Part 1: Data Loading and First Look

(Q1) There are 1250 rows and 25 columns in the dataset.

(Q2) This feature refers to the average amount of conversation a person (represented as a sample in our dataset) has daily.
This feature is ordinal as it splits the data into categories (amount of conversations per day) with a natural order (1,2,3…), but, this feature is not continuous.

```
conversations_per_day
3     218
2     204
5     179
4     168
1     108
6     107
7      94
8      54
9      42
10     29
11     16
13      8
12      7
14      6
16      5
15      3
17      1
29      1
Name: count, dtype: int64
```

(Q3)

Feature name	Description	Type
patient_id	The Identification number of each patient	Ordinal
age	The age of each patient	Ordinal (/discrete)
sex	The biological sex of each patient	categorical
weight	The body weight of each patient	Continuous
blood_type	The blood type of each patient	Categorical
current_location	The place in which each patient is living/staying in at the moment	Other (looks like it is represented by tuple of continuous values)
Num of siblings	The number of brothers and sisters a patient has	Ordinal (/discrete)
Happiness score	The rank of happiness each patient reports on his life	ordinal
Household income	The financial income of each family/household normalized to some percentage (values from 0.1 to 100).	Ordinal

sugar_levels	The Glucose level of each patient	Ordinal (/discrete)
sport_activity	How much each patient engage in physical activity	ordinal
pcr_date	The date in which the patient had the covid PCR test	Categorical (could also be converted to ordinal)
pcr_i	The results of the i'th PCR covid test.	Continuous

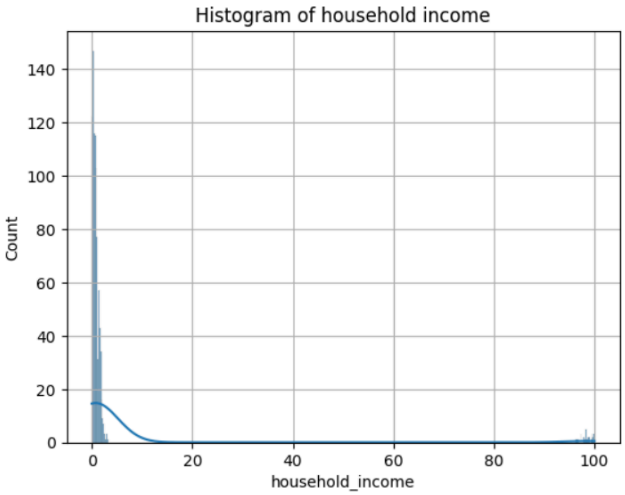
(Q4) It is important to use the exact same split for all the analyses for two reasons:

1. We want to keep a consistent analysis. If we don't keep an exact split, the consistency might be flawed as different splits might generate different and unique conditions for training, which will result in analyzing a different classifier in each analysis.
2. Preventing training the model on the test set (data leakage). If we train our model on samples from the test set, it is no longer valid to assess our model's performance on these samples.

Part 2: Missing values

(Q5) In the training set there are 86 null values in the house_income field.
In the test set, there are 23 null values in the house_income field.

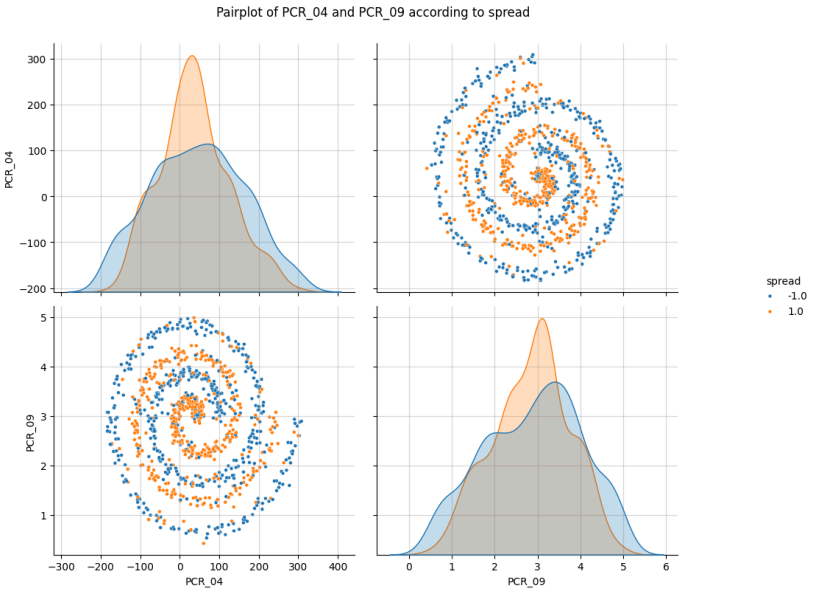
(Q6) inferring only from the histogram itself, it appears like the major mass of household income is between 0 and 5, with outliers around 100.



(Q7) The household_income **mean is ~3.85** while the household_income **median is 0.7** There is a significant difference between the mean and the median. This is because **the mean is more sensitive to extreme values** (the extreme values (outliers) directly affect the mean) compared to the median (where mostly, the **amount** of the values matter). In our case, we prefer to use the median as it will have a smaller effect on the data's distribution.

Part 3: Warming Up With k–Nearest Neighbors

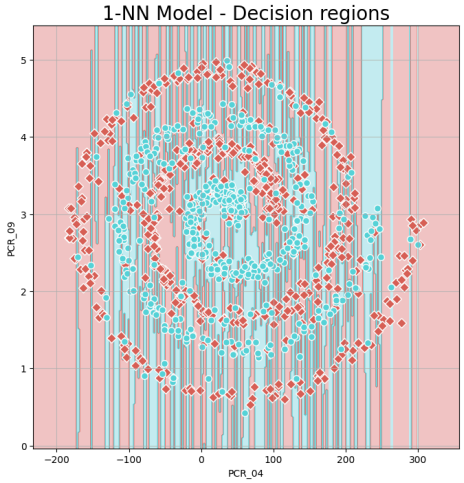
(Q8) Based on the graph we created in task B, we can say that the pair of features useful for predicting the disease's spread is {PCR_04, PCR_09}. When we look at the scatter plots that visualize the data's distribution on a 2d graph, the y-axis is the patient's PCR_04 test value and the x-axis is the patient's PCR_09 test value. The dots are colored based on the patient's disease spread. We can see 2 well-defined spirals in 2 colors - orange if the disease was spread and blue if not. The data seems to be the most separable under this pair of features.



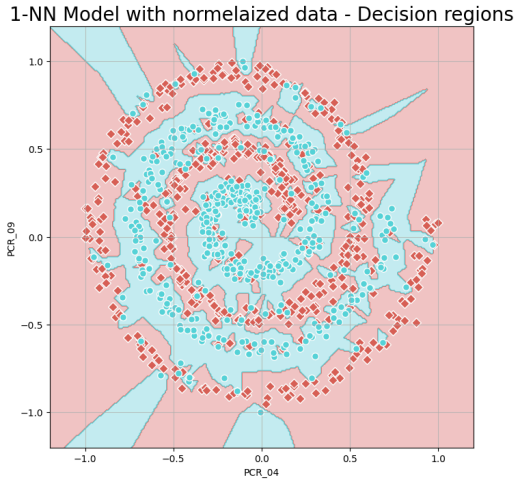
(Q9) The time complexity of the prediction function is **O(md + mloglog(m) + k) = O(md + k)**.
m = training data points
d = dimensions
k = neighbors

- Calculating the distance: O(m*d)
- Sorting the data: O(mloglog(m)) [which is O(md)]
- Chose closest k: O(k)
- **overall : O(md + mloglog(m) + k) = O(md + k)**.

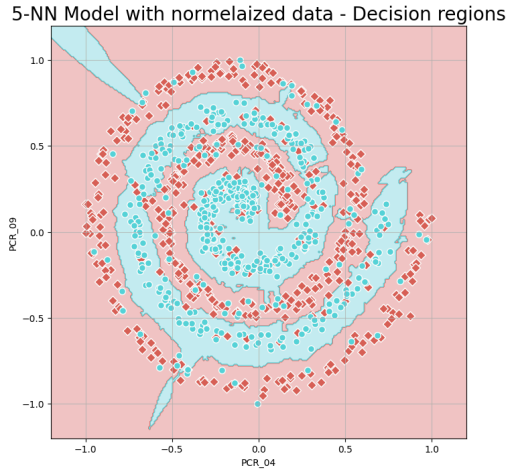
(Q10) The training accuracy of 1-NN model is 1.0, And the test accuracy is 0.6.



(Q11) The training accuracy of the 1-NN model with normalized data is 1.0,
And the test accuracy is 0.696.
The model's training accuracy remains 1.0, this is as expected since the model is already trained on these points. However, the test accuracy is better - 0.696 compared to 0.6. The normalization we did ensured that all features contributed equally to the distance and ensured that distances were calculated on a comparable scale. This allows the data points to reflect better their similarity resulting in a more accurate model, now we don't have smears in the decision region graphs.



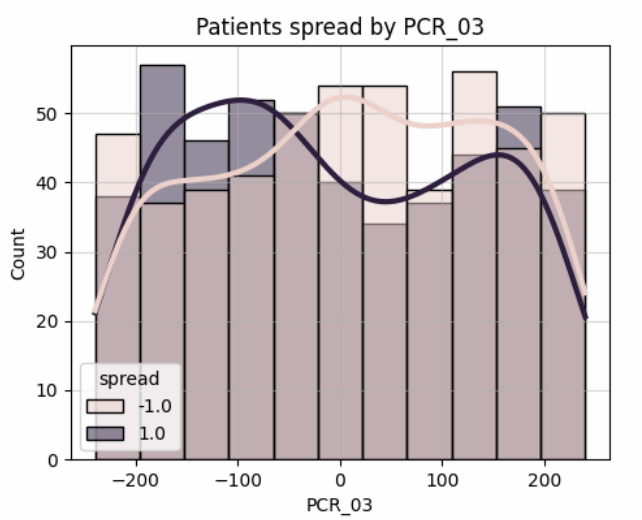
(Q12) The training accuracy of the 5-NN model with normalized data is 0.853,
And the test accuracy is 0.8.
Now, the model's training accuracy is lower 0.853 instead of 1.0, but the test accuracy is much better at 0.8 compared to 0.696.
The effect that k has on the decision areas is that for small k values the models do overfitting and as k values grow, the model takes into consideration more points in the area resulting in smoothing the borders between the decision regions.



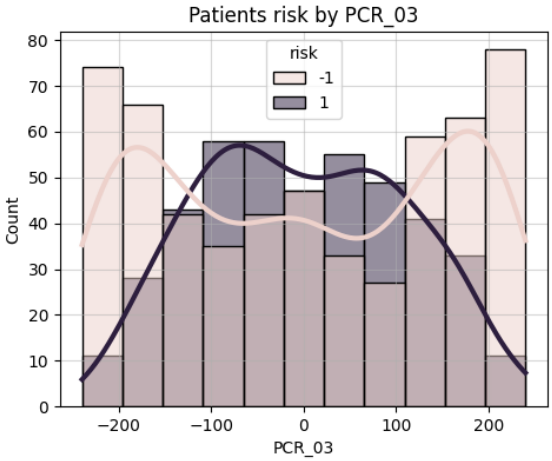
(Q13) Normalization of normally-distributed features will result in loss of information, by compacting data in order to fit in a smaller area. The data points become more similar and not accurately represented as they are represented in the original distribution.
Normalization of Chi-Squared distributed features will also result in a distortion of the distribution of this feature because it would try to fit the skewed distribution into a symmetric also values that were originally near zero could be pushed into negative territory, which might not reflect their true distributional characteristics.
Min-max normalization on both normally distributed and Chi-Squared distributed features distorts and leads to inaccurate representations of the data.

Part 4: Data Exploration

(Q14) We think that the **PCR_03 feature** is informative in predicting the spread target variable because we can see that if the PCR_03 (value) result is less than -50 then it is more likely that the spread is -1.0 and if the PCR_03 (value) result is higher then -50 it's more likely that the spread is 1.

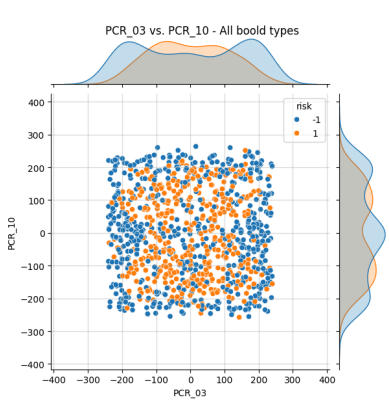


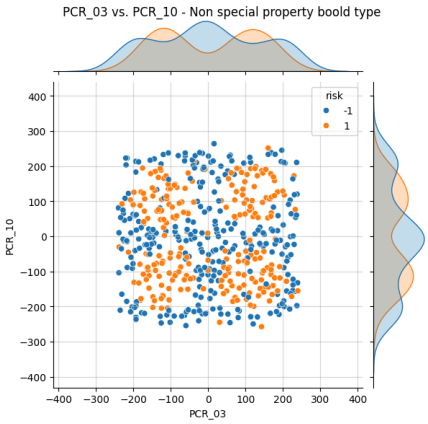
(Q15) We think that the PCR_03 feature is informative in predicting the risk target variable because we can see that the risk is more likely 1 when the PCR_03's value is between -100 to 100 and in the values above 100 or below -100 the risk is more likely -1.

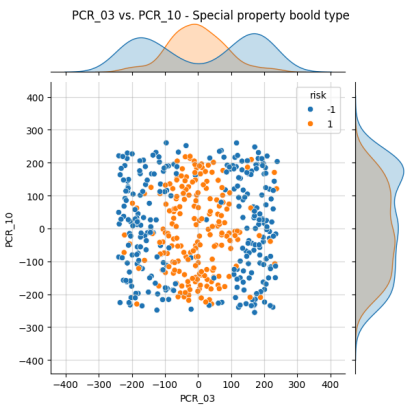


(Q16) We believe that the pair of PCR features that could be helpful for prediction is PCR_03 and PCR_10 because the data is separable.

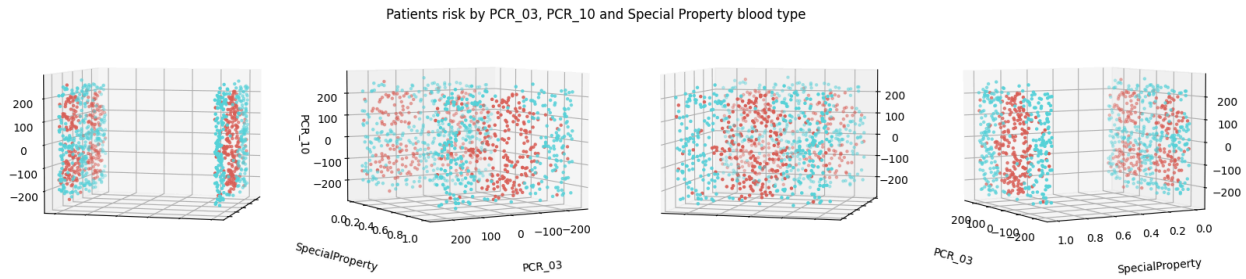
(Q17)







(Q18)



(Q19)

Max depth = 3 will be only able to **moderately fit the training data, but not well enough.** Looking at the graphs at the previous answer we can see that we need more than 3 planes to separate the data to positive vs negative Risk.

(Q20)

Max depth = 30 will highly fit the training data. This is because we get a decision tree with 2^30 leaves which means a leaf for each sample so we'll get a very high accuracy for the **training** set. (We might even say that in this depth the model will overfit on the training set).

(Q21)

A 1-NN model will moderately be able to fit the training data (probably slightly better than the max depth = 3 decision tree). We can see that there is some spatial separation of the data points for the given features, but at the "separating lines" between the groups, the model's performance will be poorer. The more features we have the greater the distance between the data points is, which means for higher dimensions we will get a better separation and fit on the training set.

Part 5: More Data Normalization

(Q22)

For (Q19 & Q21) - normalizing the data might slightly improve the accuracy for this model, but the improvement won't be major. This is because the normalization "moves" the data points in the same space, also moving the "spatial borders" between the groups with the data points. This means that the separation between different groups of data points won't improve dramatically.

For (Q20) as we anticipate an optimal accuracy for the training set, normalization should not change this answer.

Part 6: Data Preparation Pipeline

(Q23) The features we chose to normalize by the standardization have a distribution that resembles the chi-squared or the normal distribution or has outliers. This is based on Q13, where we explained why normalizing this kind of data by the min-max changes the distribution. Likewise, the features we chose to normalize by the MinMax normalization are those that have a uniform like distribution.

Feature name	Keep	New	Normalization method	Explanation
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	Explained above
PCR_02	V	X	Standard	Explained above
PCR_03	V	X	MinMax	Explained above
PCR_04	V	X	Standard	Explained above
PCR_05	V	X	Standard	Explained above

PCR_06	V	X	Standard	Explained above
PCR_07	V	X	Standard	Explained above
PCR_08	V	X	Standard	Explained above
PCR_09	V	X	Standard	Explained above
PCR_10	V	X	MinMax	Explained above