomes.

-- Calculate Interaction Terms

**ALTER** **TABLE** diabetes **ADD** **COLUMN** "BMI\_PhysActivity" **FLOAT**;

**UPDATE** diabetes

**SET** "BMI\_PhysActivity" = "BMI" \* "PhysActivity";

1. Binarization of Diabetes Type:

Diabetes types were converted into binary categories for simplified analysis, grouping them into broader classifications as necessary.

-- Convert Diabetes\_012 into Descriptive Categories

**ALTER** **TABLE** diabetes **ADD** **COLUMN** "Diabetes\_Type" **VARCHAR**(20);

**UPDATE** diabetes

**SET** "Diabetes\_Type" = **CASE**

**WHEN** "Diabetes\_012" = 0 **THEN** **'No Diabetes'**

**WHEN** "Diabetes\_012" = 1 **THEN** **'Prediabetes'**

**WHEN** "Diabetes\_012" = 2 **THEN** **'Diabetes'**

**END**;

-- Binarize Diabetes\_binary

**ALTER** **TABLE** diabetes **ADD** **COLUMN** "Diabetes\_Binary" **INTEGER**;

**UPDATE** diabetes

**SET** "Diabetes\_Binary" = **CASE**

-- No Diabetes

**WHEN** "Diabetes\_012" = 0 **THEN** 0

-- Prediabetes or Diabetes

**ELSE** 1

**END**;

1. Normalization of BMI

BMI values were normalized using min-max scaling to bring all values into a 0–1 range. This transformation facilitated consistent comparisons across variables.

-- Normalize BMI (Min-Max Scaling)

**ALTER** **TABLE** public.diabetes **ADD** **COLUMN** "BMI\_normalized" **float**;

**UPDATE** diabetes

**SET** "BMI\_normalized" = ("BMI" - (**SELECT** **MIN**("BMI") **FROM** diabetes)) / ((**SELECT** **MAX**("BMI") **FROM** diabetes) - (**SELECT** **MIN**("BMI") **FROM** diabetes));

1. Bucketing of Physical and Mental Health Values:

Physical and mental health scores were bucketed into defined categories for easier interpretation and statistical modeling. For instance:

* + Physical health: “Good,” “Moderate,” “Poor”
  + Mental health: “Good,” “Moderate,” “Poor”

-- Bucketize Mental Health Days (MentHlth)

**ALTER** **TABLE** diabetes **ADD** **COLUMN** "MentHlth\_Bucket" **integer**;

**UPDATE** diabetes

**SET** "MentHlth\_Bucket" = **CASE**

**WHEN** "MentHlth" **BETWEEN** 0 **AND** 7 **THEN** 2

**WHEN** "MentHlth" **BETWEEN** 8 **AND** 14 **THEN** 1

**WHEN** "MentHlth" > 14 **THEN** 0

**END**;

**ALTER** **TABLE** diabetes **ADD** **COLUMN** "MentHlth\_Bucket\_desc" **VARCHAR**(20);

**UPDATE** diabetes

**SET** "MentHlth\_Bucket\_desc" = **CASE**

**WHEN** "MentHlth\_Bucket" = 2 **THEN** **'Good'**

**WHEN** "MentHlth\_Bucket" = 1 **THEN** **'Moderate'**

**WHEN** "MentHlth\_Bucket" = 0 **THEN** **'Poor'**

**END**;

-- Bucketize Physical Health Days (PhysHlth)

**ALTER** **TABLE** diabetes **ADD** **COLUMN** "PhysHlth\_Bucket" **integer**;

**UPDATE** diabetes

**SET** "PhysHlth\_Bucket" = **CASE**

**WHEN** "PhysHlth" **BETWEEN** 0 **AND** 7 **THEN** 2

**WHEN** "PhysHlth" **BETWEEN** 8 **AND** 14 **THEN** 1

**WHEN** "PhysHlth" > 14 **THEN** 0

**END**;

**ALTER** **TABLE** diabetes **ADD** **COLUMN** "PhysHlth\_Bucket\_desc" **VARCHAR**(20);

**UPDATE** diabetes

**SET** "PhysHlth\_Bucket\_desc" = **CASE**

**WHEN** "PhysHlth\_Bucket" = 2 **THEN** **'Good'**

**WHEN** "PhysHlth\_Bucket" = 1 **THEN** **'Moderate'**

**WHEN** "PhysHlth\_Bucket" = 0 **THEN** **'Poor'**

**END**;

After performing the above steps, the dataset was validated to be free of missing and duplicate values, with all transformations and validations successfully applied. The dataset is now structured and enriched, ready for advanced analytical procedures.

* 1. **Feature Selection:**

Feature selection methods are essential for eliminating irrelevant and redundant attributes that do not significantly contribute to the accuracy of predictive models. In fact, such attributes can sometimes reduce the model's predictive accuracy. By applying feature selection, we can not only improve the predictive accuracy but also create a simpler, faster, and more cost-efficient model.

In this study, we employed the “CorrelationAttributeEval” feature selection technique available in Weka, setting the threshold to -1.79. This method evaluates the importance of each attribute by calculating the Pearson correlation between the attribute and the target class, as defined by Weka.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

In order to apply the feature selection method, the following steps were first performed:

1. The dataset was split into two:
   * 1. Descriptive dataset with all the descriptions and data.
     2. Classification dataset with only the data required for applying machine learning models.

-- Split into diabetes\_descriptive

**CREATE** **TABLE** diabetes\_class **AS**

**SELECT**

patient\_id,**"Diabetes\_Binary"** , **"HighChol"**, **"HighBP"**, **"CholCheck"**, **"Smoker"**, **"Stroke"**, **"HeartDiseaseorAttack"**,

**"PhysActivity"**, **"Fruits"**, **"Veggies"**, **"HvyAlcoholConsump"**, **"AnyHealthcare"**,**"NoDocbcCost"**, **"GenHlth"**, **"MentHlth"**, **"PhysHlth"**, **"DiffWalk"**, **"Sex"**, **"Age"**, **"age\_category"**,

**"BMI"**, **"BMI\_normalized"**, **"LifestyleScore"**, **"BMI\_PhysActivity"**, **"MentHlth\_Bucket"**, **"PhysHlth\_Bucket"**

**FROM** diabetes;

-- Split into diabetes\_class

**CREATE** **TABLE** diabetes\_descriptive **AS**

**SELECT**

patient\_id,**"Diabetes\_012"** ,**"Diabetes\_Type"** , **"Education"**, education\_desc , **"Income"** ,Income\_desc, **"Age"**,age\_category, age\_category\_desc, **"MentHlth"**,**"MentHlth\_Bucket"**, **"MentHlth\_Bucket\_desc"**

**"PhysHlth"**, **"PhysHlth\_Bucket"**, **"PhysHlth\_Bucket\_desc"**

**FROM** diabetes;

1. Randomized the dataset and split the dataset into train (80 percent) and test (20 percent).

-- Use a Common Table Expression to assign random values

**CREATE** **TEMP** **TABLE** randomized\_data **AS**

**SELECT** \*, **RANDOM**() **as** random\_value

**FROM** diabetes\_class;

-- Create the train table with 80% of the data

**SELECT** \*

**INTO** diabetes\_class\_train

**FROM** randomized\_data

**WHERE** random\_value <= 0.8;

-- Create the test table with the remaining 20% of the data

**SELECT** \*

**INTO** diabetes\_class\_test

**FROM** randomized\_data

**WHERE** random\_value > 0.8;

Then, feature selection method is applied to the training data. As a result, the top 10 features were identified. The features were identified in two scenarios:

1. Top 10 features including created lifestyle factors:

A screenshot of a computer

Description automatically generated

1. Top 10 features excluding created lifestyle factors:

A screenshot of a computer

Description automatically generated

New dataset for training and testing is created with the combination of the top 10 features from both the scenarios.

-- Top correlation attributes based on feature selection

**select** **"Diabetes\_Binary"**, **"HighBP"**, **"DiffWalk"**, **"HighChol"**, **"HeartDiseaseorAttack"**, **"CholCheck"**, **"Smoker"**, **"LifestyleScore"**,**"GenHlth"**, **"PhysActivity"**, **"Stroke"**, **"PhysHlth"**, **"PhysHlth\_Bucket"**, age\_category, **"MentHlth\_Bucket"**

**into** diabetes\_train\_final

**from** diabetes\_class\_train;

**select** **"Diabetes\_Binary"**, **"HighBP"**, **"DiffWalk"**, **"HighChol"**, **"HeartDiseaseorAttack"**, **"CholCheck"**, **"Smoker"**, **"LifestyleScore"**,**"GenHlth"**, **"PhysActivity"**, **"Stroke"**, **"PhysHlth"**, **"PhysHlth\_Bucket"**, age\_category, **"MentHlth\_Bucket"**

**into** diabetes\_test\_final

**from** diabetes\_class\_test;

* 1. **Data Analysis:**

We used Weka 3.8.6 for data analysis and model building. After applying feature selection to the 80 percent of the data, we used that 80 percent of the data to train the predictive models. We constructed six predictive models using algorithms such as Logistic Regression, Naive Bayes, Decision Tree, J48, Random Forest and Random Tree. After constructing these models, then we test them on the remaining 20 percent of the data. We used ROC area, precision and recall values then to compare the accuracy of these model with each other.

* 1. **Results:**

The table below shows the result of ROC area, precision and recall for each model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **ROC Area** | **Precision** | **Recall** |
| **Logistic** | 0.797 | 0.860 | 0.975 |
| **Naive Bayes** | 0.783 | 0.895 | 0.859 |
| **Decision Tree** | 0.790 | 0.854 | 0.986 |
| **J48** | 0.709 | 0.855 | 0.983 |
| **Random Forest** | 0.759 | 0.863 | 0.955 |
| **Random Tree** | 0.700 | 0.861 | 0.957 |

The results of this project highlight the effectiveness of various models in predicting diabetes based on lifestyle factors, tailored to specific needs. Logistic Regression emerged as the most balanced and reliable model, demonstrating the best ROC Area and high recall, making it ideal for general predictive accuracy. However, if the focus is on minimizing false positives, such as reducing unnecessary interventions, Naïve Bayes stands out due to its superior precision. On the other hand, for scenarios where the primary goal is to maximize the identification of positive cases, the Decision Tree model is a strong contender, thanks to its exceptional recall. These findings provide a nuanced understanding of model performance, enabling informed decisions based on the specific objectives of diabetes prediction.

* 1. **Future Work:**

This study offers significant insights into lifestyle factors influencing diabetes, but there is substantial room for expansion. Key areas for future work include:

* Expanding Lifestyle Factors: Incorporating additional lifestyle factors such as sleep patterns, stress levels, and dietary specifics (e.g., sugar and fiber intake) could provide a more comprehensive understanding of diabetes risk factors and enhance feature selection for machine learning models.
* Advanced Machine Learning Techniques: Exploring advanced machine learning algorithms such as Gradient Boosting Machines, Neural Networks, or Support Vector Machines may yield better predictive accuracy and robustness compared to the current models.
* Real-World Application: Collaborating with healthcare providers to test these models in real-world settings, integrating the predictive tools with electronic health record systems for proactive diabetes management.
* Multi-Disease Analysis: Extending the framework to include comorbid conditions such as cardiovascular diseases and hypertension, examining how shared lifestyle factors influence multiple diseases.
  1. **Conclusion:**

This project successfully evaluated the impact of lifestyle factors on diabetes prediction using various machine learning models. Logistic Regression emerged as the most reliable model, offering a balanced trade-off between precision and recall. The findings underline the importance of addressing modifiable lifestyle factors in diabetes prevention strategies.

By integrating additional lifestyle variables, refining machine learning approaches, and testing these models in real-world environments, future research can build on these results to enhance diabetes prediction and management. This iterative process will contribute to developing robust, scalable, and actionable health solutions, paving the way for proactive and personalized healthcare interventions.

**REFERENCES:**

*UCI Machine Learning Repository*. (n.d.). https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

*Diabetes Health Indicators Dataset*. (2021, November 8). Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

*CorrelationAttributeEval*. (2022, January 28). https://weka.sourceforge.io/doc.dev/weka/attributeSelection/CorrelationAttributeEval.html

*Weka-Dev-3.9.6 API*. (n.d.). https://weka.sourceforge.io/doc.dev/index.html?weka/attributeSelection/CorrelationAttributeEval.html