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,

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 | Ordinal |
| 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) | Continuous |
| 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. We do not want to accidentally train on our test set. This compromises the data

and allows us to learn from the test set which represents unseen data.

Q5.

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

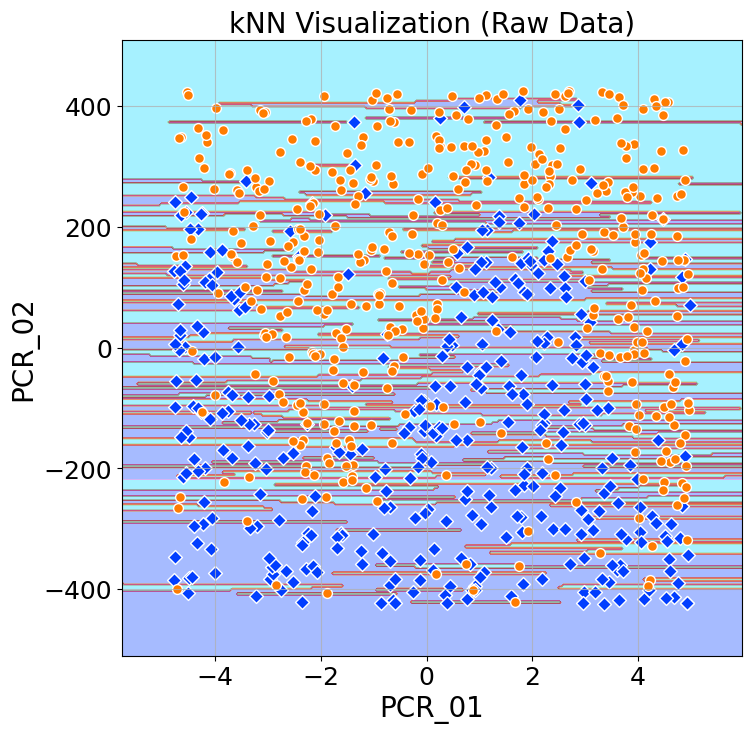
Q6.

Shape

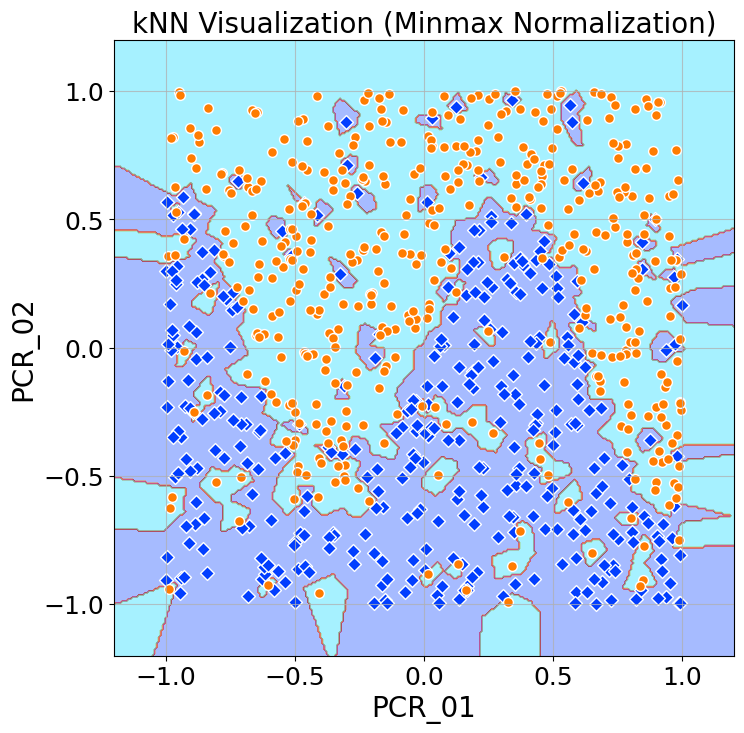
Description automatically generated

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.

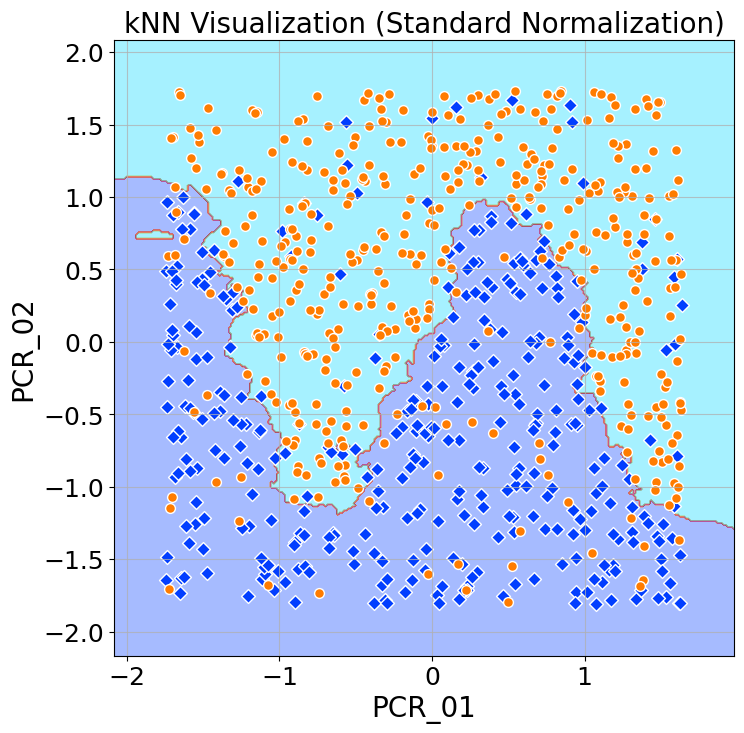


We achieved a training accuracy score of 1. This is logical since the nearest neighbor of each point is itself. We achieved a test accuracy score of 0.72.

Q8.

We achieved a training accuracy score of 1. We achieved a test accuracy score of 0.785.

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.

We achieved a training accuracy score of 0.68375. We achieved a test accuracy score of 0.65.

The increase of k has allowed for more loss in the training. This has created more contiguous decision regions and has also cause underfitting.

Q10.

The general problem with min-max scaling is that it does not deal well with outlier. 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 data is largely confined into one area of the total feature space. This greatly decreases the ability of kNN, as many of the datapoints are close even if they were not so close before the scaling.