EXPLORING THE FUTURE OF HEALTHCARE: MACHINE LEARNING-POWERED DIABETES RISK ASSESSMENT

by

Michael Eniolade

A MACHINE LEARNING PROJECT REPORT

Presented to the Faculty of

The Graduate College at McDaniel College

In Partial Fulfillment of Requirements

For the Degree of M.Sc. in Data Analytics

Under the Supervision of Professor Xuejing Duan

Westminster, Maryland

October 2023

**Abstract**:

This project focuses on the application of machine learning techniques for diabetes prediction. I employed three distinct machine learning algorithms: logistic regression, gradient boosting classifier, and random forest classifier. The project involved data preprocessing, model development, and extensive experimentation. In this report, I provide a comprehensive overview of the process and present the results obtained.

**Table of Content**

Abstract ------------------------------------------------------------------------------------ 2

Introduction -------------------------------------------------------------------------------- 4

Methodology ------------------------------------------------------------------------------- 5

Results and Discussion ------------------------------------------------------------------- 8

Conclusion --------------------------------------------------------------------------------- 10

References --------------------------------------------------------------------------------- 11

**List of Tables and Figures**

Figure 1 ------------------------------------------------------------------------------------- 6

Figure 2 ------------------------------------------------------------------------------------- 9

**Introduction**

Diabetes is a prevalent and chronic health condition that impacts the lives of millions of individuals worldwide. Early detection and proactive intervention are paramount for effectively managing diabetes. Leveraging the capabilities of machine learning, this project aims to develop a predictive model for identifying individuals at risk of diabetes. The model will utilize a diverse set of health and demographic features to enable timely and informed intervention.

*Problem Statement*

This project seeks to address the critical issue of diabetes prediction with the following objectives:

- Prediction Accuracy: Develop a robust machine learning model capable of accurately predicting the likelihood of diabetes onset.

- Early Intervention: Identify individuals at high risk of developing diabetes to facilitate early intervention, lifestyle adjustments, and improved health outcomes.

- Data-Driven Insights: Gain valuable insights into the influential factors contributing to diabetes risk, promoting a deeper understanding of this complex health condition.

*Machine Learning Technique*

To achieve these objectives, we will employ supervised machine learning, specifically focusing on classification algorithms. The model will be trained using historical health and demographic data to classify individuals into two categories: those at risk of developing diabetes (1) and those not at risk (0).

**Methodology**

*Data Collection*: Obtaining a diabetes dataset containing a range of clinical and demographic features.

*Data Cleaning*: Addressing missing values and outliers to ensure data quality.

*Feature Engineering*: Selected relevant features and performed feature scaling.

Data Source and Data Definition

The project will utilize a comprehensive dataset comprising an array of health-related variables, including:

|  |  |  |
| --- | --- | --- |
| Feature | Explanation | Values |
| Diabetes\_prediction | Diabetes | 0 = no, 1 = yes |
| HighBP | High Blood Pressure | 0 = no, 1 = yes |
| HighChol | High Cholesterol | 0 = no high cholesterol, 1 = high cholesterol |
| CholCheck | Cholesterol Check in 5 Years | 0 = no, 1 = yes |
| BMI | Body Mass Index |  |
| Smoker | Have you smoked at least 100 cigarettes in your life? | 0 = no, 1 = yes |
| Stroke | (Ever told) you had a stroke? | 0 = no, 1 = yes |
| HeartDiseaseorAttack | Coronary Heart Disease (CHD) or Myocardial Infarction (MI) | 0 = no, 1 = yes |
| PhysActivity | Physical Activity in Past 30 Days (excluding job) | 0 = no, 1 = yes |
| Fruits | Consume Fruit 1 or more times per day | 0 = no, 1 = yes |
| Veggies | Consume Vegetables 1 or more times per day | 0 = no, 1 = yes |
| HvyAlcoholConsump | Heavy Alcohol Consumption | 0 = no, 1 = yes (for adult men >=14 drinks per week and adult women >=7 drinks per week) |
| AnyHealthcare | Have any kind of healthcare coverage, including health insurance, prepaid plans such as HMO, etc. | 0 = no, 1 = yes |
| NoDocbcCost | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? | 0 = no, 1 = yes |
| GenHlth | Would you say that in general your health is: | 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor |
| MentHlth | Days of Poor Mental Health (Scale 1-30 days) | 1-30 |
| PhysHlth | Physical Illness or Injury Days in Past 30 Days (Scale 1-30 days) | 1-30 |
| DiffWalk | Do you have serious difficulty walking or climbing stairs? | 0 = no, 1 = yes |
| Sex | Gender | 0 = female, 1 = male |
| Age | 13-Level Age Category | 1 = 18-24, 9 = 60-64, 13 = 80 or older |
| Education | Education Level (Scale 1-6) | 1 = Never attended school or only kindergarten, 2 = elementary, etc. |
| Income | Income Scale (Scale 1-8) | 1 = less than 10,000, 5 = less than 35,000, 8 = $75,000 or more |
|  |  |  |

Figure 1, Depicting the data dictionary for the diabetes dataset.

**Models and Algorithms:**

*Logistic Regression*: Logistic regression is a fundamental and widely used classification algorithm in the field of machine learning. It is particularly well-suited for binary classification tasks, such as predicting whether a patient has diabetes or not. Logistic regression models the relationship between the independent variables (features) and the probability of a specific outcome. It does so by fitting a logistic function to the data, which allows it to output probabilities ranging from 0 to 1. Logistic regression is used as a baseline model. This is a common practice in machine learning projects as it provides a simple and interpretable model to start with. Logistic regression allowed me to establish a benchmark for predicting diabetes risk and served as a starting point for more complex models. While logistic regression may not capture intricate relationships in the data as effectively as ensemble methods like gradient boosting and random forest, it provides valuable insights and an initial estimate of prediction accuracy.

*Gradient Boosting Classifier*: Gradient boosting is an ensemble learning technique that excels at predictive accuracy. It works by combining multiple weak learners (usually decision trees) to create a strong predictive model. One of the key advantages of gradient boosting is its ability to capture complex and nonlinear relationships in the data. It does this through an iterative process where each new model focuses on the errors made by the previous models, thus improving the overall prediction. Gradient boosting is known for its high accuracy, making it an excellent choice for tasks like diabetes prediction. This machine learning project employed the Gradient Boosting Classifier to capture intricate and nonlinear relationships within the diabetes dataset. This model's iterative nature allows it to adapt and learn from the data, making it a powerful tool for improving prediction performance.

*Random Forest Classifier*: Random forest is another ensemble learning technique that consists of multiple decision trees. It is particularly well-suited for handling large datasets and is known for its robustness and resistance to overfitting. Random forest works by creating a forest of decision trees, each trained on a subset of the data and with a random selection of features. The final prediction is often a combination of predictions from all the trees. The Random Forest Classifier is used to enhance the model's robustness and reliability. Large datasets, especially those with numerous features, can present challenges for many machine learning algorithms. Random forest's ability to mitigate overfitting and handle such datasets makes it a pragmatic choice. It can provide accurate predictions while being less sensitive to noisy or irrelevant features in the data, which is essential for the accurate prediction of diabetes.

In summary, the logistic regression is a simple and interpretable baseline model, the gradient boosting has high accuracy and ability to capture complex relationships, and lastly the random forest is used to enhance robustness when dealing with large datasets. This comprehensive strategy ensures that the nuances of diabetes prediction are addressed to effectively provide accurate results for the machine learning project.

**Results and Discussion**

Conducting a series of experiments to evaluate the performance of the three machine learning algorithms: Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier.

Performance Metrics for Logistic Regression:

Accuracy: 0.86

Precision: 0.54

Recall: 0.15

F1-score: 0.24

AUC-ROC: 0.57

Performance Metrics for Random Forest Classifier:

Accuracy: 0.86

Precision: 0.48

Recall: 0.16

F1-score: 0.24

AUC-ROC: 0.57

Performance Metrics for Gradient Boosting Classifier:

Accuracy: 0.87

Precision: 0.59

Recall: 0.17

F1-score: 0.26

AUC-ROC: 0.57

The accuracy scores of both Logistic Regression and Random Forest Classifier models are 86%, indicating their similar predictive power. However, when we consider other metrics, such as precision, recall, and F1-score, the Gradient Boosting Classifier stands out as it achieved higher values. This suggests that the Gradient Boosting model provides a more balanced performance in terms of correctly classifying positive and negative cases.

The AUC-ROC scores for all three models are identical at 0.57, indicating their similar ability to distinguish between the two classes.

These performance metrics emphasize the effectiveness of the Gradient Boosting Classifier in predicting diabetes risk, making it the preferred choice among the three models for this task.

A graph of different colored bars

Description automatically generated

Figure 2, Bar Chart showing the performance metrics for all classification algorithms used for diabetes prediction.

**Conclusion**

This project involves the application of machine learning techniques to predict diabetes risk. The results indicate that the Gradient Boosting Classifier outperformed the other models, with an accuracy of 84%. This suggests the potential of machine learning in diabetes prediction. Early diagnosis and risk assessment can significantly impact patient care.

**References**

Diabetes Health Indicators Dataset. (2019). Kaggle. Retrieved from Dataset Link.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.