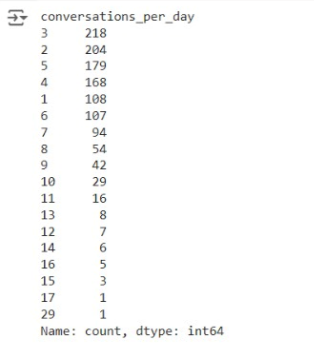
**Data Exploration and Preparation**

מגישים:

ניר הרשקוביץ 206431496

יובל רבינוביץ' 207765918

1. 1250 rows and 25 colunms.
2. 

We think this feature refers to the amount of different face to face conversations the patient made per day on average rounded.

This feature is ordinal because the values are natural integers, so it’s not continuous, and are naturally ordered.

|  |  |  |
| --- | --- | --- |
| Feature Name | Description | Type |
| Patient\_ID | The ID of the patient | Ordinal |
| Age | The age of the patient | Ordinal |
| Sex | The gender of the patient | Categorical |
| Weight | The weight of the patient | Continuous |
| Blood\_type | The blood type of the patient | Categorical |
| Current\_location | The coordinates of the patient location | Other |
| Num\_of\_siblings | The number of siblings the patient has | Ordinal |
| Happines\_score | Happiness of the patient according to a happiness test score | Ordinal |
| Household\_income | The income of the household of the patient | Continuous |
| Conversations\_per\_day | the amount of different face to face conversations the patient made per day on average rounded | Ordinal |
| Sugar\_level | The sugar level in the blood of the patient | Ordinal |
| Sport\_activity | How physically active the patient is | Ordinal |
| Pcr\_date | The date of the first Pcr test that the patient did | Ordinal |
| PCR\_01 | Part 1 of the PCR test | Continuous |
| PCR\_02 | Part 2 of the PCR test | Continuous |
| PCR\_03 | Part 3 of the PCR test | Continuous |
| PCR\_04 | Part 4 of the PCR test | Continuous |
| PCR\_05 | Part 5 of the PCR test | Continuous |
| PCR\_06 | Part 6 of the PCR test | Continuous |
| PCR\_07 | Part 7 of the PCR test | Continuous |
| PCR\_08 | Part 8 of the PCR test | Continuous |
| PCR\_09 | Part 9 of the PCR test | Continuous |
| PCR\_10 | Part 10 of the PCR test | Continuous |
|  |  |  |

1. Its important to use the same split for training and test sets, because we want to train the model only on the train set and validate it only on the test set. In that way we will ensure that the model is not fitted to the test set in any way.
2. Train has 87 missing values, all in “household\_income” feature.

Test has 22 missing values, all in “household\_income” feature.

1. A graph of income and household income

   Description automatically generated

The households that make 90-100 in household income can be considered outliers because they are far from most households by a big difference.

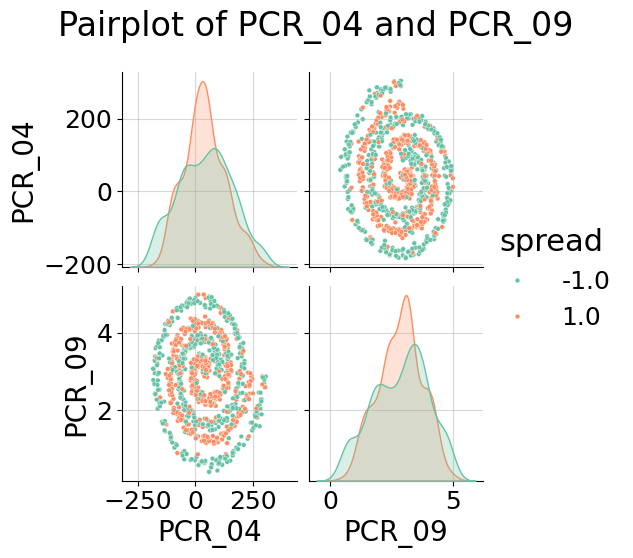
1. Household income mean: 3.64

Household income median: 0.7

The difference stems from the fact that most people have a low income, mostly between 0 to 3, so half of the people make less than 0.7. But there are a few people making much higher income, which we mentioned as possible outliers, which increases the mean value.

We think that the better method is filling with the median value because most of the patients have an income between 0 and 1, so 0.7 is a more reasonable value. There are no patients with income be

1. The pair of features useful for predicting the spread is {PCR\_04, PCR\_09} because it’s most separable pair of data. We can see that the data is distributed in two spirals who are pretty separable.



1. First we calculate the distances of m training datapoints and our single test data point. This takes for each training data point, so for m data points it will take . After that we use argpartition to find the indices of the k closest data points, which runs in time complexity . Getting all the labels of the indices from argpartition will take . The final step is to make the predictions. Here we take the mean of the k labels which will take and than take the sign of the mean . We get the following time complexity . The equality comes from the fact that

there can’t be more neighbors than training datapoints.

1. A graph of a spiral of red and blue squares

   Description automatically generated

The train accuracy was:

The test accuracy was:

1. A diagram of a spiral

   Description automatically generated

Train accuracy was:

Test accuracy was:

We can see by comparing to the figure at Q10 that normalizing the features smoothened the decision regions of the model, which led to better test accuracy. In our example PCR\_04 was on a much bigger scale than PCR\_09. Because of that PCR\_09 got more weight according to the kNN model, because the distance metric of PCR\_04 was bigger. Normalization ensures that all features contribute equally to the distance metric by placing them in the same scale. This avoids the distortion that makes a feature with smaller scale dominate the decision.

1. A diagram of a spiral

   Description automatically generated

Train accuracy was:

Test accuracy was:

We can see from the results that the train set accuracy went down and the test set accuracy went up. This happened because when we use a bigger we reduce overfitting to the training set which lowers the training accuracy but increases the test accuracy. The decision regions are less overfitted for the training data. In this way the model can generalize better from the training data to data it didn’t see before. On the other hand, if we choose a that is too big we may make decisions based on datapoints that are too far and thus lower our test accuracy.

1. Normalizing both features using min-max scaling to is a bad idea because it does not account for the differences in the distributions of the features. The first feature is distributed uniformly in the range , meaning that there is an equal probability to sample any value in that range. So after normalization the values will distribute uniformly in the range . On the other hand, according to chi-squared distribution the probability to sample a value decreases from the value 0. Therefore, there is a (small) probability to pick big values, that will be outliers and will give the Min Max normalizing a big max value. Because of that most of the samples that we will normalize will be densely concentrated near -1. This will distort the distribution of the data of the feature.
2. A graph of numbers and a diagram

   Description automatically generated

We think that “num\_of\_siblings” feature is informative because we can see from this graph that people with 3-6 siblings had a bigger chance to spread than people with 0-2 siblings.

1. A graph of a graph of risk

   Description automatically generated with medium confidence

We think that “PCR\_03” feature is informative for the risk target because the distributions of the risk are very different. In the edges of the scale of PCR\_03 the distributions indicate of a high probability for -1 and in the middle of the scale a bigger probability for 1.

1. We chose the PCR features “PCR\_03” and “PCR\_10”. We think they can be helpful for predicting the risk because for the blood type group {O+, B+} the data is separated very well with the majority of the risk=1 values in the middle of the scale and the risk=-1 values in the bottom and upper part of the scale. The data is also splitted to regions in the second blood type group’ with the risk=1 values concentrated in 4 main regions and the risk=-1 in the rest of the plot.
2. A diagram of a number of blood types

   Description automatically generated

A diagram of a number of red and green dots

Description automatically generated

A diagram of a graph

Description automatically generated

1. A diagram of a graph

   Description automatically generated with medium confidence
2. Decision tree with max\_depth=3 will not be able to fit the training data well. Like we saw in the previous questions the features PCR\_03 and PCR\_10 can be good predictors of the risk only after they are separated by the specialProperty feature. So, the first decision in the tree should be according to the specialProperty. After that we have only one decision left, and as we’ve seen the split to regions of the data of each blood type group is not simple enough to make a prediction after one decision. Especially in the case where the blood type is not O+ or B+ we can see the data is splited into many regions.
3. Decision tree with max\_depth=30 will be able to fit the training data well. The first split will be made according to the specialProperty feature and after that the next splits will be made according to PCR\_03 and PCR\_10 values. In this manner we can get a split of the plane to areas with the same risk value. However a tree with that depth may cause overfitting to the training set. We can use a shallower tree to get a good division of the plane, and continuing to split further will give too much meaning to outliers and noisy data points.
4. 1NN model will not be able to fit the training data well. The scale of the 3 features is not the same, PCR\_03 and PCR\_10 are at the scale but the specialProperty feature is Boolean and gets only the values 0 and 1. Because of that the model will give more wight to the distance in the specialProperty axis. In that way for example we may have points with the same specialProperty value but different risk values. But because the distance in the specialProperty axis is only 1 we might classify them in a wrong way because the closest neighbor with the same risk and specialProperty value is at a bigger distance than one.
5. The normalization will not change the answers of Q19-20 because the scale of the features doesn’t matter to the decision tree. The decision tree splits the data, after normalizing the split will be the same.

The answer to Q21 will change. After normalizing the data each feature will be given the same weight and the model will be able to make better predictions.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Name** | **Keep** | **New** | **Normalization Method** |
| Patient\_id | V | X |  |
| age | V | X |  |
| sex | V | X |  |
| weight | V | X |  |
| Blood\_type | X | X |  |
| specialProperty | V | V |  |
| Current\_location | V | X |  |
| Num\_of\_siblings | V | X |  |
| Happiness\_score | V | X |  |
| Household\_income | V | X |  |
| conversations\_per\_day | V | X |  |
| sugar\_levels | V | X |  |
| sport\_activity | V | X |  |
| pcr\_date | V | X |  |
| PCR\_01 | V | X | Standard Normalization |
| PCR\_02 | V | X | Standard Normalization |
| PCR\_03 | V | X | Min Max Normalization |
| PCR\_04 | V | X | Standard Normalization |
| PCR\_05 | V | X | Standard Normalization |
| PCR\_06 | V | X | Standard Normalization |
| PCR\_07 | V | X | Standard Normalization |
| PCR\_08 | V | X | Standard Normalization |
| PCR\_09 | V | X | Standard Normalization |
| PCR\_10 | V | X | Min Max Normalization |