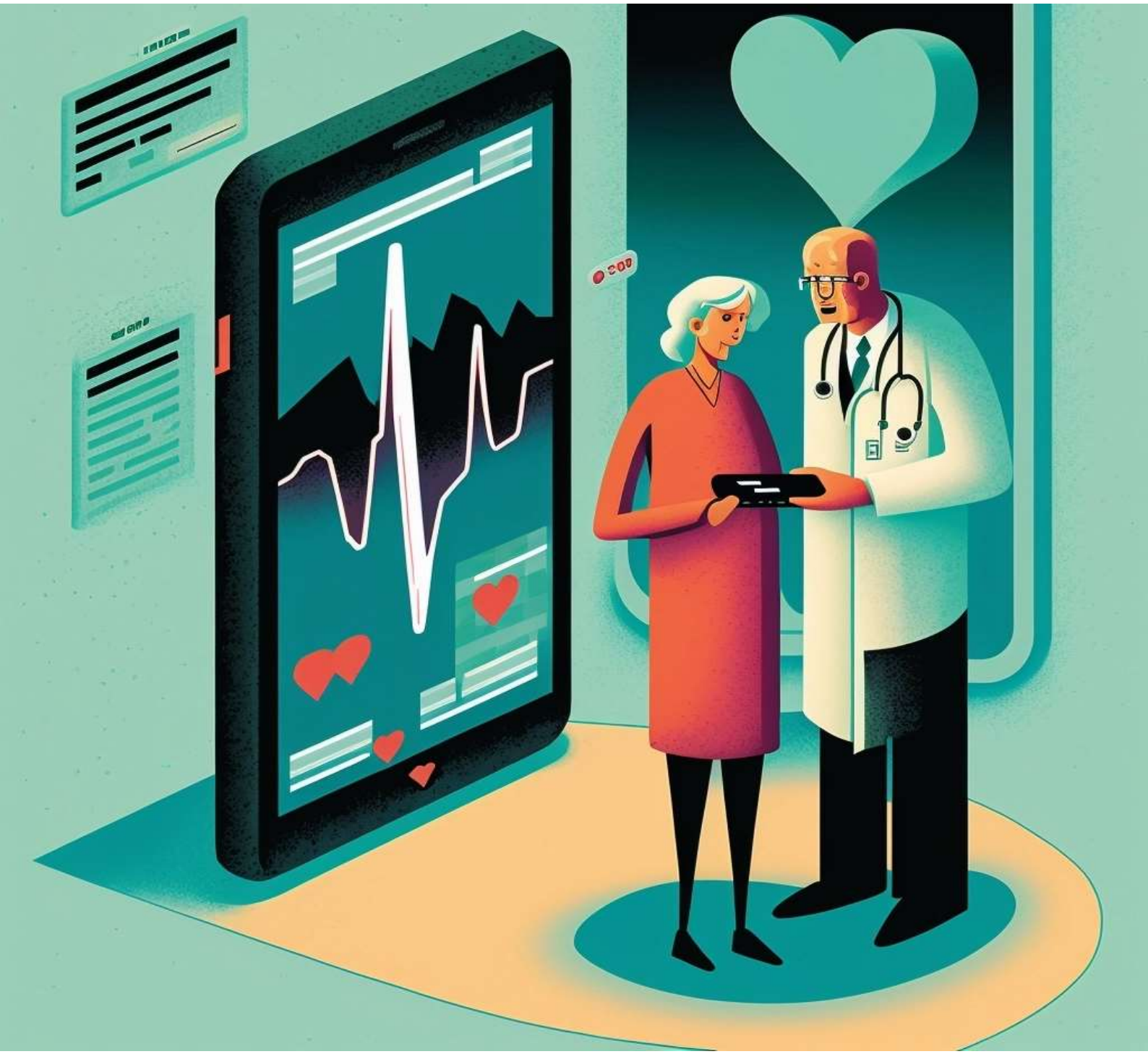
Project 1

Healthcare



March 2023

Team with no NaN
Authors: Stephan, Jiyoon, Hans, Frank

MAKE IT WORK

Table of Contents

Acknowledgments

This project would not have been possible without the guidance from our dedicated teachers: Jeroen, Ruud and Frank. Their invaluable feedback, patience and knowledge helped us immensely on our journey into machine learning.

We are also grateful for our classmates for their help and interesting discussions

Lastly we would like to thank our families and pets for their emotional support and belief.

Introduction

Machine learning and AI has great potential to increase accuracy in medical diagnosis and increase efficiency in medical processes. Medical Center Randstad wants to hold a pilot to see if they can use the data they gathered from their patients to, based on their lifestyle, predict how old someone will get. On top of that they want to give tailored advice that can help increase a patients lifespan.

This project will have the following goals:

- Gathering of the data
- Transforming the data so its suitable for modeling
- Finding out how the correlation between lifestyle and genetics affects lifespan
- Presenting the findings and a working application that can predict lifespan

I. The Pipeline

The pipeline chosen for this project consists of two parts. The first part generates the data and the second fits the regression for use in the end-user interface.

Part I: Data Model

The main philosophy behind our pipeline is versatility. At the start of the project, it was unclear to us what dataset we would be using. Three options had been presented to us for the data acquisition:

1. Pull from a rest API.
2. Pull from an SQLite database.
3. Read from a .csv file

At the start of the project there was some confusion amongst our team whether we ought to build for a static or a dynamic dataset. In an early stage of the project, we made the design choice to start building a pipeline that could import from the three sources. Because we were unsure if the supplied data would be static or dynamic, we made the choice to automate the generation of different datasets prior to the EDA conclusions what dataset to use for the final version.

The data models that made it to the next stage have an enumerated name structure df1 to df5.
For use in the next stage of the project the data model generated its output in .csv and stores it in an SQLite.db

Data analyses determined that df4 would be the model we'd be using for the final regression model. The other models are currently commented out to reduce system load and clutter.

Part II: Regression Model

The versatile design philosophy was at the forefront for our design of the regression file. The EDA made it clear to us that the path ahead led to a linear regression model. We started to craft our regression model with versatile data path in and out. For input the standard is currently set to df4 from SQLite. Import from .csv is currently commented out. With a simple filename change in the model we can use the model to run regression on our other df's.

Our regression model's output is versatile as well it is both stored in a python pickle format (.pkl) as in an SQLite database. Allowing for the diverse life expectancy interfaces our group is proud to present.

Data sources and collection methods

- Pull from a rest API.
- Pull from an SQLite database.
- Read from .csv file

Data cleaning and preprocessing techniques used

- There were very few NaNs, dropping them would not have a significant effect on results
- Duplicates were dropped (there were no duplicates)
- Few negative values that are invalid, also dropped
- Strange characters converted tot NaNs and dropped
- Outliers are possible good values in general, so we left them for further analysis

Feature selection and engineering

Mass (kg)

Maybe worthwhile to divide into separate groups and training of the model can be done based on these groups.

BMI (mass / length^2)
seems interesting to add, based on the results for mass.

- Categories and subcategories as defined by the World Health Organization seem interesting to add. Grouping on these categories give the possibility to try clustering and look for better results per category. A regression per category is interesting but out of scope, because lack of time.

Once you have worked out your BMI score, use the table below to determine your BMI rating. The rating is the same for males and females. The table shows the World Health Organization BMI classification system.

classification	**BMI (kg/m^2)**	**sub-classification**		**BMI (kg/m^2)**
		severe thinness		< 16.00
underweight	< 18.50	moderate thinness		16.00 - 16.99
		mild thinness		17.00 - 18.49
normal range	18.5 - 24.99	normal		18.5 - 24.99
		pre-obese		25.00 - 29.99
overweight	≥ 25.00	Obese (≥ 30.00)	obese class I	30.00 - 34.99
			obese class II	35.00 - 39.99
			obese class II	≥ 40.00

source: World Health Organization

Datasets
We defined different sets to test for best results, when fitting the model:

4	**Experiment Tracking Table**							
5	train, test = train_test_split(mydf, test_size=0.2, random_state=42)							
6								
7	**Experiment #**	**NaN drop**	**Dupli drop**	**Neg drop**	**object -> float**	**BMI feature**	**IQR-clip**	**IQR-drop**
8	DF1	x	x	x	x			
9	DF2	x	x	x	x	x		
10	DF3	x	x	x	x	x	x	
11	DF4	x	x	x	x	x		x

_Experiment_Tracking_Management.xlsx

IRQ method is used to clip and drop outliers.

II. Exploratory data Analysis

Source: EDA_Pipeline.ipynb

Summary statistics and visualizations

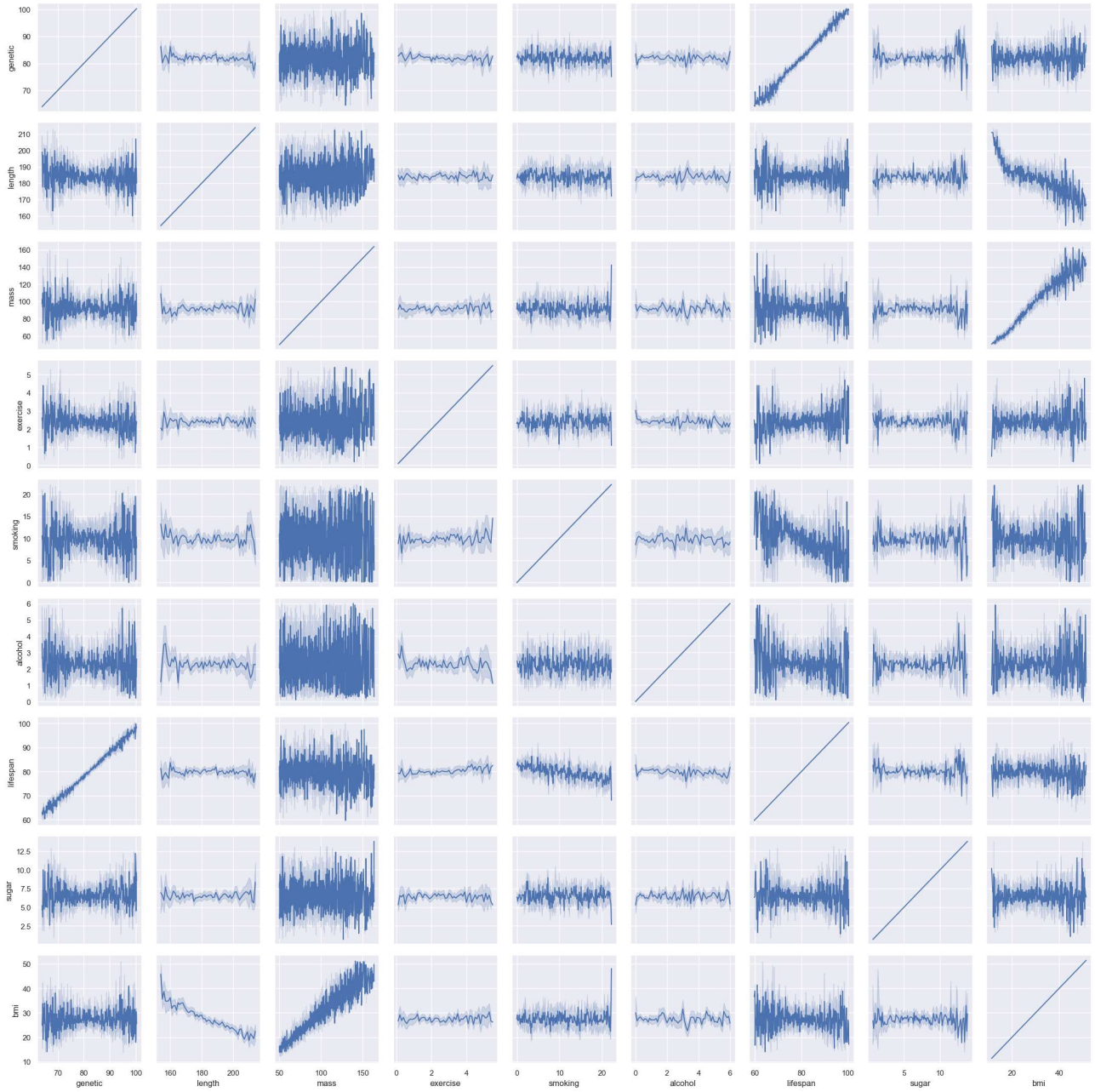
	genetic	length	mass	exercise	smoking	alcohol	lifespan	sugar	bmi
count	3977.00	3977.00	3977.00	3977.00	3977.00	3977.00	3977.00	3977.00	3977.00
mean	82.04	184.05	91.71	2.40	9.78	2.27	79.99	6.48	27,41
std	7.09	12.52	25.25	1.17	6.91	1.80	7.45	2.71	8,30
min	64.00	154.00	50.00	0.10	0.00	0.00	59.70	0.70	11,30
25%	77.60	175.00	71.30	1.50	3.30	0.60	75.00	4.40	20.90
50%	81.70	184.00	89.00	2.20	9.50	1.70	79.60	6.30	26.40
75%	86.50	193.00	110.00	3.20	15.80	4.00	84.90	8.40	33.00
max	100.30	214.00	163.60	5.50	22.20	6.00	100.40	13.80	51.50

From this table:

- we expect some outliers in mass
- smoking and alcohol with 0 values are discrete

When we plot each possible feature combination we see:

- a negative correlation between smoking, alcohol with lifespan
- a positive correlation between exercise and genetic (strong linear) with lifespan

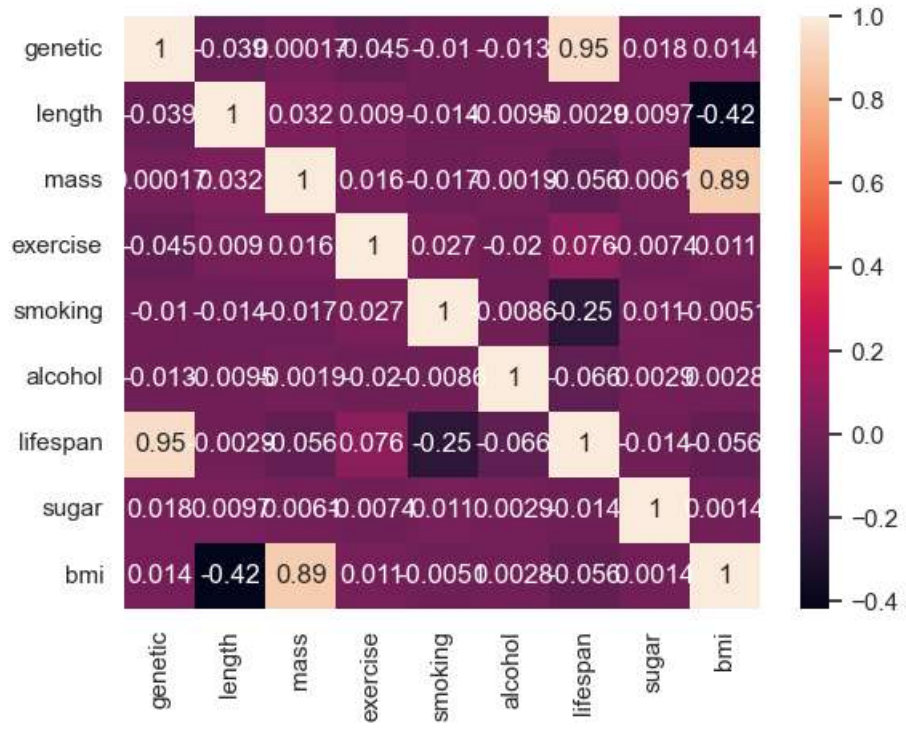


Correlation analysis and feature importance ranking

When we calculate the correlation coefficient between features (but remember correlation != causation) we see the following:

- genetic seems to be the most correlated with lifespan
- runners up are smoking (-), exercise(+) and alcohol (-)
- other values seems to have less, but negative impact on the lifespan

	genetic	length	mass	exercise	smoking	alcohol	lifespan	sugar	bmi
genetic	1.0000	-0.0388	0.0002	-0.0452	-0.0105	-0.0129	0.9474	0.0180	0.0144
length	-0.0388	1.0000	0.0322	0.0090	-0.0140	-0.0095	-0.0029	0.0097	-0.4181
mass	0.0002	0.0322	1.0000	0.0157	-0.0174	-0.0019	-0.0556	0.0061	0.8852
exercise	-0.0452	0.0090	0.0157	1.0000	0.0274	-0.0204	0.0755	-0.0074	0.0109
smoking	-0.0105	-0.0140	-0.0174	0.0274	1.0000	-0.0086	-0.2502	0.0112	-0.0051
alcohol	-0.0129	-0.0095	-0.0019	-0.0204	-0.0086	1.0000	-0.0659	0.0029	0.0028
lifespan	0.9474	-0.0029	-0.0556	0.0755	-0.2502	-0.0659	1.0000	-0.0136	-0.0563
sugar	0.0180	0.0097	0.0061	-0.0074	0.0112	0.0029	-0.0136	1.0000	0.0014
bmi	0.0144	-0.4181	0.8852	0.0109	-0.0051	0.0028	-0.0563	0.0014	1.0000

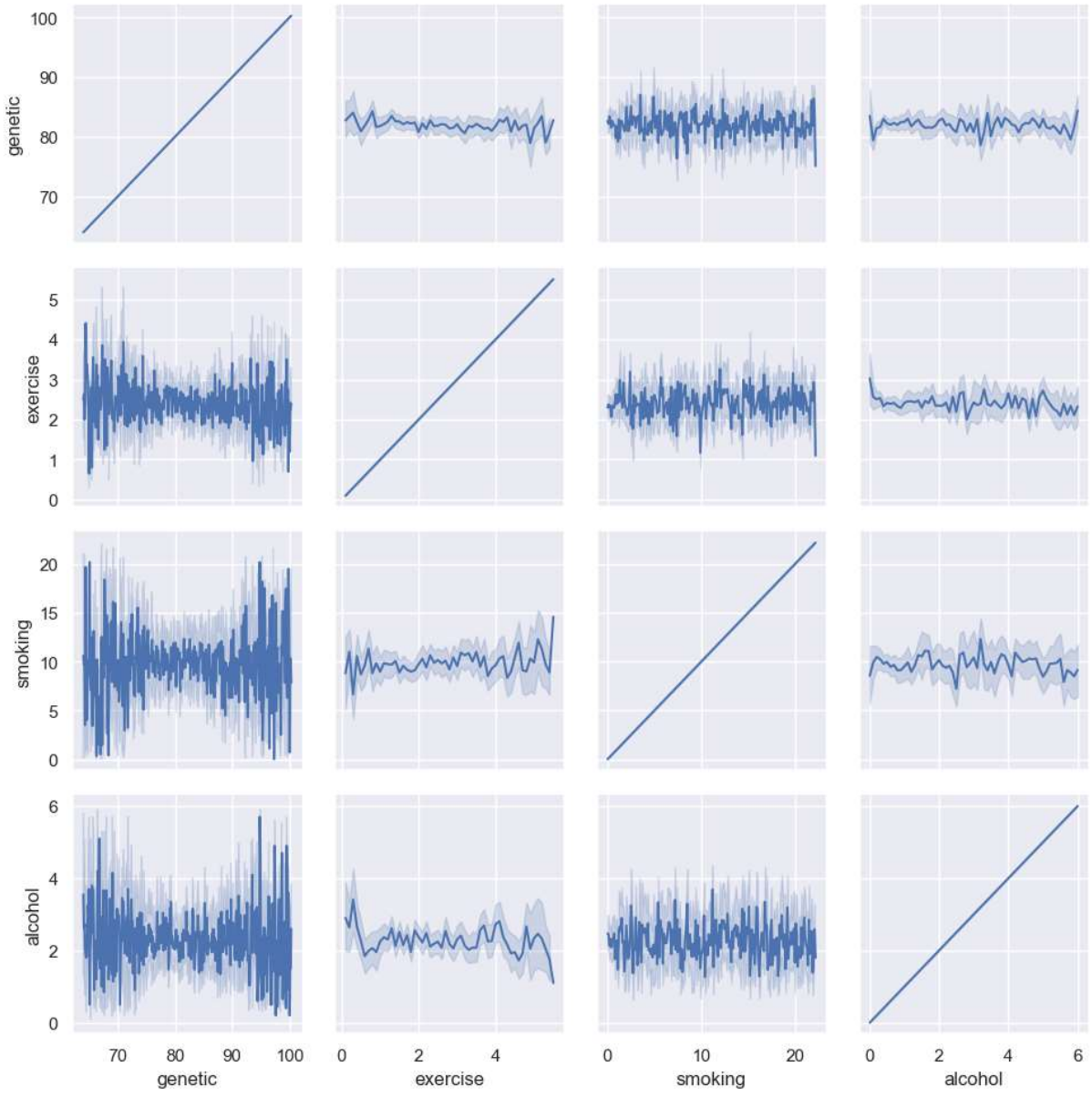


Multicollinearity check

> In statistics, multicollinearity (also collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data

From the correlation matrix we check the top 3 correlated features (apart from lifespan)

Exercise – genetic	-0.045
Exercise – smoking	0.027
Exercise - alcohol	-0,020



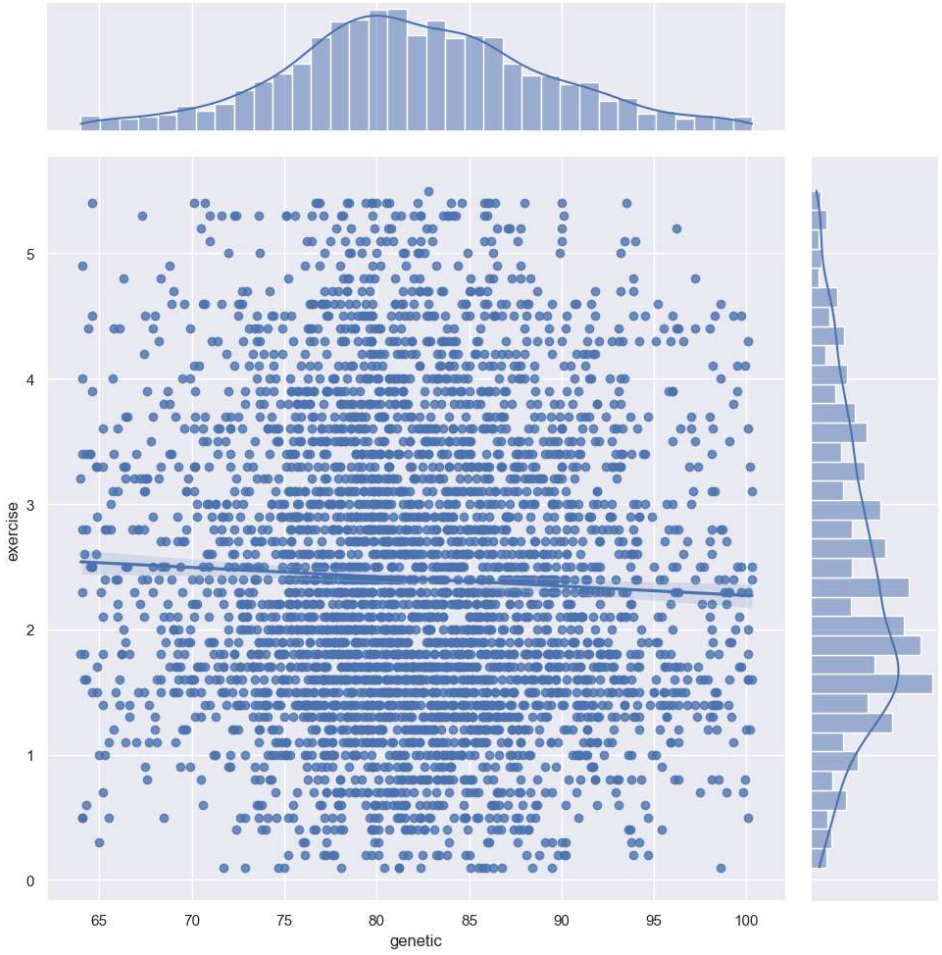
We apply the function stats.pearonr which also performs a significance test with the null hypothesis that there is no significant correlation between exercise and e.g. smoking.

When the p-value is:

- < 0.05 we reject the null hypothesis and the correlation **is** statistically significant.
- > 0.05 we fail to reject the null hypothesis and the correlation **is not** statistically significant.

```
 smoking corr: 0.0274
smoking p-val: 0.0837
alcohol corr: -0.0204
alcohol p-val: 0.1978
genetic corr: -0.0452
genetic p-val: 0.0044
```

Based on the results: we reject smoking and alcohol as statistically significant correlated to exercise, however genetic is significantly correlated to exercise, so lets graph their relation:



There is a lightly negative linear correlation between genetic and exercise, for the sake of this analysis we continue with the assumption that there is a negative linear correlation.

Greedy elimination:

If genetic and exercise are correlated then the feature with the lowest correlation to lifespan is eliminated in this case exercise with 0.076 will be removed (genetic = 0.95)

Lets quantify the difference in linear regression accuracy (with and without ‘exercise’)

```
coefficient of determination(R²) with exercise: 0.9792571387182405

RMSE: 1.085065117096135

coefficient of determination(R²)without exercise: 0.963087002111707

RMSE: 1.4474751925341771
```

Deleting the slightly correlated exercise reduced accuracy of the model therefore we conclude it improves the model and decide to keep exercise despite being multi-correlated

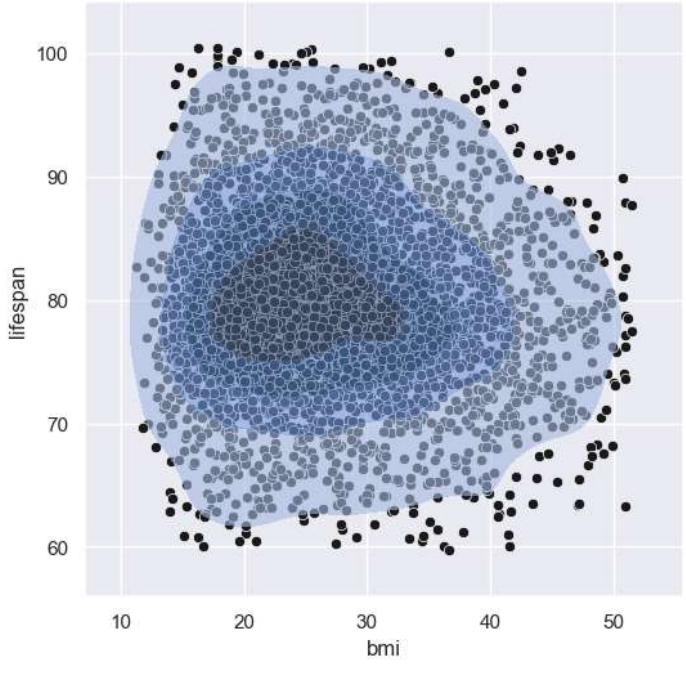
> *"The fact that some or all predictor variables are correlated among themselves does not, in general, inhibit our ability to obtain a good fit nor does it tend to affect inferences about mean responses or predictions of new observations."*
>
> Applied Linear Statistical Models, p289, 4th Edition

K-means analysis

Definition:

The k-means algorithm is an unsupervised clustering algorithm. It takes a bunch of unlabeled points and tries to group them into “k” number of clusters. It is unsupervised because the points have no external classification.

Although our data is already supervised (labeled) it would be better to apply a KNN analysis but because this realization came to us later, we show the k-means analysis as exercise:



← We start with investigating bmi-exercise because of the extensive literature already present about bmi having different categories and we expect exercise to have a big impact. Given the at initial sight random scatter plot of bmi-exercise, it might be worth it to do a k-means analysis on this feature combination.

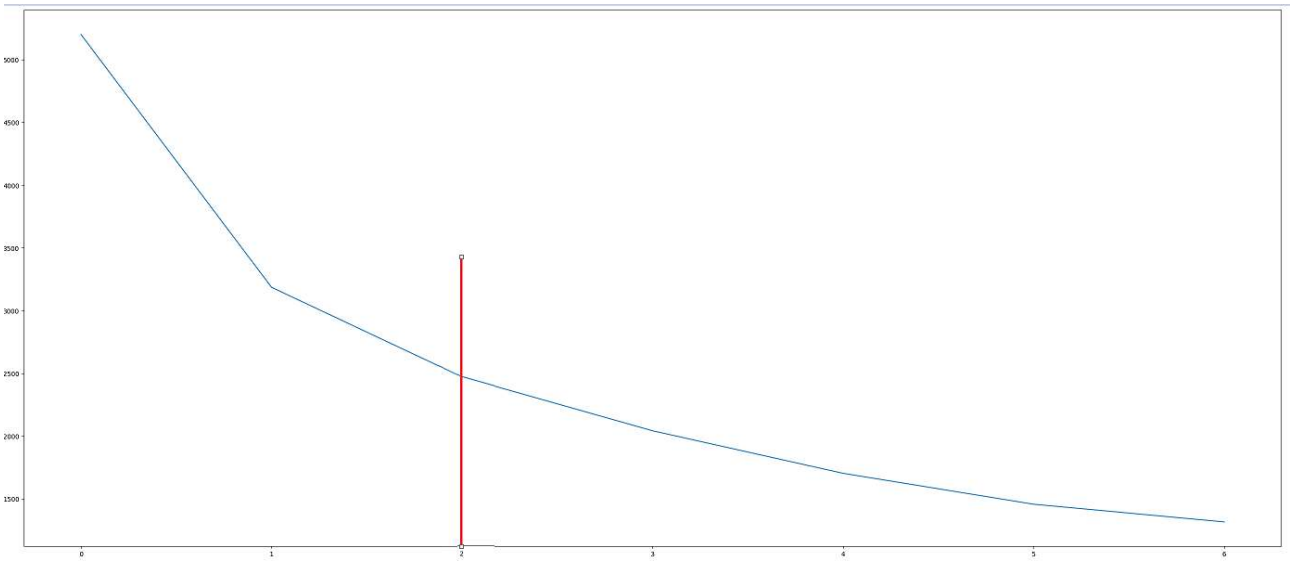
First we run a Hopkins test on the data:

> The Hopkins statistic (introduced by Brian Hopkins and John Gordon Skellam) is a way of measuring the cluster tendency of a data set. It belongs to the family of sparse sampling tests. It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed. A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.
>
> https://en.wikipedia.org/wiki/Hopkins_statistic

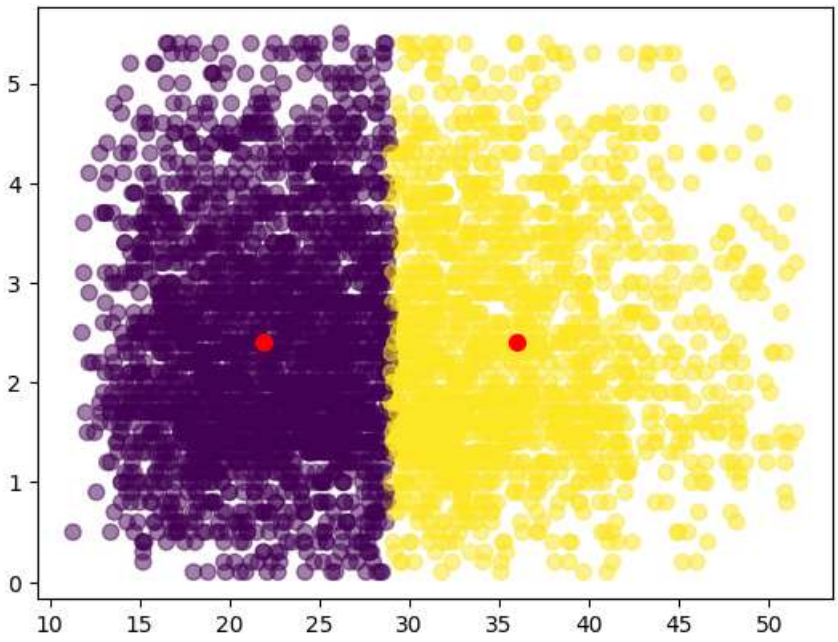
The Hopkins function is defined in exercise_Kmeans.ipnyb and gives us the following value:

```
0.7890625725275615
```

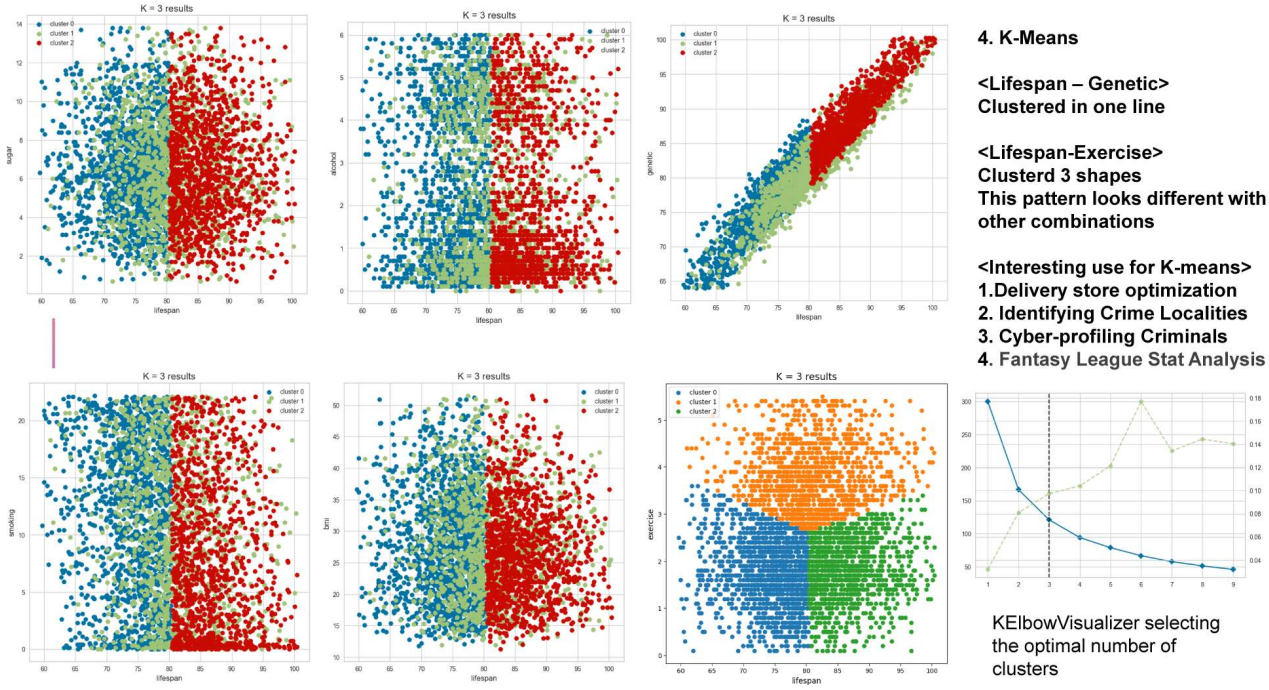
Which means the data is highly clustered (from the plot alone at least 1 big cluster). The amount of clusters will be determined using the elbow method, a rather subjective and ambigous technique since the amount still needs to be defined by sight but its simple and fast:



We identify 2 as the optimum amount of clusters for exercise-bmi (scaled) which as expected does not give us something useful



Further analysis applying K-means on features and lifespan gives us the following:



4. K-Means

<Lifespan – Genetic>
Clustered in one line

<Lifespan-Exercise>
Clusterd 3 shapes
This pattern looks different with other combinations

<Interesting use for K-means>
1.Delivery store optimization
2. Identifying Crime Localities
3. Cyber-profiling Criminals
4. Fantasy League Stat Analysis

KElbowVisualizer selecting the optimal number of clusters

Exercise-lifespan shows a nice mercedes form of clustering (K=3) but it is apparent the data is not suitable for clustering

Feature Analysis

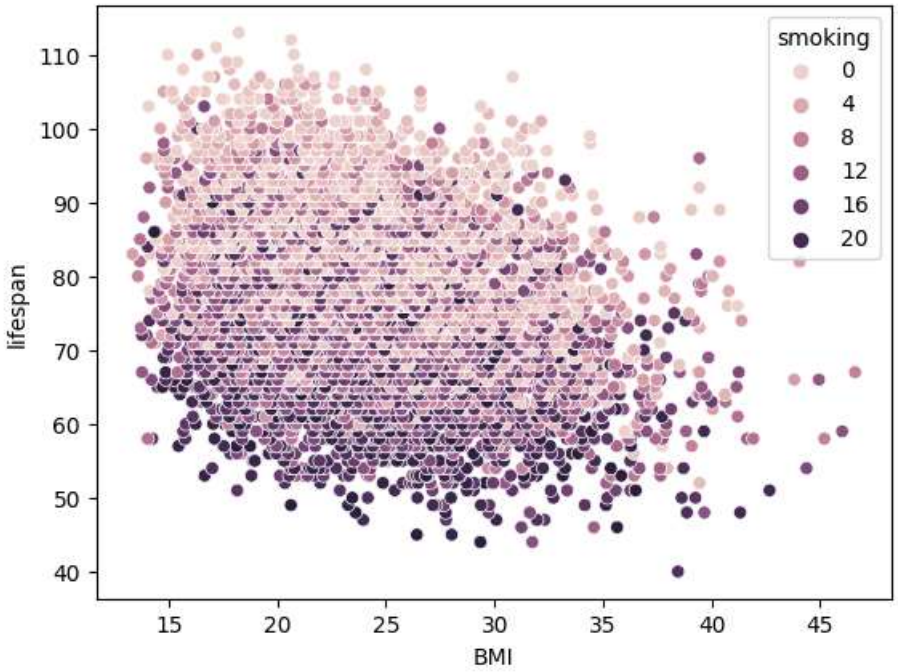
Impact of BMI:

Earlier on we decided to add BMI as a feature and now we can measure how it affects the accuracy in our linear regression model:

```
Without BMI (R²):      0.9789841911416836

With BMI (R²):         0.9792571387182405
```

This comes down to a 0.027% increase, its not a lot but since the model is already very accurate we decided to keep BMI because it has been researched extensively and we can use that knowledge (e.g. categories) to optimize our model in the future.



Feature transformation:

`sources: exercise_feature_transformations.ipynb, Mass_Sqrt_Transformation_Norm.ipynb`

Algorithms like Linear Regression, Logistic Regression, KNN or Neural Nets can be highly affected by distribution and scale of the input features. Transforming these could lead to better results.

Regarding feature scaling and linear regression:

```
Consider the model:

y=β0+β1×1+β2×2+…+ϵy=β0+β1×1+β2×2+…+ϵ.

The least squares estimator of β1,β2,…β1,β2,… are not affected by shifting. The reason
is that these are the slopes of the fitting surface – how much the surface changes if
you change x1,x2,…x1,x2,… one unit. This does not depend on location. The scaling
```

> doesn't affect the estimators of the other slopes. Thus, scaling simply corresponds to scaling the corresponding slopes.
>
> **To conclude,** technically, feature scaling does not make a difference in the regression, but it might give us some practical benefits and in further feature engineering steps.
>
> https://www.atoti.io/articles/when-to-perform-a-feature-scaling/

Continuing with only transformations:
We use the python library fitter (https://fitter.readthedocs.io) to quickly find the current best fit by comparing the features to the most common distributions:

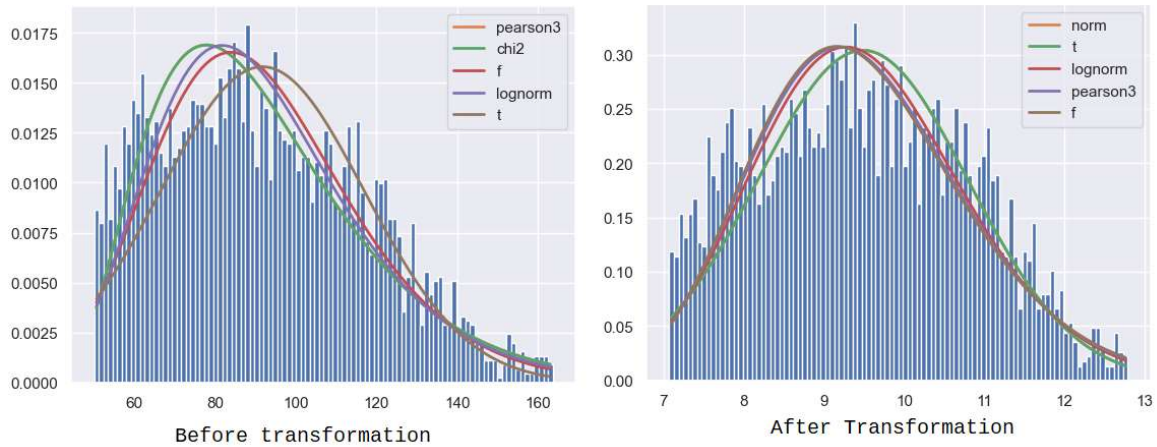
```
f = Fitter(feature, distributions=['norm','uniform','t','pearson3','loguniform', 'lognorm','chi2'])
```

Feature	Distribution
Lifespan	Normal
Genetic	Normal
Exercise	Normal
Smoking	Uniform (with outliers near 0)
Alcohol	Log-norm (but with 2 peaks)
Mass	Chi-squared
Sugar	Chi-Squared

> *A chi-squared distribution constructed by squaring a single standard normal distribution is said to have 1 degree of freedom*
>
> https://en.wikipedia.org/wiki/Chi-squared_distribution

Lets transform mass from a Chi2 distribution to a Normal distribution by applying the square root and quantify the effect on the accuracy of our linear regression model:

Before transformation:



Taking the square root of gives us a (more) normally distributed mass ✅
The effect on our regression model is as follows:

```
Rsquared: 0.9826184086652832 an improvement of: 0.05 % compared to DF4
RMSE: 0.9826184086652832
```

A minor improvement, but not big enough to change from DF4 as data for our model.

IV. Model selection and evaluation

Description of candidate models and algorithms considered

Algorithms: we considered k-means, randomforestregressor and linear regression

Random Forest Regressor
The dataset contains some features which are categorical variables and some which are continuous, Since Decision Trees are used for both regression and classification problems it might produce a better result:

```
rng_forest_regr = RandomForestRegressor(n_estimators=300)
rng_forest_regr.fit(X,y)
```

```
coefficient of determination(R²): 0.9758901067146091
```

Conclusion: random forest decision tree does not produce a better result, so we stick with linear regression.

Model evaluation metrics and criteria used

We use the following metrics

R^2

In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points.

RMSE

The root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model and the values observed.

Train-test-split seed:

```
train_test_split(df, test_size=0.2, random_state=42)
```

Model performance and results

Experiment Tracking Table

train, test = train_test_split(mydf, test_size=0.2, random_state=42)

Experiment #	NaN drop	Dupli drop	Neg drop	object -> float	BMI feature	IQR-clip	IQR-drop	R-squared	RMSE
DF1	x	x	x	x				0.980	1.118
DF2	x	x	x	x	x			0.981	1.076
DF3	x	x	x	x	x	x		0.974	1.199
DF4	x	x	x	x	x		x	0.982	1.054

_Experiment_Tracking_Management.xlsx

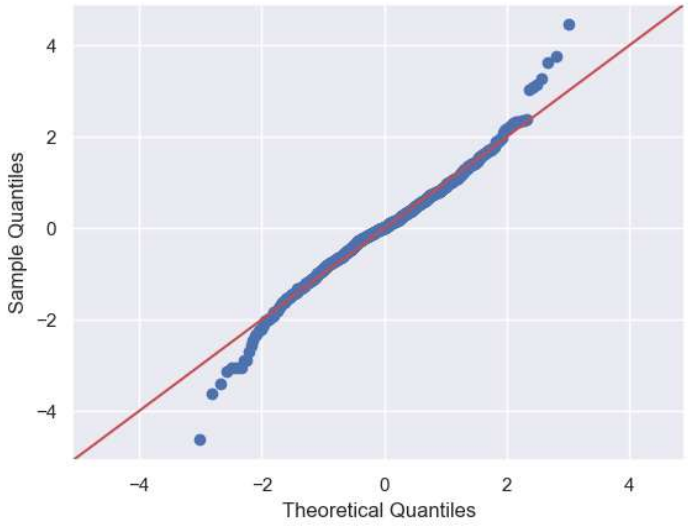
In the end df4 has the best results.
A df6 where we used the root square of mass gives a little better result, but because of the minimal improvement and lack of time df4 is used for the pipeline.

Residual Error

The residual error is the difference between the actual result and the predicted result:

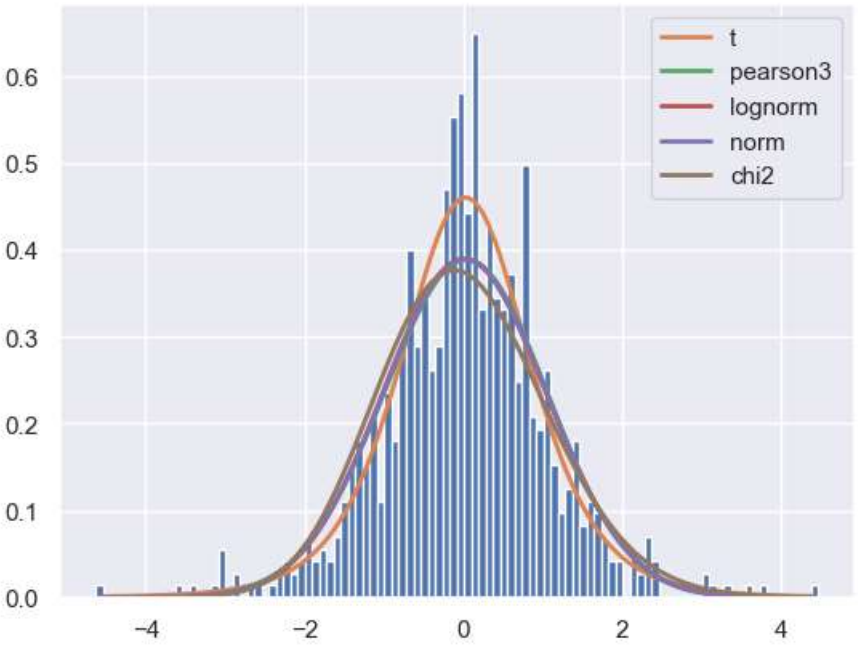
$$\varepsilon = y - \hat{y}$$

The most important assumption of a linear regression model is that *errors are independent and normally distributed.* The following figure shows a Q-Q plot:



If it were a normal distribution it the error would lie on the red line, the error distribution for our results seems to be deviating at both ends, being skewed to the left and to the right. This is a dataset with “fat tails,” meaning that compared to the normal distribution there is more data located at the extremes of the distribution and less data in the center of the distribution.

When we run the fitter library confirms that the error distribution is more like a t-distribution:



This is aligns perfectly with our previous findings since the t-distribution is a closely related distribution to the normal. It is also symmetrical and bell-shaped but it has heavier “tails” (fat) than the normal distribution. However for this analysis it suffices to assume normality (with fat tails)

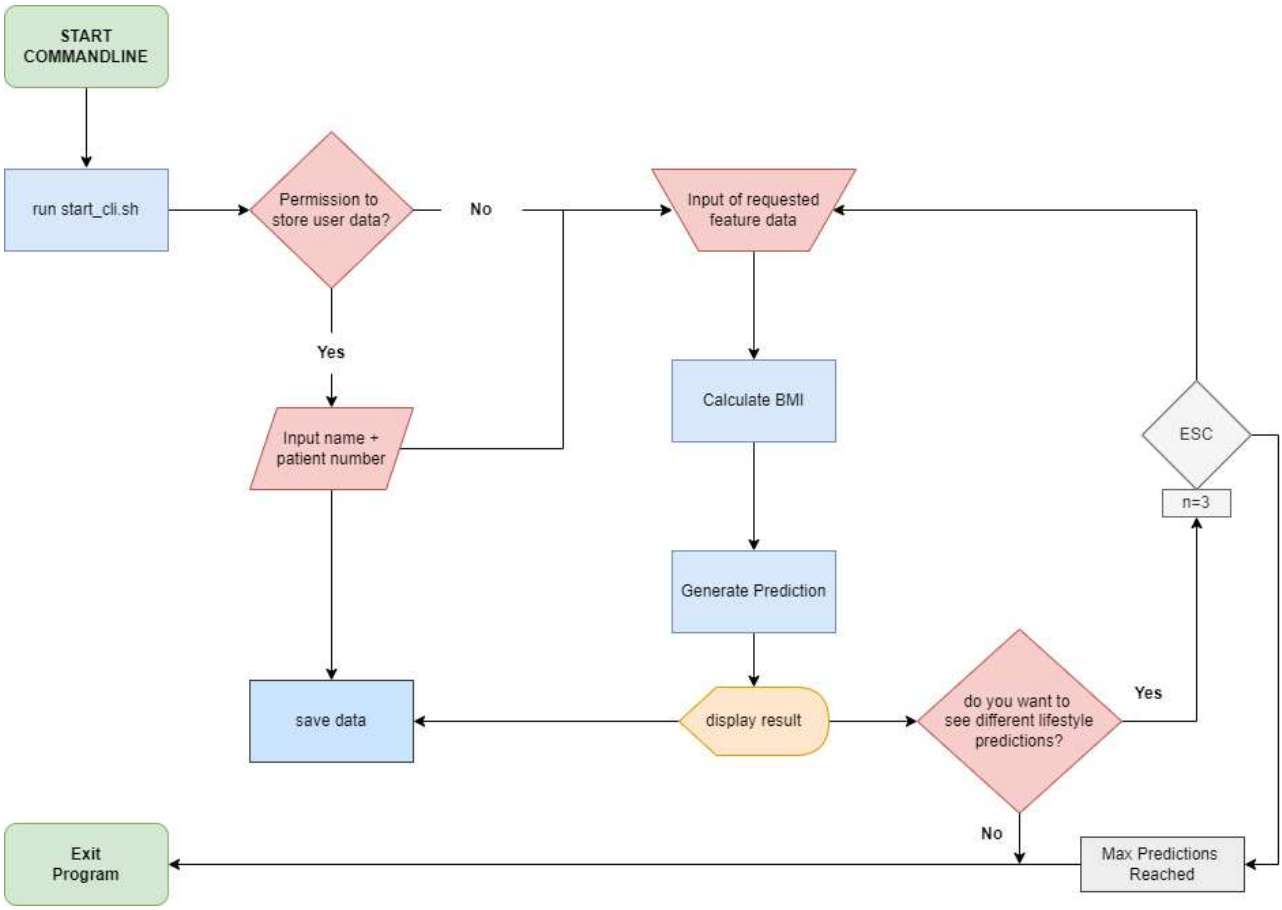
V. User Interface

Integration

The user can run both the pipeline and the cli from bash

CLI

The CLI interface follows the following flowchart:



The doctor can together with the patient input the feature data and show the patient multiple variations (up to 3) on his/her lifestyle and the impact of it on lifestyle.

VI. Discussion and Conclusion

Key take aways and recommendations

Lifespan was heavily influenced by genetics plus the fact that the initial model already scored very high so much so that any other feature ‘optimization’ yielded very little gains.

Further take-away is the importance of statistics both in understanding and knowing when to apply which method in order to get an accurate model.

It is recommended to add more features that would increase reliability and accuracy of the model greatly, such as:

- Stress
- Sense of belonging
- Ethnicity
- Immigration status
- Education
- Postal code (Socioeconomic status of the neighborhood)
- Diabetes
- High blood pressure

Ethics

Let us first establish what we mean when we talk about ethics:

Ethics are a set of principles that guide individuals to make moral and just decisions in their actions towards others and ‘society as a whole’. It involves examining what is right and wrong, and how people should behave in various situations based on moral reasoning.

During this project several main ethical questions were raised:

1. The insurance calculator uses health and genetic status, it can be argued that it is in violation of the 1st article of the constitution of the Netherlands.

2. What do we do with the client input needed for this project. Articles 6 and 9 of the EU regulation **GDPR** implemented as the **AVG** in the Netherlands. States that data regarding health status requires express consent of the subject of the data prior to storage and handling of this data. It also states that without a clear purpose defined the time of gathering.

3. When supplying data has consequences like what is the case for this project with the insurance calculator, people will manipulate the data input, in order to get the most beneficial result for themselves.

4. In the light of the discussion regarding the ‘Toeslagen Affaire’ in the Netherlands the question rises whether or not the model that leads to the results should be open source or not.

5. Keeping the GDPR in mind what was the original purpose of the dataset to be used to train the data and are we allowed to use it?

Herewith is a quick summary of the conclusions resulting from our internal discussions regarding these five points.

1. The insurance premium modifier is for a life insurance policy. Had it been for a health cost insurance policy the genetic component of the calculation could be understood as a violation of article 1 of the constitution of the Netherlands. This is not the case for our model. The law in the Netherlands does permit the use of genetic testing to determine premiums under the conditions and guidelines that autonomy and privacy of the individual is maintained.
2. Articles 6 and 9 of the EU regulation GDPR implemented as the AVG in the Netherlands. States that data regarding health status requires express consent of the subject of the data prior to storage and handling of this data. It also states that without a clear purpose defined the time of gathering. This means that we would obtain retractable informed consent in writing prior to storing, sharing or further manipulating of the data.
3. That persons are not always truthful when supplying information especially when it has consequences for them can be treated as a given.
4. Internal discussion of this point reached the conclusion that whether or not the used model should be shared as open source depends on legal legislation and the commissioner.
5. See point 2.

Our actions regarding these points:

1. Discussing this point and legal inquiry raised no objections. We build and implement the premium calculator as commissioned.
2. This is the reason of our GDPR check in the interface. The retraction of data has not been built (yet) due to time constraints.
3. For the scope of the premium calculator, we have considered this and have mentioned this during our talks with Medical Centrum Randstad. We will mention this in the documentation of the collected data so that future researchers are aware of this bias.
4. No actions within the scope of this project.
5. For the dataset used in this project there was no documentation regarding its origin and purpose. Beyond the scope of this project, we’d look into it and ask the supplier about this.

The ethical questions can be grouped in two clusters.

- The first is what are the legal implications of our actions regarding data. For this we recommend obtaining legal advice and document all decisions.
- The second is what would be the impact of your actions on data, information, and society. This second cluster can be subdivided in choices on a personal and or company level. What is ethical to do in these cases is dependent on an actor’s ethical preferences.

For this project we did our ethical discussions in the final days of the project. For a future project we would advise pushing these discussions to the start of the project because they can have a mayor implementation impact.