Diabetes Analysis & Prediction using Machine Learning

Team 3: Anjali Sebastian, Rupansh Phutela, Yesha Sharma

Introduction

CDC's findings show startling facts, that 1 in 10 Americans have diabetes (around 34 million). 1 in 3 have prediabetes, which is 88 million American adults **[1]**. Diabetes is a chronic health condition that affects how your body turns food into energy. The food you eat is broken down into glucose and released into the bloodstream. When the blood sugar goes up, insulin transfers glucose to the body's cells to use as energy. Diabetes hinders the insulin production, and more sugar goes to the cells which over time cause serious health problems, such as heart disease, vision loss, kidney disease, nerve damage, and skin disease. The complications are permanent in nature and irreversible. Diabetes can only be managed and not cured. Early detection can prevent complications and technology can aid in detection, prediction, and management of the health condition.

Problem Statement and Initial Insights

Currently, Diabetes is detected by two tests - **Hemoglobin A1c test** & **Fasting Glucose Test.** There are certain risk factors which are known to significantly increase the likelihood of an individual to get diabetes. Some of them include – weight, inactivity, family history, race or ethnicity, age, gestational diabetes, polycystic ovary syndrome, high blood pressure, abnormal cholesterol, and triglyceride levels **[2]**. We are going to use the data provided by CDC with the help of a system called BRFSS (The Behavioral Risk Factor Surveillance System **[3]**) which are health-related telephone surveys designed to collect data on health-related risk behaviors, chronic health conditions, and use of preventive services from the adult population (≥ 18 years) residing in the United States.

We will draw various insights into the dataset with the help of statistics, graphs, and other visualizations to see the correlation among various features. Also, we need to do a fair amount of data cleaning and feature engineering to solve the problem.

We intend to perform regression analysis on target variable – BMI to determine which features contribute most to above-range BMI (such as, activity level, food habits, cholesterol, etc.). Subsequently, our focus will be to predict whether the patient is having diabetic (labelled **'1'**) or not (labelled **'0'**). To achieve this, we will use supervised learning classification algorithms with tuned hyper-parameters. At the end, we will try and cluster the dataset into categories decided by machine learning clustering algorithms.

Dataset and Feature Selection

We'll use the **BRFSS 2019** data set hosted by **[4] CDC**, which interested us**.** Since the dataset is in .xpt (.sas) format, we have converted it to a csv and have rehosted it to Seattle University One Drive

- **Link to SAS File -** https://www.cdc.gov/brfss/annual_data/2019/files/LLCP2019XPT.zip
- **Link to CSV -** https://redhawks-my.sharepoint.com/:x:/g/personal/asebastian_seattleu_edu/ETcFLz6GWxlCpDhnC9j-WwkBuTaRdAGbteoV1BFu873t3g

The complete survey has information on 200+ features not all of which are relevant. We are doing extensive feature engineering to pull out a selective feature set related to the problem. The feature variables will be selected based on important risk factors for diabetes. We will map specifically these risk factors to the csv file for our analysis with the help of CDC Codebook [5]. The mapping done by us is presented in the following table.

Column	**BRFSS CSV**		**Column**	**BRFSS CSV**	
Cholestrol1	TOLDHI2	Body Parameter	Stroke	CVDSTRK3	Chronic Condition
Cholestrol2	_CHOLCH2	Body Parameter	Cardiac/Heart Disease	_MICHD	Chronic Condition
Smoke	SMOKE100	Lifestyle	Difficulty Walking	DIFFWALK	Chronic Condition
Consume Alcohol	_RFDRHV7	Lifestyle	Blood Pressure	_RFHYPE5	Chronic Condition
Vegetable Consumption	_FRTLT1A	Lifestyle	HealthPlan	HLTHPLN1	Healthcare Access
Fruit Consumption	_VEGLT1A	Lifestyle	Medical Cost (No Consultation)	MEDCOST	Healthcare Access
Physical Activity	_TOTINDA	Lifestyle	General Health	GENHLTH	General Well Being
Age	_AGEG5YR	Demographics	Physical Health	PHYSHLTH	General Well Being
Gender	SEXVAR	Demographics	Mental Health	MENTHLTH	General Well Being
Income Groups	INCOME2	Demographics	**BMI**	_BMI5	Target for Regression
Education Level	EDUCA	Demographics	**Is Diabetic?**	DIABETE4	Target for Classification

To do the mapping we have referred the work done by Alex Teboul (Kaggle) [6] who extracted information from the 2015 BRFSS dataset.

Data Cleaning, Pre-processing, and EDA

After selecting the relevant features, we will have to clean each of the 20 features and the two target variables. The total dataset has over 400,000+ (0.4 million) entries. Many of these entries have null values. BRFSS survey records these null values as 7, 8 or 77, 99 if the person did not answer or did not know. We will first have to convert them to null values and then see if we want to either drop records with null values or suitably impute values using – Mean, Median or KNN imputer. Additionally, we will need to try and identify outliers if they exist and remove if needed.

Our Data set has features with a variety of data types including Boolean, Categorical and Numerical values. After cleaning the dataset and handling null values we will need to properly typecast all the features. The range for the features is varied so we will have to normalize the data to have a uniform scale. In addition, some features (categorical nominal) may require encoding and decoding using suitable methods.

Once our data is cleaned and pre-processed, we can perform detailed exploratory data analysis which will include both descriptive and visual analysis. We will use boxplots, histograms, scatter plots and cross-tabulation to visualize the data. Scatter plots, correlation & heatmaps between features and target will need to be identified as a first step towards Regression Analysis.

The input data for the Classification will be challenging because the data is significantly imbalanced. The number of people surveyed who have diabetes is much smaller than those who do not. Since we are more concerned about detecting diabetes, the data will have to be balanced either through oversampling minority class (Diabetes = '1') or under-sampling the majority class (Diabetes = '0').

Regression

- **Linear regression** - We will verify correlation, likelihood for multiple combinations of hyper parameters to predict the BMI variable. Linear regression examines if the independent variables are significant to predict the dependent variable and measures the respective accuracy parameters.
- **Regularization** - We plan to use Lasso and Ridge methods, and plot them to find the best alpha-values
- **Normalization** - Another set of regression will be done on normalized data to quantify the impact of exceptionally high and low values on the data
- **Performance Metrics & Hyper-Parameter Tuning** - We will make use of various quality metrics such as R-square, Mean Square Error, Root Mean Squared Error, K-fold analysis cross-validation.

Classification

- **Decision Tree Classifier** – Here a model is created to predict the class or value of the target variable by learning simple decision rules inferred from training data.
- **Random Forest Classifier -** The random forest is an ensemble of decision trees, usually trained with the bagging method. The general idea of bagging is that a combination of learning models increases the overall result.
- **Entropy, Information gain, and Gini Index –** These are used to select best features at the top and traverse through the tree. This informativeness of various features is given by a measure called 'information gain.' Entropy is used to measure the impurity or randomness of a dataset. Gini Index is calculated by subtracting the sum of squared probabilities of each class from one.
- **Other Performance Metrics & Hyper-Parameter Tuning** - Both DT and RF will use confusion matrix and accuracy, precision, specificity, recall, f1-score, etc. as their performance indicators.
- **Normalization** - Another set of classification will be done on normalized data to quantify the impact of exceptionally high and low values on the data

Clustering

We will do clustering on the entire dataset using unsupervised learning algorithms like K-Means (distance-based) and/or DBSCAN (density-based) to see if any interesting patterns emerge from the dataset. Additionally, we will try to find some meaningful clusters that correspond to the supervised label used in classification.

Performance Metrics - We will make use inertia measures (elbow method) and silhouette score to determine the optimum number of clusters that maximize intra cluster similarity and minimize inter cluster similarity. Certain sensitive algorithms might need prior removal of outliers.

Group Dynamics

- **General Communication:** We are synchronizing via Microsoft Teams group chat to share ideas, Outlook to communicate with professor and Canvas portal to share important links and documents for collaboration.
- **Sharing data and code:** We will be using this *GitHub Repo* [7] to share our code and data.
- **Periodic Meeting:** We plan to meet twice a week to discuss what has been done, what we will be doing next and whether there are any roadblocks. We might increase/decrease based on requirements.
- **What Next:** This project involves EDA, data visualizations and presents an opportunity to work on prediction modeling – regression/classification/clustering. We studied above mentioned models/solution technologies and decided to pick, implement & understand at least one data science algorithm each. This will be followed by designating every member of the team few

tasks of the process without deviating from the goal. Our goal as a team is to work and help as a team without compromising one's learning, ideas, and growth. The project goal will be to gain actionable insights and predict diabetes occurrence.

Plan for Completion

- We plan to create a project file for the analysis.
- Also, we need to prepare the final report for which we will work side by side.
- We plan to finish the above tasks ahead of schedule by first week of March.

References

[1] CDC - *https://www.cdc.gov/diabetes/library/features/diabetes-stat-report.html#*

[2] Mayo Clinic- *https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444*

[3] BRFSS - *https://www.cdc.gov/brfss/annual_data/2019/pdf/overview-2019-508.pdf*

[4] CDC Dataset - *https://www.cdc.gov/brfss/annual_data/2019/files/LLCP2019XPT.zip*

[5] CDC Codebook - *https://www.cdc.gov/brfss/annual_data/2019/pdf/codebook19_llcp-v2-508.HTML*

[6] Alex Teboul - *https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset-notebook*

[7] Project GitHub Repository - *https://github.com/anjumorris/MLProject*