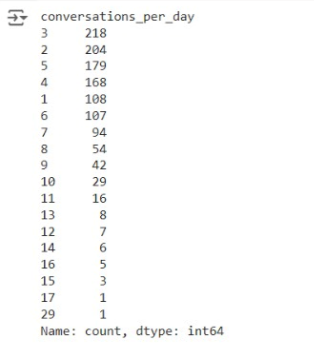
**Data Exploration and Preparation**

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 it’s categorical, the number of conversations is a natural integer it can’t be continues in the idea of the real numbers axis.

But, it has a natural order like the natural integers.

|  |  |  |
| --- | --- | --- |
| 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 | The result of the first Pcr test of the patient | Continuous |
| PCR\_02 | The result of the second Pcr test of the patient | Continuous |
| PCR\_03 | The result of the third Pcr test of the patient | Continuous |
| PCR\_04 | The result of the fourth Pcr test of the patient | Continuous |
| PCR\_05 | The result of the fifth Pcr test of the patient | Continuous |
| PCR\_06 | The result of the sixth Pcr test of the patient | Continuous |
| PCR\_07 | The result of the seventh Pcr test of the patient | Continuous |
| PCR\_08 | The result of the eighth Pcr test of the patient | Continuous |
| PCR\_09 | The result of the ninth Pcr test of the patient | Continuous |
| PCR\_10 | The result of the tenth Pcr test of the patient | Continuous |
|  |  |  |

1. **Need to answer later**
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 the majority of households.

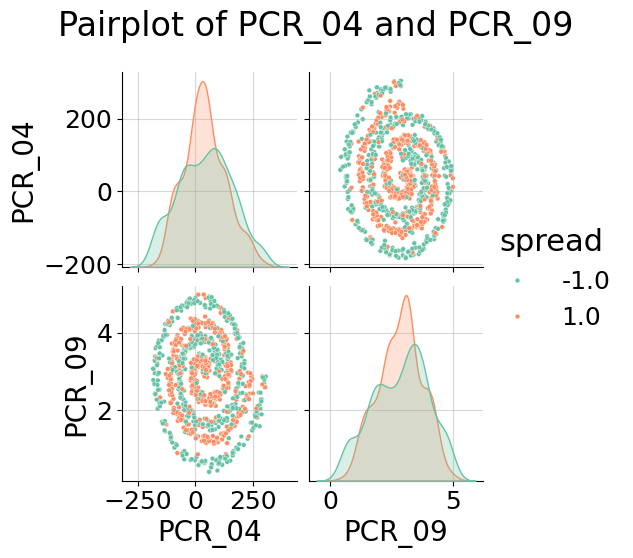
1. Household income mean: 3.64

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 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.

1. The pair of features useful for predicting the spread is {PCR\_04, PCR\_09} because it’s most separable pair of data.



1. First we calculate the distances of m training datapoints and our single test data point. This takes because the calculation of dist is in and we do it m times, for every training datapoint. After that we use argpartition to find the indices which runs in time complexity . Getting all the labels of the indices from argpartition will also 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.
2. 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 and because of this he contributed more to the distance metric. 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 big 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 because 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. Thus after normalizing it to the values will still be evenly spread across the segment. This will maintain the uniformity and symmetry of the distribution of the feature. The second feature does not distribute uniformly. In the chi-squared distribution most of the values are concentrated near the lower end of the range. After making the min-max normalization the values of the second feature will be squeezed into the segment in a non uniform manner. Most of the data points will be concentrated closer to -1 with a long tail extending towards 1. The uneven distribution can make biases in the distance metric of kNN because the distance calculations will be disproportionately influenced by the more densely populated regions of the chi-squared 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 3-4 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 pretty 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 same risk value.
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.