**(A1)** the dataset contains1250 Rows and 25 Columns

**(A2)** the number of conversations the person engages in daily.

It’s considered ‘ordinal’ because it represents a categorical variable with a natural order, where the topmost row indicates the most common amount of conversations per day(3) and the bottom row indicates the least common amount of conversations per day(29).

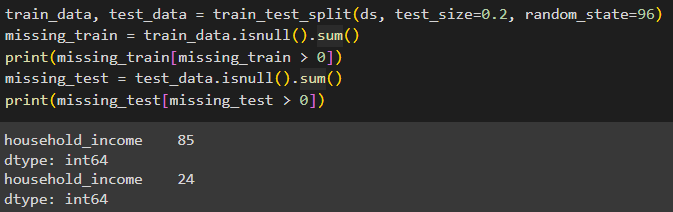
A screenshot of a computer

Description automatically generated

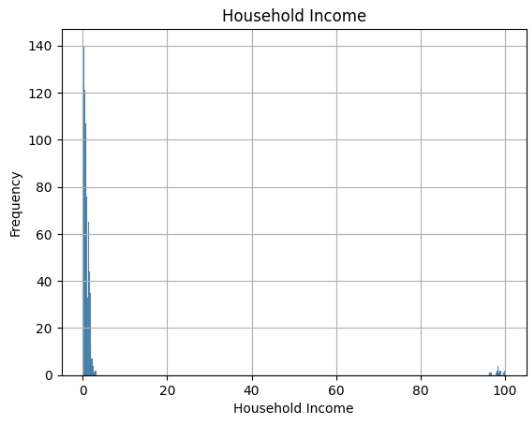
**(A3)**

|  |  |  |
| --- | --- | --- |
| ***Feature name*** | ***Description*** | ***Type*** |
| Patient\_id | The ID of each patient | Other |
| Age | The age of each patient | Continuous |
| Sex | The sex of each patient | Categorical |
| Weight | The weight of each patient in kg | Continuous |
| Blood\_type | The blood type of each patient (A, B etc.) | Categorical |
| Current\_location | The location coordinates of the patient | Other |
| Num\_of\_siblings | The number of siblings the patient has | Continuous |
| Happiness\_score | How happy the patient is | Continuous |
| Household\_income | The household income by some index | Continuous |
| Conversations\_per\_day | The number of conversations per day the person has | Ordinal |
| Sugar\_levels | Sugar level | Continuous |
| Sport\_activity | The sport activity by some index | Ordinal |
| Pcr\_date | The date of the PCR test | Other |
| PCR\_i | The i-th run result on the PCR assay | Continuous |

**(A4)** it’s important because it facilitates debugging and the comparison of different models

**(A5)**

**(A6)** the outliers are the values that are close to 100



**(A6a)**  **(A6b)**

A graph of a number of people

Description automatically generated with medium confidence

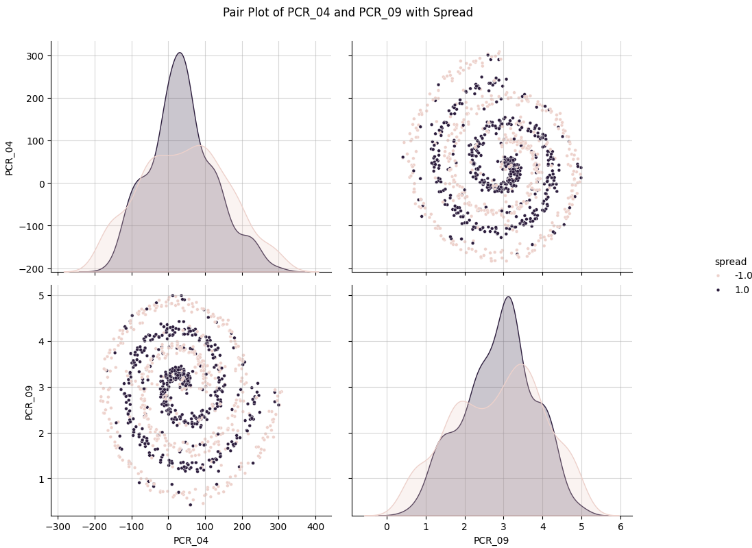
**(A7)** household Income **mean**: 3.64054

Household Income **median**: 0.7

I’d prefer using the **median** because it is more robust to outliers.

**(A8)** (PCR\_4, PCR\_9) & (PCR\_9, PCR\_4) seem to be the most useful for predicting spread,

Even though it’s not the simplest of seperations – it is still the best we got



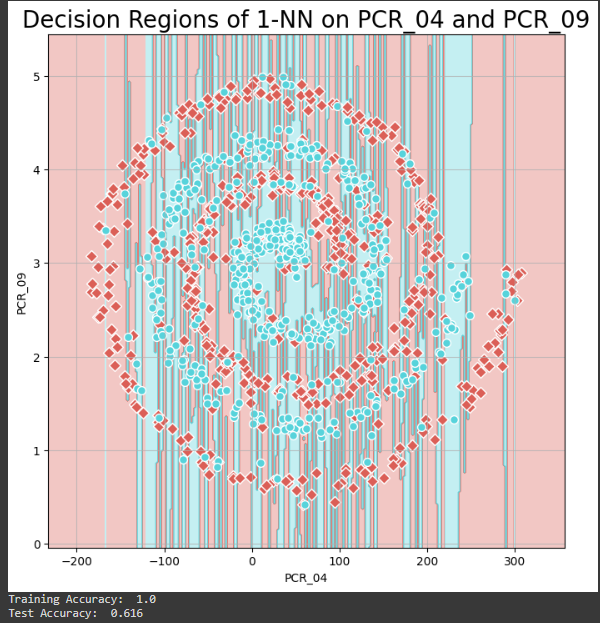
**(A9) in total**

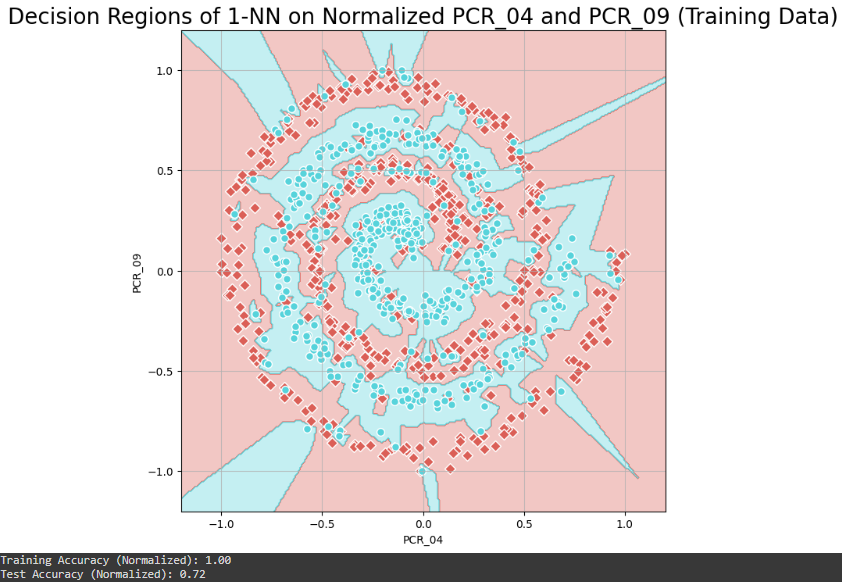
Firstly, find the new point’s distance from the existing points

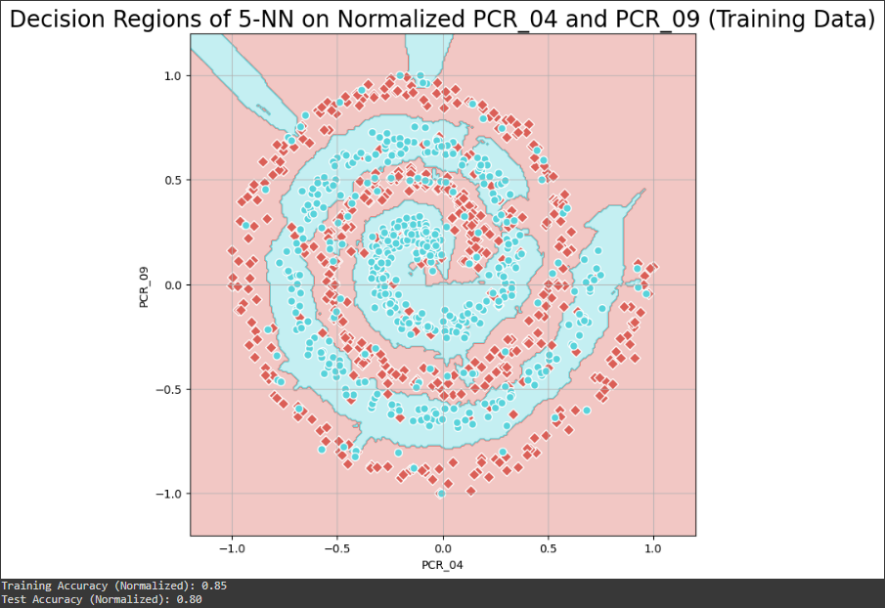
Secondly, use heap based selection approach & find k shortest distances

Thirdly, retrieve the k labels of the k nearest neighbours

Lastly, sum the k nearest neighbours and return the majority class

**(A10)**

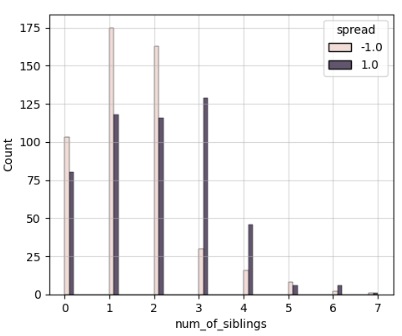
**(A11)** Normalization is important because it ensures that all the features contribute equally to the distance and that no single feature dominates the distance calculation (leading to biased results)

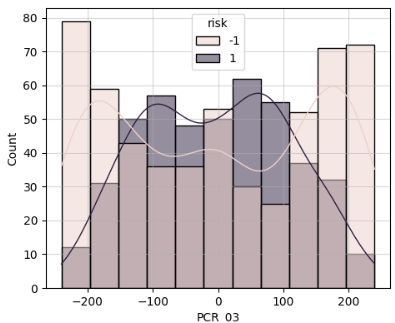
**(A12)**

**(A13)** normalizing both features using min-max scaling to [-1,1] is not ideal because it’d distort the original distributions.

The chi-squared distribution which has a specific range ( and compressing it to the range of [-1,1] would distort it significantly, especially given the non-linear nature of its spread.

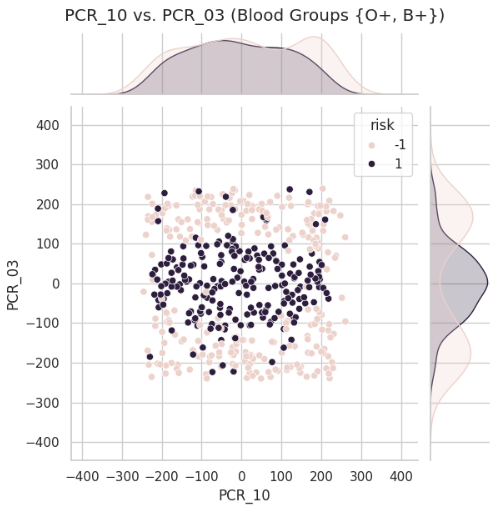
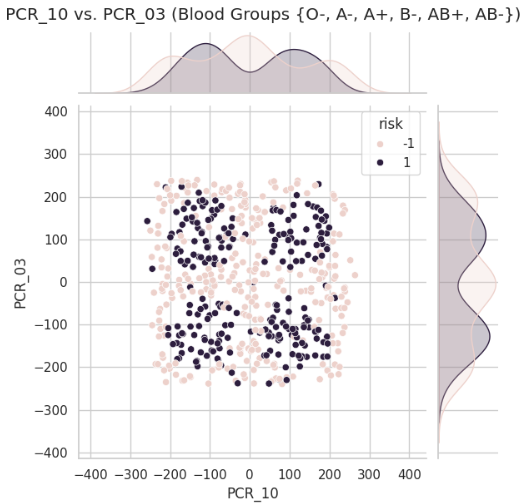
The uniform distribution has values ranging from [2,5], and transforming it to [-1,1] will not necessarily preserve the statistical properties of the uniform distribution.

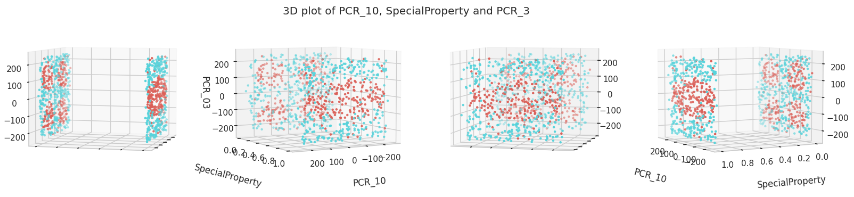
**(A14)** siblings might help in predicting ‘spread’ and there is a distinct separation in the distribution of spread levels. It can be seen in the plot that high number of siblings tends to be associated with ‘high’ spread while a low number of siblings is associated with ‘low’ spread.

**(A15)** PCR\_03 might help in predicting ‘risk’ as there is a distinct separation in the distribution of risk levels. It can be seen in the plot that high (around above 150) and low (around below 150) PCR\_3 levels tend to be associated with ‘low’ risk.

**(A16)** we chose (PCR\_3, PCR\_10) because it seems like it’s the best out of every other pair in separating both types of data.

**A graph with a number of dots

Description automatically generated(A17)**

**(A18)**

**(A19)** we think that a decision tree of max-depth=3 will work fine but not great on this data and will have the risk of ‘underfitting’, meaning it might not capture all the patterns withing the data and it won’t have enough capacity to model the relationships between the features and the target variables. A possible tree might look something like this:

has ‘Special Property’ (blood group {O+, B+})

Yes No

PCR\_3 value Here Underfitting comes in play

Yes No

Risk = 1 Risk = 0

**(A20)** we think that a decision tree of max-depth=30 will work fine but not great on this data and will have the risk of ‘overfitting, meaning it will capture too many nuances of the training pattern which might not correctly fit the test\real data. We think that a decision tree with a max depth of around 6 will work best.

**(A21) we** think that it’ll have a perfect fit on the training data, the nearest neighbor will be another point in the training set and thus predict the class correctly. On the other hand, the test data will have the risk of overfitting since the 1-NN model is extremely sensitive to outliers and noise. Some of the points have a distance greater than 1 to their nearest neighbor on the SAME plane – they will be greatly affected by their neighbor on the other plane – thus resulting in a not-so-good prediction (which ruins the separation quality of the SpecialProperty)

**(A22)**

**(A19)** all features will be on the same scale. This ensures that the tree treats all features equally when deciding on splits, potentially leading to a more balanced and generalized model. Yet the risk of underfitting still holds.

**(A20)** can help in treating features equally, will not significantly reduce the overfitting issue. The tree will still likely capture noise and nuances specific to the training data due to its depth.

**(A21)** we think that if we use min-max normalization, the data will be normalized between 0 and 1 thus the distance between the points on each plane will be relatively smaller that the distance between the different planes, thus the effects of the different planes on each other will be negated therefore the 1-NN model will fit the training data perfectly and will fit the test data better than the non-normalized data. Will also have the problem of overfitting

**(A23)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Feature name*** | ***Description*** | ***Keep*** | ***New*** | ***NormMeth*** |
| Patient\_id | The ID of each patient | V | X | - |
| Age | The age of each patient | V | X | - |
| Sex | The sex of each patient | V | X | - |
| Weight | The weight of each patient in kg | V | X | - |
| Blood Type | The Blood Type of the Patient | X | X | - |
| Current\_location | The location coordinates of the patient | V | X | - |
| Num\_of\_siblings | The number of siblings the patient has | V | X | - |
| Happiness\_score | How happy the patient is | V | X | - |
| Household\_income | The household income by some index | V | X | - |
| Conversations\_per\_day | The number of conversations per day the person has | V | X | - |
| Sugar\_levels | Sugar level | V | X | - |
| Sport\_activity | The sport activity by some index | V | X | - |
| Pcr\_date | The date of the PCR test | V | X | - |
| SpecialProperty |  | - | V | - |
| PCR\_{3,10} | The 3rd,10th runs result on the PCR assay normalized using minmax | V | - | Minmax |
| PCR\_{1,2,4,5,6,7,8,9} | The i-th run result on the PCR assay normalized using standard scalar | V | X | Standard |
| Normalized\_minmax\_PCR | Normalization of the PCRs by minmax | - | V |  |
| Normalized\_standrad\_PCR | Normalization of the PCRs by standard | - | V |  |

We chose to add the columns of the normalized PCR by both methods as from our insights on A22 they have different values for us. As for most standard will be the way to go, minmax might come in handy.