Introduction to Machine Learning

Major HW 1

Submitted by: Sofia Blyufer 321128019 and Dor Rozen 318965365

1. See code.
2. In the table below, we can see for each attribute a proposed meaning and a proposed type of attribute.

Table1 - Summary of all features in given data. Includes the name of the feature, description of our understanding of it, and its type.

Features	Meaning	Type
ID	ID of the patient	float
Address	Address of the patient	string
Age Group	The age group of the patient - ages 0-9 is 0, 10-19 is 1 and so on.	*category
BMI	BMI if the patient.	float
Blood Type	Blood group of the patient (4 groups and each has RhD positive and negative).	category
Conversations Per Day	Nr. of conversations had per day – informative of contact level.	float
Current Location	X and Y location coordinates ("GoogleMaps" coordinates) – informative of location in high spread areas.	string (was converted into float)
Date of PCR Test	Date of PCR taken.	datetime
Discipline Score	Score given to show the behavior of the patient - normally 0-10, unless bad behavior shown.	*category
Happiness Score	A score from 1 to 10 of how happy the patient normally is.	*category
Household Expense on Presents	Dollars spent on average on presents in household per month.	float
Household Expense on Social Games	Dollars spent on average on social games in household per month.	float
Household Expense Parking Tickets Per Year	Dollars spent on average on parking tickets in household per month.	float
Job	Occupation of the patient	string
Medical Care Per Year	Times received medical treatment per	float

	year - can show risk group and can indicate higher chance of getting infected at a hospital.	
Nr Cousins	Nr of cousins the patient has - may indicate a large family and thus a lot of contact with other people	*category
PCR_10 PCR_11 PCR_15 PCR_17 PCR_19 PCR_32 PCR_45 PCR_46 PCR_7 PCR_72 PCR_76 PCR_8 PCR_83 PCR_89 PCR_9 PCR_93 PCR_95	PRC is a test performed to detect Covid19 in a patient. The test is run a few times which can be the reason for the different rows. The data is continuous.	float
Self declaration of Illness Form	Symptoms the patient declared to have.	string
Sex	Sex of the patient.	*category
Social Activities Per Day	Time spent on social activities per day - may indicate infection risk.	float
Social Media Per Day	Time spent on social mead per day.	float
Sports Per Day	Time spent on sports per day.	float
Steps Per Year	We presume that the patient needed to grade his number of steps per year in a scale of 1-10. Indication of activity and movement.	*category
Studying Per Day	Time spent studying per day.	float
Virus	Patient's diagnosis	*category
Spread Level	Spread Level of covid19 at patient's area - low medium or high.	*category

Risk	Risk of spread in patient's area – low, medium or high.	*category

*Some of the categorical features weren't converted so. This is due to preferring to have as much numerical data as possible, to improve future modelling.

3. See in the table above.
4. See code.
5. See 2.
6. The strings discussed are "Job", "CurrentLocation" and "Address" and "Self_declaration_of_Illness_Form".

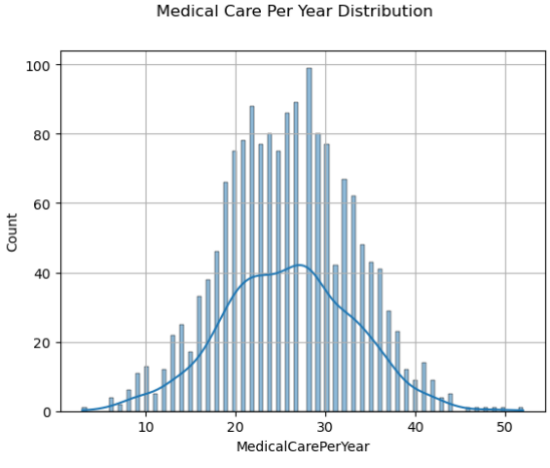
   Job – we have decided to not extract any information from this feature for two reasons. The first, there is more than 25% missing data for this attribute. The second, there are too many unique values (620 out of 3000) and is not informative enough.

   Current Location – location coordinates can be important to identify risk area. As the data already numerical, which is easiest for further analysis, we have decided to split the data into two attributes – X and Y (x and y coordinates separately).

   Address – we have seen fit to drop this attribute, as the location information is already numerically convenient in "CurrentLocation". Address is an ID feature and therefore might only harm our modelling. Using the same reason, we have decided to drop "ID" now.

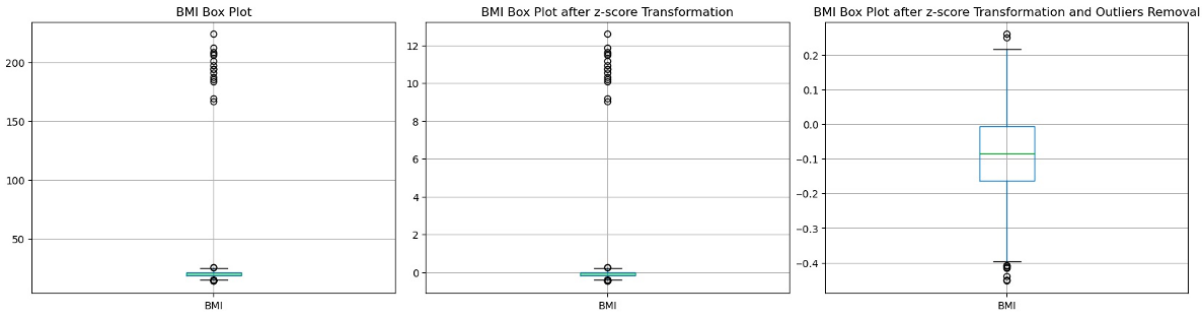
   Self-Declaration of Illness Form (symptoms) – we decided to split into the 14 different types of features, columns of 1 and 0 (symptom exists or not). We are aware it will be a lot of new information with risk of overfitting and long running times and will later find the most relevant features.
7. As previously mentioned, job, address and ID were deleted altogether. Additionally, we have dropped "PCR_11" and "PCR_15" as they both have above 80% missing values, and imputation would be too inaccurate in such a case. The rest of the features were chosen to be imputed as there is enough data to draw conclusions based on the its distribution.

   For example, the histogram below is of the attribute "MedicalCarePerYear". It can be approximated well to a normal distribution. Therefore, for such features we think imputation is relatively straightforward.

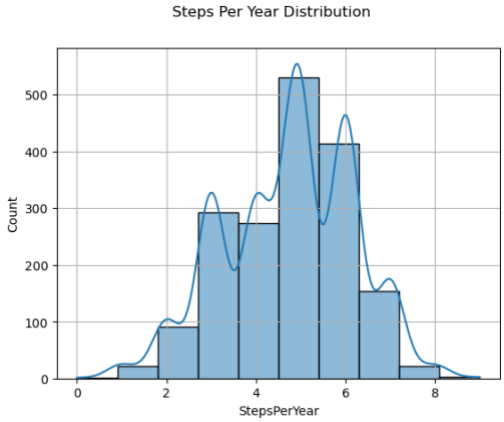


For other features that had enough data we also decided to impute, using other methods, as will be described below.

8. We have handled our features according to 3 types:
   For float type features we replaced the missing values with the mean.
   If the feature was an integer then (supposedly categorical), we replaced the missing values with the rounded mean.
   And for the categorical string features (such as blood type), we replaced the missing values by the most frequent category.
9. See code.
10. Let us show BMI as an example. First, we have transformed all values into z-score values. Then, we removed any rows that contained z-score below or above 3 (in other words $Mean \pm 3\sigma$). The process can be seen in the plots below, that show (from left to right): BMI raw train data box plot, data transformed into z-scores and data after outliers removal. For all numeric data the process was the same.



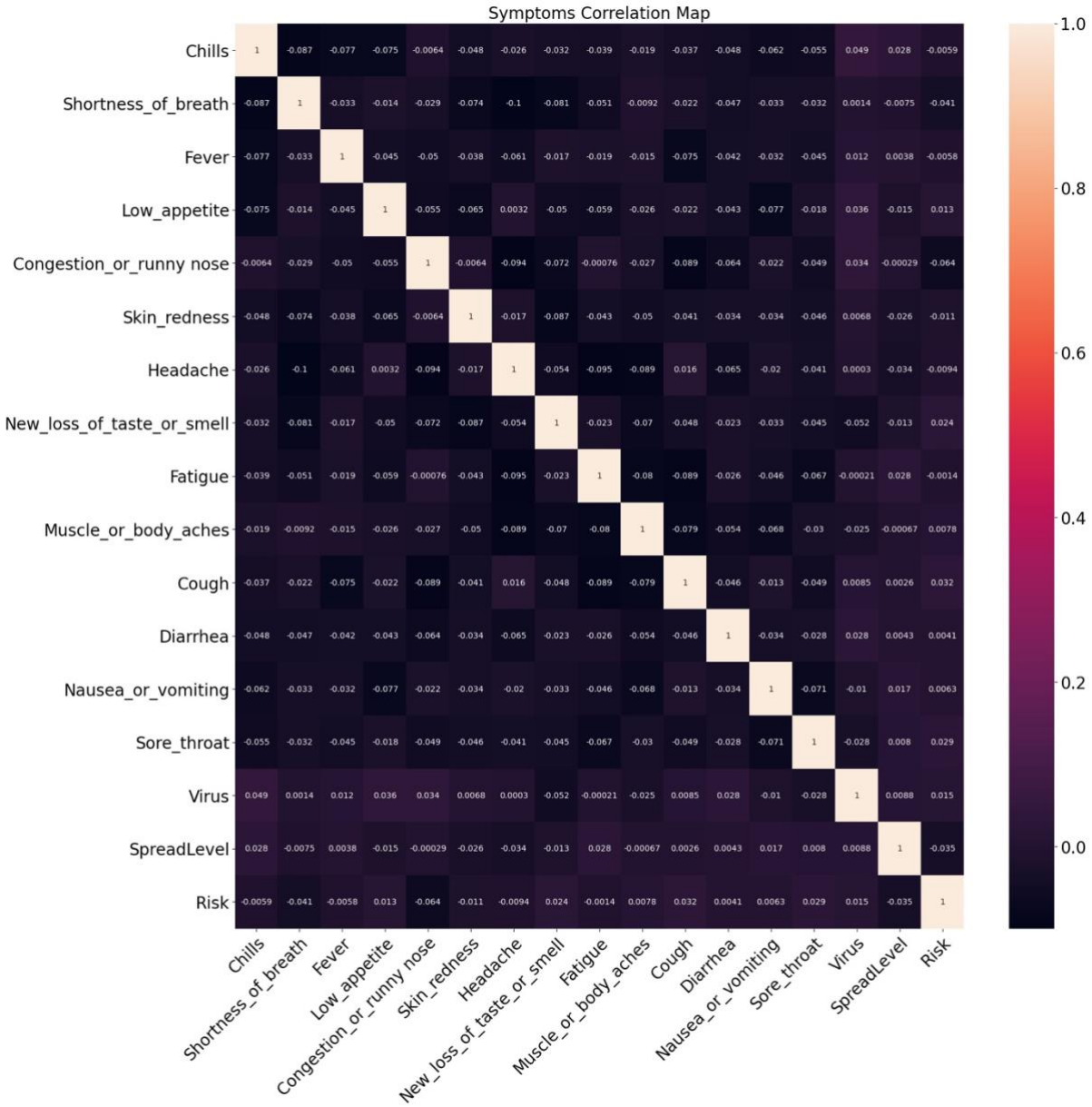
11. Step per year was normalized like all other numeric data, into z-score. The graph below represents the distribution of "StepsPerYear" and confirms a z-score transformation will be convenient (the mean is relatively representative).



12. For numeric data, we used the advantage of z-score transformation, which can normalize any type of distribution and is sensitive to outliers, unlike min/max. Additionally, for better learning, we decided to convert all categorical data to numeric

data. Henceforth, “Virus” became 0 for anything that wasn’t “covid” and 1 for “covid”. We kept a code line for another splitting, 0 for not sick, 1 for covid and 2 for other diseases, in case it might have us with learning in future assignments.
Other data, such as “Risk”, “SpeadLevel” and “BloodType” was also converted numerically. We checked that all possible categories exist in train (that we aren’t missing categories that were split only into test).

13. Before analysing all correlations, we decided to choose the most relevant symptoms amongst the 14 given. The correlation map below shows correlations for all 14 symptoms, as well our three target labels.



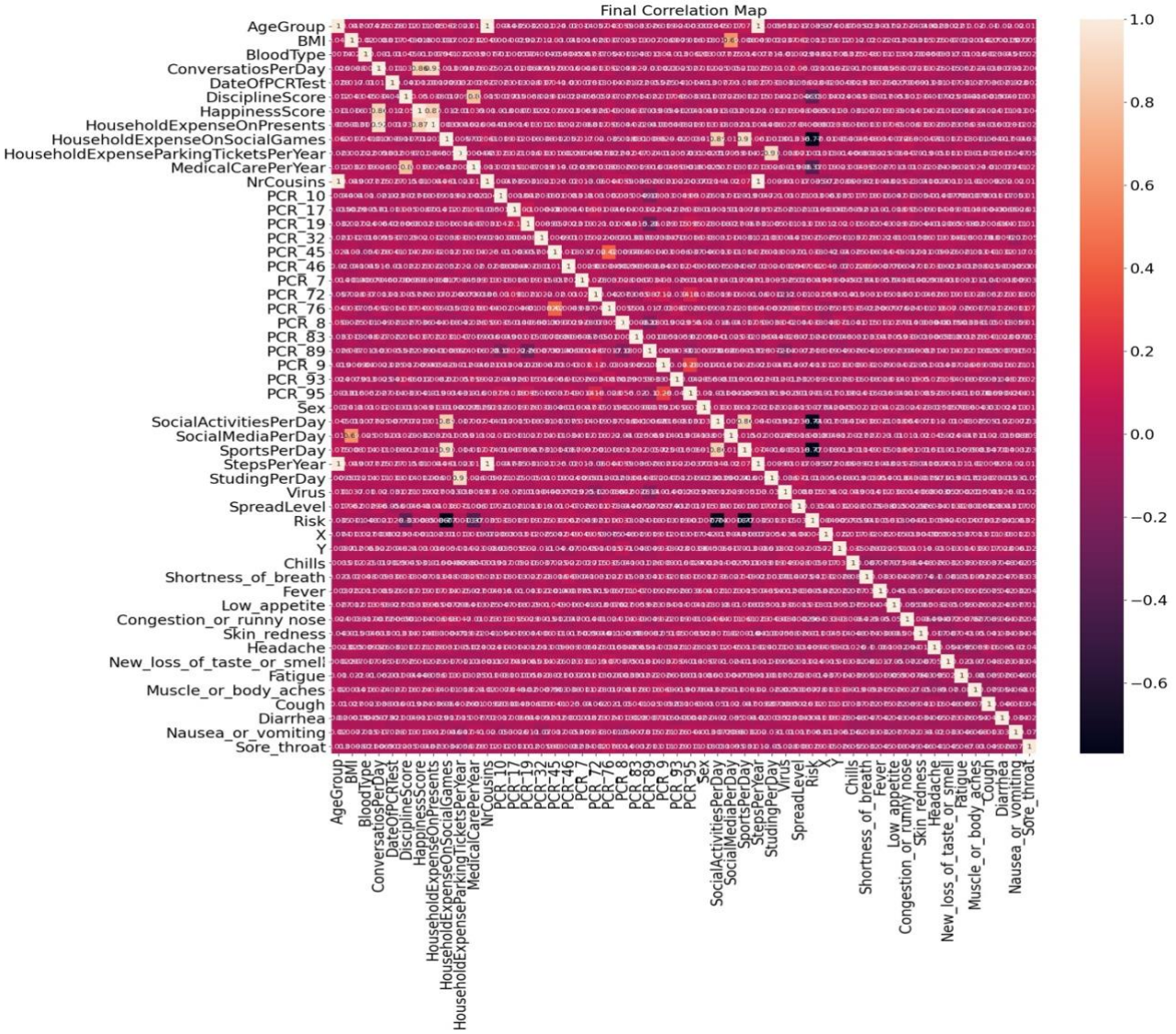
As we can see, no high correlations were generated. We decided to perform a decision tree on these features, to see if new information would arise there. The importance values can be found below. We chose “New_loss_of_taste_or_smell”,

‘Nausea_or_vomiting” and “Low_appetite”.

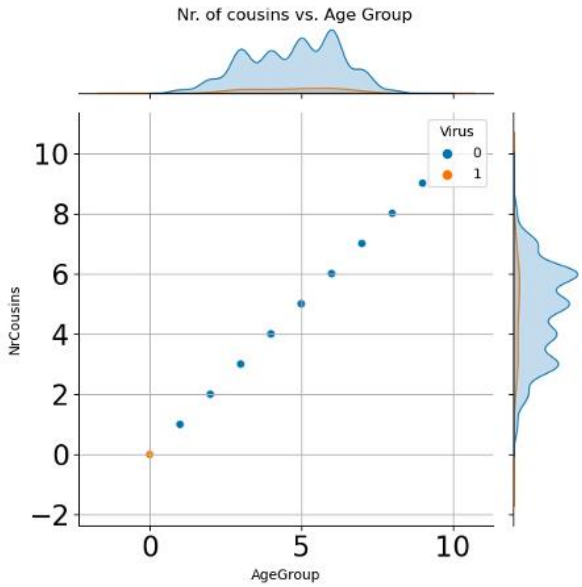
Table2 - The symptoms of train after splitting by unique. Decision tree importance scores as well as their sum is shown.

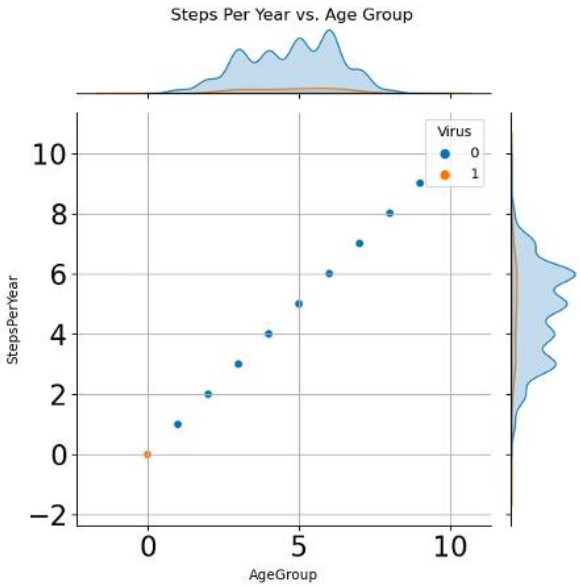
final score:	ID3 risk score	ID3 spread level score	ID3 virus score	features
0.27062	0.1255	0.13372	0.0114	**New_loss_of_taste_or_smell**
0.15505	0.04899	0.0112	0.09486	**Headache**
0.2791	0.1314	0.11	0.0377	**Nausea_or_vomiting**
0.241	0.06509	0.107	0.06891	**Cough**
0.23336	0.10016	0.062	0.0712	**Fatigue**
0.15555	0.09855	0.038	0.019	**Chills**
0.15827	0.00727	0.007	0.144	**Fever**
0.24743	0.09843	0.074	0.075	**Shortness_of_breath**
0.13406	0.03487	0.03029	0.0689	**Congestion_or_runny nose**
0.26111	0.11378	0.11033	0.037	**Low_appetite**
0.22919	0.06027	0.08982	0.0791	**Muscle_or_body_aches**
0.12379	0.01692	0.05157	0.0553	**Diarrhea**
0.25364	0.04919	0.09645	0.108	**Sore_throat**
0.25151	0.04955	0.07502	0.12694	**Skin_redness**

The correlation table of the features left is presented below, including the chosen symptoms.



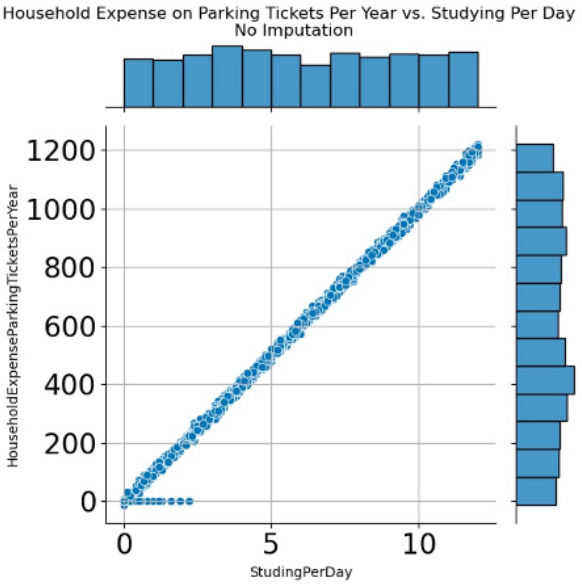
To understand the correlations, we plotted histograms and plots of highly correlated features. The pairs we decided to check were those with correlation 0.8 and above. For two pairs we've received perfect correlation: steps per year (not normalized) vs. age group and number of cousins vs. age group. The plots below confirm these numbers.

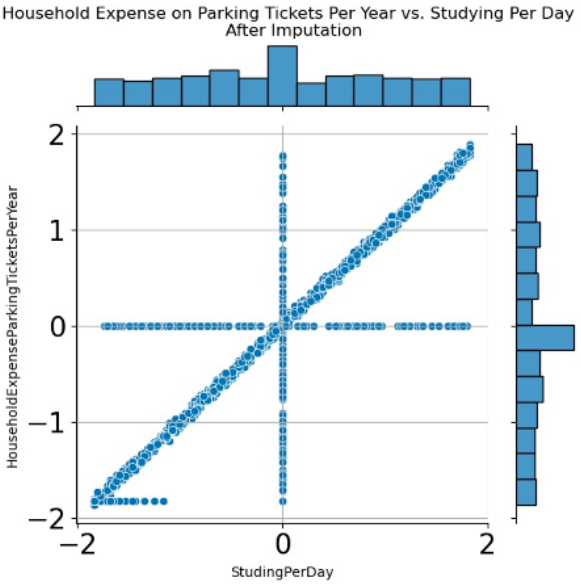




Therefore, by keeping the feature of age group, number of cousins and steps per year do not add any new data. From the plots we can also see that there is no need of manipulation and taking a combination of the data (look the same).

Before showing the next features, it should be noted that as previously explained, missing data was filled with mean. As a result, peaks in the mean are observed. Consequently, those effects slightly "ruin" correlations by creating a "cross". For example, below is a plot of household expenses of parking tickets per year vs. studying per day. On the left is the original data, and on the right is the data normalized to z-score values, after filling missing data and outlier removal. For that reason, we also look at the plots before filling the data. Nevertheless, the correlation is worse after adding the "cross", and therefore we are only being stricter. The correlation of the pair after filling the missing data is 0.91.





The other correlation groups are:

- {Conversations Per Day, Happiness Score, Household Expense On Presents}
- {Discipline Score, Medical Care Per Year}
- {Household Expense On Social Games, Social Activities Per Day, Sports Per Day}

14. In addition to the correlations explained above, we have also used the following methods for feature selection: decision tree (as wrapper method) and manual inspection of scatter plots as univariate and bivariate methods.
As we have 3 target labels, we performed 3 decision trees, one for each target. Eventually, the importance arrays were added into the table below to help choose the most important features. The best indicator for the decision tree was the sum of importance for each feature, as we need the same features for modelling for all 3 targets.

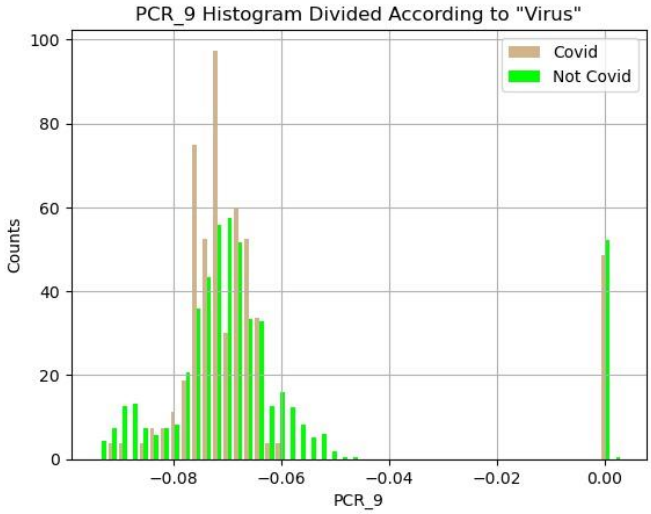
Table3 - Decision tree scores for each target: virus, spread level and risk. Final score is the sum of all 3.

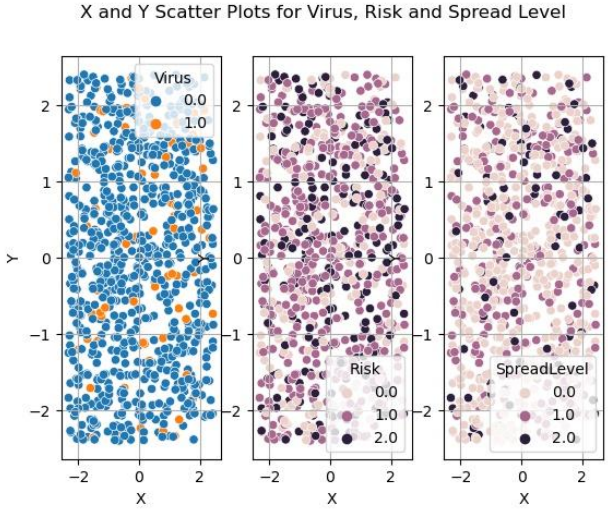
Features	Final Score	ID3 risk score	ID3 spread level score	ID3 virus score
Age Group	0	0	0	0
BMI	0.43451	0.00391	0.422	0.0086
Blood Type	0.00628	0	0.00628	0
Conversations Per Day	0.1134	0.0024	0.111	0
Date of PCR Test	0.05557	0.00467	0.0019	0.049
Discipline Score	0.0355	0.0317	0.0038	0
Happiness Score	0.0336	0	0.022	0.0116
Household Expense on Presents	0.31837	0.00337	0.3065	0.0085
Household Expense on Social Games	0.38962	0.378	0.0019	0.00972
Household Expense Parking Tickets Per Year	0.0097	0	0.0097	0
Medical Care Per Year	0.264	0.258	0.006	0
Nr Cousins	0.011	0	0	0.011
PCR_10	0.0301	0	0.0072	0.0229
PCR_17	0.03284	0.004	0.00134	0.0275
PCR_19	0.03963	0.004	0.009	0.02663
PCR_32	0	0	0	0
PCR_45	0.01468	0.00653	0.00315	0.005
PCR_46	0.00457	0.00457	0	0

PCR _ 7	0.21759	0.00449	0	0.2131
PCR_72	0.18329	0.00119	0	0.1821
PCR_76	0.00564	0	0.00564	0
PCR_8	0.05084	0.0029	0.00194	0.046
PCR_83	0.0029	0.00156	0.00134	0
PCR_89	0.200958	0.0008	0.003158	0.197
PCR_9	0.04977	0.001	0.00667	0.0421
PCR_93	0.02567	0.008	0.00467	0.013
PCR_95	0.0628	0.00275	0.00235	0.0577
Sex	0.00097	0	0.00097	0
Social Activities Per Day	0.05871	0.05578	0.00293	0
Social Media Per Day	0.0529	0.0088	0.038	0.0061
Sports Per Day	0.21093	0.19433	0.0036	0.013
Steps Per Year	0.01688	0.00298	0	0.0139
Studying Per Day	0.017457	0.001957	0.007	0.0085
X	0.015	0	0	0.015
Y	0.030498	0.010538	0.00856	0.0114
Nausea or vomiting	0	0	0	0
New loss of taste or smell	0.002	0	0.002	0
Low appetite	0	0	0	0

By eliminating correlated features, choosing them according to highest importance, we were left with features that weren't redundant correlation-wise and were not highly important on decision trees. Those we decided to inspect manually using univariate and bivariate methods. Let us first notice that high correlation of risk and some features was found but has already been noticed by the decision tree importance and therefore will not require any further treatment.

The plots below show: on the left, PCR_9 histogram with covid and not covid divided, on the right, scatter plots of X and Y with all different target labels. We saw areas where it could be informative, for example areas containing only "not covid" or more "high risk" dots. Therefore, we kept those features. We tried similar plots for other features, but those showed not enough correlation to be kept in the final table.


PCR_9 Histogram Divided According to "Virus"
Covid
Not Covid
Counts
PCR_9
100
80
60
40
20
0
−0.08
−0.06
−0.04
−0.02
0.00


X and Y Scatter Plots for Virus, Risk and Spread Level
Virus
0.0
1.0
Risk
0.0
1.0
2.0
SpreadLevel
0.0
1.0
2.0
Y
X
X
X

15. The final table:

Features	Kept	Reason
ID	no	No extra information, ID feature
Address	no	No extra information, ID feature
Age Group	no	Highly correlated to data that's taken
BMI	yes	High score on decision tree.
Blood Type	no	Extremely low score on decision tree. Univariate analysis supports no correlation.
Conversations Per Day	no	Highly correlated to data that's taken
Current Location – X and Y separately	yes, both	Showed some relevance in bivariate analysis
Date of PCR Test	yes	High score on decision tree.
Discipline Score	no	Highly correlated to data that's taken
Happiness Score	no	Highly correlated to data that's taken
Household Expense on Presents	yes	High score on decision tree.
Household Expense on Social Games	yes	High score on decision tree.
Household Expense Parking Tickets Per Year	no	Highly correlated to data that's taken
Job	no	Too much missing data and too many unique values
Medical Care Per Year	yes	High score on decision tree.
Nr Cousins	no	Highly correlated to data that's taken
PCR_10	yes	Previous knowledge + decision tree.
PCR_11	no	Too much missing data.
PCR_15	no	Too much missing data.
PCR_17	yes	Previous knowledge + decision tree.
PCR_19	yes	Previous knowledge + decision tree.
PCR_32	no	Extremely low score on decision tree.
PCR_45	yes	Previous knowledge + decision tree.
PCR_46	no	Extremely low score on decision tree.
PCR_7	yes	High score on decision tree.
PCR_72	yes	High score on decision tree.
PCR_76	no	Extremely low score on decision tree.
PCR_8	yes	Previous knowledge + decision tree.

PCR_83	no	Extremely low score on decision tree.
PCR_89	yes	High score on decision tree.
PCR_9	yes	Showed some relevance in univariate analysis
PCR_93	yes	Previous knowledge + decision tree.
PCR_95	yes	High score on decision tree.
Self declaration of Illness Form	no	Even the best features showed low correlations and low scores in decision trees.
Sex	yes	Pervious knowledge on covid19 shows relevance.
Social Activities Per Day	no	Highly correlated to data that's taken
Social Media Per Day	yes	High score on decision tree.
Sports Per Day	no	Highly correlated to data that's taken
Steps Per Year	yes	Highest score on decision tree amongst its correlated features group.
Studying Per Day	yes	Highest score on decision tree amongst its correlated features group.
Virus	yes	Target label
Spread Level	yes	Target label
Risk	yes	Target Label

16. See code.