Data mining - CDC Diabetes Health Indicators

CIS 635 - Term Project

Members:

- Bhavya Shri Meda,
- Mohamed Ndiaye,
- Raghavendra Naidu Nayanuri
- Bill Muchero

Overview of the Project

Our project assesses the relationship between lifestyle and diabetes in the US. Many factors contribute to someone having diabetes, and these factors include being overweight, physical inactivity, and poor diet, among other factors. However, behind the factors mentioned above are some underlying issues that could lead to someone being overweight or on a poor diet. For example, your level of education can determine how much you know about health and lifestyle. On the other hand, you might be educated, but if you are not doing great financially, you might not be able to afford healthier foods. Therefore, for this project, we are interested in finding out what (and if) factors can determine if someone has diabetes or not.

Our group comprises four members, Bhavya Shri Meda, Mohamed Ndiaye, Raghavendra Naidu Nayanuri, and Bill Muchero, from different backgrounds. While half of the team is from a business background, we have members with a background in the health sector. Since our data comes from the health domain, having someone on the team who can guide us with some features we can use will greatly benefit us.

In addition, we are aware that there are common causes of diabetes that many people are aware of. With that in mind, we plan on not focusing on any features that are obviously known to lead to diabetes. Our interest as a group is to look at features like education level, mental health, and income to see if those features can be used to determine whether someone has diabetes or not.

Finally, we are not aware of any projects similar to ours that might have been conducted. However, if we encounter one, we will make the necessary changes to ensure that it is not similar.

Data plan

The Centers for Disease Control and Prevention (CDC) initially funded the creation of the dataset used in this study. The data acquired was used to better understand the relationship between lifestyle and diabetes in the US. We have 21 features collected, and the target variable is whether a patient has diabetes, is pre-diabetic, or is healthy (total of 22 columns).
The link below presents the summary and characteristics of the dataset:
https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators
The dataset we will use is accessible at the link below for downloading:
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Why a dataset about diabetes?

We chose the dataset above because the study of diabetes is an exciting topic, as the disease itself has several associated factors. We think it's an excellent subject for a data mining project. The association with information about lifestyle is a motivating factor in our project. We can try to determine what elements besides medical conditions can represent a risk factor for diabetes.

Origin of the data

The dataset initially came from the Behavioral Risk Factor Surveillance System (BRFSS), a telephone survey collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventive services. The survey has been conducted since 1984. The original data from the survey was collected in 2015 by the CDC. It contains responses from 441,455 individuals and has 330 features.
For this study, we will use the *diabetes_012_health_indicators_BRFSS2015.csv* dataset, a clean dataset of 253,680 survey responses with 21 feature variables and no missing variables. This dataset has a class imbalance. The dataset was cleaned by Alex Teboul.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDise	PhysActivi	Fruits	Veggies	HvyAlcohc	AnyHealth	NoDocbcC	GenHlth
62	2	1	0	1	27	0	0	0	1	1	1	0	1	0	
63	0	1	0	1	27	1	0	0	0	0	0	0	1	0	
64	0	1	0	1	34	1	0	0	1	1	1	0	1	1	
65	0	1	1	1	30	0	0	0	1	0	0	0	1	0	
66	1	1	1	1	27	0	0	0	1	1	1	0	1	1	
67	0	1	1	1	38	1	1	1	0	1	1	0	1	0	
68	0	0	1	1	28	1	0	0	1	1	1	0	1	0	
69	0	1	0	1	25	0	0	0	0	1	0	0	1	0	

Image 1: Screenshot of the dataset

Target variable and feature variables

The target variable, named *Diabetes_012,* has 3 classes: 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes.
The feature variables we are focusing on are High Cholesterol, Smoking, Mental Health, Education, and Income.
We do not need any preprocessing data. The dataset's owner conducted data binning on the feature variable “Age.”

Implementation plan

- **Step 1 – Exploratory data analysis**
  After collecting and preprocessing, we will perform exploratory data analysis using descriptive statistics and visualize our data using bar charts, histograms, heatmaps, etc.
- **Step 2 – Validate our feature selection**
  We will validate the selected features to ensure they have predictive power and relevance to the research question.
- **Step 3 – Model building**
  We will use rule-mining and association rules between features and diagnosing diabetes by generating rules that capture the relationship between features and diagnosis and assigning high-risk labels to indicate a higher likelihood of developing the disease. We will be leveraging the libraries in Python like Pandas,apyori, pycaret, mlxtend, seaborn, matplotlib, Scikit-Learn, etc, and Colab notebooks to run our scripts in batches.

Evaluation plan

- To evaluate the model, we plan to measure its success primarily through metrics such as **accuracy, precision, recall, and F1-score**.
- We will compare the algorithm's performance against alternative methods, like Decision trees and Logistic Regression, to assess its effectiveness.
- This comprehensive evaluation approach will provide insights into the algorithm's performance and guide further refinement if needed.

Plan for group collaboration

To collaboratively implement our data mining pipeline, we will follow these steps:

- *Task Assignment*: We have broken down our goals into simpler tasks and distributed them amongst ourselves based on our strengths, backgrounds, and experiences.
- *Version Control:* We will be using GitHub for code management. Each member will work on a feature branch and create pull requests to merge into the main branch to ensure code quality and collaboration.
- *Data Management*: To ensure that all members have access to the latest data versions, we will use Google Drive.
- *Documentation*: We intend to maintain detailed documentation via Google Docs file to outline our methodology, track progress, and make updates along the way.

Meetings: We have decided to collaborate through a combination of virtual meetings via Zoom and asynchronous coordination using a messaging application called WhatsApp.

- **Zoom Meetings**: We will have twice in a week virtual meetings on Tuesdays and Fridays at 8 PM
- **WhatsApp**: We will use WhatsApp for daily communication, updates, quick questions, and collaborating asynchronously.

Timeline

Below is our timeline for a week-by-week execution of this study.

Task	Start Date	Days took/needed	End date
Data Collection	14-May	3 days	17-May
Data preprocessing	18-May	3 days	20-May
Exploratory data analysis	20-May	3 days	22-May
Feature Engineering	23-May	3 days	25-May
Model Building	25-May	3 days	27-May
Progress Report	22-May	7 days	28-May
Evaluation and Validation	25-May	3 days	27-May
Deployment	27-May	2 days	28-May
Final Report	2-Jun	10 days	12-Jun

References

- **UCI Machine Learning Repository**: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators
- **Diabetes Health Indicators Dataset Notebook**: https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook/output
- **CDC**: https://data.cdc.gov/