**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.

Table 1 - Summary of all features in given data. Includes the name of the feature, description of our understanding of it, and its type.

|  |  |  |
| --- | --- | --- |
| Type | Meaning | Features |
| float | ID of the patient | ID |
| string | Address of the patient | Address |
| \*category | The age group of the patient - ages 0-9 is 0, 10-19 is 1 and so on. | Age Group |
| float | BMI if the patient. | BMI |
| category | Blood group of the patient (4 groups and each has RhD positive and negative). | Blood Type |
| float | Nr. of conversations had per day – informative of contact level. | Conversations Per Day |
| string (was converted into float) | X and Y location coordinates (“GoogleMaps” coordinates) – informative of location in high spread areas. | Current Location |
| datetime | Date of PCR taken. | Date of PCR Test |
| \*category | Score given to show the behavior of the patient - normally 0-10, unless bad behavior shown. | Discipline Score |
| \*category | A score from 1 to 10 of how happy the patient normally is. | Happiness Score |
| float | Dollars spent on average on presents in household per month. | Household Expense on Presents |
| float | Dollars spent on average on social games in household per month. | Household Expense on Social Games |
| float | Dollars spent on average on parking tickets in household per month. | Household Expense Parking Tickets Per Year |
| string | Occupation of the patient | Job |
| float | Times received medical treatment per year - can show risk group and can indicate higher chance of getting infected at a hospital. | Medical Care Per Year |
| \*category | Nr of cousins the patient has - may indicate a large family and thus a lot of contact with other people | Nr Cousins |
| float | 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. | 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 |
| string | Symptoms the patient declared to have. | Self declaration of Illness Form |
| category | Sex of the patient. | Sex |
| float | Time spent on social activities per day - may indicate infection risk. | Social Activities Per Day |
| float | Time spent on social mead per day. | Social Media Per Day |
| float | Time spent on sports per day. | Sports Per Day |
| \*category | 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. | Steps Per Year |
| float | Time spent studying per day. | Studying Per Day |
| category | Patient's diagnosis | Virus |
| category | Spread Level of covid19 at patient's area - low medium or high. | Spread Level |
| category | Risk of spread in patient’s area – low, medium or high. | Risk |

\*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.

1. See in the table above.
2. See code.
3. See 2.
4. 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.
5. Chart, histogram

   Description automatically generatedAs 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.
6. 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.

1. See code.
2. 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 ). 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.  
   Chart, scatter chart

   Description automatically generated
3. 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).   
   Chart, histogram

   Description automatically generated
4. 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).
5. 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.   
   Graphical user interface, chart, surface chart

   Description automatically generated  
   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”.

Table 2 - The symptoms of train after splitting by unique. Decision tree importance scores as well as their sum is shown.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| features | ID3 virus score | ID3 spread level score | ID3 risk score | final score: |
| New\_loss\_of\_taste\_or\_smell | 0.0114 | 0.13372 | 0.1255 | 0.27062 |
| Headache | 0.09486 | 0.0112 | 0.04899 | 0.15505 |
| Nausea\_or\_vomiting | 0.0377 | 0.11 | 0.1314 | 0.2791 |
| Cough | 0.06891 | 0.107 | 0.06509 | 0.241 |
| Fatigue | 0.0712 | 0.062 | 0.10016 | 0.23336 |
| Chills | 0.019 | 0.038 | 0.09855 | 0.15555 |
| Fever | 0.144 | 0.007 | 0.00727 | 0.15827 |
| Shortness\_of\_breath | 0.075 | 0.074 | 0.09843 | 0.24743 |
| Congestion\_or\_runny nose | 0.0689 | 0.03029 | 0.03487 | 0.13406 |
| Low\_appetite | 0.037 | 0.11033 | 0.11378 | 0.26111 |
| Muscle\_or\_body\_aches | 0.0791 | 0.08982 | 0.06027 | 0.22919 |
| Diarrhea | 0.0553 | 0.05157 | 0.01692 | 0.12379 |
| Sore\_throat | 0.108 | 0.09645 | 0.04919 | 0.25364 |
| Skin\_redness | 0.12694 | 0.07502 | 0.04955 | 0.25151 |

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

Chart

Description automatically generated with medium confidence  
  
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.

Chart, scatter chart

Description automatically generated

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.

Chart, line chart

Description automatically generated

The other correlation groups are:

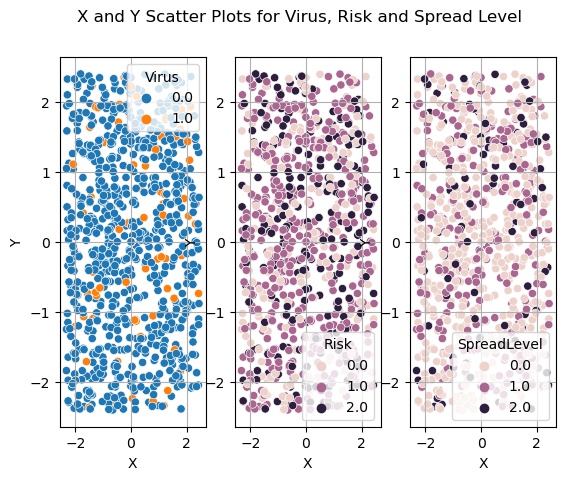
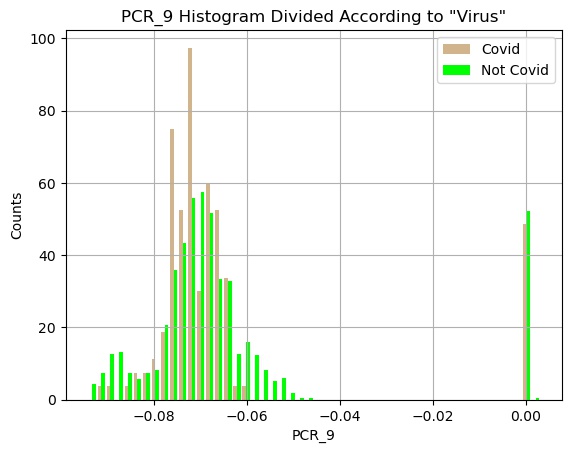
* {Conversations Per Day, Happiness Score, Household Expense On Presents}
* {Discipline Score, Medical Care Per Year}
* {Household Expense On Social Games, Social Activities Per Day, Sports Per Day}

1. 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.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID3 virus score | ID3 spread level score | ID3 risk score | Final Score | Features |
| 0 | 0 | 0 | 0 | Age Group |
| 0.0086 | 0.422 | 0.00391 | 0.43451 | BMI |
| 0 | 0.00628 | 0 | 0.00628 | Blood Type |
| 0 | 0.111 | 0.0024 | 0.1134 | Conversations Per Day |
| 0.049 | 0.0019 | 0.00467 | 0.05557 | Date of PCR Test |
| 0 | 0.0038 | 0.0317 | 0.0355 | Discipline Score |
| 0.0116 | 0.022 | 0 | 0.0336 | Happiness Score |
| 0.0085 | 0.3065 | 0.00337 | 0.31837 | Household Expense on Presents |
| 0.00972 | 0.0019 | 0.378 | 0.38962 | Household Expense on Social Games |
| 0 | 0.0097 | 0 | 0.0097 | Household Expense Parking Tickets Per Year |
| 0 | 0.006 | 0.258 | 0.264 | Medical Care Per Year |
| 0.011 | 0 | 0 | 0.011 | Nr Cousins |
| 0.0229 | 0.0072 | 0 | 0.0301 | PCR\_10 |
| 0.0275 | 0.00134 | 0.004 | 0.03284 | PCR\_17 |
| 0.02663 | 0.009 | 0.004 | 0.03963 | PCR\_19 |
| 0 | 0 | 0 | 0 | PCR\_32 |
| 0.005 | 0.00315 | 0.00653 | 0.01468 | PCR\_45 |
| 0 | 0 | 0.00457 | 0.00457 | PCR\_46 |
| 0.2131 | 0 | 0.00449 | 0.21759 | PCR \_ 7 |
| 0.1821 | 0 | 0.00119 | 0.18329 | PCR\_72 |
| 0 | 0.00564 | 0 | 0.00564 | PCR\_76 |
| 0.046 | 0.00194 | 0.0029 | 0.05084 | PCR\_8 |
| 0 | 0.00134 | 0.00156 | 0.0029 | PCR\_83 |
| 0.197 | 0.003158 | 0.0008 | 0.200958 | PCR\_89 |
| 0.0421 | 0.00667 | 0.001 | 0.04977 | PCR\_9 |
| 0.013 | 0.00467 | 0.008 | 0.02567 | PCR\_93 |
| 0.0577 | 0.00235 | 0.00275 | 0.0628 | PCR\_95 |
| 0 | 0.00097 | 0 | 0.00097 | Sex |
| 0 | 0.00293 | 0.05578 | 0.05871 | Social Activities Per Day |
| 0.0061 | 0.038 | 0.0088 | 0.0529 | Social Media Per Day |
| 0.013 | 0.0036 | 0.19433 | 0.21093 | Sports Per Day |
| 0.0139 | 0 | 0.00298 | 0.01688 | Steps Per Year |
| 0.0085 | 0.007 | 0.001957 | 0.017457 | Studying Per Day |
| 0.015 | 0 | 0 | 0.015 | X |
| 0.0114 | 0.00856 | 0.010538 | 0.030498 | Y |
| 0 | 0 | 0 | 0 | Nausea or vomiting |
| 0 | 0.002 | 0 | 0.002 | New loss of taste or smell |
| 0 | 0 | 0 | 0 | Low appetite |

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.



1. The final table:

|  |  |  |
| --- | --- | --- |
| Reason | Kept | Features |
| No extra information, ID feature | no | ID |
| No extra information, ID feature | no | Address |
| Highly correlated to data that’s taken | no | Age Group |
| High score on decision tree. | yes | BMI |
|  |  | Blood Type |
| Highly correlated to data that’s taken | no | Conversations Per Day |
| Showed some relevance in bivariate analysis | yes, both | Current Location – X and Y separately |
|  |  | Date of PCR Test |
| Highly correlated to data that’s taken | no | Discipline Score |
| Highly correlated to data that’s taken | no | Happiness Score |
| High score on decision tree. | yes | Household Expense on Presents |
| High score on decision tree. | yes | Household Expense on Social Games |
|  | no | Household Expense Parking Tickets Per Year |
| Too much missing data and too many unique values | no | Job |
| High score on decision tree. | yes | Medical Care Per Year |
| Highly correlated to data that’s taken | no | Nr Cousins |
|  |  | PCR\_10 |
| Too much missing data. | no | PCR\_11 |
| Too much missing data. | no | PCR\_15 |
|  |  | PCR\_17 |
|  |  | PCR\_19 |
|  |  | PCR\_32 |
|  |  | PCR\_45 |
|  |  | PCR\_46 |
| High score on decision tree. | yes | PCR\_7 |
| High score on decision tree. | yes | PCR\_72 |
|  |  | PCR\_76 |
|  |  | PCR\_8 |
|  |  | PCR\_83 |
| High score on decision tree. | yes | PCR\_89 |
| Showed some relevance in univariate analysis | yes | PCR\_9 |
|  |  | PCR\_93 |
|  |  | PCR\_95 |
| Even the best features showed low correlations and low scores in decision trees. | no | Self declaration of Illness Form |
|  |  | Sex |
| Highly correlated to data that’s taken | no | Social Activities Per Day |
|  |  | Social Media Per Day |
| Highly correlated to data that’s taken | no | Sports Per Day |
| Highest score on decision tree amongst its correlated features group. | yes | Steps Per Year |
|  |  | Studying Per Day |
| Target label | yes | Virus |
| Target label | yes | Spread Level |
| Target Label | yes | Risk |

1. See code.