Assignment 2: The Raven Test

In Search for Group Differences

Data Mining, University of Aveiro
2019

Filipe Pires, 85122
DETI, MSc. Informatics Engineering

João Alegria, 85048
DETI, MSc. Informatics Engineering

***Abstract*—...**
***Index Terms*—...**

I. Introduction

Lorem ipsum ...

II. Dataset & Feature Extraction

Lorem ipsum ...

III. Data Quality & Normalization

In relation to the dataset and the quality of the information present on it, through analysis of Table 1, which is the visual representation of the correlation of every possible pair of features, being the diagonal the pair composed of the feature with itself and is represented with a histogram that describe the value distribution inside that given feature, the remaining pairs are represented with scatter plots, we can take some important conclusions. The first one, which was already expected is that there was some outliers that needed to be taken care of, the second conclusion that's more important for this study is the fact that the data in all pairs of features doesn't appear to be separable, i.e., since there is two classes, the *DEI* class and the *ESEC* class there should be something resembling two groups, one for each class. That doesn't happen since in most cases we are presented with a uniform distributions, gaussian distributions or simply the grouping of all data in a unique central point.

After familiarizing better with the data and understanding it better, it was needed to process it in a way that the models that we intend to apply have the best change of finding the distinguishing factors. For that we needed to remove the already mentioned outliers and normalize the data. We started by doing the mentioned normalization, which consisted in applying the *ZSCORE* algorithm to each feature, which follows the Formula 1 and already centers the data. Following this we also removed any data entry that contained a *Null* values, since that would cause problems to the models' processing. Finally we removed the still existing outliers applying the Expression 2.

$$z_i = \frac{x_i - \mu}{\delta} \tag{1}$$

$$|x_i - \mu| < 3 \times \delta \tag{2}$$

IV. Classifiers

In this section we will present and give a brief summary of all the machine learning models used in this study.

A. Support Vector Machine

Commonly denoted as SVM, this supervised machine learning model with the supplied training set, by mapping the several data entries in the a multidimensional space in a way that the the widest possible gap exist between the data of each class. Based in that gap a model is created and a decision rule defined that when feed new data tries to assign a class to that entry by using the decision rule already mentioned.

B. MultiLayer Perceptron

This class of feedforward artificial neural network is another supervised machine learning model. MLP is sometimes strictly refers to networks composed of multiple layers of perceptrons. Perceptron is also a supervised learning algorithm capable of doing binary classifications according to a linear predictor function. The power of the MLP is the junction of several of those perceptrons that individually don't have the a good performance, but the combine performance can reach significant values.

This junction normally follows a standard architecture divided in layers, firstly a input layer, followed by one or more hidden layers and an output layer. The input and the output layers are the ones that interact with the exterior, being the first, as the name indicates teh one where the data should be injected at and the second where the result will be presented. The hidden layers is where the main learning occurs and where the most amount of fine-tunning is applied. In this entire architecture the number of layers and the number of neurons(singular cell inside a layer) can be tunning to improve the model performance.

C. Decision Tree

As the name suggests, this model uses a tree like structure to support the data classification decision. The construction of this auxiliary structure is quite simple and straightforward, consisting in assigning a condition to each new branch, causing that the structure will converge with having only one class in leaf nodes, defining the conditions required to identify the existing classes.

D. Random Forest

E. K Nearest Neighbors

V. Parameters Variation

Lorem ipsum ...

VI. Results Discussion

Lorem ipsum ...

VII. Conclusions & Future Work

Lorem ipsum ...

References

[1] A. Tomé, "Data Mining Assignment", https://elearning.ua.pt/pluginfile.php/1496406/mod_resource/content/3/ED_HCT_Raven.pdf, accessed in November 2019.

APPENDIX

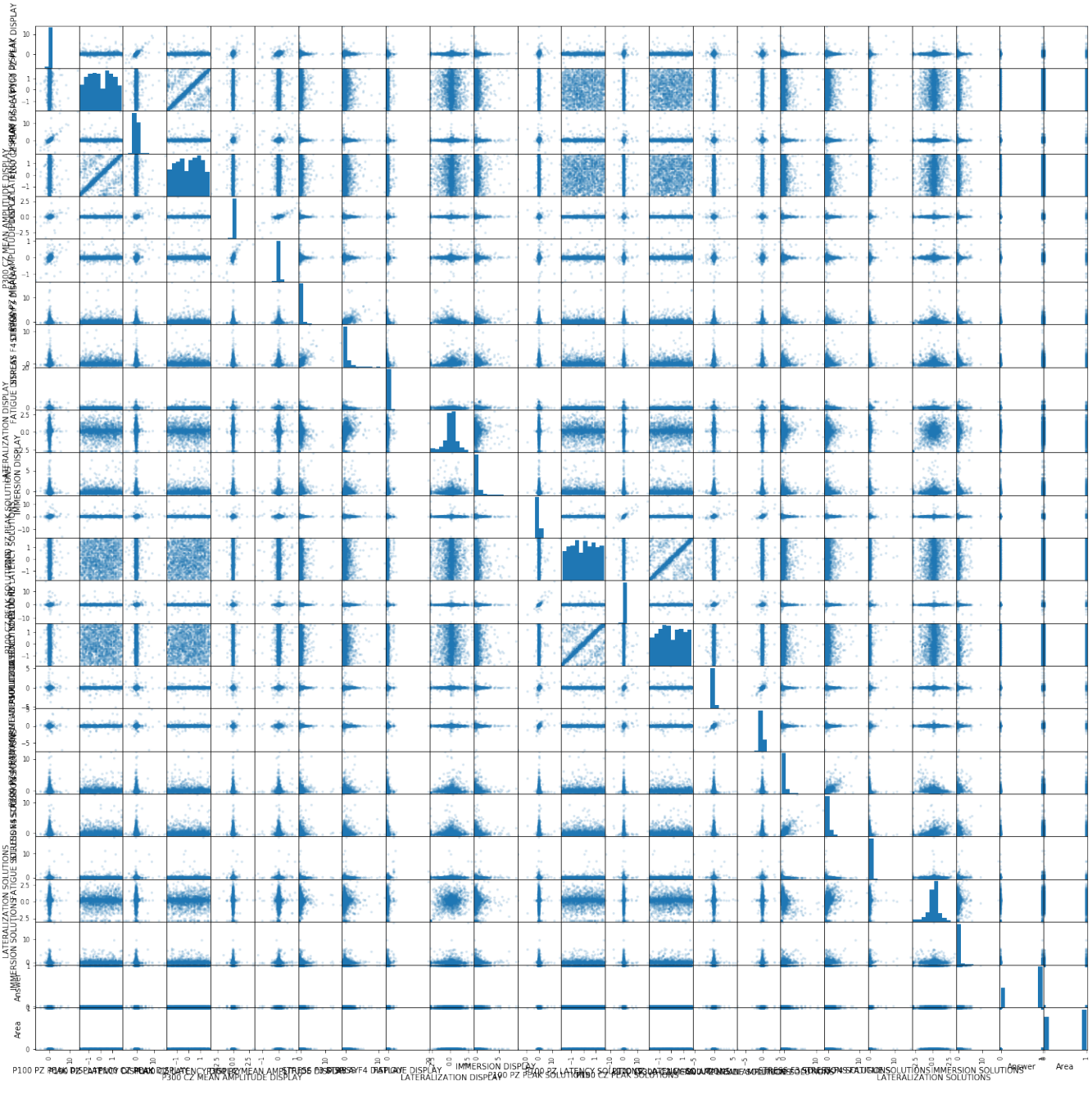

Fig. 1. Feature Correlations