Aprendizagem 2022/23

Homework II – Group 71

(ist1103471, ist1104139)

I. Pen-and-paper

1-a)

Priors: $p(y_6 = A) = \frac{4}{9}$ $\quad p(y_6 = B) = \frac{5}{9}$

PDFs: $p(y_1, y_2 | y_6)$ $\quad N(\mu_A = \begin{bmatrix} 0,275 \\ 0,52 \end{bmatrix}, \Sigma_A = \begin{bmatrix} 0,009 & 0,006 \\ 0,006 & 0,022 \end{bmatrix} | A)$

$N(\mu_B = \begin{bmatrix} 0,558 \\ 0,38 \end{bmatrix} \quad \Sigma_B = \begin{bmatrix} 0,023 & -0,016 \\ -0,016 & 0,037 \end{bmatrix} | B)$

$var(y_1|A) = \frac{0,24^2 + 0,16^2 + 0,32^2 + 0,38^2 - 4 \times 0,275^2}{4-1} \approx 0,0092$

$var(y_2|A) = 0,0224$ $\quad det(\Sigma_A) = 1,62 \times 10^{-4}$

$cov(y_1, y_2|A) = 0,0064$ $\quad det(\Sigma_B) = 5,95 \times 10^{-4}$

$var(y_1|B) = 0,02312$

$var(y_2|B) = 0,0374$ $\quad \Sigma_A^{-1} = \begin{bmatrix} 135,8 & 37,04 \\ 37,04 & 55,56 \end{bmatrix} \quad \Sigma_B^{-1} = \begin{bmatrix} 62,18 & 26,89 \\ 26,89 & 38,66 \end{bmatrix}$

$cov(y_1, y_2|B) = -0,0164$

PMFs:

$p(y_3, y_4 | y_6)$

$p(0,0|A) = 0$ $\quad p(0,0|B) = \frac{2}{5}$

$p(0,1|A) = \frac{2}{4} = \frac{1}{2}$ $\quad p(0,1|B) = \frac{2}{5}$

$p(1,0|A) = \frac{1}{4}$ $\quad p(1,0|B) = \frac{1}{5}$

$p(1,1|A) = \frac{1}{4}$ $\quad p(1,1|B) = 0$

$p(y_5|y_6)$

$p(0|A) = \frac{2}{4} = \frac{1}{2}$ $\quad p(0|B) = \frac{1}{5}$

$p(1|A) = \frac{1}{4}$ $\quad p(1|B) = \frac{3}{5}$

$p(2|A) = \frac{1}{4}$ $\quad p(2|B) = \frac{1}{5}$

1-b)

x_8: $p(0.38, 0.52, 0, 1, 0|A) =$

$= p(0.38, 0.52|A)\ p(0,1|A)\ p(0|A) =$

$= N\left(\begin{bmatrix}0,38\\0,52\end{bmatrix} \middle| \mu_1=\begin{bmatrix}0,275\\0,52\end{bmatrix}, \Sigma_A=\begin{bmatrix}0,009 & 0,006\\0,006 & 0,022\end{bmatrix}\right) \times \frac{1}{2} \times \frac{1}{2} =$

$= 5,915 \times \frac{1}{2} \times \frac{1}{2} \approx 1,479$

$p(0.38, 0.52, 0, 1, 0|B) =$

$= N\left(\begin{bmatrix}0,38\\0,52\end{bmatrix} \middle| \mu_B=\begin{bmatrix}0,558\\0,38\end{bmatrix}, \Sigma_B=\begin{bmatrix}0,023 & -0,016\\-0,016 & 0,037\end{bmatrix}\right) \times \frac{2}{5} \times \frac{1}{5} =$

$= 3,26 \times \frac{2}{5} \times \frac{1}{5} \approx 0,26$

$p(0.38, 0.52, 0, 1, 0|A) \times p(A) = 1,479 \times \frac{4}{9} \approx 0,657$

$p(0.38, 0.52, 0, 1, 0|A) \times p(B) = 0,26 \times \frac{5}{9} \approx 0,144$

$h_{MAP} = A$ x_8 classified as A //

$p(A|x_8) = \frac{0,657}{0,657+0,144} \approx 0,82$

$p(B|x_8) \approx 0,18$

x_9: $p(0.42, 0.59, 0, 1, 1|A) =$

$= N\left(\begin{bmatrix}0,42\\0,59\end{bmatrix} \middle| \mu_A=\begin{bmatrix}0,275\\0,52\end{bmatrix} \Sigma_A=\begin{bmatrix}0,009 & 0,006\\0,006 & 0,022\end{bmatrix}\right) \times \frac{1}{2} \times \frac{1}{4} =$

$= 3,812 \times \frac{1}{2} \times \frac{1}{4} \approx 0,477$

$p(0.42, 0.59, 0, 1, 1|B) =$

$= N\left(\begin{bmatrix}0,42\\0,59\end{bmatrix} \middle| \mu_B=\begin{bmatrix}0,558\\0,38\end{bmatrix}, \Sigma_B=\begin{bmatrix}0,023 & -0,016\\-0,016 & 0,037\end{bmatrix}\right) \times \frac{2}{5} \times \frac{3}{5} =$

$= 3,355 \times \frac{2}{5} \times \frac{3}{5} \approx 0,805$

$p(0.42, 0.59, 0, 1, 1|A) \times p(A) = 0,477 \times \frac{4}{9} \approx 0,212$

$p(0.42, 0.59, 0, 1, 1|B) \times p(B) = 0,805 \times \frac{5}{9} \approx 0,447$

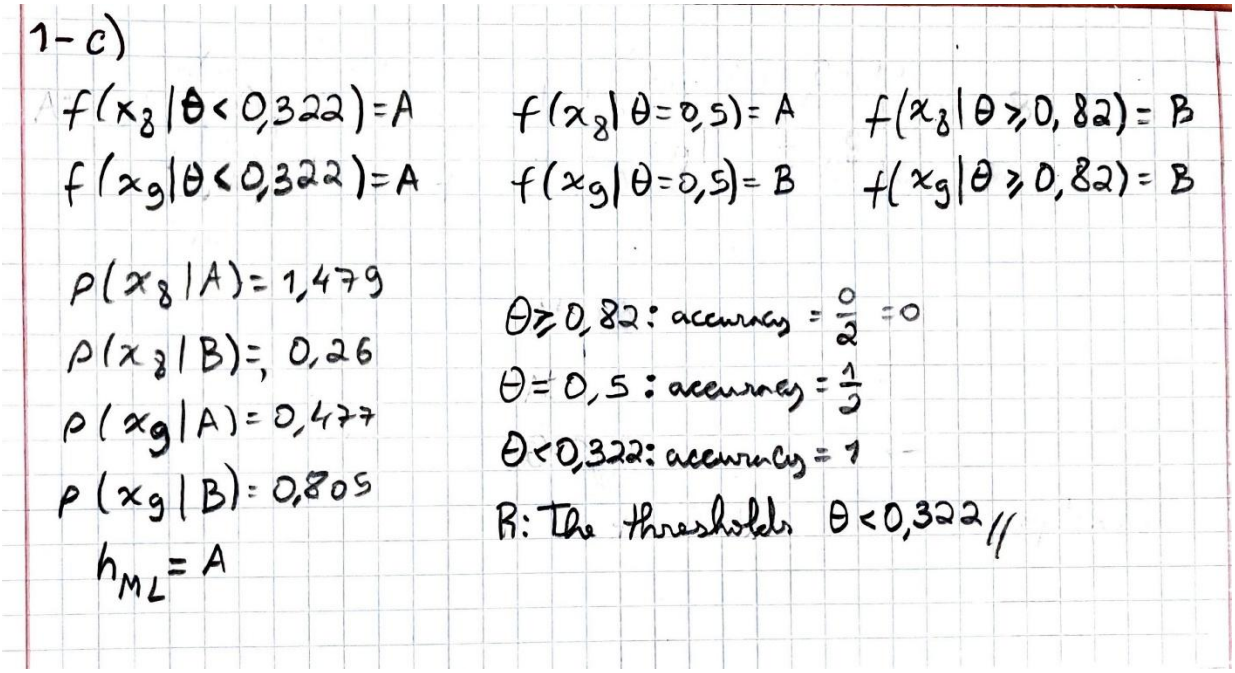
$h_{MAP} = B$ x_9 classified as B //

$p(A|x_9) = \frac{0,212}{0,212+0,447} \approx 0,322$

$p(B|x_9) \approx 0,678$

Homework II – Group 71

(ist1103471, ist1104139)



Homework II – Group 71

(ist1103471, ist1104139)

2.

a.

Y_2 [0,1]

Binarize → $Y_2'(X_i) = \begin{cases} 0 & , Y_2(X_i) \leq 0,5 \\ 1 & , Y_2(X_i) > 0,5 \end{cases}$

D	Y_1	Y_2'	Y_3	Y_4	Y_5	Y_6
X_1	0,21	0	1	1	0	A
X_2	0,16	0	1	0	1	A
X_3	0,32	1	0	1	2	A
X_4	0,54	0	0	0	1	B
X_5	0,66	0	0	0	0	B
X_6	0,76	0	1	0	2	B
X_7	0,91	1	0	1	1	B
X_8	0,38	1	0	1	0	A
X_9	0,42	1	0	1	1	B

In all folds the features are the variables Y_2'-Y_6.

Fold 3:

X_1-X_3 are the testing observations

X_4-X_9 are the training observations

Fold 2:

X_4-X_6 are the testing observations

X_1-X_3 and X_7-X_9 are the training observations

Fold 1:

X_7-X_9 are the testing observations

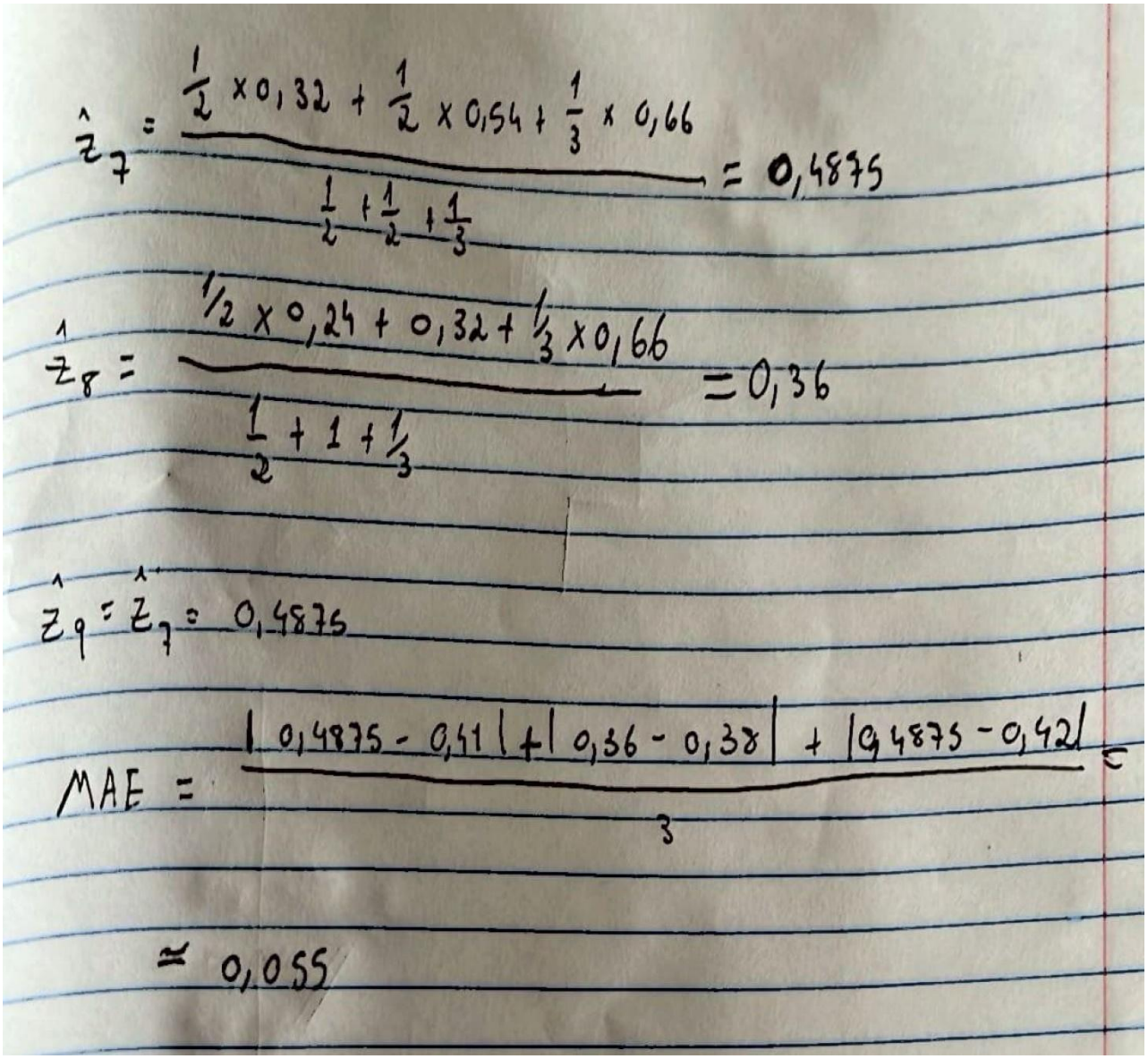
X_1-X_6 are the training observations

b)

$H(X, X_j)$	X_1	X_2	X_3	X_4	X_5	X_6
X_7	4	4	2	2	3	4
X_8	2	4	1	4	3	5
X_9	4	4	2	2	3	4

Homework II – Group 71

(ist1103471, ist1104139)



TÉCNICO LISBOA

Aprendizagem 2022/23

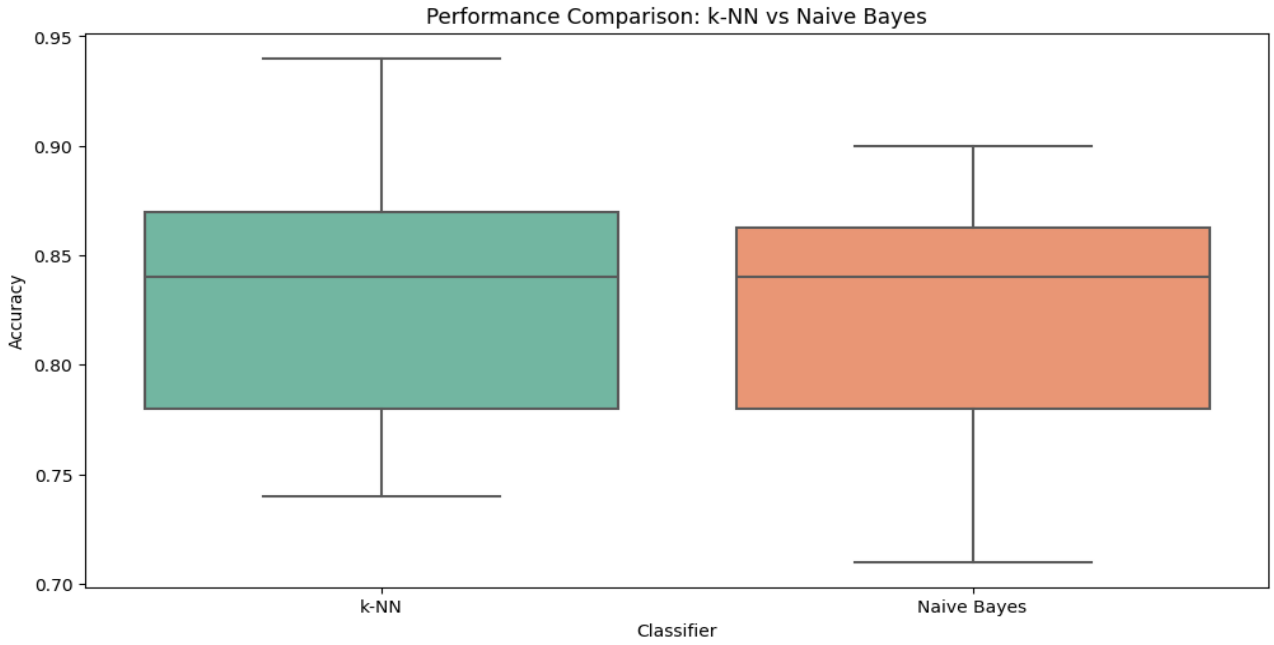
Homework II – Group 71

(ist1103471, ist1104139)

II. Programming and critical analysis

1) a)

```
knn = KNeighborsClassifier(n_neighbors=5)
nb = GaussianNB()
accuracies = {"k-NN": [], "Naive Bayes": []}

for train_k, test_k in folds.split(X, y):
    X_train, X_test = X.iloc[train_k], X.iloc[test_k]
    y_train, y_test = y.iloc[train_k], y.iloc[test_k]

    knn.fit(X_train, y_train)
    nb.fit(X_train, y_train)
    knn_pred = knn.predict(X_test)
    nb_pred = nb.predict(X_test)

    knn_accuracy = metrics.accuracy_score(y_test, knn_pred)
    nb_accuracy = metrics.accuracy_score(y_test, nb_pred)
    accuracies["k-NN"].append(round(metrics.accuracy_score(y_test, knn_pred),2))
    accuracies["Naive Bayes"].append(round(metrics.accuracy_score(y_test, nb_pred),2))


accs_df = pd.DataFrame(accuracies)
plt.figure(figsize=(12, 6))
sns.boxplot(data=accs_df, palette="Set2")
plt.title('Performance Comparison: k-NN vs Naive Bayes')
plt.xlabel('Classifier')
plt.ylabel('Accuracy')
plt.show()
```



b)

```
p_value = stats.ttest_ind(accuracies["k-NN"], accuracies["Naive Bayes"]).pvalue
print(f"P-value: {p_value}")
```

```
P-value: 0.5559851671434344
```

With the calculated p-value of approximately 0.35, we can conclude that there is not evidence enough to claim that k-NN is statistically superior to Naive Bayes, regarding accuracy based, on the observed data, thus rejecting the null hypothesis.

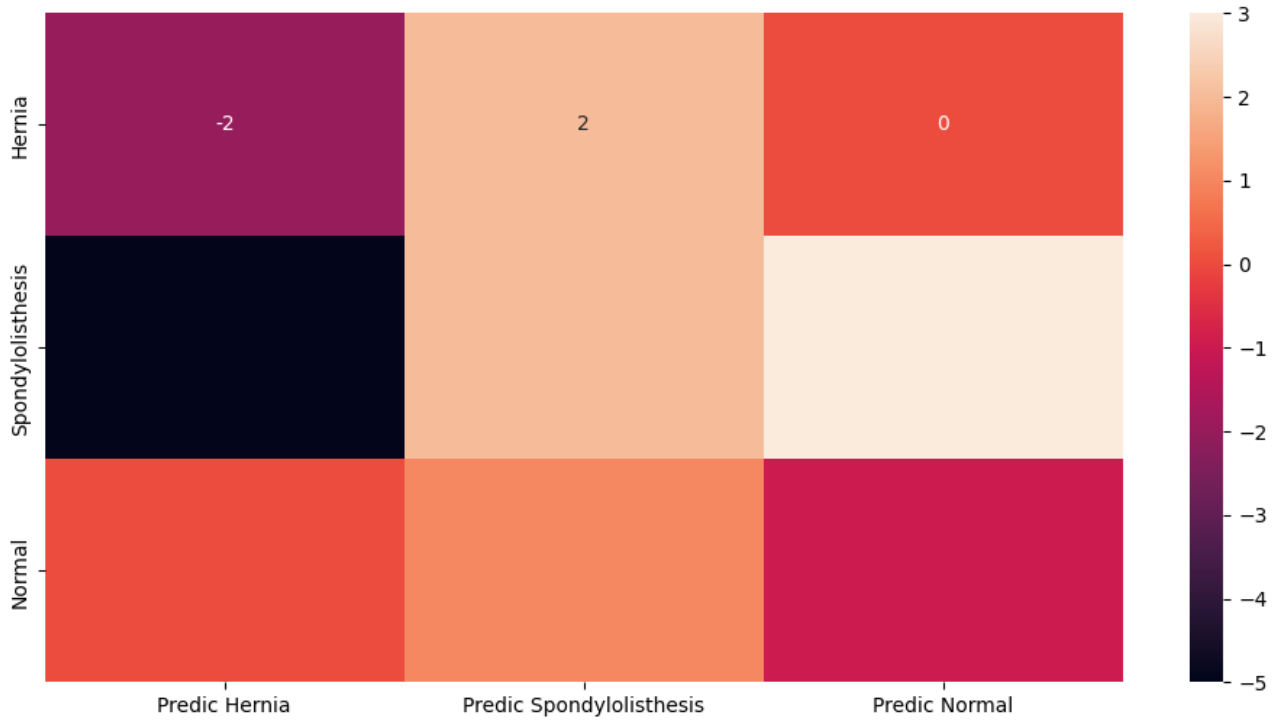
2)

```
knn_1 = KNeighborsClassifier(n_neighbors=1, weights='uniform', metric='euclidean')
knn_5 = KNeighborsClassifier(n_neighbors=5, weights='uniform', metric='euclidean')
conf_matrix_1 = np.zeros((3, 3))
conf_matrix_5 = np.zeros((3, 3))

for train_k, test_k in folds.split(X, y):
    X_train, X_test = X.iloc[train_k], X.iloc[test_k]
    y_train, y_test = y.iloc[train_k], y.iloc[test_k]

    knn_1.fit(X_train, y_train)
    knn_5.fit(X_train, y_train)

    knn_pred_1 = knn_1.predict(X_test)
    knn_pred_5 = knn_5.predict(X_test)

    conf_matrix_1 += confusion_matrix(y_test, knn_pred_1)
    conf_matrix_5 += confusion_matrix(y_test, knn_pred_5)

conf_matrix_diff = conf_matrix_1 - conf_matrix_5

plt.figure(figsize=(12, 6))
cm = pd.DataFrame(conf_matrix_diff, index=['Hernia', 'Spondylolisthesis', 'Normal'],
                  columns=['Predic Hernia', 'Predic Spondylolisthesis', 'Predic Normal'])
sns.heatmap(cm, annot=True, fmt='g')
```

Aprendizagem 2022/23

Homework II – Group 71

(ist1103471, ist1104139)



From the results given by the heatmap that represents the difference between the k-nearest neighbours method when k = 1 or 5, we can say that both predictors have a similar behaviour, because there is no big difference in any case. We can imply, that the neighbours from 1 to 5 yield similar outcomes, which could be because the data isn't very noisy nor complex.

3)

```
#1
correlations = spearmanr(X).statistic
print(correlations)
print("\n")

#3
Q1 = X.quantile(0.25)
Q3 = X.quantile(0.75)
IQR = Q3 - Q1
z_scores = zscore(X)
outliers = (X < (Q1 - 1.5 * IQR)) | (X > (Q3 + 1.5 * IQR))

outliers_count_per_feature = pd.DataFrame(outliers, columns=X.columns).sum()
print(outliers_count_per_feature)
print(len(X))
```

```
[[ 1.          0.65315204  0.77849937  0.80082955 -0.26654025  0.67077822]
 [ 0.65315204  1.          0.41874285  0.12405479 -0.00691726  0.39676684]
 [ 0.77849937  0.41874285  1.          0.69280319 -0.13418023  0.68362926]
 [ 0.80082955  0.12405479  0.69280319  1.         -0.36562411  0.5327264 ]
 [-0.26654025 -0.00691726 -0.13418023 -0.36562411  1.         -0.1742733 ]
 [ 0.67077822  0.39676684  0.68362926  0.5327264  -0.1742733   1.        ]]


pelvic_incidence            3
pelvic_tilt                13
lumbar_lordosis_angle       1
sacral_slope                1
pelvic_radius              11
degree_spondylolisthesis   10
dtype: int64
310
```

As we can see from the spearman correlation values between the features of column_diagnosis, there are high correlations values, meaning not all features are independent from each other, which is one of Naive Bayes method flaws, because it assumes that all features are independent.

Variables are not normally distributed, which is a Gaussian Naive Bayes prediction problem, because it assumes that all data is normalized.)

Finally, the Gaussian Naive Bayes, is sensitive to outliers, which by the values calculated aren't a lot, so it really isn't a difficulty for this data set. Nonetheless, we find important to check this numbers, due to the influence that they have in the statistics that generate the predictions. This happens because the calculations made in this method rely on mean and variance values.

END