How Habits and Genetics may Induce Risk for Diabetes

Anna Maria Thum

Abstract

For this analysis the Diabetes dataset 2019 was used.

The main part of this project is to predict if a person has diabetes or not, as well as how a habits and genetics (like gender, regular intake of medicine, habit of smoking…) correlate.

After the data was cleaned (more about data cleaning on slide 5), the method heatmap from seaborn was used to visualise the correlations and for the diabetes prediction decision trees were used.

Motivation

As diabetes affects more and more people worldwide , with China having currently the highest number of diabetics, this presentation deals with the different factors, that may increase ones risk of having or getting diabetes. Diabetes can lead to serious health complications, such as strokes, cardiovascular disease and is under the top ten leading causes of death worldwide.

This issue might be helpful for everyone, struggeling with this illness or trying to avoid this problem.

Dataset(s)

This dataset was gathered by Neha Prerna Tigga and Dr. Shruti Garg of the Department of Computer Science and Engineering, BIT Mesra. It can be found under https://www.kaggle.com/tigganeha4/diabetes-dataset-2019.

The dataset consists of 952 samples and 18 variables. The observations are binary (diabetic, gender, smoking, …), numeric (pregnancies, sleep, BMI, …) and grouped (age: less than 40, 50-59,… or time of physical activity).

Data Preparation and Cleaning

As preparation for the analysis the data was modified as followed:

- Drop all NA values with .dropna()
- In column “Age” values varied between “yes”, “no” and “ no”. To get a binary result “ no” was changed to “no”. Similar to that, in column “BPLevel” (Blood Pressure Level) were values “High”, “high”, “normal”, “normal “, “Low” and “low”. These values were grouped to either “high”, “normal” or “low”.
- In column “Regular Medicine” there were “yes”, “no” and “o”. As “o” could not be interpreted, these rows were dropped. Same in “Pregnancies” as “high”, “normal” and “low” could not be interpreted as numbers and were deleted.

Data Preparation and Cleaning

As preparation for the analysis the data was modified as followed:

- Whole columns where deleted as they were almost everywhere 0, or could not be explained.
- As all values where of datatype object, it was not possible to perform correlations or decision trees. So dummy variables where introduced.
- The remaing data were split into training and testing set (67% of the data for training with n = 599, the remaining 33% for testing with n = 296)

Research Questions

Questions to be answered in this project:

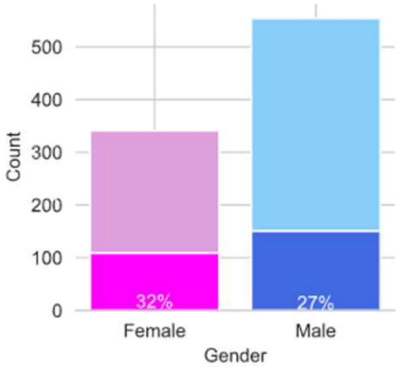
1. How do different factors influence the risk for having diabetes? What are the interactions between them?
2. Computation of the risk one might be affected by diabetes and how well can the risk of having or getting diabetes be predicted from the dataset?

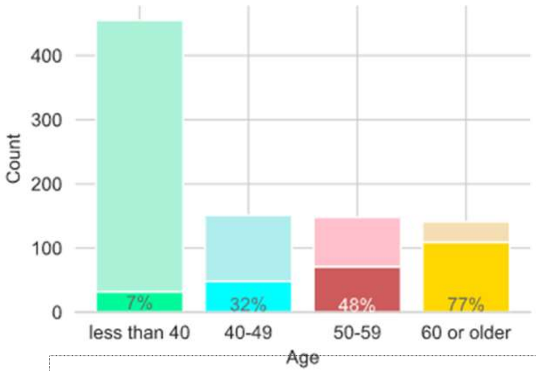
Methods

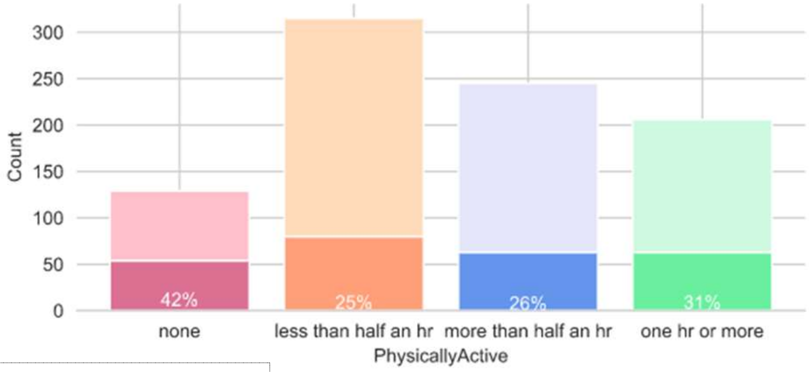
1. At first the dataset was inspected to find out, for example how diabetes is distributed between different features. Here the correlation between each factor was visualised with a correlation heatmap from seaborn.

2. To predict the chances for diabetes the data was split randomly in 67% training and 33% test set. After that the machine learning process decision tree was fitted to the training set and checked with the test set.

Findings: Diabetes Distribution by different factors







Lower and darker bars represent the amount of diabetics in each observation.

- There are more men (n=554) in this dataset than women (n=341).

- The diabetes ratio is roughly the same in both genders (females with 32% and males with 27%).

- This dataset consists of 50% of young people under 40 (n=455). The other age groups have roughly the same number (n=[151,148,141], in correct order).

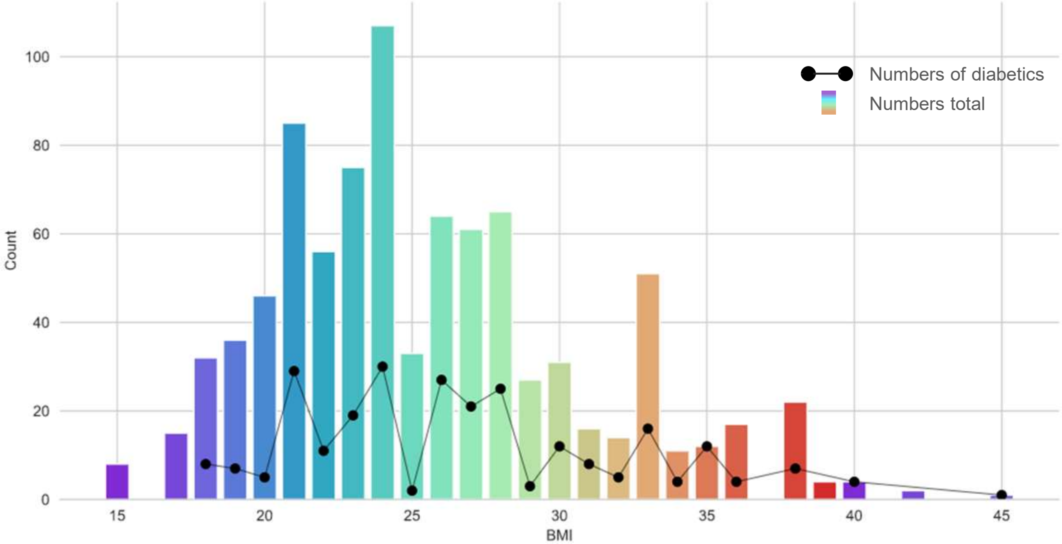
- The older the people the more people have diabetes (from people less than 40 with 7% up to people 60 or older with 77%).

- 42% of people that do no sports have diabetes.

- People being less or more than half an hour active have the same amount of diabetics (around 25-26%).

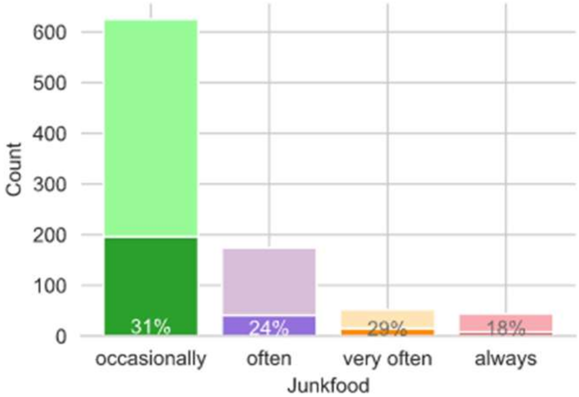
- 31% of active people have diabetes.

Findings: Diabetes Distribution by different factors



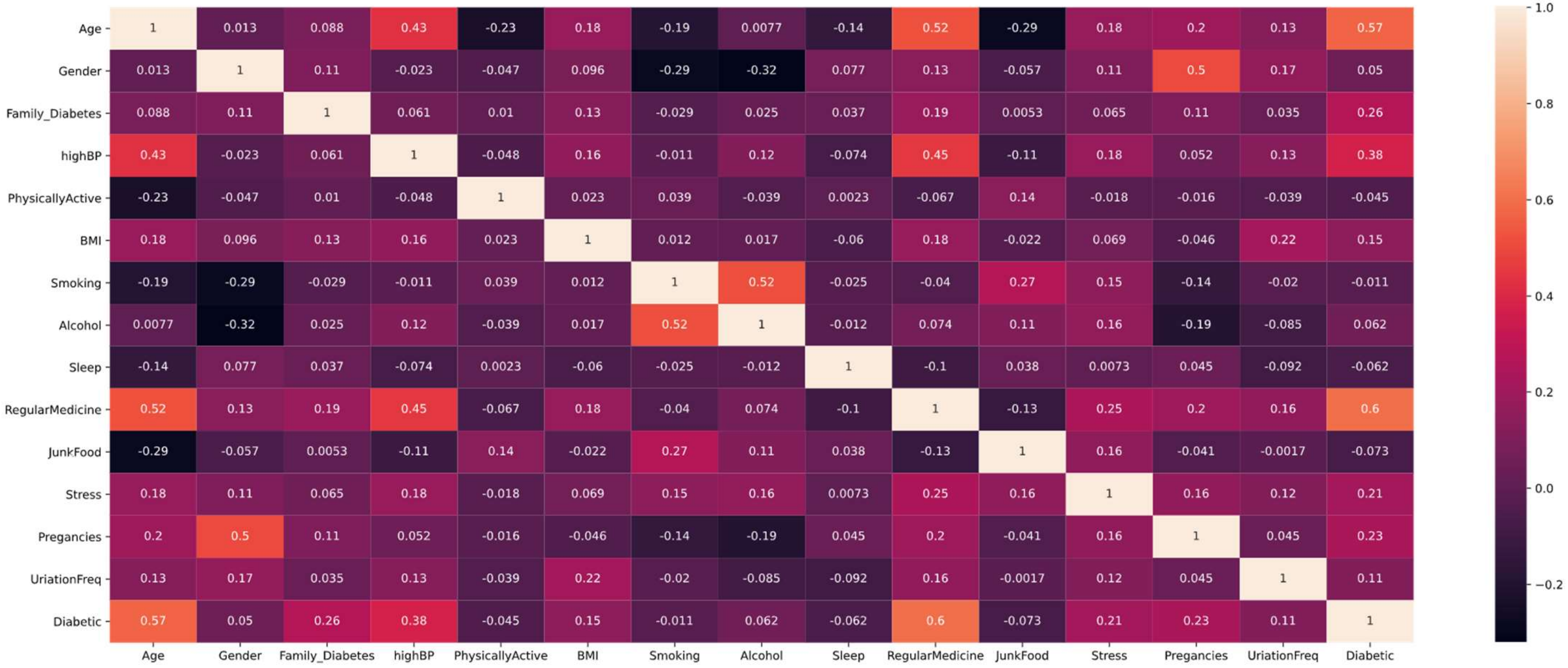
- The most people in this data set have a healthy BMI between 19 and 24 (n=405).
- People with lower BMI (n=55) have the lowest number of diabetics (14%). The higher the BMI the higher are the numbers of diabetics:
    - normal BMI: 25% with n=405
    - high BMI: 34% with n=435

Lower and darker bars represent the amount of diabetics in each observation.



- Most people eat occasionally junkfood (n=625), after that people eat in descending order often (n=174), very often (n=52) and always (n=44) junkfood.
- People eating occasionally junkfood have the highest amount of diabetics. Followed by people eating very often (29%), often (24%) and always (18%) junkfood.

Findings: Correlationmatrix



	Age	Gender	Family_Diabetes	highBP	PhysicallyActive	BMI	Smoking	Alcohol	Sleep	RegularMedicine	JunkFood	Stress	Pregancies	UriationFreq	Diabetic
Age	1	0.013	0.088	0.43	-0.23	0.18	-0.19	0.0077	-0.14	0.52	-0.29	0.18	0.2	0.13	0.57
Gender	0.013	1	0.11	-0.023	-0.047	0.096	-0.29	-0.32	0.077	0.13	-0.057	0.11	0.5	0.17	0.05
Family_Diabetes	0.088	0.11	1	0.061	0.01	0.13	-0.029	0.025	0.037	0.19	0.0053	0.065	0.11	0.035	0.26
highBP	0.43	-0.023	0.061	1	-0.048	0.16	-0.011	0.12	-0.074	0.45	-0.11	0.18	0.052	0.13	0.38
PhysicallyActive	-0.23	-0.047	0.01	-0.048	1	0.023	0.039	-0.039	0.0023	-0.067	0.14	-0.018	-0.016	-0.039	-0.045
BMI	0.18	0.096	0.13	0.16	0.023	1	0.012	0.017	-0.06	0.18	-0.022	0.069	-0.046	0.22	0.15
Smoking	-0.19	-0.29	-0.029	-0.011	0.039	0.012	1	0.52	-0.025	-0.04	0.27	0.15	-0.14	-0.02	-0.011
Alcohol	0.0077	-0.32	0.025	0.12	-0.039	0.017	0.52	1	-0.012	0.074	0.11	0.16	-0.19	-0.085	0.062
Sleep	-0.14	0.077	0.037	-0.074	0.0023	-0.06	-0.025	-0.012	1	-0.1	0.038	0.0073	0.045	-0.092	-0.062
RegularMedicine	0.52	0.13	0.19	0.45	-0.067	0.18	-0.04	0.074	-0.1	1	-0.13	0.25	0.2	0.16	0.6
JunkFood	-0.29	-0.057	0.0053	-0.11	0.14	-0.022	0.27	0.11	0.038	-0.13	1	0.16	-0.041	-0.0017	-0.073
Stress	0.18	0.11	0.065	0.18	-0.018	0.069	0.15	0.16	0.0073	0.25	0.16	1	0.16	0.12	0.21
Pregancies	0.2	0.5	0.11	0.052	-0.016	-0.046	-0.14	-0.19	0.045	0.2	-0.041	0.16	1	0.045	0.23
UriationFreq	0.13	0.17	0.035	0.13	-0.039	0.22	-0.02	-0.085	-0.092	0.16	-0.0017	0.12	0.045	1	0.11
Diabetic	0.57	0.05	0.26	0.38	-0.045	0.15	-0.011	0.062	-0.062	0.6	-0.073	0.21	0.23	0.11	1

Findings : Correlationmatrix

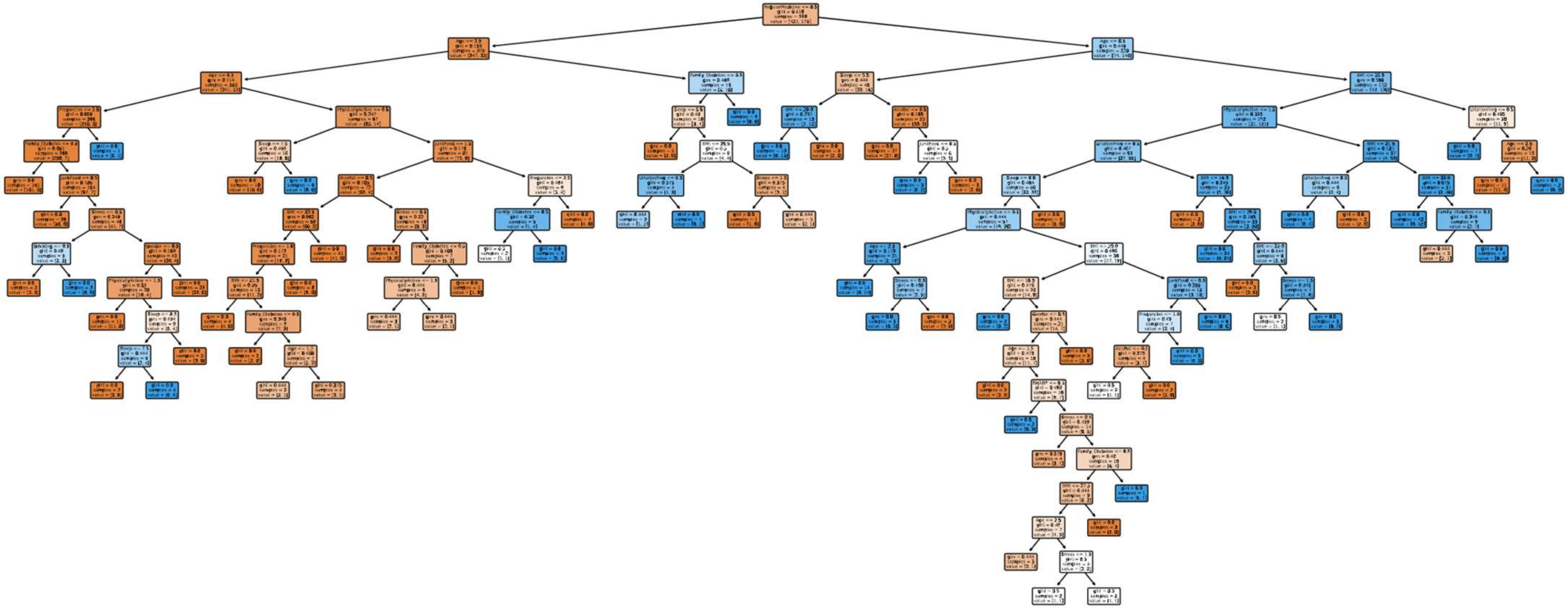
- Almost all risk factors are not correlated with each other.
- Gender is weak negativ correlated with Alcohol and Smoking, that means that more more men are smoking (R^2 = -0.29) and drinking alcohol (R^2 = -0.32) . Smoking and drinking alcohol are weak positive correlates (R^2 = 0.52).
- Age is negativ correlated with junkfood habits (R^2 = -0.29) . Meaning that younger people eat more junkfood than older generations.
- Age, regular medicine intake, diabetes, and high blood pressure are positive correlated (for example age and regular medicine with R^2 = 0.52).
- Interestingly gender and number of pregnancies are only weak correlated with R^2 = 0.5. After further investigations, it turns out that 12 men had more than 0 pregnancies. This would need more questioning.

Findings: Machine Learning with Decision Tree

For this project a decision tree was built to predict if a person has diabets or not. Features were:
Age, Family_Diabetes, highBP, PhysicallyActive, BMI, Smoking, Alcohol, Sleep, RegularMedicine, JunkFood, Stress, Pregnancies, UrinationFreq and Diabetic.

According to these observations the decision tree was 96% accurate with the test data.

Findings: Visualisation of the Decision Tree



Limitations

Since 952 people were asked, these findings are not representative, but offer guidelines.

This dataset does not distinguish between diabetes type 1 and diabetes type 2.

The age groups were quite big for people under 40 and over 60.

Conclusions

The risk factors mostly do not correlate. Some numbers would need further exploration (see men having multiple pregnancies).

The prediction if a person has diabetes is well modelled with a decision tree at 96% accuracy.

Acknowledgements

The data was found under:
https://www.kaggle.com/tigganeha4/diabetes-dataset-2019

Informations about diabetes and BMI:

- https://www.statista.com/statistics/271464/percentage-of-diabetics-worldwide/
- https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/

References

This work was done by me and I received no feedback.