Technical University of Denmark

Introduction to Machine Learning and Data Mining (02450)

Project 1 Report

"Data: Feature extraction, and visualization"

Group 390

Elias Kallioras (s232069)

Nikolas Xiros (s231920)

19 September 2023

The following table describes portion of the sections every student worked on.

Section	Elias Kallioras-s2320	Nikolas Xiros-s231920
Section 1	60	40
Section 2	40	60
Section 3	65	35
Section 4	35	65
Exam questions	50	50

Table 1: Project tasks split

1 Description of the data set

The **HTRU2** dataset contains measurements of 8 different metrics used to classify an interstellar object as a *pulsar* or *non-pulsar*. It consists of 1639 pulsar and 16259 non-pulsar candidates that were obtained during an analysis of HTRU Medium Latitude data. In the original paper [1] it is stated that until recently candidate selection was a predominately manual task. More modern approaches integrate computational resources to solve this problem, known as the "candidate selection problem". However advances in telescope receiver technology and the computational power significantly increased the number of candidates produced by modern pulsar surveys. With this as motivation the authors turn to a Machine Learning approach to classify candidates into pulsars or non-pulsars.

The authors of the original paper used several methods for classifying the candidates. Amongst these were Decision trees, the Naive Bayes Classifier, the Multilayer Perceptron and Support Vector Machines. The best performance in classifying candidates was shown by the "Gaussian Hellinger Very Fast Decision Tree algorithm" which deals with the bias created by the inequality between pulsar and non-pulsar candidates in the dataset. This classifier achiaved an accuracy of **0.978** in the HTRU2 dataset, with the other classifiers following with accuracies that do not fall below the 0.94 threshold. In our case this dataset will be analysed using similar **Classification** techniques which will be used to predict weather a new candidate is a pulsar or not based on the values of the other attributes. We expect some values to influence this decision more than others, but none are intuitively obvious without any further analysis. We will also apply **Regression**, with which we will predict the value of a continuous attribute, based on knowledge from the other attributes of an entry. Similarly, it is difficult to predict which attributes directly or indirectly influence others without further analysis, although the original authors state that attributes like the mean, excess kurtosis, and skew of the integrated pulse profile *"exhibit strong correlations* ($> |0.5|$)*"*, so we expect these to reveal some information about each other. In order to carry out these tasks, we need to "center" the data by subtracting their mean, and normalize them by dividing with the standard deviation (more on this in Section 3). We also added the class labels to the original .csv file, since they were missing.

2 Detailed explanation of the attributes of the data

2.1 Detailed attribute characteristics

The data set shared here contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples.These examples have all been checked by human annotators.

Each candidate is described by 8 continuous variables. The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency (see Section 3 for more details). The remaining four variables are similarly obtained from the Dispersion Measure - Signal to Noise Ratio (DM-SNR) curve (again see Section 3 for more details). These are summarised below:

The table below briefly explains each attribute of the data set, and a statistical quantitative description of the category features is also given. The D/C column specifies if the corresponding attribute is discrete or continuous.

Feature	Explanation	**Discrete/Continuous**	**Type**
Prof.μ	Mean of the integrated profile	Continuous	Ratio
Prof.σ	Standard deviation of the integrated profile	Continuous	Ratio
Prof.s	Skewness of the integrated profile	Continuous	Ratio
Prof.k	Excess kurtosis of the integrated profile	Continuous	Ratio
DM_μ	Mean of the DM-SNR curve	Continuous	Ratio
DM_σ	Standard deviation of the DM-SNR curve	Continuous	Ratio
DM_s	Skewness of the DM-SNR curve	Continuous	Ratio
DM_k	Excess kurtosis of the DM-SNR curve	Continuous	Ratio
class	Pulsar/Non Pulsar	Discrete	Nominal

Table 2: Summary of features

2.2 Data issues.

There are no missing values in our selected data and also no corrupted data but we found some high percentages of outliers using the box plot method which will we analyze in depth in the next section of the report(Data visualization section).

2.3 Basic summary statistics of the attributes.

We extracted some of the basic summary statistics of each attribute of our data set .In addition we now have a better idea of the actual values of our data based on the standard deviation and their mean/median values.
We show some representative results of the attribute Statistics:

Feature	Mean	Standard Deviation	Median	Range
Prof.μ	111.07996	106.5145	83.0645	1192.97781
Prof.σ	46.5495	6.8431	46.9474	74.0068
Prof.s	0.4778	1.0640	0.2232	9.9455
Prof.k	1.7702	6.1679	0.19871	69.8935
DM_μ	12.6143	29.4728	2.8018	223.1789
DM_σ	26.3265	19.4705	18.4613	103.2717
DM_s	8.3035	4.5060	8.4335	37.6791

Continued on next page

Table 3 – *Continued from previous page*

Feature	Explanation	Measurement	Range	
DM_k	104.8577	106.5145	83.0645	223.1789

Table 3: Basic summary statistics

We also observed that there is a big difference in the values of the attributes between pulsars and non pulsars samples.This also consists a reason for outliers in the data set as we will analyze later based on these extracted statistics.Below,we show some examples for some representative attributes.The statistics below can be visualized and understood better in the plot of section 3.2.

Feature	Mean	Standard Deviation	Median	Range	Label
DM_μ	12.6143	29.4728	2.8018	223.1789	Both
DM_μ	8.8632	24.4114	2.6354	223.1789	Non Pulsar
DM_μ	49.8259	45.2879	33.4949	199.0911	Pulsar
DM_σ	26.3265	19.4705	18.4613	103.2717	Both
DM_σ	23.2879	16.6514	17.618	103.2717	Non Pulsar
DM_σ	56.4689	19.7310	59.3670	101.9967	Pulsar

Table 4: Basic summary statistics separated for Pulsars/Non Pulsars samples

3 Data visualization based on suitable visualization techniques including a principal component analysis (PCA)

3.1 Potential outliers

In order to analyse the possibility of outliers existing in our data set, we plotted the box plots (Figires 1, 2 and 3). First of all many outliers exist because of the big difference in the values between Pulsar and Non-Pulsar samples. This can be seen in Section 3.2 plots. As a result, outliers are generated that should not be excluded from our dataset since they represent the actual difference that we need for the classification. Though, as we can see in the box plots (Figures 1 and 2) there are some outliers also in the samples belonging in the same class, especially regarding the Non-Pulsars samples. The existing outliers of the dataset included in the Pulsar samples may have an negative impact to our results.

3.2 Data distribution

To check weather the data is normally distributed we construct a histogram for each of the attributes.

In Figure 4 we see that all attributes appear to follow a normal distribution with varying mean and standard deviation values. An exception to this rule would maybe be the DM skewness and DM kurtosis, but they are not too far from a normal distribution. Notice that we also create a histogram for the class attribute. This is done to bring attention to the fact that the pulsar candidates are far outnumbered by the non-pulsar candidates in the dataset, and this may affect performance when using the common classification techniques. Further analysing this, we can see that the mean values of some of the attributes, as seen in Figure 5 are very different for

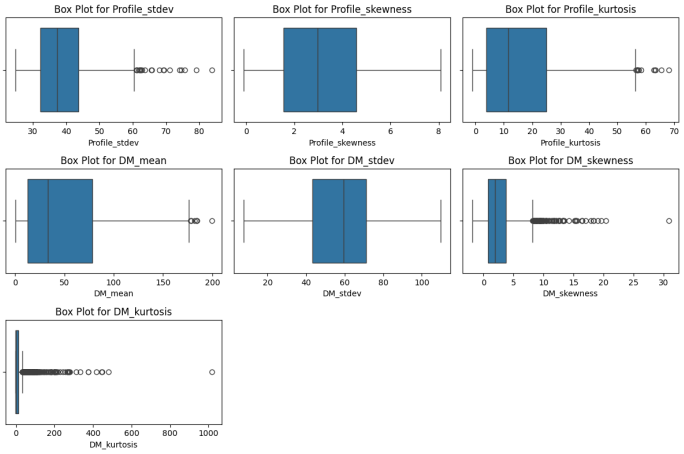

Figure 1: Box plot of the Pulsar class.

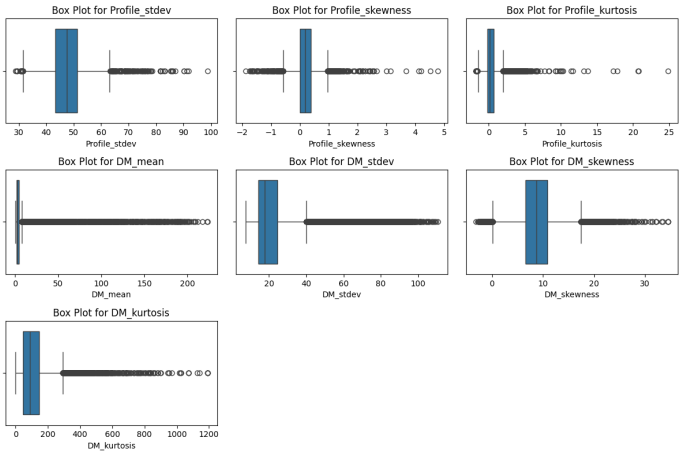

Figure 2: Box plot of the Non-pulsar class.

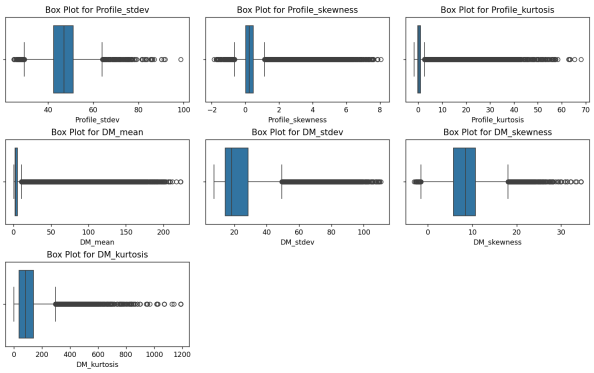

Figure 3: Box plot of the whole dataset.

the two classes. Taking into account the fact that the Non-pulsar class heavily outnumbers the Pulsar class, we can conclude that the mean value of the total dataset is mostly influenced

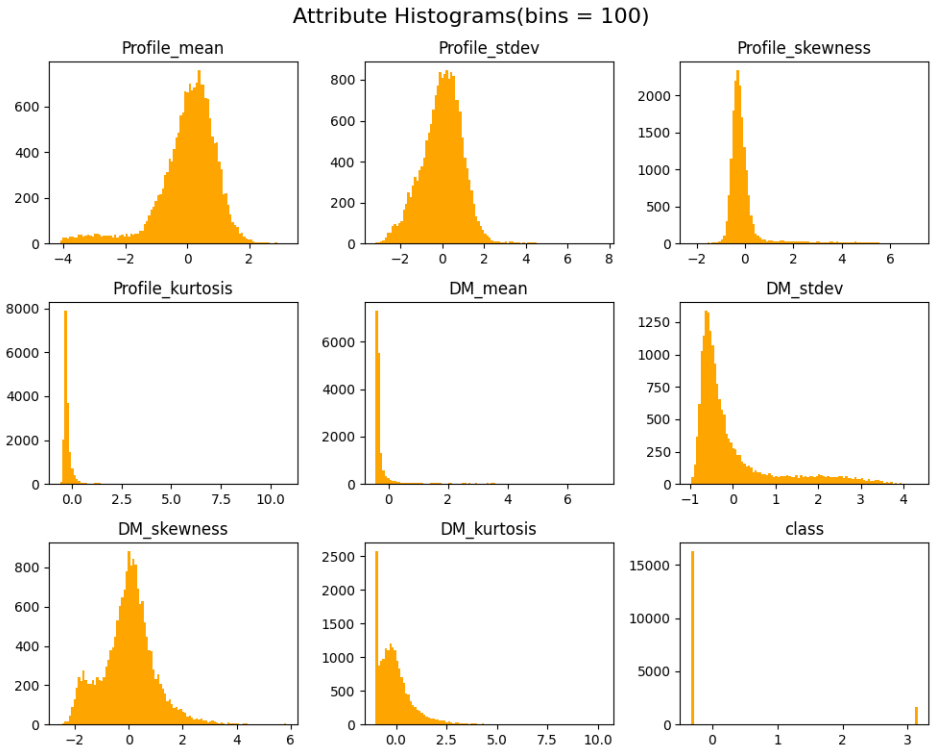

Figure 4: Histograms of attributes.

by the Non-pulsar candidates. This in turn makes almost every pulsar candidate seem like an outlier in the data when common techniques like the IQR method or the Z-score are used.

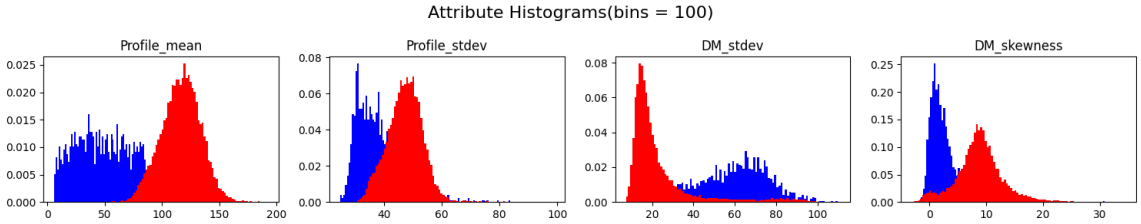

Figure 5: Histograms of the two classes for some attributes. The histograms are scaled so that they have the same height for easy comparison.

To make the above clear we present an attribute whose mean value is heavily influenced by the Non-pulsar class.

3.3 Attribute correlation

In this section we analyse the correlation between the variables that define a candidate in our dataset. The correlation of the valieables can be studied using a *correlation matrix*. In addition, we use a heatmap to visualize the values of the correlation.

In Figure 7 we see that some attributes have very high correlation. For example the skewness and kurtosis measures seem to be **highly correlated** in every case. For example profile skewness

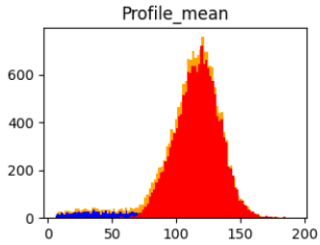

Figure 6: Orange: Histogram of the attribute on the whole dataset, Red: Histogram of the attribute on the Non-pulsar class, Blue: Histogram of the attribute on the Pulsar class

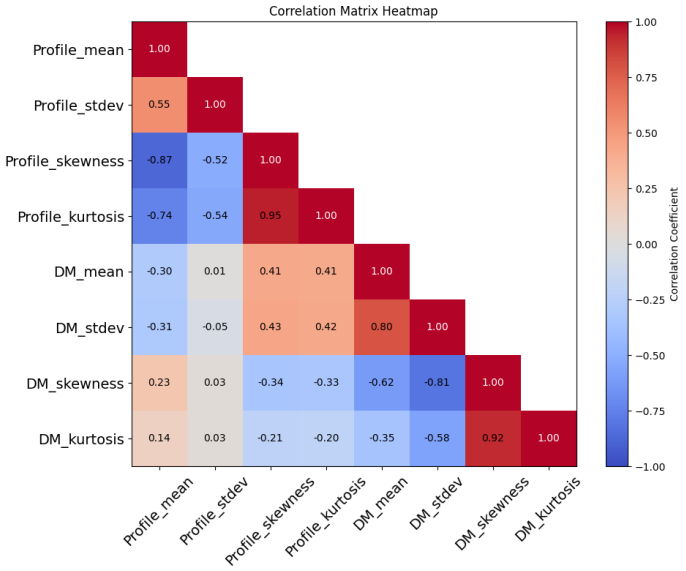

Figure 7: Heatmap correlation matrix of attributes.

and profile kurtosis as well as DM skewness and DM kurtosis have a correlation of > 0.9. We also notice a high correlation between DM standard deviation and DM mean value. We notice, however, that the correlation between attributes regarding the "Profile" and attributes regarding the "DM" have a low correlation, which means they are well seperated. This can be cross-validated with the original paper [1], where the authors state that attributes were chosen so that they have a high separation.

3.4 Primary Machine Learning Modeling Aim

The feasibility of the primary machine learning modeling aim comes down to weather the two classes of our dataset can be seperated or not. To come to a conclusion about the feasibility of our goal, we need to conduct a PCA analysis to the data. This section focuses on this matter.

As a first step in our PCA analysis, we will look into how much of the explained variance each PCA component offers. We set a threshold of 0.9 and we demand that in order to preserve the characteristics of the data we need variance that is above our threshold. In Figure 8 we can see the explained variance as a function of the number of PCA components. We can see that 4 PCA components suffice in order to achieve the above.

It is also in our favour to know the principal directions which make up the necessary principal

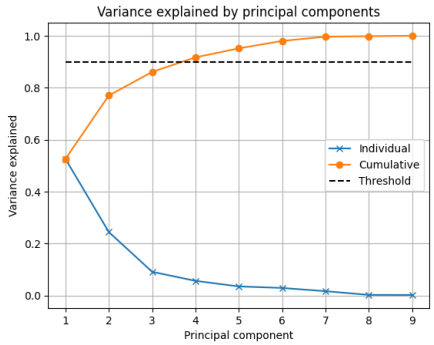

Figure 8: Explained variance as a function of the number of Principal Components.

components. In order to visualize the principal directions of the first 4 Principal Components, we create the plot shown in Figure 9.

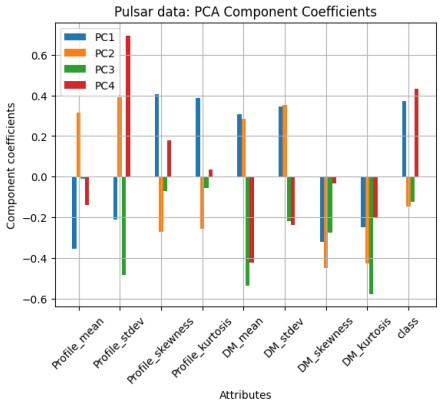

Figure 9: Coefficients of the first 4 Principal Component vectors.

Finally, we can visualize what the projection of the data in these sub-spaces would look like. It is more intuitive to visualize the projected data in 2D sub-spaces, so that is what we present first. In Figure 10 we can see the projection of the data on the first 2 principal components, which preserve roughly 80% of the explained variance. In this projection we can already see that the 2 sets of classes have some degree of separation.

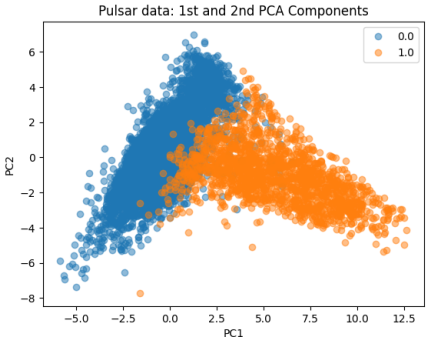

Figure 10: Projection of data on the subspace of 1st and 2nd principal components.

For comparison, we also project the data onto the subspaces created by all of the combinations of the first 4 principal components in Figure 11.

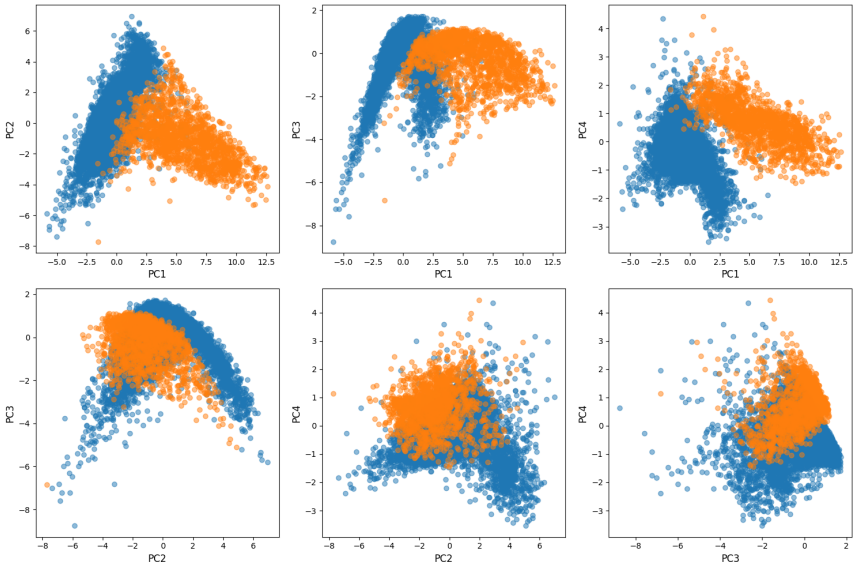

Figure 11: Projection of data on all the subspaces the first 4 principal components.

We can see that the degree of separation in the projection on each subspace is decreasing as we project the data on the higher order principal components. For example, we can see that the projection made on the 3$^{\text{rd}}$ and 4$^{\text{th}}$ principal components much less separation on the two classes compared to the projection on the 1$^{\text{st}}$ and 2$^{\text{nd}}$ principal components.

For visualization purposes, we also project the data on the subspace constructed by the 3 first principal components in Figure 11. This is a 3D subspace, and thus it is harder to visualize the way the data can be separated. In this projection the preserved explained variance account for roughly 85% of the total explained variance.

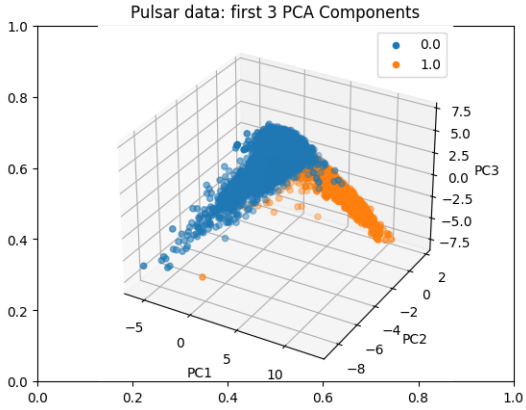

Figure 12: Projection of data on the subspace created by the 1$^{\text{st}}$, 2$^{\text{nd}}$ and 3$^{\text{rd}}$ principal components.

4 What we have learned about the data

In summary, all the attributes in the data seem to roughly follow the shape of a normal distribution, and seem to be separated in 2 categories. One category includes all attributes that refer to the Profile and the other category includes the attributes that are associated with the DM. These two are not largely correlated with each other. Moreover, the dataset is not balanced, meaning a big percentage of entries are non-pulsars, and only a small portion of candidates are pulsars. This might affect performance when using standard classification methods. This also affects the detection of possible outliers. From the PCA analysis conducted in 3.4 we can see that using the first 4 principal components we preserve $> 90\%$ of the explained variance. When projecting the data in the subspace created by the first 2 principal components, which preserves roughly $> 80\%$ of the explained variance, we can see that there is a decent amount of seperation between the classes. Based on this, it is fair to conclude that when using the first 4 principal components, which preserve even more variance, the classes will be well separated and the classification task will be possible. Correlations between some attributes, as shown in 3.3 also point to the achievability of the regression task.

5 Problem solutions

• Question 1

Option D:
We will analyze each attribute one by one:

- $x1$ is nominal because it is used to 'split' the day in intervals of 30minutes that have measurable distances between each other.
- $x2$ and $x7$ are ratio because they describe that are a count for something(Traffic lights/Running over accidents) and also zero means the absence of this attribute.
- y is ordinal because there is a clear ranking between the values (1-4) of this attribute from less to more congestion

• Question 2

Option A:
For $p = \infty$, we know that the p-norm distance is equal to

$$\max\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|\}$$

which in our example is equal to $26 - 19 = 7$.

• Question 3

Option A:
We can compute that the variance explained from each principal component:

$$\text{i's varinace} = \frac{\sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$$

so we get:
First principal component variance =0.288

Second principal component variance =0.232
Third principal component variance =0.196
Fourth principal component variance =0.15
Fifth principal component variance =0.133
and so A answer is correct as the first 4 variances add up to 0.866

• Question 4

Option D:

Option D is the correct one because we can see that the fifth PCA component has positive coefficients in the Time of day,Broken Truck,Defects and negative values in the Accident victim and Immobilized bus attributes.As a result, this will give a positive projection on the second PCA component for observation with a low value of Time of day, a high value of Broken Truck, a high value of Accident victim, and a high value of Defects because of the positive coefficients in the high attributes,negative coefficients in the low one and a very low coefficient in the 4th attribute(Immobilized bus) which means that this attribute will not affect the component projection.

• Question 5

Option A:
We will use the formula of Jaccard similarity:

$$J(x, y) = \frac{f_{11}}{K - f_{00}}$$

where f_{11} equals to the number of words that exist in both sentences which is 2
and the f_{00} equals to the number of words that do not exist in neither of two sentences which is 19.987(K(total number of attributes) minus the total number of words existing in the two sentences
Doing the calculation we get 0.153846 which corresponds to option A.

• Question 6

Option B:

We will use the rule:

$$p(y) = \sum_{x} p(y, x)$$

which in our case applies as

$$p(x_2 = 0|y = 2) = p(x_2 = 0, x_7 = 0|y = 2) + p(x_2 = 0, x_7 = 1|y = 2)$$

and so doing the calculations we get result 0.84

References

[1] R. Lyon, "HTRU2," UCI Machine Learning Repository, 2017, DOI: https://doi.org/10.24432/C5DK6R.