# Feature Extraction Analysis

Some of the features have anomalies that skew the results greatly and therefore we truncated these features.

Looking at the bevaioral/physiological features:

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Time on Social Media and Discipline Score have major outliers so those were truncated.

The bmi score had data that was anomalous because it was physically impossible because bmi = weight/ (height)^2. Therefore, any bmi above 45 was truncated.

We changed the classification to multi label with labels for risk, type of disease, and whether the person spreads the disease. This is important because some features like Syndrome Class indicate whether a person spreads or doesn’t but isn’t informative about the other labels.

We observed the location to check whether it can give any reliable information about the type of disease. Maybe if people are congregated closely to each other then there is an outbreak. Yet observing the data there isn’t a clear case of that:

Cases of Flue are in black and white not-flue:

Chart, scatter chart

Description automatically generated

Cases of covid in black and non-covid in white:

Chart, scatter chart

Description automatically generated

Graphs of other diseases are similar and show no specific area or zones of outbreak. Therefore the location feature is redundant.

Also, the likelihood of spread or risk are not related to the location of the person, but rather to physiological, behavioral traits.

From general knowledge, the address, patient ID, date of test, number of cousins are redundant as well. Also, the professions have too many types in the dataset and from human knowledge are irrelevant.

For scaling we tried to use Quantile Transformer from sklearn but because it truncated too much of the data the results were flawed. Eventually we used standard scaling of (x- (avg))/(std. deviation).

We used PCA, filter algorithm , on the pcrResults mainly because they were they had correlation among them. Then we also used Sequential Backward Selection with the data from the PCA, and we implemented by ourselves the SBS , with k-nearest neighbors’ classifier as scoring function

Plotting SyndromeClass vs. the TestResultsCode we observed the following:

Chart

Description automatically generated

Where the right-hand side is the TestResultsCode and the top side are the syndrome classes. We saw that non-spreading people mostly didn’t exhibit syndromes 1 and 4. Then we chose to keep the feature.

We chose to remove AvgHouseholdExpenseOnPresents, AvgHouseholdExpenseOnSocialGames, AvgHouseholdExpenseParkingTicketsPerYear, AvgMinSportsPerDay, AvgTimeOnSocialMedia and AvgTimeOnStuding.

By Occam’s razor principle the learning will be more efficient excluding them, because the simpler model that takes into account the most relevant features only will likely be better performing. They are immaterial to communicable diseases, because they focus on the household and activities that don’t pose much risk to catching diseases.

We used Mutual Information to assess SelfDeclarationOfIllnessForm and the occurrence of diseases. We

found very low mutual information correlation between a syndrome reported and one of the diseases in the test results, but there was some correlation if the person was spreading. as seen :

A picture containing chart

Description automatically generated