Fusão de Informação em Análise de Dados - Projeto 2

Mariana Lopes Paulino[1 [2020190448]] and Rui Alexandre Tapadinhas[2 [2018283200]]

Faculdade de Ciências e Tecnologias da Universidade de Coimbra
{marianapaulino,rui}@student.dei.uc.pt

Abstract. This report outlines the project on the stratification of COVID-19 patients, conducted in the Data Fusion in Data Analysis class. The primary objective is to develop a decision model to assist health professionals in determining whether a patient with suspected COVID-19 should be hospitalized or sent home based on initial admission data. The decision model employs a Bayesian approach to integrate various patient data, including gender, age, marital status, vaccination status, breathing difficulty, heart rate, blood pressure, and temperature. Additionally, clinical guidelines are incorporated into the model to enhance decision accuracy. The report covers the dataset used, the methodology for data fusion, and the evaluation of the model's performance. Key considerations include the suitability of input variables, the distinction between discrete and continuous variables, and the assumptions of conditional probabilities.

Keywords: Data Fusion · Information Fusion · Bayes Fusion · Bayesian Inference · Multi Sensor Fusion · COVID-19 · Patients.

1 Introduction

In this project, we aim to predict which patients have to remain at the hospital for further examination or treatment after being admitted to the emergency room just with some variables/parameters acquired when admitted that are low cost and simple to obtain. By using only these variables, we aim to avoid expensive exams and keep only the patients needing care at the hospital.

2 Dataset

The dataset used has data and measurements about 599 patients admitted to the emergency room with suspected COVID-19. The features present in the dataset are gender, age, marital status, if the patient is vaccinated or not, breathing difficulties, heart rate, blood pressure, temperature, the target feature (Decision about returning home or stay at the hospital) and a feature called clinical guideline that is defined considering a rule which is the following:

```
IF
    Breathing Difficulty >= Moderate (2)
AND
    Temperature > 37.8
THEN
    Stay at hospital
```

Table 1. Dataset Attributes

Variable	**Name**	**Range**	**Description**
X1	Gender	0,1	Female, Male
X2	Age	[34, 99]	Age of the patient in years
X3	Marital Status	0,1	Single, Married
X4	Vaccinated	0,1	No, Yes
X5	Breathing Difficulty	0,1,2,3	None, Some, Moderate, High
X6	Heart Rate	[38, 272]	Unit: bpm
X7	Blood Pressure	[115, 164]	Unit: mm Hg
X8	Temperature	[36.00, 38.98]	Unit: ºC
X9	Clinical Guidelines	0,1	A rule based on Breathing Difficulty and Temperature.
T	Decision	0,1	Return Home, Stay at Hospital

2.1 Visualizations of the Dataset

Before starting implementing and calculating the probabilities and Bayes Inference in our project, we made some visualizations that clearly helped us understand the distribution of the **continuous** variables. We only made the visualizations for the continuous variables because the discrete ones only take two to three values and wouldn't make an interesting visualization, and we can't conclude anything on the distributions of those variables.

Table 2. Discrete and Continuous Variables

Discrete	**Continuous**
Gender	Age
Marital Status	Heart Rate
Vaccinated	Blood Pressure
Breathing Difficulty	Temperature
Clinical Guidelines	

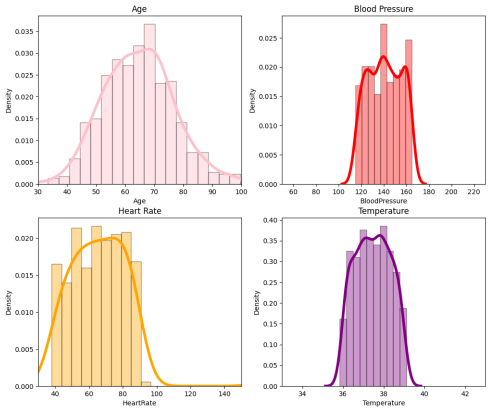

Fig. 1. Distribution of the Continuous Variables

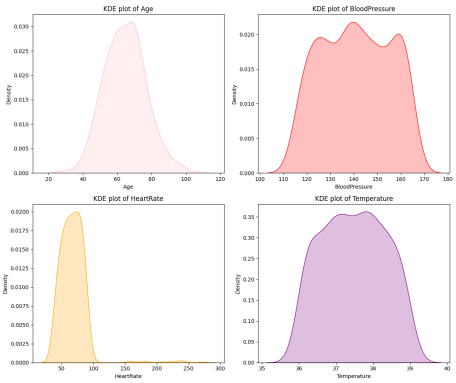

Fig. 2. KDE Graphic of the Continuous Variables

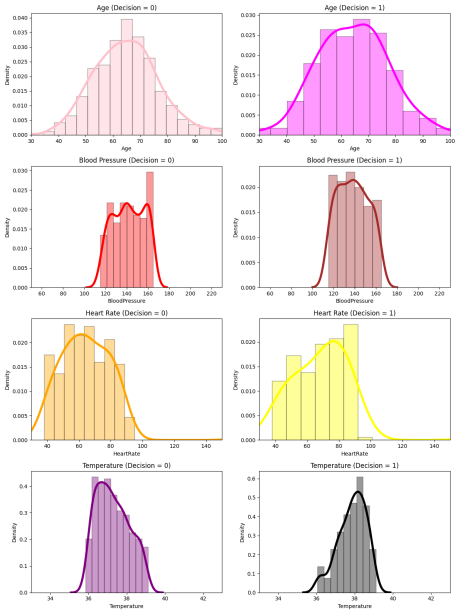

Fig. 3. Distribution of the Continuous Variables by Decision

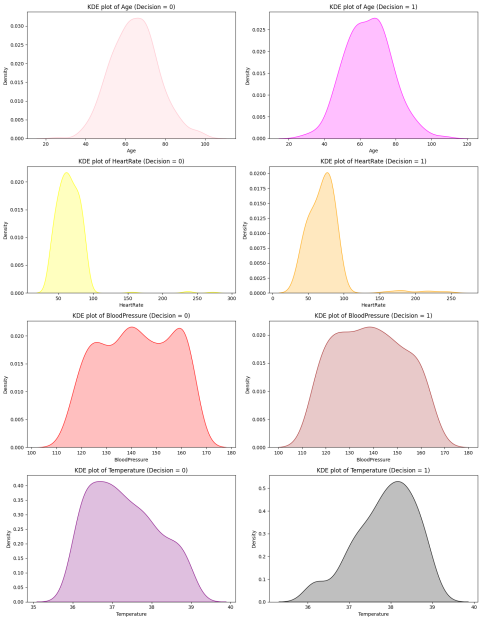

Fig. 4. KDE Graphic of the Continuous Variables by Decision

The two figures on the first line in the page above (**Figure 1** and **Figure 2**) represent the distribution of the continuous variables along their range. To take visualizations further along, we decided to see the differences in the distribution of those variables but having a separation in the decision class whether it is classified as **0 (Return Home)** or **1 (Stay Hospitalized)**. In the second row of figures, (**Figure 3** and **Figure 4**) want to describe the distributions of the classes of the **decision** variable. In those graphics, we separated the left side for when the decision was equal to 0, and the right side for when the decision was equal to 1. Those are the decompositions of the distributions above.

3 Implementation

To implement a Bayes Decision Model, some requirements are:

3.1 Model Formulation

Defining the Decision Variable (T - Decision): The target variable T is binary, indicating whether the patient should return home (0) or stay in the hospital (1).

Bayesian Framework: Use Bayes' theorem to calculate the posterior probability of the decision variable T using the formula:

$$P(T \mid X) = \frac{P(X \mid T)P(T)}{P(X)}$$

where X represents the vector of input features (gender, age, etc.).

3.2 Parameter Estimation

Prior Probabilities: Estimate the prior probabilities P(T=0) and P(T=1) based on historical data or domain knowledge, which, in this dataset are P(T=0) = 0.6778 and P(T=1) = 0.3222 .

Likelihood: Calculate the likelihood P(X | T) for each feature. For continuous variables, we assume a normal distribution and estimate the mean and standard deviation. For categorical variables, probabilities are estimated from the frequency of occurrences in the training data.

3.3 Posterior Calculation:

Compute the posterior probabilities P(T=0 | X) and P(T=1 | X) for each patient using the prior and likelihood estimates.

3.4 Model Evaluation

Table 3. Metrics Evaluation

Class	Precision	Recall	F1-Score	Support
0	0.82	0.91	0.86	406
1	0.76	0.58	0.65	193

After the **Classification Report** presented above, the only missing information is the **Accuracy**, which had a value of 80% approximately. The **Confusion Matrix** also gave us some important information for this project. Having 111 **True Positives**, 82 **False Positives**, 370 **True Negatives**, 36 **False Negatives**.

With these results, we could easily calculate the metrics we got before using the Classification Report because, the **Accuracy** is the sum of all the True values divided by all the values. The **Precision** is the number of True Positives divided by all the positives (False and True) combined. **Recall** is the number of True Positives divided by the sum of True Positives and False Negatives, which means the correctly classified. Finally, the **F1-Score** is equal to two times the multiplication of the Precision by the Recall divided by the sum of the Precision and Recall.

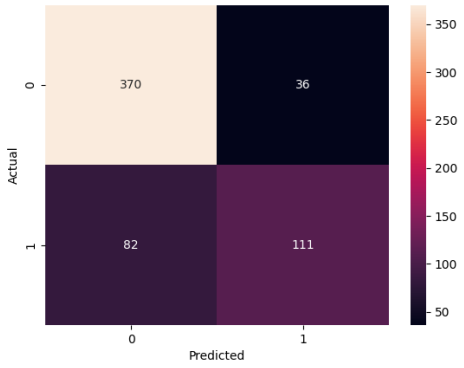

Fig. 5. Confusion Matrix Results

3.5 Model Evaluation Removing Features

Table 4. Removing Features Impact on Accuracy - 1 Feature

Feature Removed	**Accuracy**
Gender	0.650
Age	0.708
Marital Status	0.683
Vaccinated	0.667
Breathing Difficulty	0.667
Heart Rate	0.700
Blood Pressure	0.675
Temperature	0.700
Clinical Guidelines	0.817

When removing only one variable, we could conclude that the most important attributes were **Gender, Vaccinated, Breathing Difficulties and Blood Preassure** attributes because they were the variables that when removed most changed the accuracy of the decision prediction.

Table 5. Removing Features Impact on Accuracy - 1 to 8 Features

Number of Features Removed	**Min. Accuracy**	**Max. Accuracy**
1	0.650	0.817
2	0.642	0.850
3	0.617	0.867
4	0.592	0.875
5	0.600	0.867
6	0.608	0.850
7	0.633	0.800
8	0.633	0.700

Since the dataset has 9 features, we can keep removing features until there is only one feature left to see a significant drop in the maximum accuracy when trying all combinations of features. About the minimum accuracy when removing features, we can conclude that removing only one important feature already makes a big difference, since the model accuracy drops to 0.650 when removing only the **Gender** attribute.

4 Conclusion

After developing this work, we can understand how a Bayes decision model works, how the model accuracy changes when removing the most and least important features and the importance of combining features that work well together so that the results of the application of the combinations in the models is the best possible and the outcomes improve as much as they can. **Bayesian Inference** is fundamental for **Data Fusion** and with this project, we had the opportunity to work with it and experience its characteristics like the incorporation of prior knowledge, or even the combination of multiple sources of data and most importantly its Adaptability that was a very important characteristic we wanted in our model.

To sum up, this work about Bayesian Inference, has made us learn a lot about this theme and made us realize the importance of this theme for the field of Data Fusion.

References

1. Seeing Theory - Chapter 5 - Bayesian Inference, https://seeing-theory.brown.edu/bayesian-inference/index.html. Last accessed 13/jun/2024