DIABETES PREDICTION

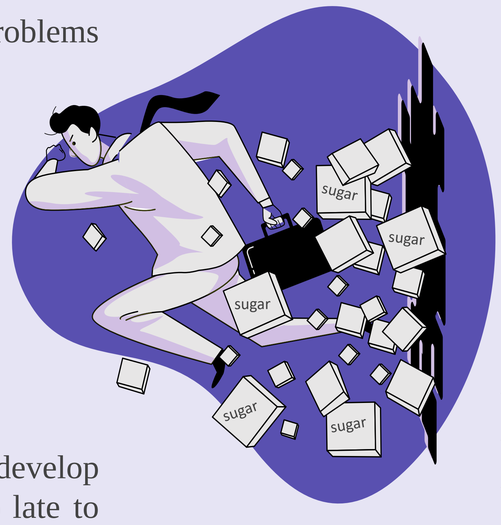
Kayahan Kaya
Magnus Karlson
Ekaterina Sedykh

INTRODUCTION

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

What Is Diabetes?

Diabetes is a serious medical condition where your body has problems with regulating blood sugar (glucose), therefore, blood glucose level is too high. **Insulin** is the key component in the process of this regulation.



There are **2 types** of diabetes:
Type 1 diabetes - you can't make any insulin at all.
Type 2 diabetes - 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.
The main problem with type 2 diabetes is that symptoms develop slowly over the years and you can notice them when it's too late to prevent the disease.

Objectives

- Using machine learning algorithms, find which factors are directly caused by diabetes
- Finding machine learning model that gives the best accuracy result

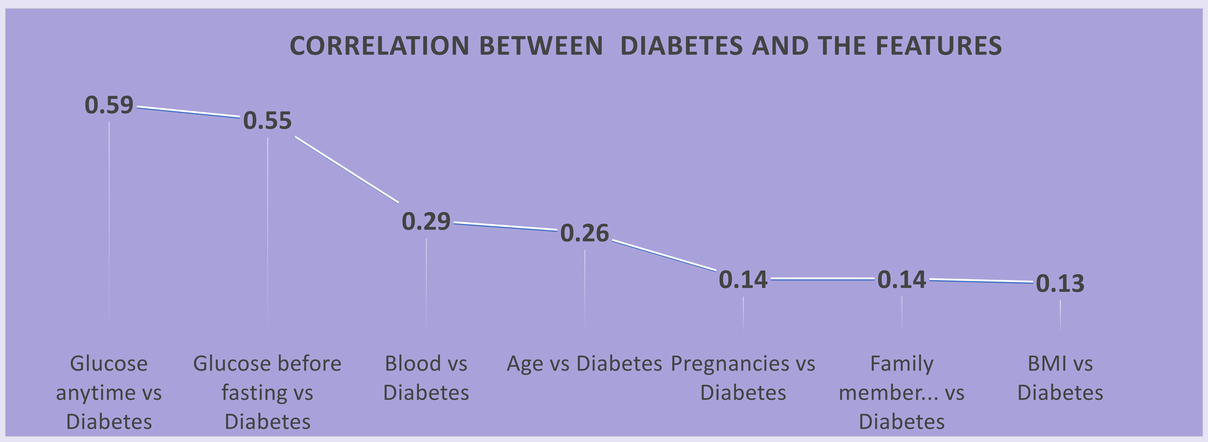
Overall goal: to help people be aware of their condition and seek help if needed

DATA ANALYSIS

In our dataset – public dataset from Kaggle, there are 8 features which are the consequences(sugar level) or prerequisites factors of diabetes, 1 feature of dataset is the probability of occurring diabetes along with the 15251 observations.

Features of the dataset:

Glucose Before fasting	Glucose Anytime	Age	Sex	Blood Pressure	Family member with diabetes past or present	BMI	Pregnancies	Percentage of occurring diabetes
156	211	63	Male	0	1	17	0	69.25
85	139	60	Male	0	0	18	0	14.5
117	156	46	Male	1	1	29	0	52.25
151	286	38	Female	1	1	20	0	70
106	277	56	Male	0	0	35	0	53.75



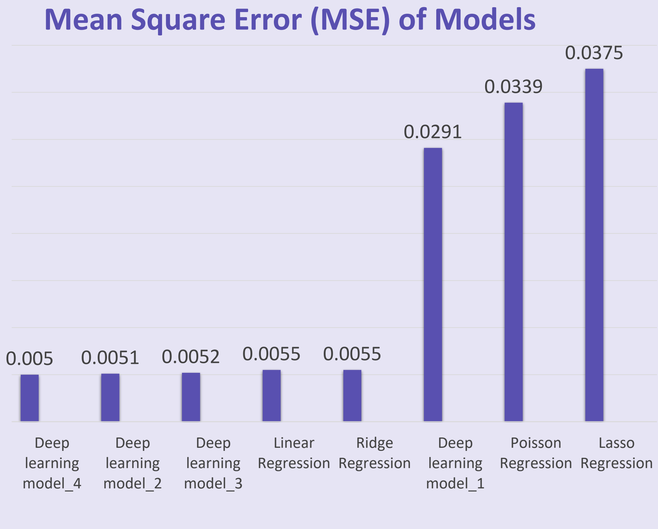
MODEL DESIGN

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 to use regression models since the value that we want to predict is continuous. Also we used deep learning models with different amount of layers, neurons and different activation functions.

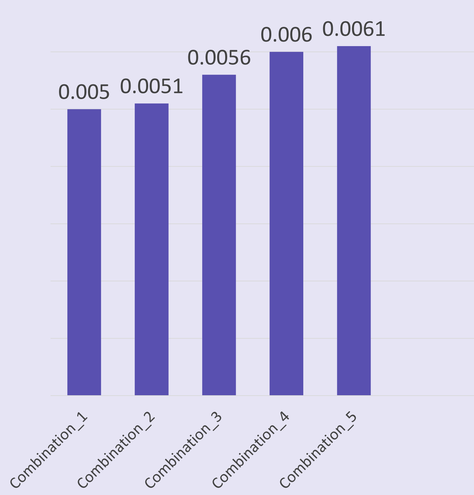
1. Deep learning model_1
2. Deep learning model_2
3. Deep learning model_3
4. Deep learning model_4
5. Linear Regression
6. Lasso Regression
7. Ridge Regression
8. Poisson Regression

RESULTS

Mean Square Error (MSE) of Models



MSE of Data Combinations



Combination_1: Glucose Before Fasting, Glucose Anytime, Age, Blood Pressure, Family member with diabetes past or present, Pregnancies
Combination_2: Glucose Before Fasting, Glucose Anytime, Age, Blood Pressure, Family member with diabetes past or present, BMI, Pregnancies
Combination_3: Glucose Before Fasting, Glucose Anytime, Age, Sex, Blood Pressure, Family member with diabetes past or present, BMI
Combination_4 : Glucose Before Fasting, Glucose Anytime, Age, Sex, Blood Pressure, Family member with diabetes past or present, Pregnancies
Combination_5: Glucose Before Fasting, Glucose Anytime, Age, Blood Pressure, Family member with diabetes past or present, BMI

CONCLUSIONS

- We used **8 different models** and various **data combinations** to obtain a lower MSE score and find the factors that play the important role in diagnosing people with diabetes
- We applied regression algorithms to all the combinations of data with the number of features greater than 5
- We found the top 5 data combinations that have lower MSE scores to apply neural network algorithm. They are consistent with our initial research about diabetes.
- 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 layers and neurons. We also observed that when the deep learning algorithm has more neurons and more layers, it 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 with diabetes past or present, Pregnancies