Sweet Secrets

Nikhil Chakka
CS
CU Boulder
Boulder, CO
nich4416@colorado.edu

Vraj Patel
CS
CU Boulder
Boulder, CO
vrpa3077@colorado.edu

Kavin Ramesh
CS
CU Boulder
Boulder, CO
kara9783@colorado.edu

Introduction

We plan to use the Diabetes Health Indicators Dataset. In the United States itself, millions of Americans are impacted each year by diabetes as it is one of the most prevalent chronic diseases in the United States. It is a serious chronic disease where the ability to regulate glucose levels in the blood is lost. This leads to reduced quality of life and life expectancy. The inability to regulate glucose is characterized by either an insufficient amount of insulin being released, or the body being unable to effectively use the insulin as needed.

Those affected by diabetes can be subject to complications such as heart disease, vision loss, lower-limb amputation, and kidney disease. While there is no cure for diabetes, there are treatments and strategies like losing weight and eating healthy and more. Making predictive models based on research behind diabetes is especially important and can be used by public health officials for improving current treatments and strategies.

The scale of this problem also needs to be addressed. As of 2018, the Centers for Disease Control and Prevention found that 34.2 million Americans have diabetes and 88 million have prediabetes. This risk is unknown to around 1 in 5 diabetics, and around 4 in 5 prediabetics. The most common type of diabetes is type II diabetes, and it affects people differently across different ages, educations, incomes, locations, races, and other social determinants of health. Economically, diabetes is still a massive burden. Diagnosed diabetes costs reach $327 billion dollars and total costs with undiagnosed diabetes and prediabetes reach $400 billion dollars annually.

It is especially important to analyze this data to not only further our understanding of it, but also take steps forward and implement improvements on currently existing treatments. By creating a model, we can identify risk factors early and enable targeted interventions for high-risk individuals. These models help personalize treatment by analyzing glucose trends and patient-specific data, optimizing medication plans and lifestyle adjustments for more effective management. Additionally, data models can detect complications, such as retinopathy or cardiovascular issues, at an early stage, helping to reduce long-term health risks. By automating diagnosis and treatment recommendations, data models improve healthcare efficiency, reduce human error, and offer scalable solutions, transforming diabetes care and patient outcomes.

1 Related Work

As expected, a lot of related work and research has already been conducted on diabetes, especially in recent years. Machine learning models, such as decision trees and neural networks, have been used to predict the onset and progression of diabetes based on patient data like demographics and lifestyle. Continuous glucose monitoring (CGM) and wearable devices offer real-time insights to optimize insulin dosing and treatment plans. Studies also focus on predicting diabetes complications, like retinopathy and heart disease, using analytics on factors such as HbA1c and cholesterol levels.

Genomic research, including genome-wide association studies (GWAS), identifies genes linked to diabetes, aiding in predicting disease risk. Biomarker analysis, such as C-peptide and insulin resistance markers, provides further insights into disease progression. Lifestyle interventions, especially diet and exercise, have been proven effective in preventing Type 2 diabetes. Additionally, telemedicine and AI-powered tools help tailor personalized treatments and improve healthcare delivery. These innovations are shaping the future of diabetes management and prevention.

2 Proposed Work

In our project, we plan to split our tasks into four separate sections: data preprocessing, exploratory data analysis, split/scale data, training model. Our dataset is derived from the CDC surveys with over 22.74 MB, 22 features, and 254,000 datapoints. These features entail: diabetes (0, I, II), High Blood Pressure (I, II), High Cholesterol (0, I), BMI, smoking history, stroke history, heart disease/attack history, physical activity, fruits, vegetables, alcohol consumption, health care, financial issues, general heath, mental health, physical health, difficulty walking, sex, age, education, and income.

2.1 Data Preprocessing

Data preprocessing is important to ensure the quality and reliability of our models. The initial step involves organizing our data into a structured format, which includes addressing missing values. While one approach is to drop all rows with missing values or use placeholders, both methods have their own drawbacks. Dropping rows with missing values can lead to a substantial loss of correlated data, adversely affecting model training. And although placeholders can mitigate this issue, they do not necessarily enhance correlation and may only reflect patterns in the missing data itself. Therefore, we plan to employ K-Nearest Neighbors (KNN) imputation to substitute missing values. This method involves identifying k closest neighbors based on the similarity of other features and averaging those values to fill in the missing data. Additionally, we must also transform our categorical features into numerical values. We can do this by using one hot encoding, which encodes our features such as sex or education as numbers, enabling our models to interpret them.

Another step of preprocessing is to normalize our data as our features use different scales. The BMI has a range of 12-98, mental health has a range of 0-30, physical health has a range of 0-30, education has a range of 1-6, income has a range of 1-8, etc. Moreover, there are many boolean features. To be able to effectively compare these features, we plan to use Standard Scalar to normalize the values.

2.2 Exploratory Data Analysis

Data visualization plays a critical role in our analysis by creating charts and plots (i.e. histograms, scatter plots) to visualize data distribution and relationships. This helps in identifying patterns, trends, and anomalies in the data. Effective visualizations can provide intuitive insights that are not immediately apparent from raw data, guiding further analysis and model development.

To achieve this, we will utilize several types of visualizations:

- Heatmaps: to show correlations between distinctive features
- Box Plots: to visualize the distribution and identify outliers
- Line Charts: to observe trends over time, especially useful for time-series data

We will employ tools and libraries such as Matplotlib and Seaborn for creating static and attractive statistical graphics. By incorporating these elements, our data visualizations will not only support the identification of significant patterns and trends but also facilitate a deeper understanding of the dataset, contributing to more informed decision-making in our research.

2.3 Split/scale data

To evaluate the accuracy of our model, we will employ k-folds cross validation. This technique involves partitioning the dataset into k equally sized folds. In each iteration, one-fold is designated as the test set, while the remaining k-1 folds are used for training the model. This process is repeated k times, with each fold serving as the test set exactly once. By iterating through all folds, we can assess the model's performance on different subsets of the data, thereby simulating its behavior on unseen data. The results from each iteration are then averaged to provide a comprehensive measure of the model's overall accuracy. This approach ensures that our evaluation is robust and less prone to overfitting, offering a more reliable estimate of the model's predictive capabilities.

2.4 Training model

We plan to test our data against multiple different training models to assess which one has the highest classification accuracy.

The first model we plan to utilize is logistic regression. Logistic regression uses a sigmoid function that takes a specific set of inputs and returns a probability of whether a specific sample belongs to a certain class (in our case diabetes).

The second model we plan to utilize is a decision tree. The decision tree consists of decision nodes which branches into more decision nodes based on a specific feature and metric (gini impurity, etc.) and leaf nodes that determine the final classification (in our case diabetes) of the samples.

The third model we plan to utilize is a Support Vector Machine (SVM). An SVM tries to find the best hyperplane that separates the points that are in different classes. It finds the hyperplane that maximizes the distance between the classes.

The fourth model we plan to utilize is a random forest. Random Forest constructs multiple decision trees using random subsets of data and features, reducing overfitting and enhancing prediction accuracy. It aggregates the results of all trees through voting for classification tasks or averaging for regression tasks, leading to stable and precise outcomes.

The last model we plan to utilize is a neural network. Neural networks consist of multiple layers (input, hidden layers, and output), where the layers perform comlpex computations (multiplying by weights, etc.) to eventually map it to an output.

3 Evaluation

We plan to use precision, recall, and F1 scores to evaluate our model. Precision determines the proportion of true positives out of all samples the model classified as positive. Recall, however, determines the proportion of true positives out all samples the model should have classified as positive. And lastly, the F1 score is a combination of both the recall and precision accuracies. We are looking for values close to 1 for all these measurements.

4 Milestones

We plan to have the initial preprocessing done by Week 7 (all the preprocessing steps outlined above), so we can have the data ready to be input into the models. Though, we may have to modify the preprocessing steps later depending on if we see a need.

By Week 9, we plan to finalize data analysis using logistic regression (and its associated metrics).

By Week 10, we plan to finalize data analysis using a decision tree (and its associated metrics).

By Week 11, we plan to finalize data analysis using a SVM (and its associated metrics).

By Week 12, we plan to finalize data analysis using a random forest (and its associated metrics).

By Week 13, we plan to finalize data analysis using a neural network (and its associated metrics).

After this, we plan to focus on the final project/presentation.

ACKNOWLEDGMENTS

Sources

Gozhulovskyi, Andrii. “Choosing a Model for Binary Classification Problem.” *Medium*, 1 Sept. 2022, medium.com/@andrii.gozhulovskyi/choosing-a-model-for-binary-classification-problem-f211f7a4e263.

“Which Metrics Are Used to Evaluate a Binary Classification Model’s Performance?” Pi.exchange, 2021, www.pi.exchange/knowledgehub/metrics-to-consider-when-evaluating-a-binary-classification-models-performance. Accessed 27 Sept. 2024.

GeeksforGeeks. “Understanding Logistic Regression.” GeeksforGeeks, 9 May 2024, www.geeksforgeeks.org/understanding-logistic-regression/.

Abhishek Sharma 44 , et al. “Decision Tree in Machine Learning.” *GeeksforGeeks*, 15 Mar. 2024, www.geeksforgeeks.org/decision-tree-introduction-example/.

“Random Forest Algorithm in Machine Learning.” *GeeksforGeeks*, GeeksforGeeks, 12 July 2024, www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/.

Geeksforgeeks. “Neural Networks | a Beginners Guide.” GeeksforGeeks, 17 Jan. 2019, www.geeksforgeeks.org/neural-networks-a-beginners-guide/.