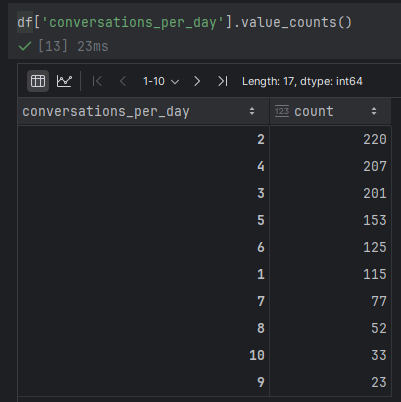
Part 1: Data Loading And First Look

1) 1250 rows and 25 columns

2)



I assume that it indicates the conversations per day the test subject has.

It’s ordinal because it’s categorical with a sort of natural order (less conversations correlates to higher count).

3)

|  |  |  |
| --- | --- | --- |
| Feature name | Description | Type |
| patient\_id | Patient identification number | Ordinal |
| age | Age of the patient | Ordinal |
| sex | Gender of the patient | Categorical |
| weight | Weight of the patient | Continuous |
| blood\_type | Patient’s blood type | Categorical |
| current\_location | Patient’s location | Categorical |
| num\_of\_siblings | Patient’s number of siblings | Ordinal |
| happiness\_score | Patient’s self-reported happiness score | Categorical |
| household\_income | Patient’s income on some scale. | Continuous |
| conversations\_per\_day | Num of conversations the patient has a day | Ordinal |
| sugar\_levels | Patient’s tested sugar levels | Continuous int |
| sport\_activity | How active the patient is | Categorical |
| pcr\_date | Date of the pcr test | Datetime |
| PCR\_01 | One result feature from the test | Continuous |
| PCR\_02 | Another result feature from the test | Continuous |
| PCR\_03 – PCR\_10 | Same here for all |  |

Partitioning the data:

4) It’s important to use the exact same split because we analyze the data to draw conclusions and we wouldn’t want them to be affected from changes in the underlying data but from our intentional processing steps.

Part 2: Missing Values and Outliers

5) train missing values:

Household\_income: 110

PCR\_02: 63

Test missing values:

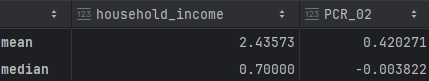
household\_income: 29

PCR\_02: 11

6) outliers are potential irregularities in the data. Usually, it’s defined as data that is not in the range of the lower bound and upper bound when they are:

This region should contain ~99.3% of the data if it’s normally distributed.

It seems there are outliers in the fields I plotted.

7)

There is a difference between the mean and median because the mean is more affected by the outliers and ‘pushed’ by them. I would prefer fillings the NANs with the median value as it closer to the majority of the data.

Part 3: Warming up with k-Nearest Neighbours

8)

A screenshot of a computer screen

Description automatically generated

It seems PCR\_01 and PCR\_08 make the data pretty seperetable.

9)



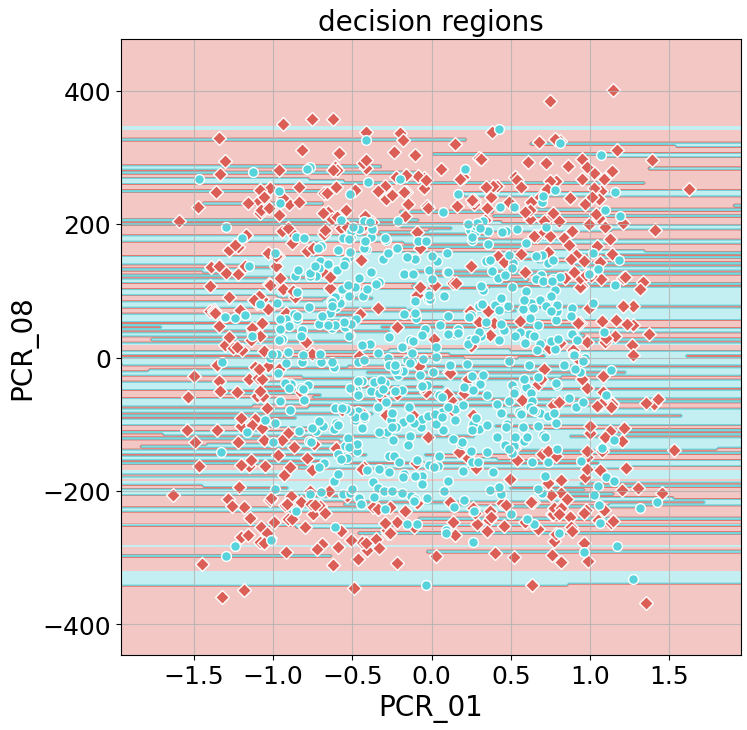
It doesnt contridict as each variable alone doesn’t need to corralate to the target spread (if it did we could come up with an easy classifier). The data can be sepertable in 2d and not in 1d.

10) The complextity for each datapoint will be

a. First of all we will compute the euclidian distance to every of the m data points, each point will take steps.

b. Then we will sort to the k-th elements and get the respective labales.

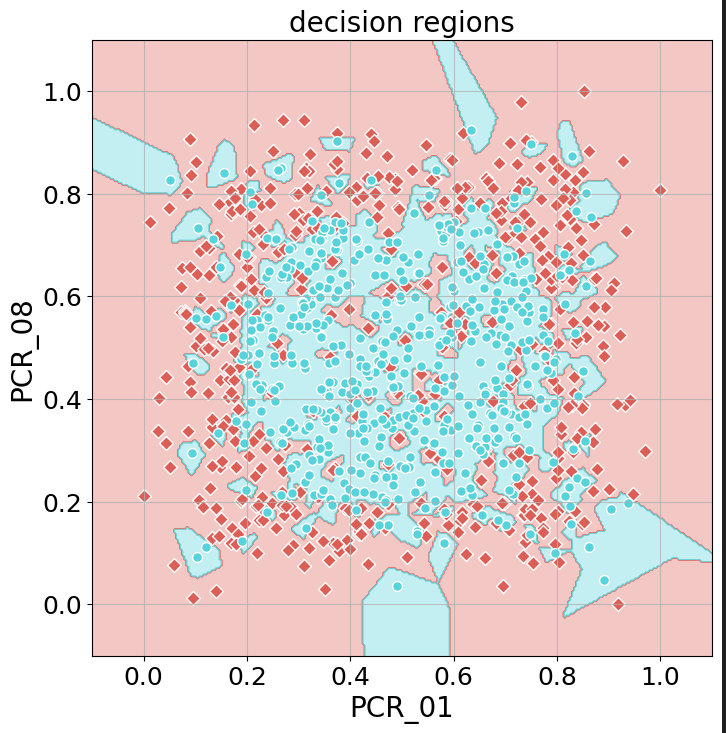
c. summing the data will take

11) decision regions-

Train score – 1

Test score – 0.564 (🥲)

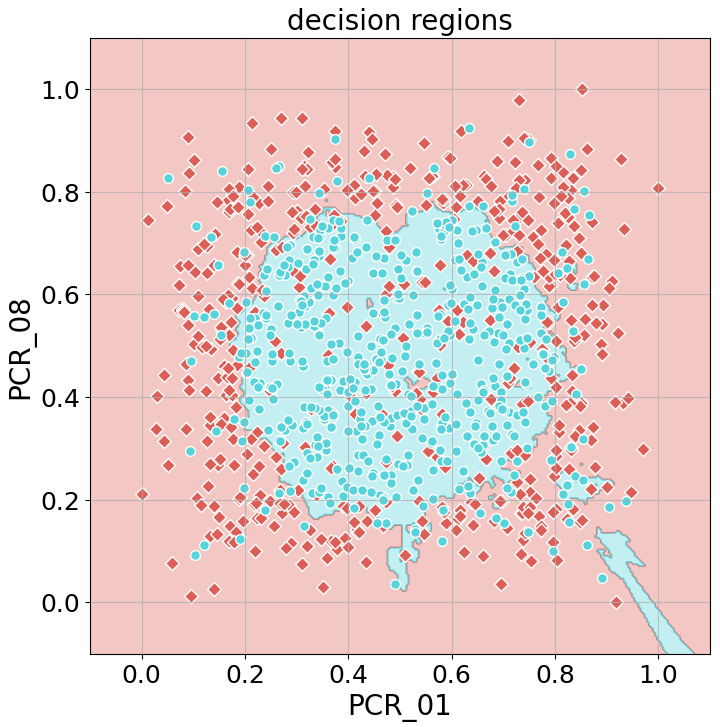
12) After normalization decision regions



Train score – 1

Test score – 0.66

It seems normalization is important to keep the decision regions nice and compact because now each feature has the same ‘weight’ in the distance calculation.

13)



We can see that the training score went down but test score went up so we can say we got better generalization.

With (reasonably) higher k it seems the decision region shift toward more certainty, ie every decision reason is made with higher certainty because there are more points to infer the classification from.

14) when normalizing the uniform feature it retains its shape and behavior but when doing so with the chi feature higher x value will loss it’s correct ‘weight’. For example a change in 0.4 in the normalized uniform feature should not count (according to L2 norm) like a 0.4 change in the normalized chi feature as it’s less probable.

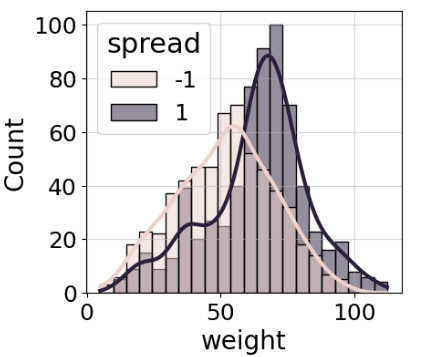
The result will be probably not classifying correctly near the tail of the chi feature.

Its easier to see this when thinking about the following classification task: given the features classify to which distribution a point belongs.

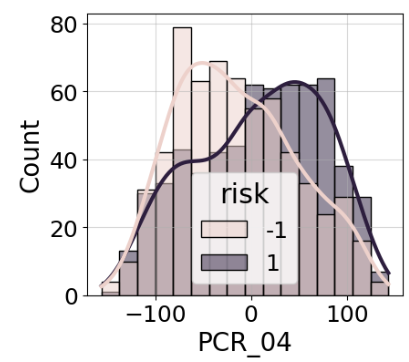
Without the normalization we would be able to easily identify a point from the ‘tail’ of the chi distribution (x>5), but after the normalization points from the uniform distribution will overshadow the once ‘tail’ part and the classification will probably be incorrect.

Part 4: Data Exploration

? 15) No one feature seems very informative. If I had to choose, I would choose the weight feature as it has the most separatable distributions.



16) For the same reason I would choose PCR\_04.



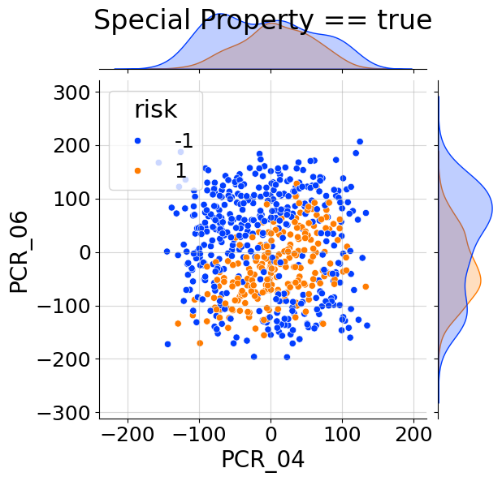
Bivariate Analysis

17) I chose PCR\_04 and PCR\_06 because they seem separable in both special property’s cases

A graph with orange and blue dots

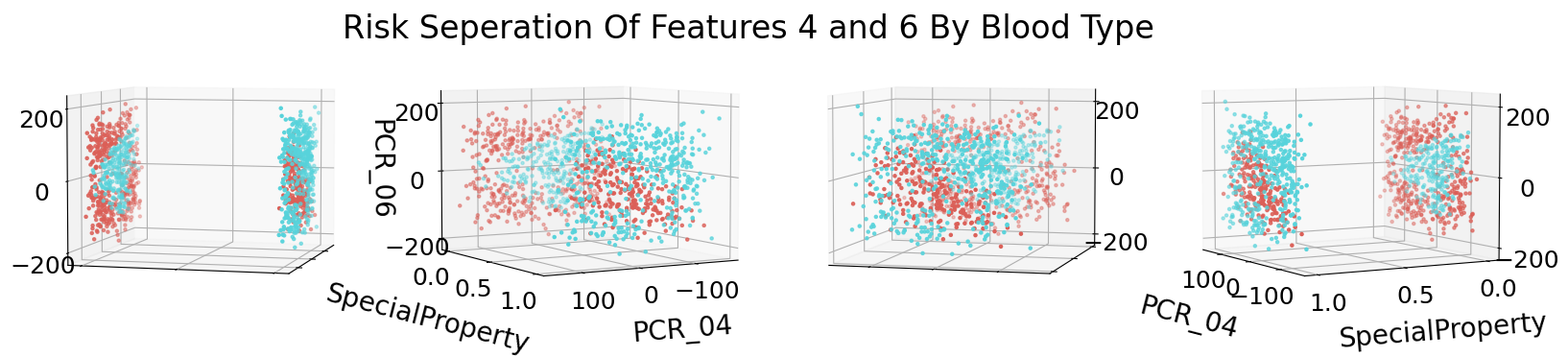
Description automatically generated18)

A diagram of different types of data

Description automatically generated

The risk target feature

19)



20) A decision tree of depth 3 will probably score poorly. After the initial separation on blood type, we would have 2 more decisions to try to create a ‘bounding box’ for the data, and it’s not so easily bounded.

21) A decision tree of depth 30 will probably score pretty good as it would be able to create a good enough ‘bounding polygon’ for the classification.

22) 1-NN model will probably score poorly because the it’ll miss the clear separation we see based on the blood type because of the scale of the features (and the fact we use Euclidian).

The special property has a range of [0,1] while the pcr’s range is in the hundreds, so the KNN is practically blind to the special property’s distance.

Part 5: More Data Normalization

23) Q20 and Q21 won’t change much as the only thing that will change is the values for the bounding polygons. Q22 will change because after the normalization the special property distance will matter and the data will be more separable.