Diabetes Prediction

Kayahan Kaya
Ekaterina Sedykh
Magnus Karlson
University of Tartu

introduction

In this project, we used machine learning algorithm to predict the probability of occurring diabetes based on information about the patient such as blood pressure, body mass index (BMI), age, glucose before fasting, glucose etc.

What is Diabetes?

Diabetes is a serious condition where your blood glucose level is too high. It can happen when your body doesn't produce enough insulin, or when you can't produce any at all. There are 2 types of diabetes;

When you've got type1 diabetes, you can't make any insulin at all. If you've got type 2 diabetes, it's a bit different. The insulin you make either can't work effectively, or you can't produce enough of it. They're different conditions, but they're both serious.

Objectives

- Using machine learning algorithms to find which factors are directly caused by diabetes.
- Finding which machine learning model give the best accuracy result.

Data Analysis

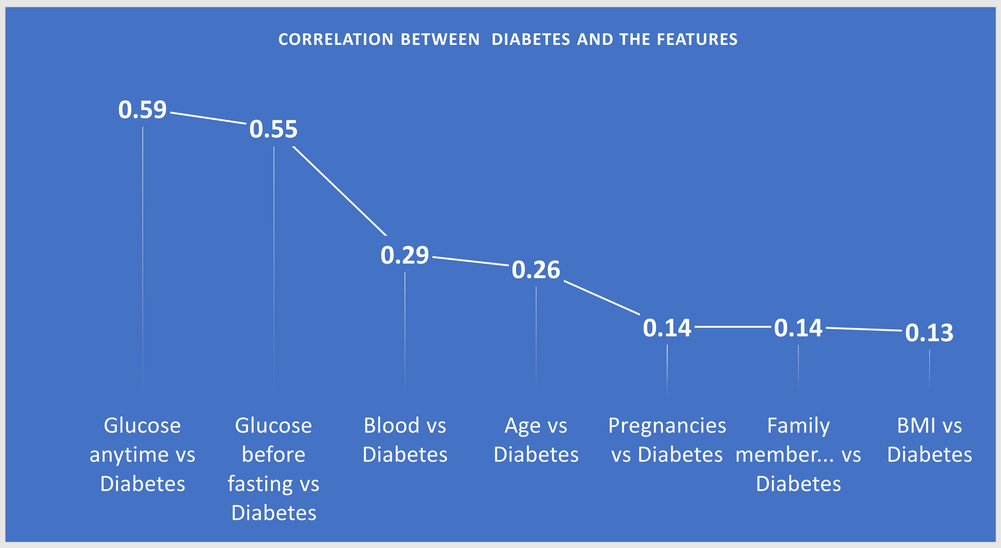
Understanding of the data is crucial step in solving complex, real world problems.

In this dataset, There are 8 features which are the factors of diabetes, 1 feature of dataset is the probability of occurring diabetes along with the 15251 observations.

The Features of the dataset;

- Glucose Before Fasting
- Glucose Anytime
- Age
- Sex
- Blood Pressure
- Family member with diabetes past or present
- BMI (Body Mass Index)
- Pregnancies
- Percentage of occuring diabetes

Here is the correlation relationship between diabetes and features;

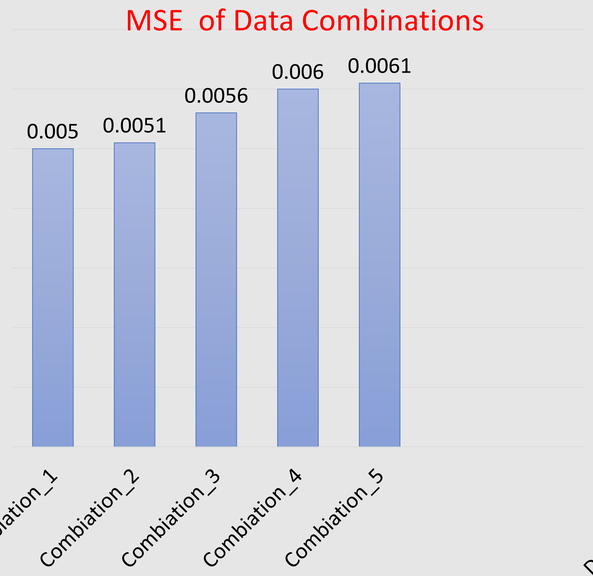


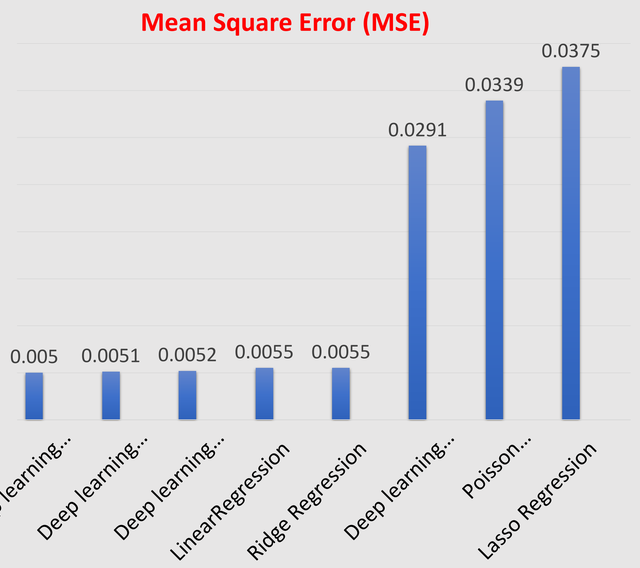
Model Design

We used 8 different machine learning models with different combinations of features (in total 368) to observe which model gives the best result with which features of the data. We decided o use regression models since the value that we want to predict is continuous.
Here is the machine learning models that we used;

1. Deep learning model_1
2. Deep learning model_2
3. Deep learning model_3
4. Deep learning model_4
5. Linear Regression
6. Lasso Regression
7. Ridge Regression
8. Poisson Regressor

Results





Combination_1: ['Glucose Before Fasting', 'Glucose Anytime', 'Age', 'Blood Pressure', 'Family member Diabetes past or present', 'Pregnancies']
Combination_2: ['Glucose Before Fasting', 'Glucose Anytime', 'Age', 'Blood Pressure', 'Family member Diabetes past or present', 'BMI', 'Pregnancies']
Combination_3: ['Glucose Before Fasting', 'Glucose Anytime', 'Age', 'Sex', 'Blood Pressure', 'Family member Diabetes past or present', 'BMI']
Combination_4 : ['Glucose Before Fasting', 'Glucose Anytime', 'Age', 'Sex', 'Blood Pressure', 'Family member Diabetes past or present', 'Pregnancies']
Combination_5: ['Glucose Before Fasting', 'Glucose Anytime', 'Age', 'Blood Pressure', 'Family member Diabetes past or present', 'BMI']

Conclusion

We used 8 different models and various data combinations to obtain a lower MSE score and the most important factors which play an important role to diagnose people whether they have diabetes or not.

Firstly, we applied regression algorithms to all the combinations of data with the number of features greater than 5 then we found the top 5 data combinations that have lower MSE scores to apply neural network algorithm.

After implementing 4 different types of neural network algorithm we got a lower MSE score than regression algorithms except for the one which has less layer and neuron. We also observed that when the deep learning algorithm that has more neurons and more layers gives better results.

As a result, Based on 368 different combinations of data trained with 8 different machine learning algorithms. The best data combination is;

['Glucose Before Fasting', 'Glucose Anytime', 'Age', 'Blood Pressure', 'Family member Diabetes past or present', 'Pregnancies'].