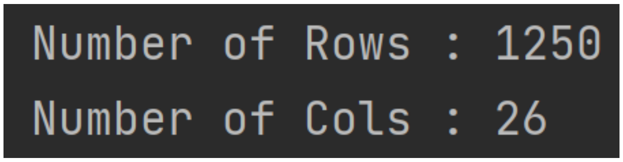
Ori Zohar - 205960750
Omer Taub - 316497122

Intro to ML – Wet1

Part 1 – Data Loading and First Look

(Q1).

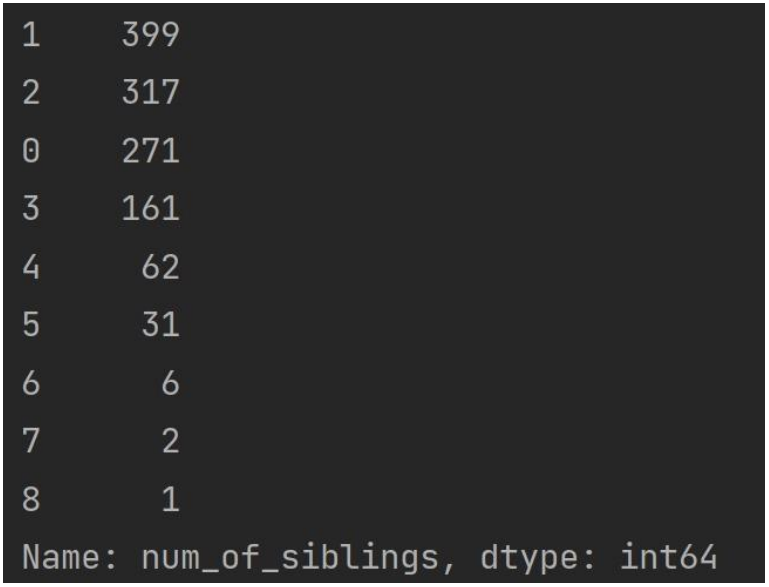
Number of columns is 26 and number of rows is 1250.



```
Number of Rows : 1250
Number of Cols : 26
```

(Q2).

Obtained output :



```
1    399
2    317
0    271
3    161
4     62
5     31
6      6
7      2
8      1
Name: num_of_siblings, dtype: int64
```

Num_of_siblings refers to the number of brothers and sisters one's has.
This feature's type is "ordinal" because there are finite numbers of unique values, and number of siblings that certain person can have is in natural order.

(Q3).

Feature name	Description	Type
patient_id	The id of the patient.	Continuous
age	The age of the patient.	Ordinal
sex	The sex of the patient.	Categorical
weight	The weight of the patient.	Continuous
blood_type	The blood type of the patient.	Categorical
current_location	The location that the patient lives.	Other
num_of_siblings	Number of brothers and sisters the patiens has.	Ordinal
happiness_score	How happy the patient in his daily life	Ordinal
household_income	How much money the patient family gets.	Continues
conversations_per_day	How much conversation the patient does in one day.	Continuous
sugar_levels	The patient's blood sugar levels.	Ordinal
sport_activity	The level of activity the patient.	Ordinal
symptoms	Which symptoms the patient has.	Other
pcr_date	When the patient did the PCR.	Other
PCR_01	The result the patient got in the PCR number 1.	Continuous
PCR_02	The result the patient got in the PCR number 2.	Continuous
PCR_03	The result the patient got in the PCR number 3.	Continuous
PCR_04	The result the patient got in the PCR number 4.	Continuous

PCR_05	The result the patient got in the PCR number 5.	Continuous
PCR_06	The result the patient got in the PCR number 6.	Continuous
PCR_07	The result the patient got in the PCR number 7.	Continuous
PCR_08	The result the patient got in the PCR number 8.	Continuous
PCR_09	The result the patient got in the PCR number 9.	Continuous
PCR_010	The result the patient got in the PCR number 10.	Continuous

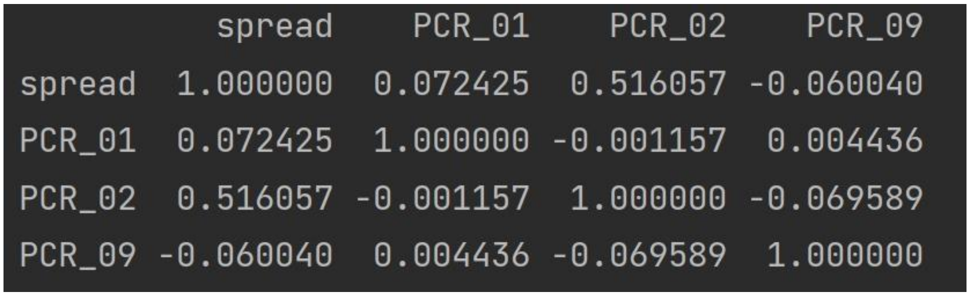
(Q4).

It Is important that we use the exact same split for all our analyses because when we want to decide which model has the best accuracy performance, we want the data we used to train which model to be the same.

Part 2 – Warming up with k-Nearest Neighnors :

(Q5).

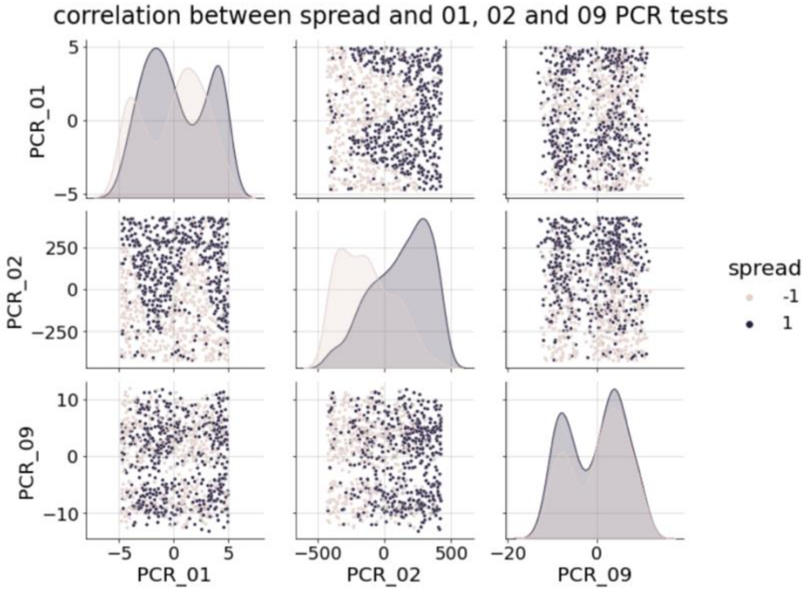
The correlations between the spread feature to 01,02 and 09 feature are:



```
           spread    PCR_01    PCR_02    PCR_09
spread   1.000000  0.072425  0.516057 -0.060040
PCR_01   0.072425  1.000000 -0.001157  0.004436
PCR_02   0.516057 -0.001157  1.000000 -0.069589
PCR_09  -0.060040  0.004436 -0.069589  1.000000
```

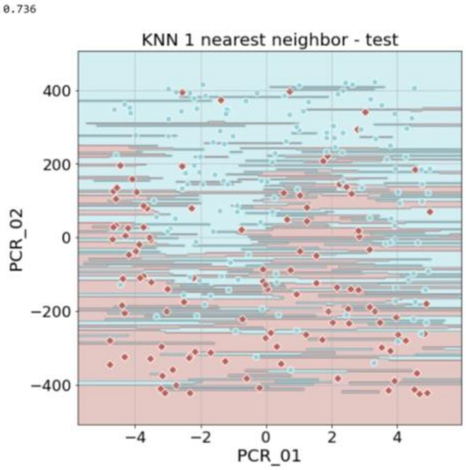
(Q6).

The 2 features that are most useful to predict are PCR_01, and PCR_0

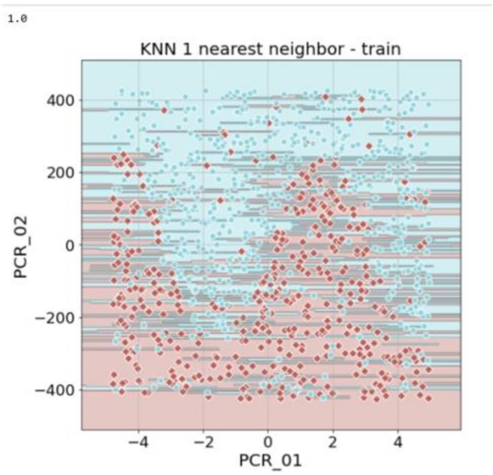


correlation between spread and 01, 02 and 09 PCR tests

(Q7).



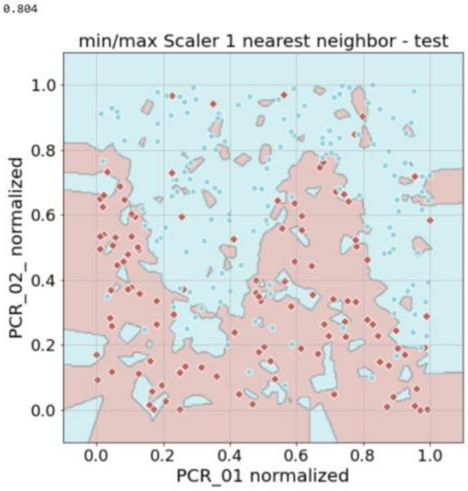
Accuracy – 0.736



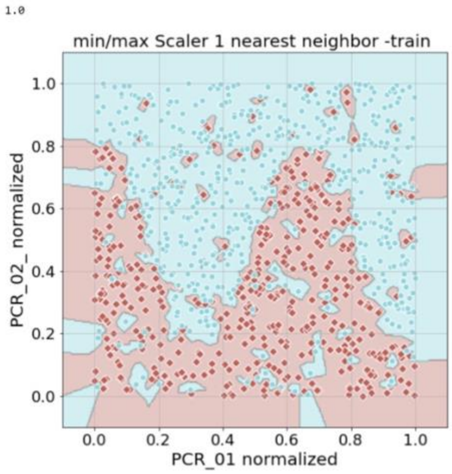
Accuracy - 1

(Q8).

In Q7, the scale of PCR_02 is much larger than the scale of PCR_01, therefore in the calculation of the distances in kNN, PCR_02 will have much greater effect on the distance and as a result the prediction will be affected (based mostly on 1 feature than on 2 we already know).
As a result of the normalization, the two factors can be used to calculate the predication and the accuracy increase.



Accuracy – 0.804

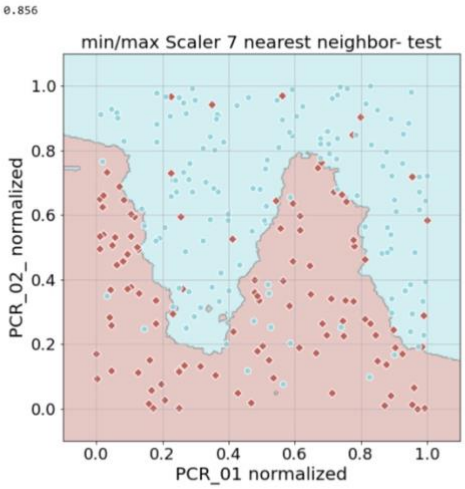


Accuracy - 1

In general, normalization is important for kNN because it gives each feature equal weight in the distance calculation.

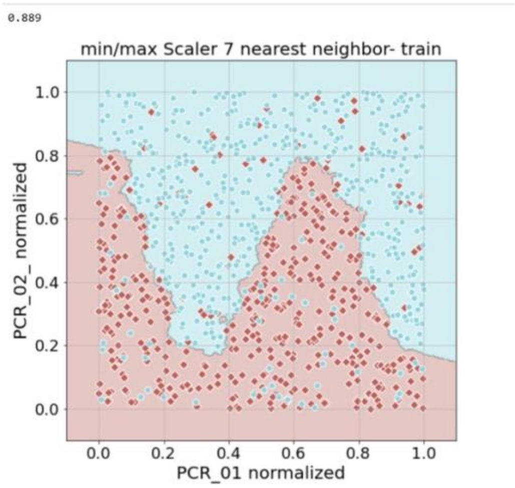
(Q9).

K value is essential requirement to create decent and accurate kNN model.
If k selected to be too low the model becomes too specific and fails to generalize well (overfitting).
If k selected to be too large the model becomes too generalized and fails to accurately predict the data points in both train and test sets.
(underfitting).



```
0.856
```

Accuracy – 0.856



```
0.889
```

Accuracy – 0.889

(Q10).

Normalizing this normally distributed feature using min-max scaling is a bad idea because In this case, most of the points are in the center, when we use minimum and maximum we will move the points found in the center to the edges, thereby increasing the influence of the edges and affecting the reliability of the model

Part 3 –Data Exploration:

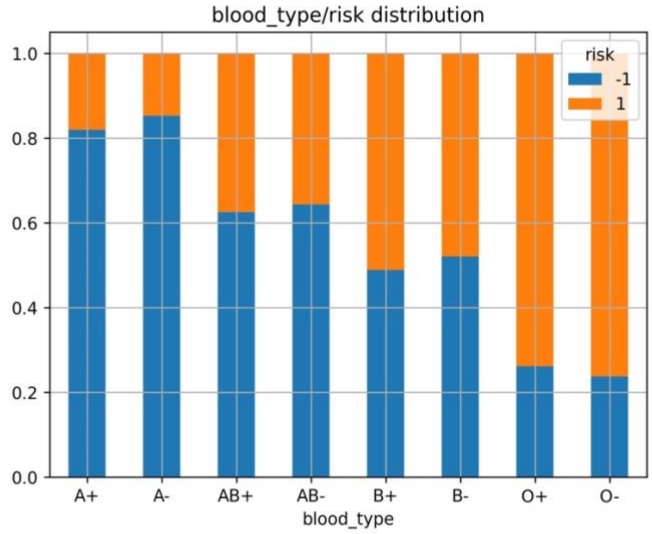
(Q11).

Number of features are needed to create OHE representation is 8.

(Q12).

Three group should be :
Group_A= {A+,A-}
Group_B={AB+,AB-,B+,B-}
Group_C={O+,O-}

The reason for dividing the groups is that in each group the ratio between risk and non risk is the closest of all other divisions.

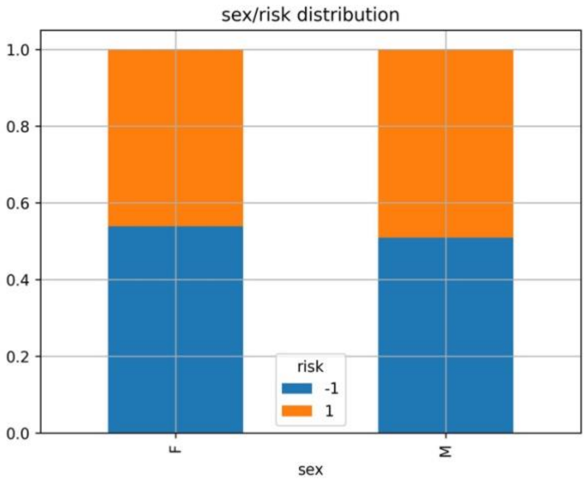


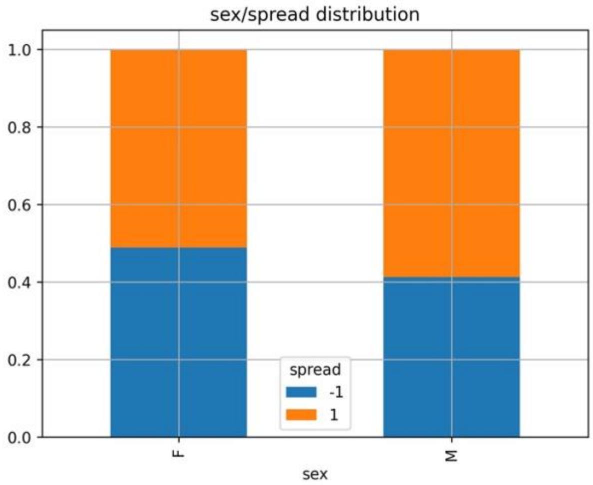
(Q13).

We can extract from the symptoms feature by turning it into categorical, for example by using the OHE method, creating a Boolean feature for each symptom(low_apetite, cough, shortness_of_breath, fever, sore_throat) and filling the rows according to the substrings found in the symptoms.

We can extract essential information from the current location by splitting the coordinate into x and y feature and then each one is continuing.

We can't extract essential information from the sex feature because there is no correlation between sex to risk and spread. Furthermore, patient_ feature dropped as well for the same reason.



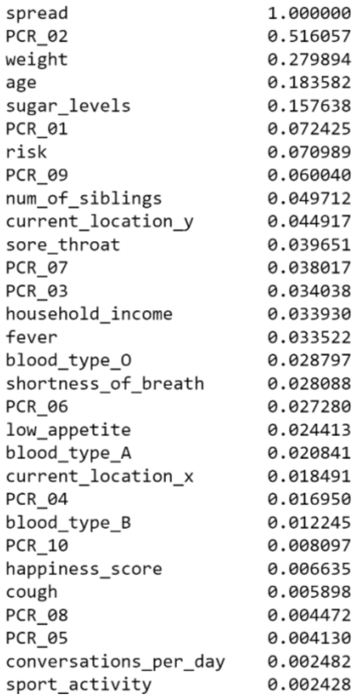


(Q14)

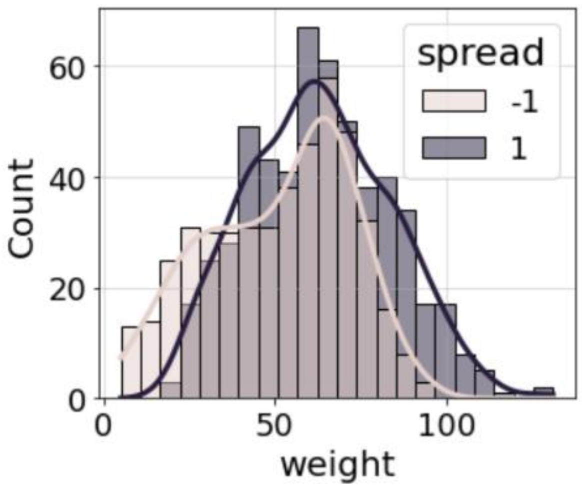
We think that the feature which has the most information on the spread target variable is weight for the following reasons:

- We can see, in the following graph that after the value of 80 in weight, one is most likely spreading the disease. And below 30 it is most likely that the person is not spreading.
  The change between those distributions is at 40~

- The correlation between weight feature and the spread feature is big in relation to the other features (in abs measure):



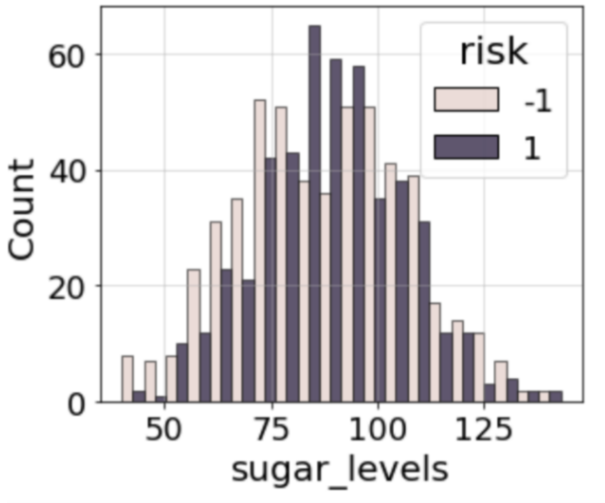
```
spread                  1.000000
PCR_02                  0.516057
weight                  0.279894
age                     0.183582
sugar_levels            0.157638
PCR_01                  0.072425
risk                    0.070989
PCR_09                  0.060040
num_of_siblings         0.049712
current_location_y      0.044917
sore_throat             0.039651
PCR_07                  0.038017
PCR_03                  0.034038
household_income        0.033930
fever                   0.033522
blood_type_O            0.028797
shortness_of_breath     0.028088
PCR_06                  0.027280
low_appetite            0.024413
blood_type_A            0.020841
current_location_x      0.018491
PCR_04                  0.016950
blood_type_B            0.012245
PCR_10                  0.008097
happiness_score         0.006635
cough                   0.005898
PCR_08                  0.004472
PCR_05                  0.004130
conversations_per_day   0.002482
sport_activity          0.002428
```



Q(15)

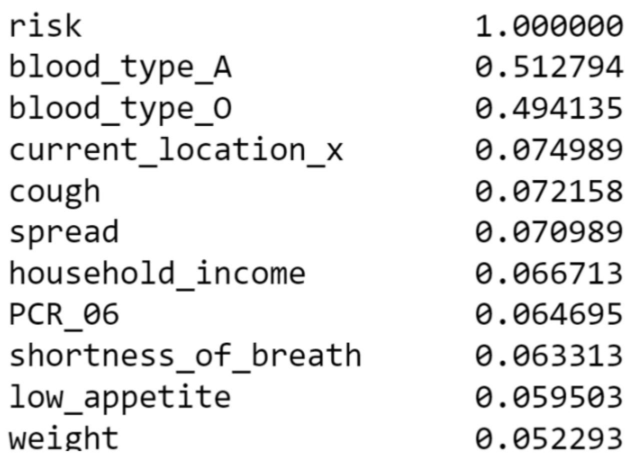
The feature we thought we the most informative to predict the 'risk' target variable is the sugar_levels feature for the following reasons:

- As we can see in the graph, when person has less than 70~ in sugar_levels, he most likely not to be at risk. And when person has between 80-90~ sugar_levels there is high chance that he is in risk.



Q(16)

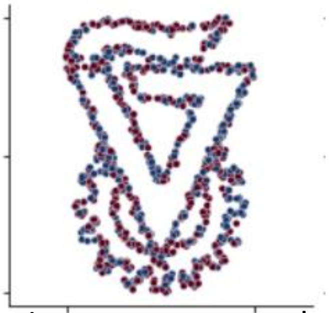
The 10 most correlated features to risk are:



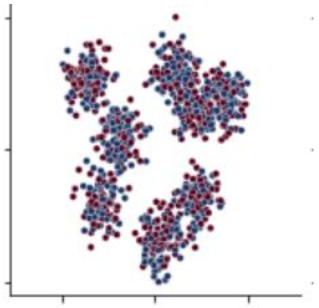
```
risk                     1.000000
blood_type_A             0.512794
blood_type_O             0.494135
current_location_x       0.074989
cough                    0.072158
spread                   0.070989
household_income         0.066713
PCR_06                   0.064695
shortness_of_breath      0.063313
low_appetite             0.059503
weight                   0.052293
```

Q(17)

The features PCR_03 and PCR_04 are performing very interesting structure->



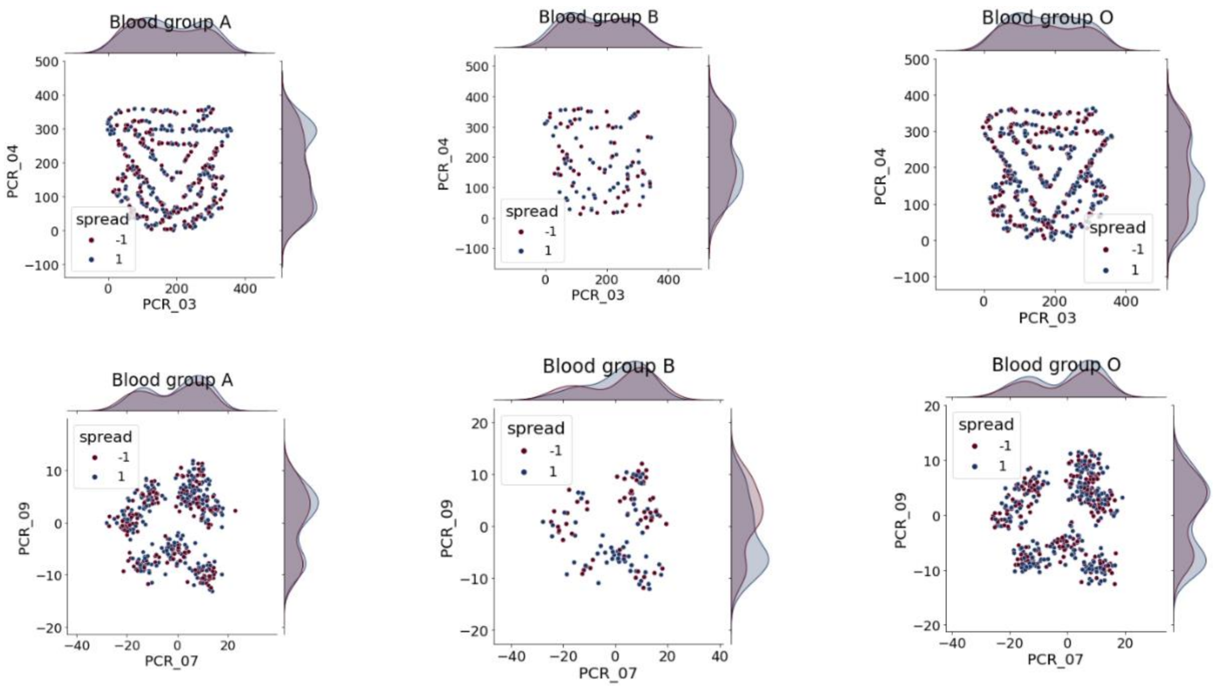
And another pair with interesting structure are the features: PCR_07 with PCR_09->



Neither one of the pairs gives us an indicative information. The distribution of the points on the graph on the doesn't have correlation to the target variable.

Q(18)

he three joint plots for each pair of chosen features for each blood group we created previously:



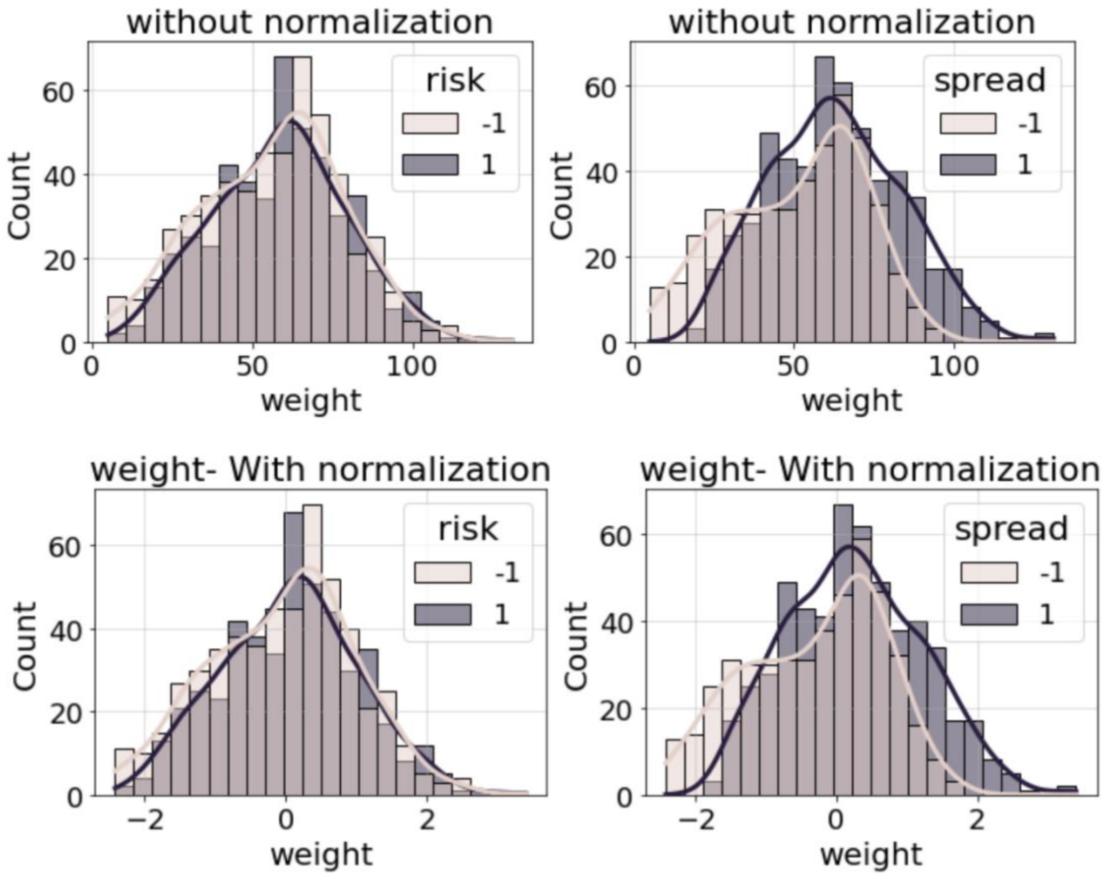
Q(19)

As you can see from the last graphs, there is high correlation between Type_blood_A (and even blood_type_O) to the risk chance. We MENTIONED it when we analyzed the features for each of the target variable.

Q(20)

We think due to our analyzation and experiments that the most suitable model to predict the risk is Decision Tree model.
As we mentioned in class KNN doesn't perform well when we have a lot of features in relation to the number of samples.
Furthermore, linear model won't make high accuracy either because there are not enough features correlation when we saw at the graph that we can use linear method.
Therefore, decision tree model is flexible enough and can use the big number of categorical features effectively to perform high accuracy to predict the risk.

Q(21)



Part 5 – Feature Selection :

(Q22).

In forward selection only d2 affects the complexity of the number of the models we have to train because d2 is the number of features we would add to our subset.
In backward selection both d1 and d2 affects the complexity of the number of the models we have to train because d2-d1 is the number of features we would remove from our subset.

a. In forward feature selection, we try to extract the most d2 informative features therefore we will do d2 iterations. Each iteration will run on decreasing number of features. The first iteration will go on d1 features the next one will go over d1-1 features and so on..
   The complexity series:

$$\sum_{i=1}^{d_2}(d_1-(i-1)) = \sum_{i=1}^{d_2}(d_1+1-i) = d_1d_2 - d_2 - \sum_{i=1}^{d2} i = d_1d_2 - \frac{d_2^2}{2} + \frac{d_2}{2} = \frac{d_2}{2}\left(2d_1 - d_2 + 1\right)$$

b. In backward feature selection we will operate d1-d2 iteration. Like the forward feature selection, in each iteration the method goes over d1-i+1 features. In different to the forward method, on each iteration the algorithm looks for the least informative feature.
   The complexity series:

$$\sum_{i=1}^{d_1-d_2}(d_1-(i-1)) = \sum_{i=1}^{d_1-d_2}(d_1+1-i) = d_1^2 - d_1d_2 + d_1 - d_2 - \sum_{i=1}^{d_1-d_2} i =$$

$$d_1^2 - d_1d_2 + d_1 - d_2 - \frac{d_1-d_2}{2}(1+d_1-d_2) = d_1^2 - d_1d_2 + \frac{d_1-d_2}{2} + \frac{-d_1^2+2d_1d_2-d_2^2}{2} = \frac{d_1^2-d_2^2+d_1-d_2}{2}$$

(Q23).

Three features that the SFS algorithm found are: weight, PCR_01 and PCR_02.
The feature we chose manually in Q14 is 'weight' and the features we chose in Q6 are PCR_01 and PCR_02 .

(Q24).

It is important to perform the normalization step before performing sequential feature selection because we want to get data into reasonable bounds and in this way every feature will have an equal chance to be selected.

(Q25).

The choice of a learning algorithm is matter in a sequential feature selection process because each learning algorithm focused on different values of the data for prediction (e.g., distance, variance) and we would like the data to be as adapted to the algorithm as possible and thus improve the prediction in the best possible way.

(Q26).

Feature name	Keep	New	Normalization method	explanation
Patient_id	X	-	-	patient_id doesn't share any information on our target variables
sex	X	-	-	We removed Sex because we saw no correlation to target variable using crosstab
age	V	X	MinMax	Age doesn't have normal distribution therefore MinMax scaler helps more than std scaler
weight	V	X	Standard	The distribution of weight feature is similar to normal distribution therefore we do std normalization. Found high correlation between the feature to target variable
Current_location	X	X	-	Use OHE to extract relevant data as follows:
Current_location_x	V	V	Standard	The location on X axis is similar to normal distribution therefore we used standard scaler. High risk and spread correlation
Current_location_y	V	V	MinMax	The location on X axis is not similar to normal distribution therefore we used MinMax scaler
Num_of_siblings	V	X	MinMax	Num_of_siblings doesn't have normal distribution therefore MinMax scaler helps more than std scaler

Conversations_per_day	V	X	MinMax	The right method of scailing categorical data is using MinMax scaler.
Sugar_levels	V	X	Standard	Sugar_levels is continuous feature therefore we chose Std scaler
Sport_activity	V	X	MinMax	The right method of scailing categorical data is using MinMax scaler.
Household_income	V	X	Standard	Houshold_income is continuous feature therefore we chose Std scaler. High risk correlation
Happiness_score	V	X	MinMax	The right method of scailing categorical data is using MinMax scaler.
PCR_01	V	X	MinMax	The distribution of PCR_01 is not similar to normal distribution therefore we chose MinMax scaler
PCR_02	V	X	MinMax	The distribution of PCR_02 is not similar to normal distribution therefore we chose MinMax scaler
PCR_03	V	X	MinMax	The distribution of PCR_03 is not similar to normal distribution therefore we chose MinMax scaler
PCR_04	V	X	MinMax	The distribution of PCR_04 is not similar to normal distribution therefore we chose MinMax scaler
PCR_05	V	X	MinMax	The distribution of PCR_05 is not similar to normal distribution therefore we chose MinMax scaler.

PCR_06	V	X	Standard	The distribution of PCR_06 is similar to normal distribution therefore we chose std scaler
PCR_07	V	X	MinMax	The distribution of PCR_07 is not similar to normal distribution therefore we chose MinMax scaler
PCR_08	V	X	Standard	The distribution of PCR_08 is similar to normal distribution therefore we chose std scaler
PCR_09	V	X	MinMax	The distribution of PCR_09 is not similar to normal distribution therefore we chose MinMax scaler
PCR_10	V	X	Standard	The distribution of PCR_10 is similar to normal distribution therefore we chose std scaler
symptoms	X	X	-	Extracted new data from this string/categorical feature as follows:
Low_appetite	V	V	-	We didn't scaled binary feature
cough	V	V	-	We didn't scaled binary feature
Shortness_of_breath	V	V	-	We didn't scaled binary feature
Fever	V	V	-	We didn't scaled binary feature
Sore_throat	V	V	-	We didn't scaled binary feature
Pcr_date	X	-	-	We found pcr_date no valuable to predict risk and spread target variables
Blood_type	X	X	-	We extracted new feature from blood type categorical feature into 3 feature:
Blood_type_A	V	V		We didn't scaled binary feature
Blood_type_B	V	V		We didn't scaled binary feature
Blood_type_O	V	V		We didn't scaled binary feature

- All PCR features have correlation in some way to target variable
- Extracted symptoms have high correlation and found them to help us to predict risk and spread.
- The blood type features, as you can see In Q16, have high correlation and we think this features will help us at the predicting assignment