Abstract

The aim of this study is to extract knowledge from the final researches of the Mumbai University Science Faculty. Five classification models were applied: Vector Support Machines, Neural Networks, Decision Tree, Random Forest and Powering; considering the Experiment Design and Multivariate Analysis Lines. Results showed that for the Experiment Design line, the most accurate model was Random Forest with 71.48% predictions that are correct respecting to the total. Regarding the Multivariate Analysis line, there was no significant difference in overall accuracy, fluctuating by 97%.

© 2019 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Peer-review under responsibility of the Conference Program Chairs.

Keywords: data mining education; data mining techniques; education indicators; classification.
1. Introduction

Data mining is a process that allows to extract knowledge from databases [1], its objective is to discover anomalous and / or interesting situations, trends, patterns and sequences in the data. One of its results is the classification [2, 4], which tries to obtain a model that allow to assign a case of unknown line to a specific line [2]. Given a database of $D = \{t_1, t_2, \ldots, t_n\}$ of records (final researches) and a set of lines $\{C = C_1, C_2, \ldots, C_m\}$, the problem of classification is to find a function $f: D \rightarrow C$ such that each $t_i$ is assigned in a line $C_j$. Two lines were considered for the study: Experimental Design and Multivariate Analysis, and as function $f$, the classification models were taken:

- Vector Support Machines (VSM): A technique that allows to extract relevant information from datasets and build efficient and fast algorithms. It is based on finding a hyperplane to separate the lines in the largest possible margin, in this way, the decision border is established by the patterns that highlight the distribution of lines [5].

- Neural Networks: These networks are self-adaptive dynamic systems [6] conformed by a set of interconnected nodes and weighted bindings. The output nodes are the sum of each of the input values according to the weights of their links.

- Decision Tree: It is a supervised classification technique [7] which allows to determine the decision to be taken following the conditions that are met from the root to one of its leaves [8]. This recursive technique considers the criterion of the highest proportion of information gain, i.e. chooses the attribute that best ranks the data [9].

- Random Forest: It is a combination of predictive trees which works with a collection of incorrigible trees and averages them. Each tree depends on the values of a random vector of the sample and with the same distribution of all trees in the forest [7].

- Powering: This model takes a random sample of the original data and applies a qualifying method on it, then increases the weight (power) to poorly classified individuals so that, in the next application, the classification method focus more on these poorly classified individuals, improving their classification, and so on.

2. Method

The method applied was proposed by Williams Graham in his work entitled "Data Mining with Rattle and R, The Art of excavating data for knowledge discovery" where the researcher shows the package created in R and its applications [10]. It consisted of using a percentage of data to test the models: Vector Support Machines, Neural Networks, Decision Tree, Random Forest, and Powering, leaving the rest to validate them. Its performance was assessed using global, positive, and negative accuracy, using the ROC curve to graphically show whether or not the models are suitable.

For the research, a database corresponding to 25720 observations was used, which consists of 1054 final researches [11] in the following careers: Biochemistry and Pharmacy, Biophysics, Engineering in Environmental Biotechnology, Engineering in Computer Statistics, Chemistry Engineering, and Environmental Education at Mumbai University Science Faculty. The final researches submitted during the years 2015 – 2018 were selected. The variables used were the 12 programs: Sustainable Biodiversity, Alternative Energies, Air, Water and Soil Management and Treatment, Environmental Bioremediation, Development of Software Applications for public and private management processes. Education, Implicative and Computational Statistical Analysis, Assessment of the food security state, Human consumption to improve nutrition and health conditions, health and nutrition assessment, health services administration, nutrition and food, biophysics applied to medicine, and development of phytopharmaceuticals" [12].

The application was done using Rattle, for the evaluation of the models a learning table was required, the same used to train (learn) the prediction model, that is, $f$ is calculated from this table; and a testing table, which allows to validate the model and is internally selected by the software, which verifies that the results in individuals who did not participate in the construction of the model are acceptable [13]. The study used 75% of the data for the learning table and 25% for the testing table. To evaluate the performance of the models, the measures were determined: Global Accuracy, which shows the total number of predictions that are correct to the total, Positive Accuracy (Sensitivity), which is the proportion of positive cases that were correctly identified, and Negative Accuracy (Specificity) which indicates the proportion of negative cases that were correctly identified [14]. These measures were calculated from the confusion matrix containing information about the predictions made by a model or classification system,
comparing the learning or testing table for the set of final researches, with the given prediction versus the line to which they belong. The Table shows the confusion matrix for a two-line classifier.

A graph showing the performance of the models was the ROC curve [15]. It compares the false positive rate with true positives, placing it on the Y Axis = Sensitivity and the X=1 – Specificity. It is used to select optimal models in binary classification problems. This method considers the area under the curve, in which, if it is 1, the model is ideal for the dataset, if it is greater than 0.5, it is optimal, and the model is negligible if its value is less than 0.5.

Table 1. Confusion matrix.

<table>
<thead>
<tr>
<th>Real Value</th>
<th>Prediction</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>Vn (True Negative)</td>
<td>Fp (False Positive)</td>
</tr>
<tr>
<td>Positive</td>
<td></td>
<td>VP (True Positive)</td>
</tr>
</tbody>
</table>

3. Results and Analysis

By Rattle, the confusion matrix was determined for each classifier. A particular case is displayed for the Vector Support Machines model when the Experiment Design line was applied, Table 2.

Table 2. Confusion Matrix, VSM; experiment design line

<table>
<thead>
<tr>
<th>Actual Value</th>
<th>Prediction</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>It is not an experimental design</td>
<td>63</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>It is an experimental design</td>
<td>33</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>

Performance measures were calculated using the equations Global Accuracy P (Accuracy), Positive Accuracy (Sensitivity) and Negative Accuracy (Specificity), (1), (2) and (3), respectively, as follows:

\[
P = \frac{V_N + V_P}{V_N + F_P + F_N + V_P} \quad (1)\]

\[
P_P = \frac{V_P}{F_N + V_P} \quad (2)\]

\[
P_N = \frac{V_N}{V_N + F_P} \quad (3)\]
The results of these indicators for the five models, applying the two lines, are shown in Table 3. Due to the Rattle version, the confusion matrix could not be determined for the Powering model and the precisions were not calculated.

From Table 3, for the Line: Experiment Design, the highest overall accuracy was presented in the Neural Networks model with 71.48% predictions that are correct respecting to the total, followed by Random Forest with 68.59%. Among the positive cases that were correctly identified, Neural Networks was the model that reached the highest percentage with 67.24% and the lowest was the Decision Tree with 50.86%. The model with the highest proportion of correctly identified negative cases was Random Forest with 77.77%, while VSM reached the lowest percentage at 73.81%. Table 3 shows the results of the ROC curve for each of the models with the two lines. Figure 1 shows the graph of this curve for the VSM model with the Experiment Design classifier.

### Table 3. Results from Performance Measures

<table>
<thead>
<tr>
<th>LINE: Experimental Design</th>
<th>CLASSIFIED R (MODEL)</th>
<th>Positive Accuracy: Sensitivity (P%:PP%)</th>
<th>Negative Accuracy: Specificity (PN%)</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector Support Machines</td>
<td>64.46</td>
<td>54.31</td>
<td>73.81</td>
<td>0.72</td>
</tr>
<tr>
<td>Neuronal Networks</td>
<td>71.48</td>
<td>67.24</td>
<td>75.39</td>
<td>0.76</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>64.05</td>
<td>50.86</td>
<td>76.19</td>
<td>0.69</td>
</tr>
<tr>
<td>Random Forest</td>
<td>68.59</td>
<td>58.62</td>
<td>77.77</td>
<td>0.77</td>
</tr>
<tr>
<td>Powering</td>
<td></td>
<td></td>
<td></td>
<td>0.78</td>
</tr>
<tr>
<td>LINE: Multivariate Analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vector Support Machines</td>
<td>97.93</td>
<td>99.08</td>
<td>86.95</td>
<td>0.96</td>
</tr>
<tr>
<td>Neuronal Networks</td>
<td>97.11</td>
<td>97.26</td>
<td>95.65</td>
<td>0.97</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>97.93</td>
<td>99.09</td>
<td>86.95</td>
<td>0.93</td>
</tr>
<tr>
<td>Random Forest</td>
<td>97.93</td>
<td>99.09</td>
<td>86.95</td>
<td>0.99</td>
</tr>
<tr>
<td>Powering</td>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
</tr>
</tbody>
</table>

The ROC curve showed that the appropriate models for this line are Potentiation and Random Forest with an area under the curve of 0.78 and 0.77 respectively, while VSM has the smallest area (Fig. 1). For the Multivariate Analysis line, it was shown that accuracy (97.93%) and Specificity (86.95%) are the same for VSM and Sensitivity (99.09%) for Decision Tree and Random Forest. The ROC curve showed that the models with the largest area under the curve were Random Forest and Potentiation with a value close to 1. It should be noted that all models showed areas greater than 0.90.

Fig. 2 shows the Decision Tree for the Multivariate Analysis line, the main node is the Human Consumption indicator to improve nutrition and health conditions (NHC), for which there is an 89% probability that the data is 0, so it is NOT a Multivariate Analysis and the remaining is 1. If the NHC score is less than 2, then the final research is a Phytopharmaceutical Development Program (PDP), where there is a 94% chance that it is not a Multivariate Analysis and 0.06% that it is. If the PDP indicator scores less than 2, the research topic is predicted to be a Computational Implicative Statistical Analysis (PAEIC) program with 92% of the data.
4. Conclusions

The classification methods made it possible to identify relevant information about educational indicators when the Experiment Design and Multivariate Analysis lines were applied. Rattle identified the appropriate models for the
dataset. Decision Trees visualized representative variables that can help predict future events. Rattle is a package with an easy-to-use interface and has no drawbacks. It is important to consider that the database should be saved with CSV extension and, at the time of uploading the file, the signs used as a data separator and decimal places should be identified. For the study, the data file was loaded, then the percentages were selected for the testing and learning table, then the classification models were applied, and the performance measures were calculated using Excel. This study determined the trend of future final researches, since depending on the career it is possible to identify whether they will apply Experiment Design or a Multivariate Analysis.

References