Machine Learning in Medical Diagnostics

# Introduction

Getting the right diagnosis is probably the most important aspect of health care: It explains a patient’s health problem and provides the basis for subsequent health care decisions. But medical diagnoses are known to be extremely complex: Going through a patient’s medical history, correctly detecting diseases and recommending the optimal treatment are not only time-consuming but also highly subjective, as those tasks greatly depend on the physician’s expertise level. Because of this, diagnostic errors contribute to approximately ten percent of patient deaths and continue to represent a blind spot in the delivery of quality health care. To address these challenges modern research is turning to artificial intelligence. With rapid progress made in the field of machine learning, it is possible to train algorithms to automatically detect patterns of certain diseases within a patient’s medical record and inform clinicians of any anomalies. The use of machine learning can therefore improve medical diagnostics significantly.

In this project, we want to give two examples on how machine learning can be applied to medical data for the purpose of detecting diseases and supporting clinicians in their decision-making process. We have worked on two very different sets of data and subsequently implemented different machine learning methods to diagnose the data. The first half of the project deals with the prediction of diabetes based on demographic, questionnaire data and laboratory examinations. Diabetes is a chronic disease that directly causes an estimated 1.6 million deaths per year. The number of people affected by the disease has risen rapidly over the past years, amounting 422 million in 2014. With early detection, the burden of diabetes can be significantly reduced or even effectively prevented. On our dataset, we implemented random forest, support vector machine and multilayer perceptron algorithms to predict whether a person is “not diabetic”, “at risk for diabetes” or “diabetic”.

The second dataset we worked on contains images of blood smears, half of which represent healthy samples while the other half shows malaria infected samples. Malaria is a life-threatening disease caused by parasites that are transmitted to people through mosquitoes. Though it is preventable and easily curable, it still causes around 500.000 deaths per year. The most widely used method for malaria detection is examining thin blood smears under a microscope, and visually searching for infected cells. Especially in countries that carry a high share of the global malaria burden and where most people do not have regular access to health care, this approach is too time-consuming. Opposing the traditional method, we implemented a convolutional neural network as well as a support vector machine algorithm to classify healthy and malaria infected blood smears.

# Diabetes

The idea of this part of the project was to find a dataset with lifestyle data (nutrition, physical activity, stress levels) and health examination data and try to predict health condition based on the lifestyle of a person. Our motivation is the following: to show that health and wellbeing is a direct consequence of our everyday choices. We have chosen diabetes for two reasons: first, because it is a widely distributed chronic disease and, second, because it (at least the type 2 diabetes mellitus) is usually developed in adulthood due to unhealthy diet, obesity and little physical activity.(<https://scholar.google.com/scholar_lookup?title=Metabolic+syndrome+and+development+of+diabetes+mellitus%3A+application+and+validation+of+recently+suggested+definitions+of+the+metabolic+syndrome+in+a+prospective+cohort+study&author=D.+E.+Laaksonen&author=H.+M.+Lakka&author=L.+K.+Niskanen&author=G.+A.+Kaplan&author=J.+T.+Salonen&author=T.+A.+Lakka&publication_year=2002>) The primary question was: is it possible to predict diabetes using a set of demographic, body measurements and lifestyle data? Secondary questions were: How good can we distinguish between three classes (“no diabetes”, “pre-diabetes”, “diabetes”)? How good can we distinguish between two classes (“no diabetes”, ‘’pre-diabetes or diabetes” merged together)? Which features have the most predictive power? How do we interpret the accuracy of predictions? How do different models perform on the same task?

The code was written in Python 3 using its libraries including pandas, numpy, pyplot, seaborn, scikit-learn and other. The Jupyter notebook including some descriptions, step by step code and graphs can be found on GitHub under <https://github.com/ElviraMingazova/PRML_Project>.

## 2.1 Dataset and preprocessing

The dataset on the lifestyle and healthfrom 2013-2014 was found on Kaggle. (<https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey>) It originates from the program of studies designed to assess the health and nutritional status of adults and children in the United States and is called National Health and Nutrition Examination Survey (NHANES). The dataset consists of 6 tables, however for our analysis we used information from 4 of them (demographic.csv, questionnaire.csv, labs.csv and examination.csv). After merging these 4 tables together we got one with a total of 9813 entries and 1645 attributes. In the case of this dataset, feature selection and preprocessing was the most time-consuming part. Also it was something new, since we didn’t do it much during our practical course.

First, we studied the dataset and the description of the columns trying to figure out what could be relevant for the diabetes prediction. Initially the following set of attributes was selected:

* From the demographics data

Id, age, gender, educational level, annual family income

* From the questionnaire data

Minutes sedentary activity, vigorous recreational activities, minutes vigorous recreational activities, moderate recreational activities, minutes moderate recreational activities, frequency of meals from fast-food or pizza places, number of frozen meals/pizza in the past 30 days, trouble sleeping or sleeping too much, feeling tired or having little energy, how do you consider your weight

* From the examination data:

BMI (Body Mass Index) and Waist Circumference

* From the labs.csv table:

Blood manganese

Demographics data was taken to look at the basic statistics and distribution of the disease among people from different groups. The questionnaire data was used to assess physical activity and nutritional habits of the respondents. BMI and waist circumference were reported to be major obesity markers associated with diabetes (<https://scholar.google.com/scholar_lookup?title=Comparison+of+body+mass+index%2C+waist+circumference%2C+and+waist%2Fhip+ratio+in+predicting+incident+diabetes%3A+a+meta-analysis&author=G.+Vazquez&author=S.+Duval&author=D.+R.+Jacobs&author=K.+Silventoinen&publication_year=2007)>. Some studies also reported that diabetic patients have lower blood levels of manganese so it was also included in this analysis (https://scholar.google.com/scholar?hl=en&as\_sdt=0%2C5&q=Copper%2C+Chromium%2C+Manganese%2C+Iron%2C+Nickel%2C+and+Zinc+Levels+in+Biological+Samples+of+Diabetes+Mellitus+Patients&btnG=). For the units of measurement and value code description please refer to the notebook ( <https://github.com/ElviraMingazova/PRML_Project>).

Second, we decided to split the data into three categories based on the glycohemoglobin measurements since this was the only test performed on almost all the study participants. This test does not distinguish between type 2 and type 1 diabetes mellitus, hence further on we do not talk about the types. The split was done based on the laboratory procedure manual found on the official NHANES website where it is written that values between 4 and 6% are normal and results higher than 6,5% are indicative of diabetes (<https://wwwn.cdc.gov/nchs/data/nhanes/2013-2014/labmethods/GHB_H_MET_GLYCOHEMOGLOBIN.pdf>). The values in between were marked by us as “prediabetic”.

For the analysis, we only took people older than 25. This was chosen arbitrarily as we wanted to look at adult people who had time to accumulate consequences of their lifestyle. After transforming some features and cleansing the data from NA-values and responses like “refused” and “don’t know” we ended up with 1579 rows in our table. In the Figure 2.1 we see the total counts for categories by gender. We had approximately 200 people in the risk zone for diabetes, 200 diabetic and 1200 people without diabetes.

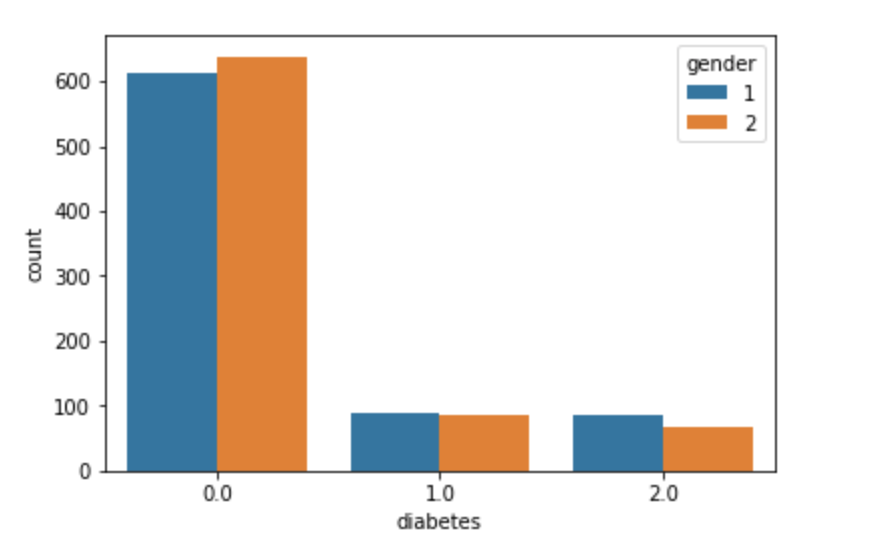


Figure 2..1 Total counts in three categories ( 0 - no diabetes, 1 - prediabetes, 2 - diabetes) by gender (1 – male, 2 - female).

As next we looked at the distribution of different features over three categories (see Figure 2.1.2 and Figure 2.1.3 for an example).

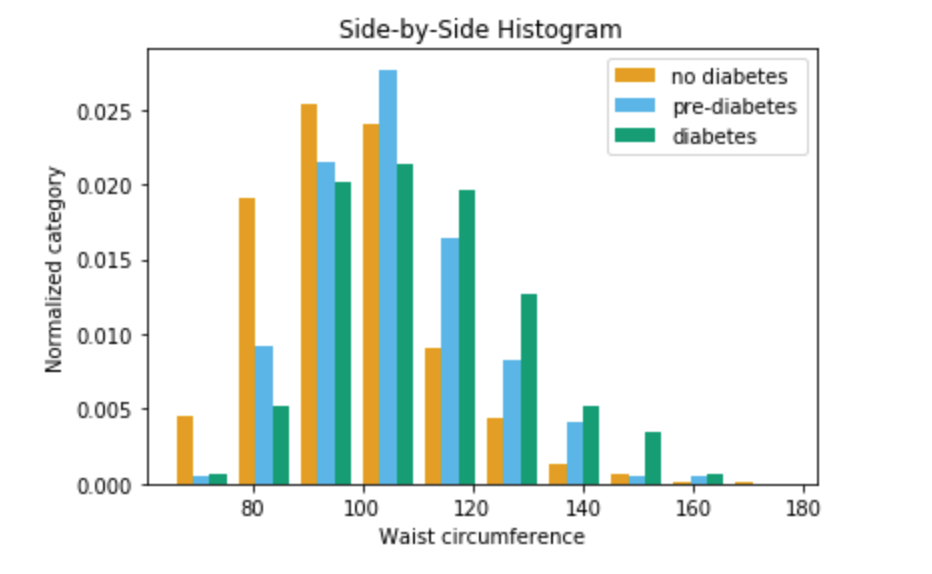


Figure .1.2 Distribution of waist circumference for three classes

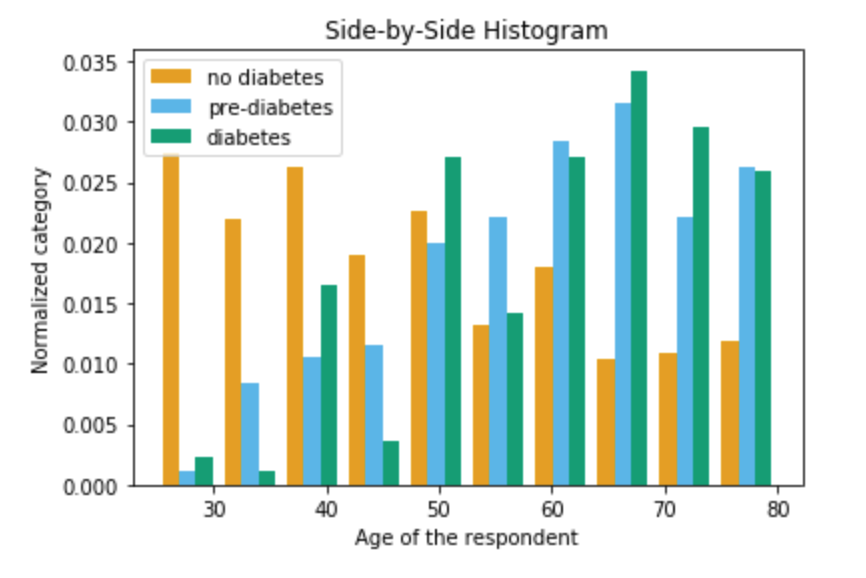


Figure 2.1.3 Distribution of the age for three classes

In both figures we see that prediabetic and diabetic people tend to have a distribution skewed to the right compared to people with negative diagnosis. The conclusion: both attributes can be helpful in training a predictive model. At the same time it is clear that it will be hard to distinguish between prediabetic and diabetic class, more on that in the results section.

## 2.2 Models

After the dataset was ready for use for our predictive purposes, we looked at the performance of several models. Here we would like to present three of them: random forest, support vector machine and MLP.

Theorie?

Parameters MLP

parameter: {'activation': 'relu', 'alpha': 0.0001, 'batch\_size': 'auto', 'beta\_1': 0.9, 'beta\_2': 0.999, 'early\_stopping': False, 'epsilon': 1e-08, 'hidden\_layer\_sizes': (49,), 'learning\_rate': 'constant', 'learning\_rate\_init': 0.001, 'max\_iter': 50, 'momentum': 0.9, 'n\_iter\_no\_change': 10, 'nesterovs\_momentum': True, 'power\_t': 0.5, 'random\_state': 0, 'shuffle': True, 'solver': 'adam', 'tol': 0.0001, 'validation\_fraction': 0.1, 'verbose': False, 'warm\_start': False}

## 2.3 Results

All three models have reached approximately the same accuracy on the test set – around 79% (see Table 2.3.1). It looked good at the first sight, but the confusion matrix and recall rates showed that the algorithms learned the simplest strategy to classify almost all the entries as ‘not diabetic’ and with that reaching the accuracy rates below.

We also looked at the feature importance graph for random forest (see Figure 2.3.1). The first thing we saw was the fact that from all the attributes only age, BMI and waist circumference had some predictive power.

These unsatisfactory results made us rethink the design of the classification. First, looking at the histograms in the Figures 2.1.2 and 2.1.3, we see that values for two classes “diabetes” and “prediabetes” go hand in hand. So, using the set of chosen features it was impossible to separate them from each other. As the number of the people without diagnosis made almost 80% of the dataset, the most efficient way to assign entries to the classes was the naïve strategy just mentioned. One possible solution to that problem was to merge “diabetes and prediabetes” into one single class and perform a binary classification.

Table 2.3.1 Accuracy of diabetes prediction in the 3 classes case.

| **Model** | **accuracy** |
| --- | --- |
| RF | 0.791139 |
| SVM | 0.791139 |
| MLP | 0.793349 |

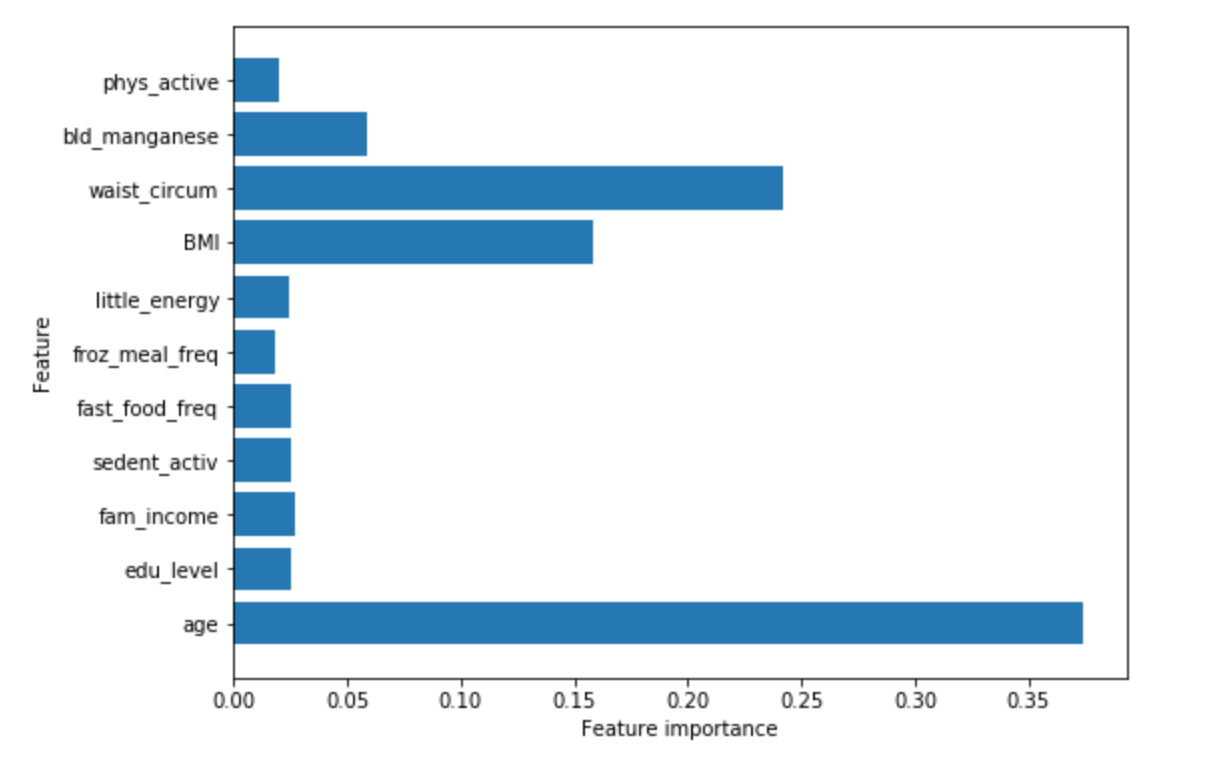


Figure 2.3.1 Feature importance in RF

In the final version of the binary classification we included only 4 features from the initial set, because it gave us best results. We had a total of 4644 entries, 1092 of them belonging to the “diabetic or at risk” class. Prediction accuracy did not get higher than it was in the three classes case, but the algorithm started to distinguish between the classes and changed the naïve strategy to a different one. All three algorithms performed in a similar way, the accuracy reaching 0.784 for random forest, 0.77 for SVM and 0.782 for MLP. In the Figure 2.3.2 you see an example decision tree from RF algorithm, that classified the samples as “diabetic or at risk” when the person was older than 46 years and had a waist circumference bigger than 112.65 cm.

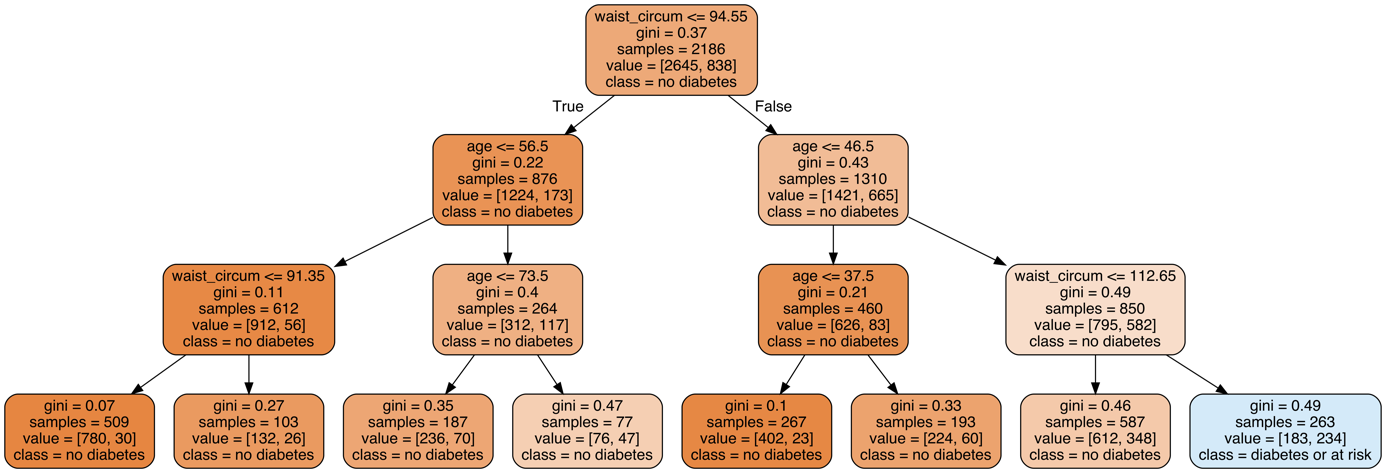
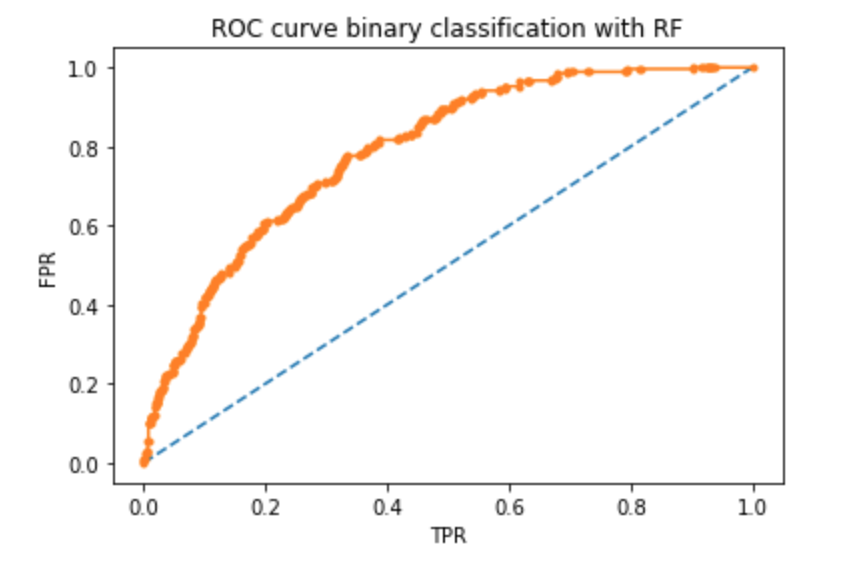


Figure 2.3.2 One of the decision trees from the RF algorithm

Below you see the confusion matrix for the RF algorithm and the ROC curve giving a better idea of the performance.

Table 2.3.2 Confusion matrix RF

|  |  |  |
| --- | --- | --- |
| Predicted  True | 0 | 1 |
| 0 | 869 | 19 |
| 1 | 232 | 41 |



AUC = 0.791

Figure2.3. ROC Curve RF

The model cannot distinguish between non-diabetic and diabetic people perfectly, but the accuracy of almost 80% is still quite good. An interesting question is, why actually features like ‘physical activity’ or ‘fast food frequency’ did not help in the classification? One possible reason is that the behavior of people changes one’s they get a positive diagnosis for diabetes. They probably start reducing unhealthy products in their diet and start to do more exercise. Mostly it doesn’t help to cure the disease, but at least it slows down its progress. If so, then we will find healthy people eating little fast food and sick people eating the same little amount and fail to separate them in the classification. If we would have values over a long enough period with nutritional behavior and physical activity data and could follow the onset of diabetes depending on that, then it maybe would be possible to predict the risks for getting sick given the lifestyle. Considering the features that did help to separate the classes, we can see that indirectly they also mirror people’s lifestyle. High Body Mass Index or large waist circumference means obesity. Obesity is in turn very often a result of a certain lifestyle and everyday choices. With that in mind we should become conscious about putting ourselves under high risk of diabetes if we don’t pay attention to the wellbeing of our bodies. The older we get, the higher is the risk.

# Malaria

## Dataset and preprocessing

## Models

## Results

# Discussion

# Conclusion

To diabetes part:

It was interesting to go from a rough idea of the project through a search of an appropriate dataset, its preparation and feature selection to the classification using different ML algorithms. Analysis of the results revealed a couple of initial mistakes in the expectation logic and gave us space to adjust the approach.