# Algorithm for predicting diabetes based on lifestyle factors and health statistics

Kristina Katarina Kaljumäe, Karen Roht
https://github.com/KristinaKatarina/ID2023

HW 10

## Task 2 - Business understanding

- **Identifying your business goals**
  Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. The World Health Organization estimates that about 422 million people in the world have diabetes and that it accounts for roughly 1.5 million deaths each year. Diabetics who don't get diagnosed and treated properly have a high risk of premature heart disease and stroke, blindness, limb amputations and kidney failure. For these reasons, it's vital for diabetics to get a diagnosis as soon as possible so they can get the proper treatment. Our business goal is to create a model that will be able to predict if a person is diabetic so they can reach out to a healthcare professional and get proper help. We can measure our success by checking how often it correctly identifies people with diabetes. This is important because our main goal is to make sure our model helps improve people's health.

- **Assessing your situation**
  Our main resource is the dataset we use to train and test our machine learning algorithm. We are using the CDC Diabetes Health Indicators dataset that we originally found from UC Irvine Machine Learning Repository but it is also available on Kaggle. We will be using Jupyter Notebook for writing all the code concerning this project. We are assuming that the dataset we are using has columns with features that play a role in a person's likelihood of having diabetes. If the features are not correlated with the diabetes diagnosis then we most likely can't develop a good model. The success of our algorithm is also constrained by how many features we have to work with. Similarly, the main risk of our project is that we will not be able to create an accurate enough model based on our data. There isn't any field-specific terminology in our project to further explain. The most prominent benefit of this project will probably be the opportunity for us to learn basic machine-learning principles and how to apply them.

- **Defining your data-mining goals**
  Our goal is to create an accurate machine learning algorithm that can predict how likely a person is to be diabetic based on their lifestyle factors and health indicators. We also want to find out what are the primary factors that contribute to a person being

diabetic. The success of our model will be measured by achieving a minimum prediction accuracy of 85% on a test dataset, indicating the model's ability to accurately identify individuals with diabetes.

## Task 3 - data understanding

- **Gathering data**

Our data is obtained from Kaggle and is available in the form of a csv file. The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related phone survey collected annually by The Centers for Disease Control and Prevention. The survey collects responses from over 400,000 Americans on health-related risk behaviors and chronic health conditions. For this dataset, results from 2015 were used.

- **Describing data**

Our dataset contains 21 feature variables and is not balanced. The target variable is Diabetes_binary, which has 2 classes: 0 is for no diabetes, and 1 is for prediabetes or diabetes. About 85% of people in this survey don't have diabetes and the rest do. The dataset contains 253,680 survey responses. All of the columns have float64 type data, there are no missing values.

The data includes:

- **Health indicators** that could potentially influence a person's risk of diabetes.
- **Lifestyle factor**s and **access to healthcare,** which could also impact a person's risk of diabetes.
- Information about a respondent's general, mental, and physical **health**.
- **Demographic variables** can provide additional context.

- **Exploring data**

- Diabetes_binary: From our data, it appears that the majority of respondents (86%) do not have diabetes or prediabetes, with a count of 218,334 falling within the 0.00 - 0.02 range. The mean of 0.14 suggests that about 14% of respondents have prediabetes or diabetes.
- HighBP: Approximately 43% of respondents have high blood pressure.
- HighChol: About 42% of respondents have high cholesterol.
- CholCheck: The majority (96%) have had a cholesterol check.
- BMI: The mean BMI is 28.4, categorizing respondents as overweight. BMI values range widely from 12 to 98, with a median value of 27.
- Smoker: Around 44% of respondents are smokers, with a standard deviation of 0.5.
- Stroke: Approximately 4% of respondents have had a stroke.

- HeartDiseaseorAttack: About 9% of respondents have heart disease or have had a heart attack.
- PhysActivity: The mean of 0.76 suggests that a majority (76%) of respondents have had physical activity.
- Fruits: Approximately 63% of respondents consume fruit daily.
- Veggies: The majority (81%) of respondents consume vegetables daily.
- HvyAlcoholConsump: About 6% of respondents are heavy alcohol consumers.
- AnyHealthcare: The mean of 0.95 indicates that the majority (95%) of respondents have healthcare coverage.
- NoDocbcCost: About 8% of respondents reported facing difficulty seeing a doctor in the past 12 months due to cost.
- GenHlth: The mean score of 2.51 suggests that, on average, respondents perceive their health as somewhat between very good and good.
- MentHlth: Respondents reported an average of 3.18 days of poor mental health in the past 30 days.
- PhysHlth: On average, respondents experienced 4.24 days of physical illness or injury in the past 30 days.
- DiffWalk: Approximately 17% of respondents reported having serious difficulty walking or climbing stairs.
- Sex: The dataset is fairly balanced, with 44% identified as male and 56% as female.
- Age: The mean age of respondents is 8.03, falling within the age range of 60-64.
- Education: The majority of respondents (107,325) have completed education up to the level of "5.90 - 6.00," indicating some college education or an associate degree.
- Income: The income distribution shows a mean score of 6.05, corresponding to an income range of $35,000 to $49,999.

Based on our analysis, we believe that factors such as High Blood Pressure, High Cholesterol), BMI, and age could be the most significant contributors to an individual's likelihood of having diabetes. This is because the frequency of these features is higher among individuals with diabetes. Diabetes seems to be more prevalent among older individuals.

- **Verifying data quality**

There are no missing values in the dataset. However, some duplicates have been identified, which may require attention during subsequent data cleaning steps. Other than that, the data seems to be in good condition.

Task 4 - project plan

- Making sure that there are no null or Nan values and that all values are numeric in the dataset. (Karen -1h)
- Seeing if there are any columns that do not have a correlation with the diabetes_binary column. (Karen- 2h, Kristina - 2h)
- Seeing if balancing the dataset is necessary and what is the best way to do it. (Karen- 3h)
- Finding the best train and test data proportions for splitting the data. (Karen- 2h)
- Seeing if using feature selection techniques can help us make a more accurate model. The feature selection techniques we plan to use are recursive feature elimination, feature importance and univariate selection**.** (Kristina -3h)
- Comparing different machine learning algorithms and selecting the best one for us. We plan to use RandomForestClassifier and logistic regression but this might change based on our earlier findings. (Kristina - 3h)
- Designing and creating the final poster. (Karen- 4h, Kristina - 4h)