A comparison between the Naïve Bayes and the C5.0 Decision Tree Algorithms for Predicting the Advice of the Student Enrollment Applications

Yannick Kiffen:
Tilburg University
Yannick.kiffen@gmail.com

Francesco Lelli
Tilburg University
f.lelli@tilburguniversity.edu
https://francescolelli.info

Omid Feyli
Tilburg University
O.Feyli@tilburguniversity.edu

Abstract

In this paper we introduce a dataset containing students enrolment applications combined with the related result of their filing procedure. The dataset contains 73 variable. Student candidates, at the time of applying for study, fill a web form for filing the procedure. A committee at Tilburg University review each single application and decide if the student is admissible or not. This dataset is suitable for algorithmic studies and has been used in a comparison between the Naïve Bayes and the C5.0 Decision Tree Algorithms. They have been used for predicting the decision of the committee in admitting candidates at various bachelor programs. Our analysis shows that, in this particular case, a combination of the approaches outperform a both of them in term of precision and recall.

Introduction

Before the data pre-processing begins there will be a short description of the dataset used in this experimental research. The dataset consisted of the data that is obtained by the Department of Marketing and Recruitment during the enrollment application for students which is called

"MYSAS". The data is in an excel format and this data consists out of 74 variables including the nationality, education country, diploma etc. The data was collected from 2016 till 2020 and consisted of both Master and Bachelor applications. There were some missing values for all the variables. After discussion with Mr. Feyli the variable "Advice" was chosen to be the target variable of the experimental research.

The research question of the thesis is: "What is the best classification technique for predicting the advice of the student enrollment applications for the bachelor programs at Tilburg University: a comparison between the Naïve Bayes and the C5.0 decision tree algorithms?" the goal of the experimental research is to be able to answer this research question. This will be done by performing a comparison between the two classification algorithms. The two algorithms will be compared using different evaluation metrics, such as accuracy, recall and precision.

The conclusion of the research is that a combination of the two algorithms leads to the best performance on the evaluation metrics. The recommendations are to integrate the combined model for Tilburg University and to use this research as a framework for other higher education institutions to find their best performing algorithm.

Method

Pre-processing

In this section there will be a step-by-step description of the processes involved in the experimental research. The total process is based on papers comparing different classification techniques and describing the different steps of data mining. In the paper of Anil et al. (2019) a framework for the different steps is introduced, starting with the first step Data pre-processing.

After discussion with Mr. Feyli the variable "Advice" was chosen to be the target variable of the experimental research. The Pre-processing of the data began with making sure that there were no observations with a missing target variable value. After the first step and after discussion with Mr. Feyli, the dataset was filtered on the variable "programmetype" and only the Bachelor programs were selected. This was done, because the rules for bachelor programs are more straightforward and include less interpretation of an enrollment officer. After the relevant variables were selected, which will be explained in the next section. These variables have to be converted to other data formats. For example, the "datereceived" was first a string variable, but was transformed into a "date" format. Furthermore, for both the diploma, the education country and the combination of the two, observations which occurred less than five times in the whole dataset, were deleted. This was done to secure the privacy of individuals and to make sure there were no mayor imbalances in the classes. All these transformations were performed by using STATA.MP 16. There were two remaining values for the variable advice: "Ok" and "Not ok", the other values were deleted due to imbalance problems. After the data was structured correctly, there were a few instances in which the advice was missing while the variable "FinalDecision" had the value "Refused" or "Admitted", after discussion with Mr. Feyli the values of advice for these instances were changed accordingly. This was done for observations that missed the deadline or that were admitted, because of the VWO regulations at Tilburg University. Furthermore, a new variable called AdviceBin was created that represent the outcome of the advice variable in Binary values, 1 for Ok and 0 for Not ok. This transformation was necessary for a calculation of an ROC curve, this last transformation was done using SPSS Statistics. All the processing codes can be found in Appendix A of the original thesis.

Variables

After examination of the dataset relevant features were selected this led to the following variables with the content described:

Table 1: Overview of Variables

Variable	**Content**
Advice (target variable)	Categorial variable, consist out of Ok/Not ok
Nationality	
Programmename	f.e. Economics, Global Law etc.
Facultyname	The name of the faculty at Tilburg University
Datereceived	The date the first time the application was received
Educationcountry	Country in which the student received their diploma
Diploma	
Obtained	If the diploma is obtained or not
ConditionalDiplomarequired	If there is the need for a conditional diploma next to the already obtained diploma
ConditionalDiplomaname	
ConditionalLanguagerequired	If there is the need for a conditional language exam
ConditionalLanguagename	

These variables were selected out of the original 74 variables in the following process:

- Firstly, the variables has to be filled in by the student and not by the admission officer or another employee of Tilburg University, variables that were excluded are f.e. items missing, sent for advice and statement to student.

- Secondly, variables with too many distinct values were not selected, because these will complicate the C5.0 decision tree too much, in this category belong f.e. High School, Test score and correspondence city/country.
- Lastly, variables with too many missing values were not selected, in this category belong f.e. specialization name and second nationality.

Data integration

The dataset with the variables described in section 1.2 is imported into SPSS Modeler, the data is partitioned into a training and testing set using the holdout method. After the partitioning the different variable types are selected, most of the time SPSS Modeler classifies the variables within the right type, but this has to be checked and corrected using the "Type" Field Ops.

After the data is correctly standardized, the partitioned data is integrated into the Naïve Bayes and C5.0 Decision tree algorithm. For the C5.0 algorithm Cross validation is not used since the dataset is large enough and using Cross validation does not result in other classification outcomes. There will be a distinction in the models in which boosting is used and in which not this will be represented in the tables. The number of trials of boosting will be equal to 10 for all the models in which boosting is used. The C5.0 algorithm uses the split criterium Gain ratio. The C5.0 algorithm builds a number of different trees with different accuracy performances, this is done by using the REP pruning method. In the end the tree with the highest accuracy over the testing set becomes the final decision tree.

For the Naïve Bayes algorithm, SPSS Modeler gives the option to select only complete records, which is automatically turned on. Furthermore, the algorithm builds the model for each split, which means that for all the different splits in the training and testing dataset the Naïve Bayes model is built. When choosing the structure type there is the choice between the Markov Blanket (MB) or the Tree Augmented Naïve Bayes (TAN). These two structures are described and compared in the paper of Chiroma et al. (2014). Both structures are tested for the MYSAS database and the results are displayed in the tables. At the end of this step both models are built using the training data and can be displayed in SPSS Modeler.

After separately testing the two classification techniques there will be a combination of firstly the best two Naïve Bayes models using both the TAN/MB structure types are combined. This Naïve Bayes model is then combined with the best C5.0 model and produces the combined model of the best performing techniques. This is done Using SPSS Modeler in a similar way as in the article of Laudy (2015). When voting is tied between the two classification techniques, a value is selected using het highest confidence. An overview of the stream created in SPSS Modeler is presented in appendix B of the original thesis.

When comparing the classification techniques, SPSS modeler does have the function of performing an analysis. This produces a confusion matrix and an accuracy for both the training data and testing data. Using this confusion matrix, the other relevant evaluation metrics can be calculated, such as recall or precision.

Results

Tables and figures

In the original thesis the results of different models with different training/testing splits for both the algorithms are presented. In this manuscript the table that contains the comparison between the best model for each algorithm and the table of the combination of the different algorithms are presented.

Table 2: Overview of Two Best Performing Models

	Accuracy (%)	Precision	Recall	Specificity	F-Measure	Geometric Mean

Naïve Bayes 85/15 split with MB growing technique	80.75	**0.9339**	0.9294	**0.7150**	0.9316	**0.8152**
C5.0 80/20 split with boosting	**89.80**	0.9250	**0.9705**	0.6116	**0.9472**	0.7522

Table 3: Overview of Combined Models

Models	Accuracy (%)	Precision	Recall	Specificity	F-Measure	Geometric Mean
Naïve Bayes*	89.04	0.9372	0.9383	0.7023	0.9377	**0.8118**
Naïve Bayes + C5.0**	**90.29**	**0.9388**	**0.9452**	**0.7027**	**0.9420**	0.8120

* = Combination of 85/15 MB Naïve Bayes + 80/20 TAM Naïve Bayes

** = Combination of Naïve Bayes* + 80/20 C5.0 with boosting

Results in contrast to earlier literature

In this section the findings of this research are compared with the finding of existing literature, that was discussed in section 3.5. When looking at the paper of Ragab et al. (2014), where the goal was to find the best classification technique for predicting the number of students that could enroll to their desired study and college, it can be concluded that the results are somewhat similar in this thesis. However, in the study of Ragab et al. (2014) the C4.5 algorithm performed better on all the evaluation metrics, Specificity and Geometric mean were not included. In this research the Naïve Bayes models are performing better on some evaluation metrics when compared to the C5.0 decision tree models, but in general the performance of the C5.0 models are better.

When the results of this thesis are compared to the paper of Lestari et al. (2019), the results in the paper are all in favor of the C4.5 algorithm. In the paper the goal is to find the best classification technique to predict a student GPA and based on that GPA their enrollment possibility. There is a comparison between the Naïve Bayes and the C4.5 decision tree algorithm in which both are evaluated using Precision, Recall and Accuracy. For all the different training/testing splits the three evaluation metrics are higher for the C4.5 decision tree models. When compared to the results in this thesis the Accuracy and Recall are higher however, the Precision is not.

Where the above-described papers stop at comparing two classification techniques separately, this thesis has also compared a combined model of the two techniques with the original separate models. In the paper of Singh et al. (2010) a combination of the ID3 and the Naïve Bayes algorithm was performed and evaluated. The ID3 algorithm is the precursor of the C5.0 algorithm and is discussed in section 3.2.2.3. in this paper it is found that a combination of the two algorithms leads to better performance of the metrics; Accuracy, Balanced detection rate and False positives. This is in conformity with the findings in this thesis, where the performance of the model increased after combining the two algorithms.

In the paper of Farid et al. (2014) an evaluation of a combined algorithm was made, it is concluded that the evaluation metrics Accuracy, Precision and Sensitivity/Recall all increased when there was a combination of the Naïve Bayes and C4.5 algorithm. The Specificity decreased; all these metrics were tested using 10 diverse datasets. This is almost in total agreement with the results in this thesis. The Accuracy, Precision and Sensitivity all increased in the combined model. The Sensitivity only increased when compared to the separate C5.0

decision tree model, but the Sensitivity was at its highest value in the separate Naïve Bayes model.

Discussion

The scope of this thesis limits itself to the dataset that contains the potential bachelor student for Tilburg university. Possible future research can be to investigate whether the results that this research produced are also relevant for master students or potential students at a different Higher education institution. As earlier explained the current model predicts the 'first' advice of Tilburg University, in the future there can be made a variable in which students need to fill in their grades and mathematics courses, these variables can result in a successful prediction of the final decision. This will lead to even better performance and an improvement of time efficiency when compared to the current prediction. However, a change in the student enrollment application process is therefore required. Once this is achieved new research could look at the best classification model for the final decision.

The tested algorithms are limited to the Naïve Bayes, C5.0 and a combination of the two. These two classification techniques are chosen for reasons that are explained in section 1.4. However, there are also reasons to assume that there is a possible better performing classification algorithm. In future research there can be more algorithms tested or newly designed classification techniques can be compared to the ones tested in this research.

Conclusion

In this preprint we introduced a dataset containing students enrolment applications combined with the related result of their filing procedure. The dataset is suitable for data analysis and algorithmic comparison. As part of this study we compared the performance of a Naïve Bayes and the C5.0 Decision Tree Algorithm for Predicting the Advice of the Student Enrolment.

References

Anil, B., Pasha, A., Aman, K. S., & Aditya, K. S. (2019). Multiple Machine Learning Classifiers for Student's Admission to University Prediction. *International Journal of Engineering and Advanced Technology (IJEAT)*, *8*(5), 192–198. https://www.ijeat.org/wp-content/uploads/papers/v8i5S/E10400585S19.pdf

Chiroma, H., Gital, A. Y., Abubakar, A., & Zeki, A. (2014). Comparing Performances of Markov Blanket and Tree Augmented Naïve-Bayes on the IRIS Dataset. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, *1*(1), 1–4. http://www.iaeng.org/publication/IMECS2014/IMECS2014_pp328-331.pdf

Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, *41*(4), 1937–1946. https://doi.org/10.1016/j.eswa.2013.08.089

Laudy, O. (2015, 1 augustus). *Blending/stacking model ensembles in IBM SPSS Modeler*. LinkedIn. https://www.linkedin.com/pulse/blendingstacking-model-ensembles-ibm-spss-modeler-olav-laudy

Lestari, M., Darmawan, A., Septiani, N. W. P., & Trihapsari, A. (2019). Non-Written Enrolment System using Classification Methods. *Journal of Physics: Conference Series*, *1175*, 012055. https://doi.org/10.1088/1742-6596/1175/1/012055

Ragab, A. H. M., Noaman, A. Y., Al-Ghamdi, A. S., & Madbouli, A. Y. (2014). A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining. *IDEE '14: Proceedings of the 2014 Workshop on Interaction*

Design in Educational Environments, 106–113. https://dl-acm-org.tilburguniversity.idm.oclc.org/doi/pdf/10.1145/2643604.2643631

Singh, D. M., Harbi, N., & Zahidur Rahman, M. (2010). Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection. *International journal of Network Security & Its Applications*, *2*(2), 12–25. https://doi.org/10.5121/ijnsa.2010.2202

-