Aprendizagem 2023/24

Homework I – Group 004

(ist102695, ist102703)

I. Pen-and-paper

1)

1-

a) $Y_6 = A \quad \vec{\mu} = \begin{bmatrix} 0,24 \\ 0,52 \end{bmatrix} \quad var(Y_1) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1|n} = \frac{\sum (x_i)^2 - n\bar{x}^2}{n-1/n} = \frac{0,24^2 + 0,16^2 + 0,32^2 - 3 \times 0,24^2}{2} = 0,0064$

$var(Y_2) = \frac{0,36^2 + 0,48^2 + 0,72^2 - 3 \times 0,52^2}{2} = 0,0336$

$cov(Y_1, Y_2) = \frac{\sum (Y_1 - \bar{Y}_1)(Y_2 - \bar{Y}_2)}{n-1} = \frac{(0,24-0,24)(0,36-0,52) + (0,16-0,24)(0,48-0,52) + (0,32-0,24)(0,72-0,52)}{2} = 0,0096$

$\Sigma_A = \begin{bmatrix} 0,0064 & 0,0096 \\ 0,0096 & 0,0336 \end{bmatrix}$

$Y_6 = B \quad \vec{\mu} = \begin{bmatrix} 0,5925 \\ 0,3275 \end{bmatrix}$

$var(Y_1) = \frac{0,54^2 + 0,66^2 + 0,76^2 + 0,41^2 - 4 \times 0,5925^2}{3} = 0,022892$

$var(Y_2) = \frac{0,11^2 + 0,39^2 + 0,28^2 + 0,53^2 - 4 \times 0,3275^2}{3} = 0,031492$

$cov(Y_1, Y_2) = \frac{(0,54 - 0,5925)(0,11 - 0,3275) + (0,66 - 0,5925)(0,39 - 0,3275)}{3} = -0,009758$

$\Sigma_B = \begin{bmatrix} 0,022892 & -0,009758 \\ -0,009758 & 0,031492 \end{bmatrix}$

$P((x_1, x_2) | A) = N\left(\mu = \begin{bmatrix} 0,24 \\ 0,52 \end{bmatrix}, \Sigma = \begin{bmatrix} 0,0064 & 0,0096 \\ 0,0096 & 0,0336 \end{bmatrix}\right) \qquad P(x_5 | A) \rightarrow P(0|A) = 1/3 \quad P(1|A) = 1/3 \quad P(2|A) = 1/3$

$P((x_1, x_2) | B) = N\left(\mu = \begin{bmatrix} 0,5925 \\ 0,3275 \end{bmatrix}, \Sigma = \begin{bmatrix} 0,022289 & -0,00976 \\ -0,00976 & 0,03149 \end{bmatrix}\right) \qquad P(x_5 | B) \rightarrow P(0|B) = 1/4 \quad P(1|B) = 1/2 \quad P(2|B) = 1/4$

$P((Y_3, Y_4) | A) \rightarrow P(0,0|A) = 0 \quad P(1,0|A) = 1/3 \quad P(0,1|A) = 1/3 \quad P(1,1|A) = 1/3$

$P((Y_3, Y_4) | B) \rightarrow P(0,0|B) = 1/2 \quad P(1,0|B) = 1/4 \quad P(0,1|B) = 1/4 \quad P(1,1|B) = 0$

b) $(x_8) \qquad N(\vec{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi \times det(\Sigma)}} \times e^{-(\vec{x} - \vec{\mu})^T \times \Sigma^{-1} \times (\vec{x} - \vec{\mu})}$

$Y_6 = A \qquad P(Y_6 = A | x_8) = P(Y_6 = A) P(x_8 | Y_6 = A) =$

$= P(Y_6 = A) \times P(Y_1 = 0,38, Y_2 = 0,52 | Y_6 = A) \times P(Y_3 = 0, Y_4 = 1 | Y_6 = A) \times P(Y_5 = 0 | Y_6 = A) =$

$= \frac{3}{7} \times N\left(\begin{bmatrix} 0,38 \\ 0,52 \end{bmatrix} \middle| \mu = \begin{bmatrix} 0,24 \\ 0,52 \end{bmatrix}, \Sigma = \begin{bmatrix} 0,0064 & 0,0096 \\ 0,0096 & 0,0336 \end{bmatrix}\right) \times 1/3 \times 1/3 =$

$Y_1 \times Y_2 \in \mathbb{R}^2$ is normally distributed

$= 1/21 \times \frac{1}{2\pi \sqrt{det(\Sigma)}} \times e^{-[((0,38 \ 0,52) - (0,24 \ 0,52)) \Sigma^{-1} ((0,38 \ 0,52) - (0,24 \ 0,52))]} =$

$= 0,04689$

$Y_6 = B$

$P(Y_6 = B | x_8) = P(Y_6 = B) P(x_8 | Y_6 = B) =$

$= P(Y_6 = B) \times P(Y_1 = 0,38, Y_2 = 0,52 | Y_6 = B) \times P(Y_3 = 0, Y_4 = 1 | Y_6 = B) \times P(Y_5 = 0 | Y_6 = B) =$

Homework I – Group 004

(ist102695, ist102703)

$= 4/7 \times N\left(\begin{bmatrix} 0,38 \\ 0,52 \end{bmatrix} \middle| \mu = \begin{bmatrix} 0,5925 \\ 0,3275 \end{bmatrix}, \Sigma = \begin{bmatrix} 0,02289 & -0,00976 \\ -0,00976 & 0,03149 \end{bmatrix} \right) \times 1/4 \times 1/4 = \frac{1}{28} \times 1,9625 = 0,07009$

$det(\Sigma_B) = 0,0006255 \qquad \Sigma_B^{-1} = \begin{bmatrix} 50,3398 & 15,6023 \\ 15,6023 & 36,5914 \end{bmatrix}$

→ $P(y_6 = B \mid x_6)$ é superior a $y_6 = A$ logo, por MAP assumption, a classe é B

x_9

$y_6 = A$

$P(y_6 = A \mid x_9) = P(y_6 = A)\, P(x_9 \mid y_6 = A) =$

$= P(y_6 = A) \times P(y_1 = 0,42, y_2 = 0,59 \mid y_6 = A) \times P(y_3 = 0, y_4 = 1 \mid y_6 = A) \times P(y_5 = 1 \mid y_6 = A) =$

$= \frac{3}{7} \times N\left(\begin{bmatrix} 0,42 \\ 0,59 \end{bmatrix} \middle| \mu = \begin{bmatrix} 0,24 \\ 0,52 \end{bmatrix}, \Sigma = \begin{bmatrix} 0,0064 & 0,0096 \\ 0,0096 & 0,0336 \end{bmatrix} \right) \times 1/3 \times 1/3 = 0,01919$

$y_6 = B$

$P(y_6 = B \mid x_9) = P(y_6 = B)\, P(x_9 \mid y_6 = B) =$

$= P(y_6 = B) \times P(y_1 = 0,42, y_2 = 0,59 \mid y_6 = B) \times P(y_3 = 0, y_4 = 1 \mid y_6 = B) \times P(y_5 = 1 \mid y_6 = B) =$

$= 4/7 \times N\left(\begin{bmatrix} 0,42 \\ 0,59 \end{bmatrix} \middle| \mu = \begin{bmatrix} 0,5925 \\ 0,3275 \end{bmatrix}, \Sigma = \begin{bmatrix} 0,02289 & -0,00976 \\ -0,00976 & 0,03149 \end{bmatrix} \right) \times 1/4 \times 2/4 = 0,12348$

→ $P(y_6 = B \mid x_9)$ é superior logo, por MAP assumption a classe é B

Homework I – Group 004

(ist102695, ist102703)

c) Segundo uma Maximum Likelihood assumption podemos dizer que $P(c|x) = P(x|c)$

$$P(y_6 = A | x_8) = P(x_8 | y_6 = A) = \frac{0,04689}{3/7} = 0,10941$$

$$P(y_6 = B | x_8) = P(x_8 | y_6 = B) = \frac{0,07009}{4/7} = 0,1227$$

$\rightarrow \sum P(h|x) = 1$

normalizando:

$$\frac{P(y_6 = A | x_8)}{P(y_6 = A | x_8) + P(y_6 = B | x_8)} = 0,47 \rightarrow P(y_6 = A | x_8)$$

$$\frac{P(y_6 = B | x_8)}{P(y_6 = A | x_8) + P(y_6 = B | x_8)} = 0,53 \rightarrow P(y_6 = B | x_8)$$

} normalizar

$$P(y_6 = A | x_9) = P(x_9 | y_6 = A) = 0,0448$$

$$P(y_6 = B | x_9) = P(x_9 | y_6 = B) = 0,2161$$

$\rightarrow \sum P(h|x) = 1$

normalizando:

$$\frac{P(y_6 = A | x_9)}{P(y_6 = A | x_9) + P(y_6 = B | x_9)} = 0,17 \rightarrow P(y_6 = A | x_9)$$

$$\frac{P(y_6 = B | x_9)}{P(y_6 = A | x_9) + P(y_6 = B | x_9)} = 0,83 \rightarrow P(y_6 = B | x_9)$$

} normalizar

- x_8 deve ser classificado como A e x_9 como B
- Se $\theta \in]0,17; 0,17[$ a accuracy seria de 100%, ou seja, x_8 seria classificado como A porque $P(y_6 = A | x_8) > \theta$ e x_9 seria classificado como B porque $P(y_6 = A | x_9) < \theta$

Homework I – Group 004

(ist102695, ist102703)

2)

se $y_2 < 0.5 \rightarrow y_2 = 0$

se $y_2 > 0.5 \rightarrow y_2 = 1$

Data fold 1

	y_1	y_2	y_3	y_4	y_5	y_6
x_1	0.124	0	1	1	0	A
x_2	0.16	0	1	0	1	A
x_3	0.32	1	0	1	2	A

Data fold 2

	y_1	y_2	y_3	y_4	y_5	y_6
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

Homework I – Group 004

(ist102695, ist102703)

Data fold 3

	y_1	y_2	y_3	y_4	y_5	y_6
x_7	0.41	1	0	1	1	B
x_8	0.38	1	0	1	0	A
x_9	0.42	1	0	1	1	B

b)

① x_7

$d(x_7, x_1) = 4$ $\quad 3NN = (2, 2, 3)$

$d(x_7, x_2) = 4$

$d(x_7, x_3) = (2)$ $\quad \hat{z} = \frac{1/2 \times 0.32 + 1/2 \times 0.54 \times 1/3 \times 0.66}{1/2 + 1/2 + 1/3}$

$d(x_7, x_4) = (2)$

$d(x_7, x_5) = (3)$ $\quad = 0.4875$

$d(x_7, x_6) = 4$

Homework I – Group 004

(ist102695, ist102703)

② x_8

$3NN = (2, 1, 3)$

$d(x_8, x_1) = 2$

$d(x_8, x_2) = 4$

$d(x_8, x_3) = 1$

$d(x_8, x_4) = 4$

$d(x_8, x_5) = 3$

$d(x_8, x_6) = 5$

$$\hat{z} = \frac{\frac{1}{2} \times 0.24 + \frac{1}{1} \times 0.32 + \frac{1}{3} \times 0.66}{\frac{1}{2} + \frac{1}{1} + \frac{1}{3}} = 0.36$$

③ x_9

$3NN = (2, 2, 3)$

$d(x_9, x_1) = 4$

$d(x_9, x_2) = 4$

$d(x_9, x_3) = 2$

$d(x_9, x_4) = 2$

$d(x_9, x_5) = 3$

$d(x_9, x_6) = 4$

$$\hat{z} = \frac{\frac{1}{2} \times 0.32 + \frac{1}{2} \times 0.54 + \frac{1}{3} \times 0.66}{\frac{1}{2} + \frac{1}{2} + \frac{1}{3}} = 0.4875$$

Homework I – Group 004

(ist102695, ist102703)

$$z - \hat{z} = (0.41 - 0.4875, 0.38 - 0.36, 0.42 - 0.4875)$$

$$= (-0.0775, 0.02, -0.0775)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |z_i - \hat{z}_i|$$

$$= \frac{1}{3} (|-0.0775| + |0.02| + |-0.0775|)$$

$$= 0.0583$$

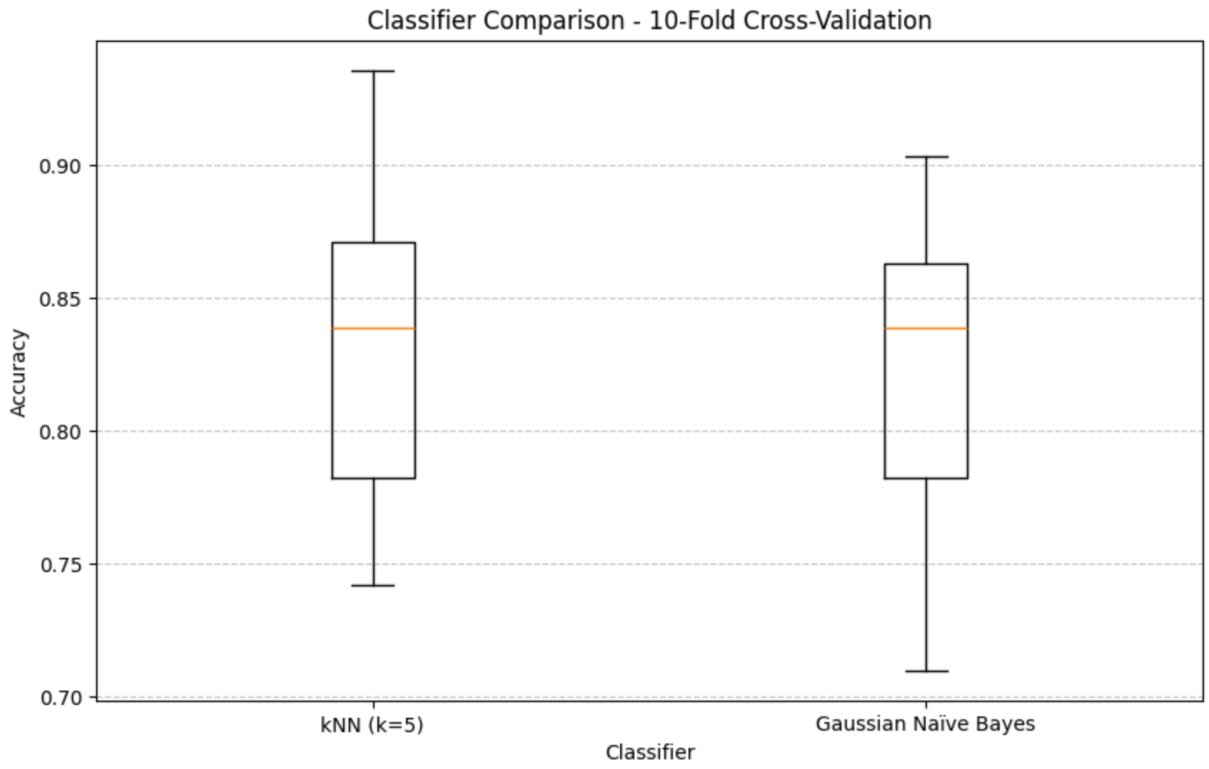
II. Programming and critical analysis

1)

a.

```
#1
#a)
import pandas as pd
from scipy.io.arff import loadarff
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import StratifiedKFold
from sklearn.preprocessing import LabelEncoder
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score

data = loadarff('column_diagnosis.arff')
df = pd.DataFrame(data[0])
df['class'] = df['class'].str.decode('utf-8')

X = df.drop('class', axis=1)
y = df['class']

kf = StratifiedKFold(n_splits=10, shuffle=True, random_state=0)

knn_classifier = KNeighborsClassifier (weights="uniform", n_neighbors =5, metric="euclidean")
nb_classifier = GaussianNB ()

knn_accuracy_scores = []
nb_accuracy_scores = []

for train_index, test_index in kf.split(X, y):
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]

    knn_classifier.fit(X_train, y_train)
    knn_y_pred = knn_classifier.predict(X_test)
    knn_accuracy = accuracy_score(y_test, knn_y_pred)
    knn_accuracy_scores.append(knn_accuracy)

    nb_classifier.fit(X_train, y_train)
    nb_y_pred = nb_classifier.predict(X_test)
    nb_accuracy = accuracy_score(y_test, nb_y_pred)
    nb_accuracy_scores.append(nb_accuracy)

plt.figure(figsize=(10, 6))
plt.boxplot([knn_accuracy_scores, nb_accuracy_scores], labels=["kNN (k=5)", "Gaussian Naïve
Bayes"])
plt.title("Classifier Comparison - 10-Fold Cross-Validation")
```

```
plt.ylabel("Accuracy")
plt.xlabel("Classifier")
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



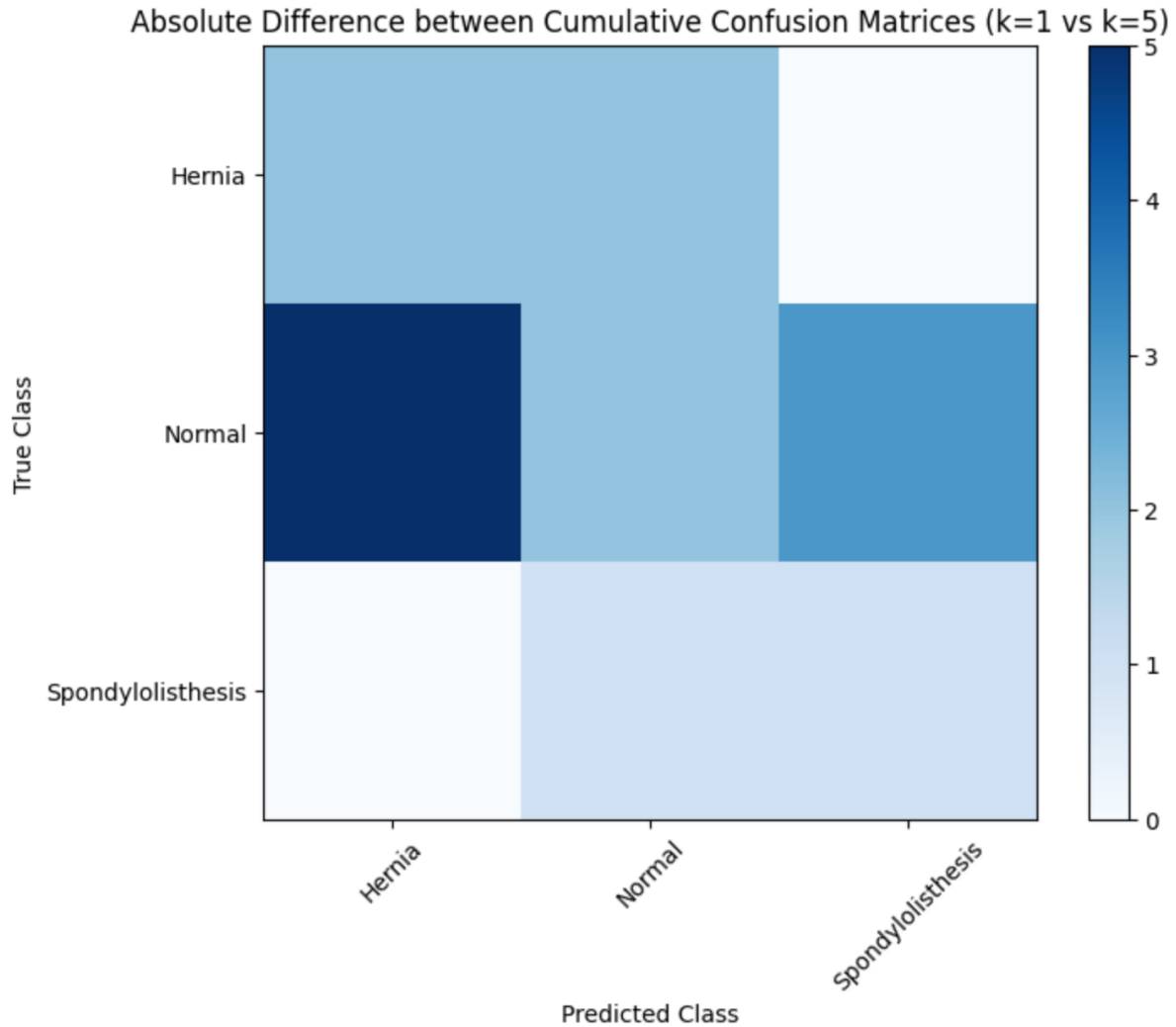

```
#b)
from scipy import stats

res = stats.ttest_rel(nb_accuracy_scores ,knn_accuracy_scores, alternative ="greater")

alpha = 0.05

if res.pvalue < alpha:
    print("Rejeitar a hipótese nula: kNN é estatisticamente superior a Naive Bayes em relação a accuracy.")
else:
    print("Falha ao rejeitar a hipótese nula:não há uma diferença significativa entre a accuracy de kNN e Naive Bayes.")
```

```
Falha ao rejeitar a hipótese nula:não há uma diferença significativa entre a accuracy de kNN e Naive Bayes.
```

2)

```
#2
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import StratifiedKFold
from scipy.io import arff
from sklearn.preprocessing import LabelEncoder

# Load ARFF file
data, meta = arff.loadarff('column_diagnosis.arff')

# Extract features and labels from the loaded ARFF data
X = np.array([list(data[i])[:-1] for i in range(len(data))])
y = np.array([data[i][-1].decode('utf-8') for i in range(len(data))])

# Initialize label encoder and encode the class labels
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)

# Initialize k-NN classifiers with k=1 and k=5
knn1 = KNeighborsClassifier(n_neighbors=1)
knn5 = KNeighborsClassifier(n_neighbors=5)

# Get the number of unique classes
num_classes = len(np.unique(y_encoded))

# Initialize StratifiedKFold with 10 folds and shuffling
stratkf = StratifiedKFold(n_splits=10, shuffle=True, random_state=0)

# Initialize cumulative confusion matrices
cumulative_cm1 = np.zeros((num_classes, num_classes))
cumulative_cm5 = np.zeros((num_classes, num_classes))

# Perform 10-fold cross-validation
for train_index, test_index in stratkf.split(X, y_encoded):
    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = y_encoded[train_index], y_encoded[test_index]

    # Train the classifiers
    knn1.fit(X_train, y_train)
    knn5.fit(X_train, y_train)

    # Make predictions
    y_pred1 = knn1.predict(X_test)
    y_pred5 = knn5.predict(X_test)

    # Calculate confusion matrices
```

```
    cm1 = confusion_matrix(y_test, y_pred1)
    cm5 = confusion_matrix(y_test, y_pred5)

    # Update cumulative confusion matrices
    cumulative_cm1 += cm1
    cumulative_cm5 += cm5

# Compute the absolute difference between cumulative confusion matrices
diff_confusion_matrix = np.abs(cumulative_cm1 - cumulative_cm5)

# Plot the difference matrix
plt.figure(figsize=(8, 6))
plt.imshow(diff_confusion_matrix, interpolation='nearest', cmap=plt.cm.Blues)
plt.title('Absolute Difference between Cumulative Confusion Matrices (k=1 vs k=5)')
plt.colorbar()
plt.xlabel('Predicted Class')
plt.ylabel('True Class')
plt.xticks(np.arange(num_classes), label_encoder.classes_, rotation=45)
plt.yticks(np.arange(num_classes), label_encoder.classes_)
plt.show()
```



2. Comentário:

Existe um desequilíbrio no True Class para a classe "Normal" pois conseguimos ver que existem bastantes amostras. Isto é visível pela cor azul mais escura. Achamos que ambos os modelos têm uma performance parecida, qual se queira escolher irá depender somente de qual é a finalidade: se é menos importante o erro de encontrar mais pessoas que estejam doentes e na verdade estão normais ou o contrário. Ou seja, se o custo de falsos positivos é alto, então a precisão pode ser uma métrica mais importante a ser considerada.

3)

3.

Suposição de recursos independentes:

Naive Bayes assume que todos os recursos são independentes uns dos outros, dado o rótulo da classe. Em conjuntos de dados do mundo real, especialmente em contextos médicos ou biológicos, as características podem frequentemente ser correlacionadas. Por exemplo, neste conjunto de dados, certas condições como as pelvic_incidence, pelvic_tilt, lumbar_lordosis_angle, sacral_slope e pelvic_radius podem estar relacionadas. No entanto, se essas características não forem verdadeiramente independentes, isso viola a suposição de Naive Bayes, levando a previsões tendenciosas.

Distribuição desequilibrada de classes:

Se as classes no conjunto de dados estiverem desequilibradas, o que significa que algumas classes têm significativamente mais instâncias do que outras, isso poderá distorcer os resultados da previsão. Em conjuntos de dados médicos, a distribuição desequilibrada de classes é comum porque algumas condições são mais raras que outras. Previsões tendenciosas podem ser problemáticas, especialmente em diagnósticos médicos onde todas as classes deveriam idealmente ser tratadas com igual importância.

Sensível a recursos irrelevantes:

Naive Bayes considera todas as características igualmente importantes durante a classificação. Se houver recursos irrelevantes no conjunto de dados, eles poderão impactar negativamente a precisão da classificação. Neste conjunto de dados, se existirem características irrelevantes relacionadas com a demografia do paciente, por exemplo, podem não contribuir significativamente para o diagnóstico, mas podem introduzir ruído.

END