Major HW 1 – Data Exploration and Preparation

Ofir Manor, 316084623, [ofir.manor@campus.technion.ac.il](mailto:ofir.manor@campus.technion.ac.il)

Orad Barel, 311288203, [oradbarel@campus.technion.ac.il](mailto:oradbarel@campus.technion.ac.il)

Q1. The Virus Data dataset contains 1250 rows and 26 columns.

Q2. The output of value\_counts of num\_of\_siblings is:

1 399

2 317

0 271

3 161

4 62

5 31

6 6

7 2

8 1

Name: num\_of\_siblings, dtype: int64

This refers to the number of siblings each datapoint (patient) has in the dataset.

The number of siblings each datapoint has can only be ordered integers, thus the data

is ordinal.

Q3.

|  |  |  |
| --- | --- | --- |
| Feature Name | Description | Type |
| patient\_id | Number used to identify  the patient | Other |
| age | Age of the patient | Ordinal |
| sex | Sex of the patient  (M or F) | Categorical |
| weight | The wight of the patient  in Kg | Continuous |
| blood\_type | The blood type of the  patient | Categorical |
| current\_location | The coordinates of the  Patients (latitude and  longitude) | Other |
| num\_of\_siblings | The number of siblings  the patient has | Ordinal |
| happiness\_score | A general happiness  score of the patient  out of 10 | Ordinal |
| household\_income | The income of all  members of the patient’s  household | Continuous |
| conversations\_per\_day | Amount of conversations the patient has with others per day | Ordianl |
| sugar\_levels | Concentration of glucose  In the patient’s blood  (mg\dL) | Ordinal |
| sport\_activity | Feature describing how  the amount of sport  the patient partakes in | Ordinal |
| symptoms | A list of symptoms  exhibited by the patient | Other (list) |
| pcr\_date | The date of the patient’s  PCR tests | Other (date) |
| PCR\_01 | The results of the  patient’s PCR 01 test | Continuous |
| PCR\_02 | The results of the  patient’s PCR 02 test | Continuous |
| PCR\_03 | The results of the  patient’s PCR 03 test | Continuous |
| PCR\_04 | The results of the  patient’s PCR 04 test | Continuous |
| PCR\_05 | The results of the  patient’s PCR 05 test | Continuous |
| PCR\_06 | The results of the  patient’s PCR 06 test | Continuous |
| PCR\_07 | The results of the  patient’s PCR 07 test | Continuous |
| PCR\_08 | The results of the  patient’s PCR 08test | Continuous |
| PCR\_09 | The results of the  patient’s PCR 09 test | Continuous |
| PCR\_10 | The results of the  patient’s PCR 10 test | Continuous |

Q4. Throughout the exploration and preparation of the data we conduct several tests to see the effectiveness of different features, normalization, and strategies in learning the data. We do so in a trial-and-error method where we make slight changes and run our scripts again and again. During all of this we do not want to be susceptible to our test dataset. Otherwise, we will make decision based on the test dataset, and this will ruin the credibility of our final results when we use our model on said test set.

Q5.

|  |  |
| --- | --- |
| PCR Test | Correlation with spread |
| PCR\_01 | 0.08301882503960864 |
| PCR\_02 | 0.4788563994550575 |
| PCR\_09 | -0.04155539143381472 |

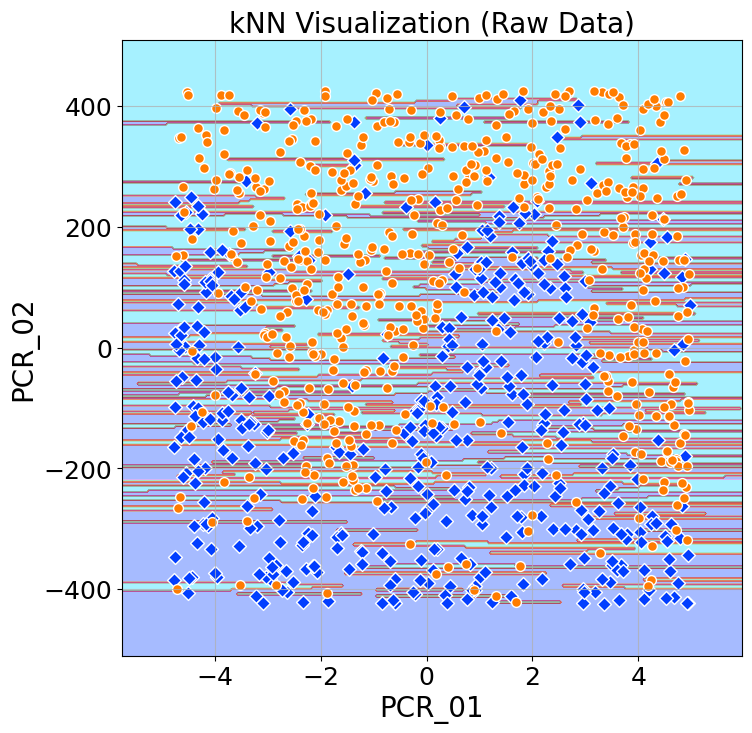
Q6.

Shape

Description automatically generated

According to the pairplot we can see that PCR\_01 and PCR\_02 create the most separable plot, creating an almost sinuous line that separates the spread from the not spread.

Q7.



We achieved a training accuracy score of 1. This is logical since the nearest neighbor of each point is itself. We achieved a test accuracy score of 0.704.

Q8.

Chart, scatter chart

Description automatically generated

We achieved a training accuracy score of 1. We achieved a test accuracy score of 0.832.

This shows why normalization is important for kNN. In the previous example we saw that the range of PCR\_02 was far greater than the range of PCR\_01. This means that the value of PCR\_02 had far greater effect on the nearest neighbors, because two datapoints could be much farther away on PCR\_02 than they could PCR\_01. After normalization, both PCR\_01 and PCR\_02 share the same range (-1,1), so that each has the same “power” over the distance from its neighbors.

Q9. Chart, scatter chart

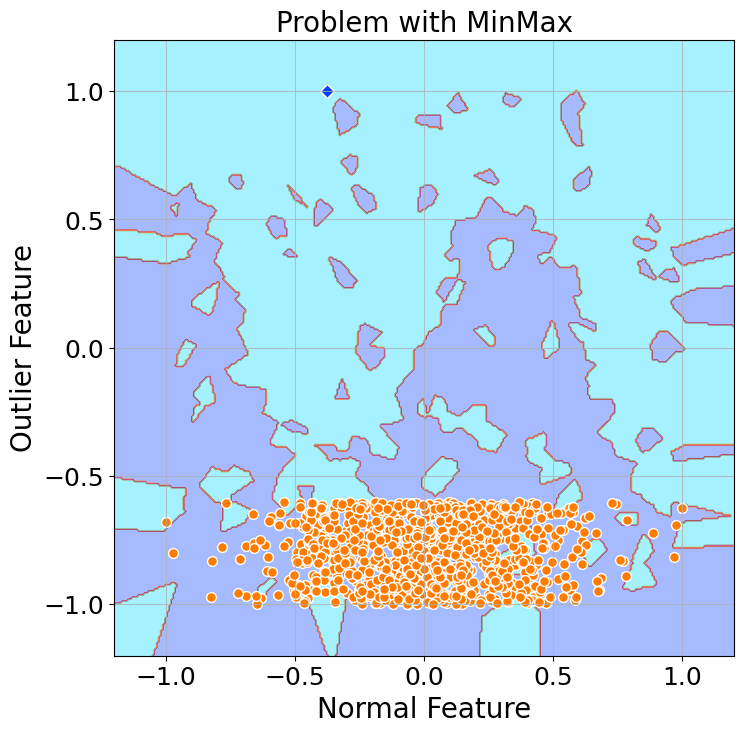
Description automatically generated

We achieved a training accuracy score of 0.869. We achieved a test accuracy score of 0.928.

The increase of k has allowed for more loss in the training. This has created more contiguous decision regions and led to better results on the test through less overfitting to the training data.

Q10.

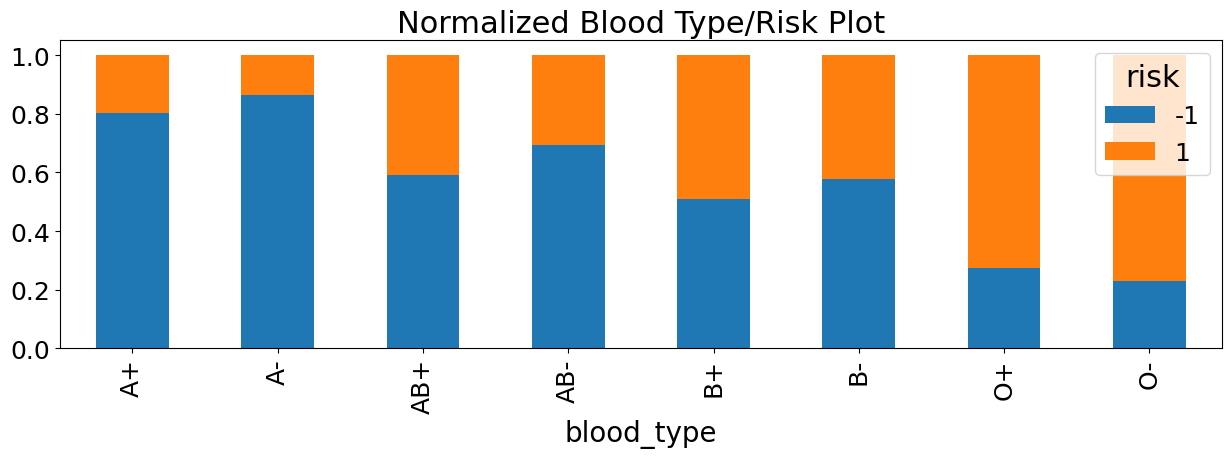
The general problem with min-max scaling is that it does not deal well with outliers. If say a feature had a maximum that was far above the rest of the datapoints that maximum would then become 1 on the scale and the rest would squish into the lower end of the normalized scale. If we have a feature that is normally distributed, then min-max scaling would do very little to change the relative positions of the datapoints in that scale. Hence, if we use two features, one normally distributed and the other with an outlier, we will find that the normalized data utilizes the entire scale ([-1,1] in out case) while the data with the outliers mostly utilizes a smaller area. This will mean that the datapoints will be closer together on the axis with the feature with the outlier than the axis with the normally distributed feature. This kNN will rely more heavily on the feature with the outlier. To represent this, we have provided a graph:



Q11.

There are 8 blood types in out data, hence we would need 8 Boolean features to accurately represent the blood type feature.

Q12.



We divided the blood type into three groups based on the proportions of to in them. This is to attempt and group those with similar probability of risk.

These group are:

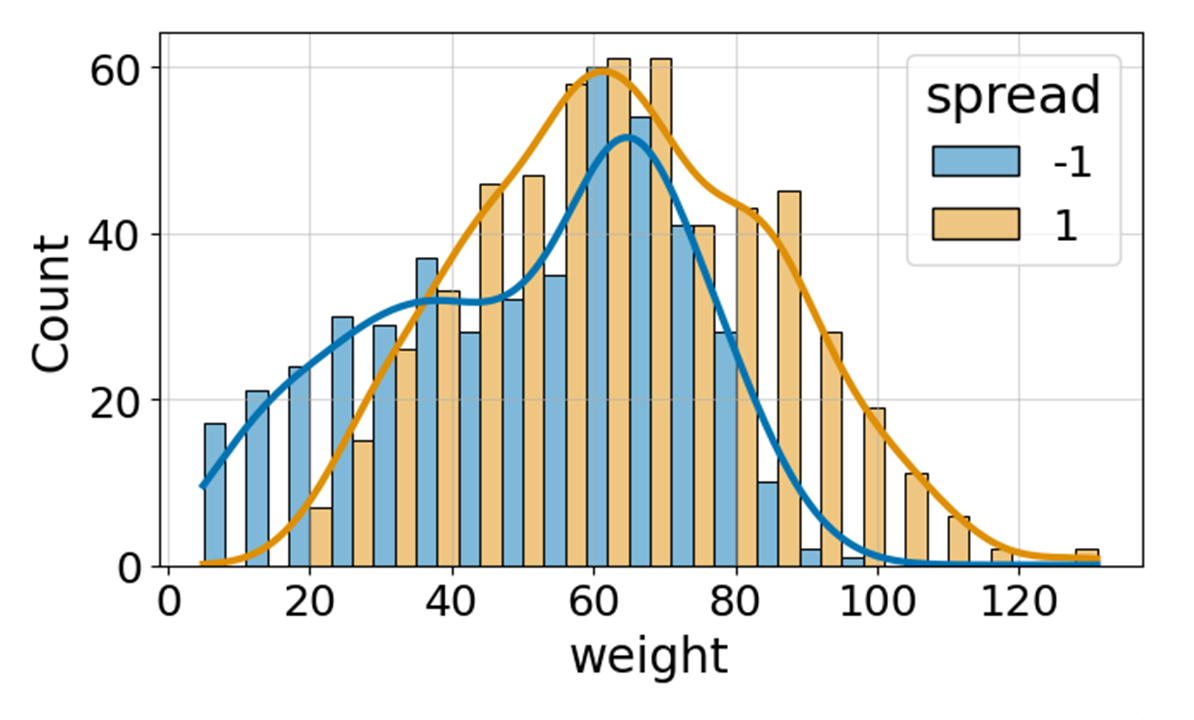
* Blood Group A with a low proportion of risk
* Blood Group AB or B with a moderate proportion of risk
* Blood Group O with a high proportion of risk.

Q13.

We believe that the symptoms feature holds useful information we can extract. We replaced the list of symptoms with Boolean features, each stating whether the patient’s displays that symptom or not. That way we can see exactly which patient carries which symptoms, without losing any information and giving each symptom an equal value while not allowing a combination of some symptoms to necessarily represent something more than some other symptoms.

Q14.

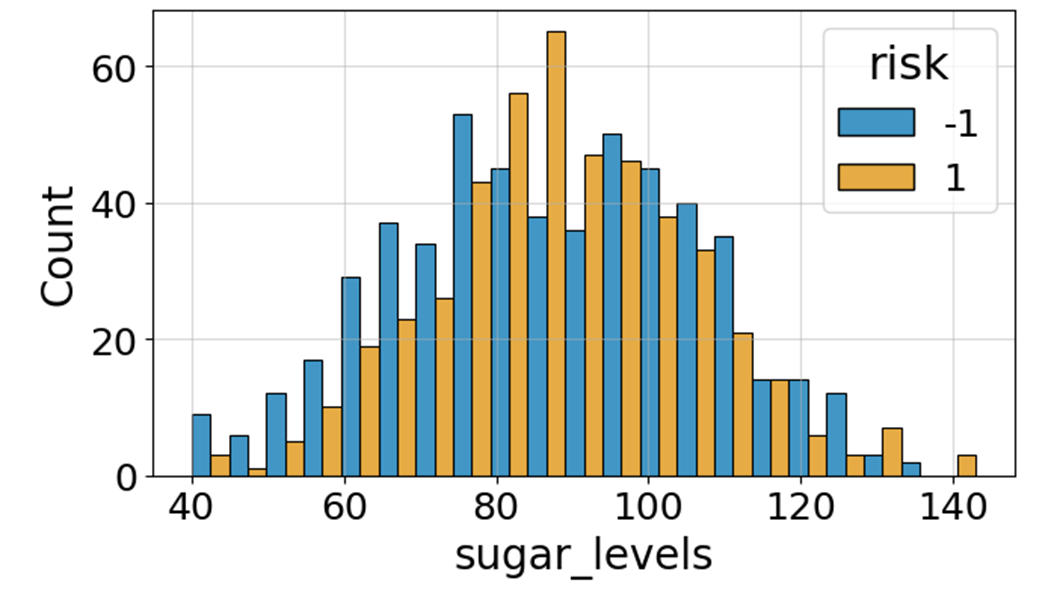
Weight seems to be an informative feature for prediction spread.



As we can see, the probability of spreading is higher on people with more weight. People with over 80 kg are far more probable to have while people below 30 kg hare far more probable to have .

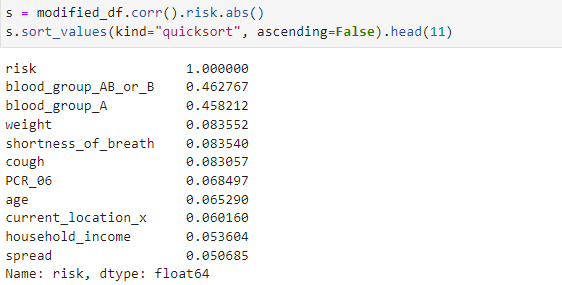
Q15.

We find that the sugar levels feature seems informative in predicting risk.



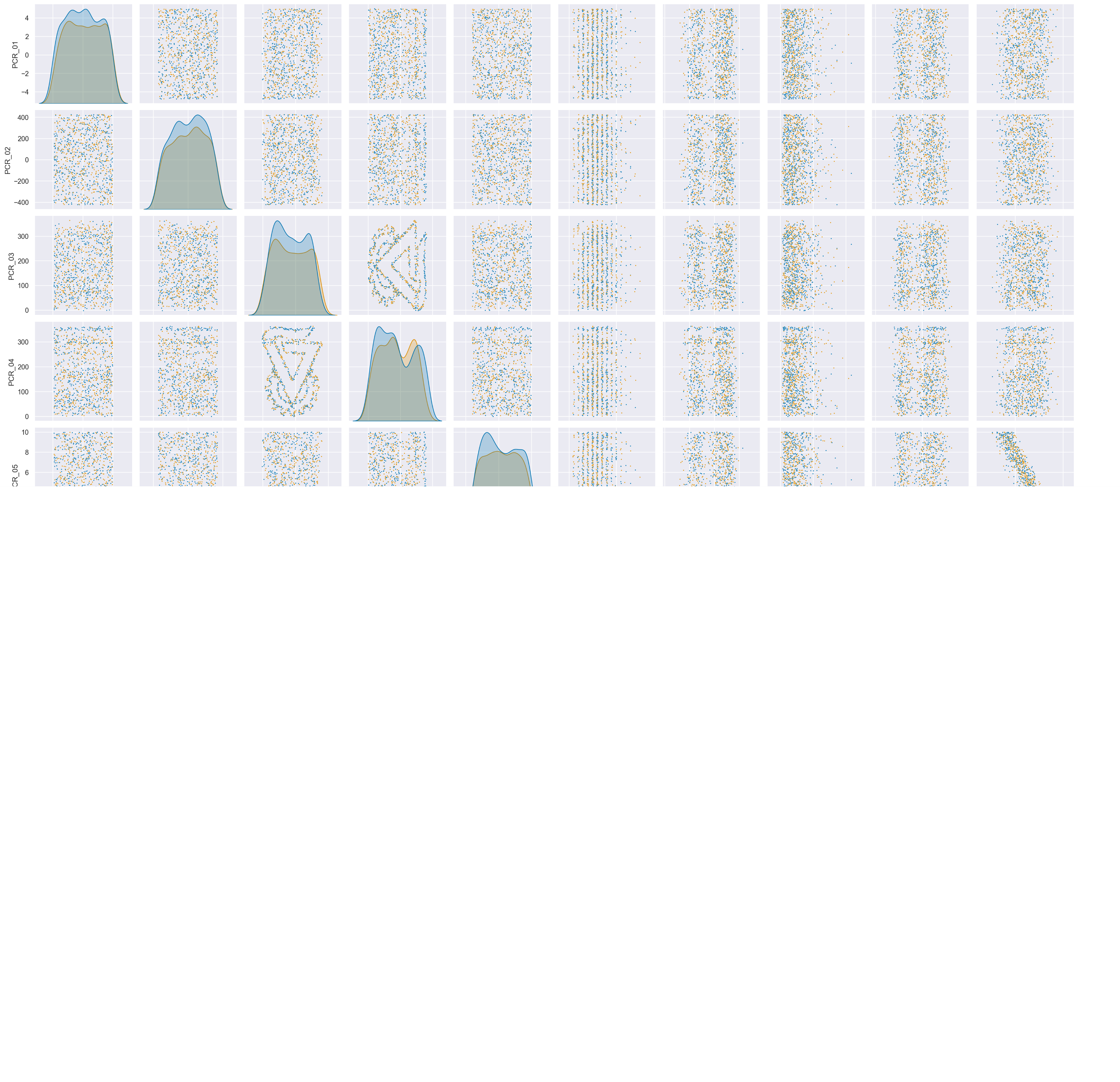
That is because people with lower sugar levels (below 80 mg\dL) are far less at risk while those at sugar levels between 80 and 100 mg\dL are at far higher risk.

Q16.

The following output shows the 10 most correlated features with risk

Obviously risk it perfectly correlated with itself, but that provides not additional information. Spread is also a target variable, but we have decided to keep it in the top ten correlated features as it show the correlation between them (something that might help in predicting risk based on our prediction of spread in the future).

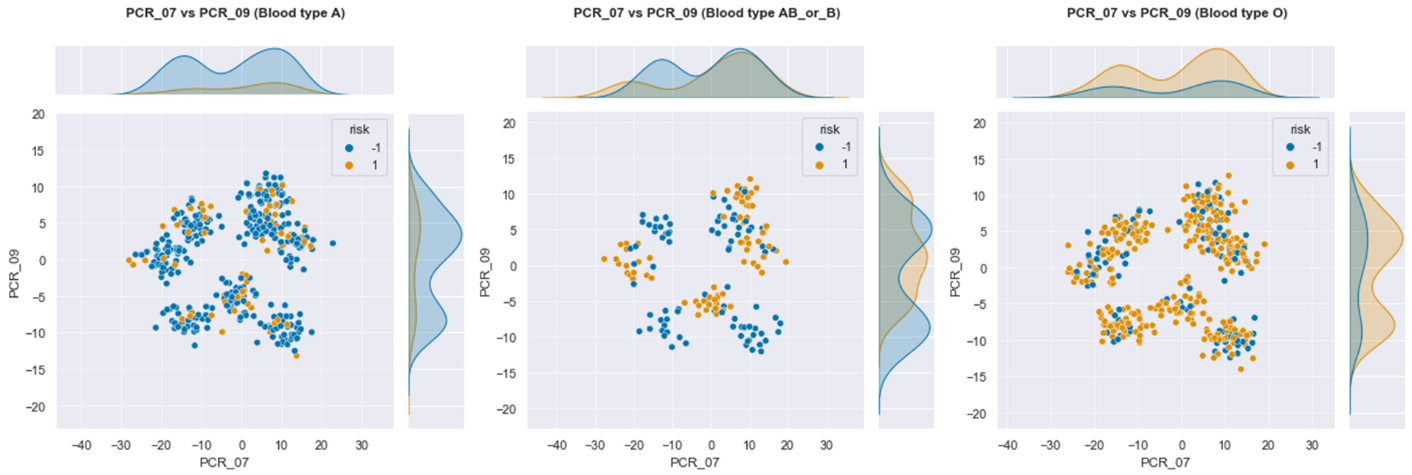
Q17:



PCR\_04 & PCR\_03 PCR\_09 & PCR\_07

Neither of them seem to explain risk to any large extent by themselves.

Q18.



Q19.

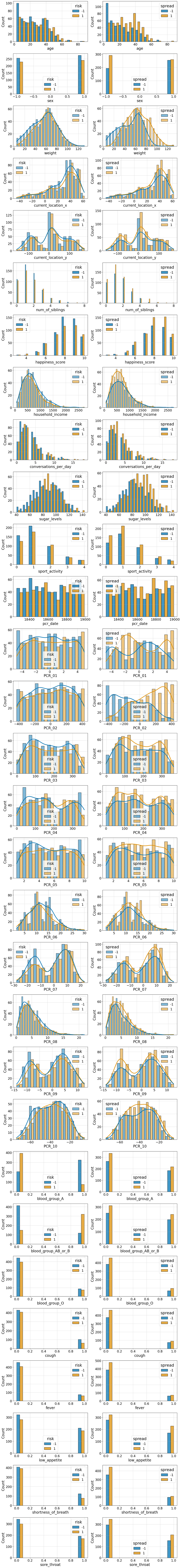
The plot that looks the most informative for predicting risk is the jointplot of PCR\_07 and PCR\_09 with only blood type AB or B. These features did not exhibit “high” correlation with regards to risk. This is because individually they did not. But together, they represent a richer environment. Also, the plot we got does not simply divide the area into two, one of high risk and one of low, but separates it into several areas of high concentration of high risk and low risk. This will not show accurately on correlation.

Q20.

The most suitable model for predicting risk with the informative plot is kNN. Looking at the plot, there is no line we can uses to separate the datapoint where to the datapoints where . So, in that case a linear model would simply fail. A decision tree is possible if we do not allow it to go to its maximum depth, seeing as there are areas where there are small number of datapoints around far more datapoint. Still the datapoints are not divided nicely amongst the axis, so the tree would be complicated still. Yet, using kNN with a relatively small n (around 5 or 7) would produce the best results. As most datapoints with the same label are largely clumped together, this suggests that certain areas in the feature space are more likely to be of high risk or low risk.

Q21.

Before



After



Q22.

For the forward pass, we first check all features, pick the best one, add it to the features we plan the use, and then check the remaining features. We continue to util we have reached features. So, the complexity of the number of models we’d have to check is:

In the backwards pass, we first check models, each containing all features except for 1. We pick the feature whose absence improves the model the most and then check models, each without one of the remaining features. We do so until we have only features remaining. So, the complexity of the number of models we’d have to check is:

Q23.

The three features that were found are PCR\_01, PCR\_02, and weight. Where PCR\_01 and PCR\_02 are the features we chose in Q6, and weight is the feature we chose in Q14.

Q24.

It is important to preform normalization before sequential feature selection because of the model used in sequential feature selection. In our SNS we use a kNN model, a model known to be highly perceptible to unnormalized data. This is because where one feature is on a scale of [-1,1] while another is on a scale of [-250,250] the latter will have much larger distances on datapoints that are relatively are closer to one another than on the former.

Q25.

The choice of a learning algorithm affects the sequential features selection. In general, different algorithms use different ways to score the utility of selecting a certain feature and work differently on the data plane. For example, in a forward pass, decision trees and knns will work differently. A decision tree on the first feature can only divide the data into 2 groups depending on the feature, so it will select such a feature that will give it the greatest information gain. A knn model, works with Euclidian distances so it very well might find a different feature that gives it the best utility.

Q26.