Diabetes Prediction

|  |  |  |
| --- | --- | --- |
| 1st Kanika Saxena ksaxena2@buffalo.edu | 2nd Rajashree Shanmuganathan rajashre@buffalo.edu | 3rd Alankriti Dubey [adubey2@buffalo.edu](mailto:adubey2@buffalo.edu) |

1. Logistic Regression

# I. PROBLEM STATEMENT

This report aims to investigate the relationship between demographic and medical factors and diabetes risk, identifying key patterns that can inform early intervention and prevention strategies. Specifically, the analysis focuses on understanding how obesity, age, and lifestyle choices like smoking contribute to the onset and progression of diabetes, and how these insights can be used to target at-risk populations more effectively.

## **Justification for Algorithm Choice:** We have chosen the Logistic Regression algorithm based on the following reasons –

## **1. Nature of the Problem**

## The problem at hand—**predicting diabetes based on medical and demographic factors**—is fundamentally a binary classification task. Logistic regression is well-suited for such tasks as it models the probability of a binary outcome, making it a natural choice for predicting the presence or absence of diabetes (0 or 1).

## **2. Interpretability**

## One of the critical advantages of logistic regression is its interpretability. The model coefficients can provide insight into the relationship between each predictor and the likelihood of diabetes, allowing for meaningful medical and demographic interpretations. This is crucial in healthcare, where understanding risk factors can inform treatment and prevention strategies.

## **3. Handling of Predictor Types**

## Logistic regression effectively handles both continuous predictors (like age and BMI) and categorical predictors (like gender). The ability to include both types of variables without extensive preprocessing makes it a practical choice for this analysis.

## B. **Model Training and Tuning**

## **1. Data Preprocessing**

## Before training the model, essential preprocessing steps were taken, including:

## **Handling Missing Values**: Any missing data in the dataset was addressed to ensure a complete dataset for training.

## **Handling Data Inconsistency**: Removing data inconsistency by making all string values in title case.

## **Encoding Categorical Variables**: Gender and smoking history were encoded using techniques such as one-hot encoding to convert categorical data into a numerical format suitable for logistic regression.

## **2. Model Training**

## **Splitting the Data**: The dataset was split into training and test sets (typically an 80-20 split) to evaluate model performance effectively.

## **Fitting the Model**: The logistic regression model was trained using the training set. Hyperparameters such as regularization were adjusted using cross-validation to prevent overfitting and enhance generalization.

## **3. Tuning and Validation**

## Hyperparameter tuning was conducted through techniques like **Grid Search** to optimize parameters like the regularization strength, which can significantly affect model performance. The final model was validated using k-fold cross-validation to ensure robustness.

## C. **Effectiveness of the Algorithm**

## **1. Performance Metrics**

## Several metrics were utilized to evaluate the effectiveness of the logistic regression model:

## **Accuracy**: Measures the overall correctness of the model. Given the class distribution in the dataset, accuracy alone might be misleading.

## **Precision and Recall**: These metrics are crucial in healthcare applications where false negatives (missed diabetes cases) can have serious consequences. High precision indicates a low false positive rate, while high recall indicates a low false negative rate.

## **2. Model Performance Results**

## Upon evaluation, the logistic regression model achieved:

## **Accuracy**: Approximately 95.68%, suggesting that the model performs well in classifying diabetes cases.

* **Macro Average**:

## **Precision**: 0.91 indicates the average precision across classes, treating all classes equally, indicating that when the model predicts a patient has diabetes, it is correct 91% of the time.

* **Recall**: 0.80 indicates the average recall, again treating all classes equally, which means the model correctly identifies 90% of actual diabetes cases.

## **F1-Score**: 0.85 is the average F1-score across classes, , indicating a balanced trade-off between precision and recall.

* **Weighted Average**:
* **Precision**: 0.95, reflecting the average precision while considering the number of instances in each class.
* **Recall**: 0.96, indicating a strong overall recall.
* **F1-Score**: 0.95 shows a strong balance between precision and recall when accounting for class imbalance.

## D. Insights Gained from the Algorithm

1. **Key Predictors**: The analysis revealed that **higher BMI**, **older age**, and **female gender** are significantly associated with an increased risk of diabetes. This supports public health initiatives focusing on obesity and age as critical factors in diabetes management.
2. **Impact of Lifestyle Choices**: While the dataset indicated that a majority of patients are non-smokers, the presence of diabetes in current and former smokers emphasizes the need for continued awareness of lifestyle choices.
3. **Data-Driven Interventions**: The insights from the predictive model allow for tailored interventions, such as:
   * Targeted screenings for older populations and individuals with higher BMI.
   * Implementing educational programs focused on weight management and smoking cessation.
   * Developing personalized treatment plans that take into account individual risk profiles based on demographic and medical factors.
4. **Future Research Directions**: The analysis opens avenues for further research, such as investigating the impact of other lifestyle factors (e.g., diet, physical activity) on diabetes risk. It may also provide a basis for exploring more complex models (like ensemble methods) to see if predictive performance can be improved.