Diabetes prediction

Group: A6
Members: Magnus Karlson, Ekaterina Sedykh, Kayahan Kaya
Repository: Diabetes Prediction
Idea: Diabetes Prediction Kaggle Dataset

Business understanding

Diabetes is a serious health problem, reaching high numbers of people, affected by it. In the United States, approximately 10% of the population has this disease[1].

If not treated in the early stages it can increase cardiovascular problems (heart attack, stroke), vision degradation, nerve damage. These consequences can be diminished with proper lifestyle and treatment. Our goal is to help people be aware of their condition and seek help.

Usually, if you are younger than a certain age, doctors in some countries will not diagnose it as diabetes type 2 because it is used to be known as a disease for middle age - elderly people.

Type 2 diabetes happens when the body becomes resistant to insulin. The exact factors why it's happening are unknown. People at risk usually are overweight, having an inactive lifestyle, and having a family member with diabetes increases your chances of getting it. Also with age the probability of having this disease increases, especially after 45. An easy check for diabetes is measuring glucose blood level, but it's not the only affecting factor.

Implementation in real life is very simple. Most features from the dataset are known for people (i.e. BMI calculated from weight and height), except sugar level, which you can check at home if you have a glucometer or at your family doctor – it takes 30 seconds to get your result. But you can't always go to the doctor for a regular check-up (in coronavirus pandemic for example). If you know you might be prone to diabetes, you can control it with your diet until getting more professional help.

The success of this project, firstly, can be measured by the accuracy of prediction. Data results can be checked with doctors, who can give their verdict whether they would diagnose diabetes or not with these health parameters.

[1] https://www.cdc.gov/diabetes/data/statistics-report/index.html

✓ Inventory of resources: 3 people, familiar with diabetes issues and able to do programming.

✓ Requirements, assumptions, and constraints: The deadline for all of the deliverables – December 14th. Our dataset is public (found on Kaggle)

✓ Risks and contingencies: Finding enough time combining with other studies.

✓ Terminology:

BMI – body max index = m/h^2, where m – weight in kilograms, h – height in meters

✓ Costs and benefits: as it is a mini-project, which is a part of the course, there's no budget

Deliverables for this task include two reports:

✓ Data-mining goals:

Cleaned dataset; source code; classifier models; 3-minute video with a short description of the project; poster

✓ Data-mining success criteria: >90% accuracy on the test part of the dataset. Finding the most important features, contributing to a positive outcome.

Data understanding

Gathering Data

Gathering data is one of the most crucial parts of the project. We are aware that most successful projects start with the good plan of data collection. Because having good quality of data allows you to make good decisions for further research and study.

Outline data requirements

We first started to search and collect information about what diabetes is. As a result of this information, we defined what factors provide information about early diagnosis of diabetes. Then we started to search suitable datasets based on the factors that we defined. After a deep searching process, we found many databases which are associated with diabetes. But the problem started from that point because most of the databases only covered the level of glucose in the blood and blood pressure. We know that those are the primary factors of diagnosing diabetes but not enough to make predictions precisely. Here is the other factors should take consideration before finding suitable database; Pregnancy, Sex, Age, BMI etc.

Verify data availability

I would like to verify that we found the dataset which contains all the requirements to get a predicted percentage of occuring diabetes.
Here is the link of our dataset: https://github.com/magnuskarlson/diabetes-prediction

Define selection criteria

The selection criteria of the database based on the the most primal factors of diabetes which are level of glucose and blood pressure but in order to make precise prediction we need define more features. For example if we only focus on the level of glucose in the blood and blood pressure during the prediction process we can't not predict well the range of ages. Because the level of those factors might be different in different ages.In order to get more accurate and consistent prediction. Here is the list of parameters that we have in our database;

1) Glucose Before fasting (Quantitative Data)

2) Glucose Anytime (Quantitative Data)

3) Age (Quantitative Data)

4) Sex (Qualitive Data)

5) Blood Pressure (Qualitative Data)

6) Family member with Diabetes past or present (Qualitative Data)

7) BMI (Quantitative Data)

8) Pregnancies (Qualitative Data)

Verifying Data Quality

I would like to indicate that the data we use in the project verifies quality conditions. Here is the some requirements that our data meet;

1) Accuracy (All the information stored in the data are correct and precise)

2) Completeness (We have all the information we need for all records. Just clean some null values)

3) Data uniformity (Similar information is presented in similar ways across all over the records)

5) Uniqueness (There are no duplicate rows in the dataset.)

6) Timeliness (The information stored in our dataset is up to date. They are the most important parameter of diagnosing diabetes.)

The parts of describing and exploring data are attachment to the documents.

Project plan

Tasks

1. Data cleanup. Kayahan, 5 hours
2. Research about diabetes. Kayahan, 10 hours
3. Creating 10 different models for predicting. Magnus, 10 hours
4. Creating different features combinations. Magnus, 5 hours
5. Predicted results analysis. Ekaterina, 15 hours
6. Creating posters. Ekaterina, Magnus, Kayahan, 10 hours everyone.
7. Creating video. Ekaterina, Magnus, Kayahan, 5 hours everyone.

Tools

Jupyter is used for everything related to project code.
GitHub is used for storing everything.
Dataset is used for making diabetes predictions.

Methods

Writing code means creating models, cleaning data.
Dataset analyzing means analyzing features combinations.
Prediction analyzing means analyzing, how accurately a model is making predictions.