

Performance Evaluation of ML-based Classifiers for HEI Graduate Entrants

Khrystyna Zub^a, Pavlo Zhezhnych^a

^a Lviv Polytechnic National University, S. Bandera str., 12, Lviv, 79013, Ukraine

Abstract

The development of intelligent decision support systems for admission to higher education institutions (HEI) is an essential task for both the institution, particularly for the selection of the best entrants, and for the entrant - to assess their chances of admission to the chosen HEI. The efficiency of such systems is largely based on the accuracy of the intelligent components underlying the system. This article investigates the effectiveness of machine learning (ML) based classifiers in solving the task of predicting the entry of entrants in the HEI. The simulation was performed using Orange software and a real data set. The task relates to binary classification in the case of an unbalanced data set. The simulation was performed by selecting the optimal operating parameters of each studied classifier and running it 100 times on a randomly generated data sample. This approach ensured the reliability of the results. Comparison of the accuracy of different classifiers was performed based on total accuracy, F-measure, Precision, and Recall measures. It has been experimentally established that Support Vector Machine (SVM) based classifiers demonstrated the highest accuracy in all four performance indicators among the considered methods. Receiver operating characteristic (ROC) curves in both classes also confirmed the highest accuracy of its work. This makes it possible to apply it in practice.

Keywords ¹

classifiers, performance evaluation, machine learning, university, HEI, graduate admissions, information systems, binary classification, imbalance dataset.

1. Introduction

An important point during the admission campaign, both for the educational institution and for those interested in admission, is the choice of specialty and HEI to obtain an education and qualification level. The effectiveness of the decision to choose a specialty made can directly affect both the activities of HEI and the further educational and professional trajectory of a potential student. Therefore, one of the critical factors influencing the entrant's choice is the assessment of their chances of admission.

Obviously, in order to assess their chances of admission, a student needs to consider many factors. Uncertainty of such a factor can lead to the fact that the student will eventually enter a HEI with a low rating while deserving of a more prestigious. Given the complexity of such a set and the incomplete awareness of entrants, it is unlikely that this can be determined independently in a proper way. In addition to the risk of choosing the wrong specialty, there are other difficulties. Applying for admission is, in any case, a cost of financial and time resources of the applicant. Also, admission to ranking HEI is accompanied by high competition between entrants, which makes the decision-making process for applying more challenging. Therefore, if we assume that the entrant has decided on a major, assessing the chances of admission to a particular HEI is a critical task.

Existing studies aimed at providing support to applicants in assessing the chances of admission, indicate the feasibility of using ML methods to solve this task.

This study [1] aims to provide entrants the probability to be admitted by the university. The gradient boosting regressor model was deployed using the data of student's academic performance and university rating. The study showed effective statistical results fetched by graduate admission chance prediction model.

Applying different types of artificial intelligence (AI) algorithms, authors [2] proposed Graduate

Admissions Prediction framework. Besides that, a user interface to interact with a user to see the result was proposed. Though from the proposed work, users are able to identify chances to get a seat without the possibility to get a list of universities in which they can obtain admission.

The authors of [3] claim that the disadvantage of existing admission prediction systems is using only if/else methods. They emphasize the need to use ML algorithms to solve this task. This study aimed to classify whether a student can get admission to a particular HEI. The dataset included previous years' entered student data based on specific attributes or parameters, which profoundly affect the class, attribute or have a high-value dependency. In addition, the purpose was to predict the number of potential students that have ready to enroll in the current HEI. This was made to help the education institution management work on these students interested in getting accepted/enrolled in the HEI. Authors use classification methods/algorithms in supervised learning such as Decision Trees, Support Vector Machines, k-Nearest Neighbors, Random Forest classifier, Naive Bayes classifier. Estimating classifiers help authors to choose one model by measuring the accuracy of each mode.

This study [4] presented a ML approach to predict the student's chances to be admitted helping them to recognize and target the universities which are best suitable for their profile. This paper evaluates these a few models to define the one that will give the highest accuracy rate and the least error. Authors proposed regression strategies to predict the university rate given the students' profile; namely, Linear Regression, Decision Tree (Tree), and Logistic Regression model. Logistic Regression model shows the most accurate prediction.

Additionally, there are commercial software solutions, which are regularly used by educational institutions, and aims to maintain the admission process. However, just a small part reflects supporting decision-making from the entrant's perspective. They provide wide enough functionality, but the problem of privacy, data security, and the high price of purchase and support process become challenging for many universities [5].

We should mention that today's rapid development of information technologies, makes HEIs find and implement the most effective technological solutions [6, 7]. Improving such technologies will increase the efficiency of the task of supporting applicants during the admission campaign [8]. A clear increase in the number of studies of ML methods confirms the relevance of their application in the context of the enrollment campaign in the HE. The Scopus search engine was used for analysis Search request: TITLE ((enroll* OR entran* OR admiss*) AND "Machine Learning") AND (LIMIT-TO (SUBJAREA , "COMP")) in the number of publications in 2021 is due to the date of the search query - September 2020. This search result confirms the interest of the scientific community in the application ML methods in the context of the admission process.

Therefore, given the relevance and effectiveness of the approaches described above to solve the prediction task, this study focuses on the methods of ML methods. The primary purpose of this work is to study and experimentally analyze the effectiveness of existing ML-based classifiers in resolving the task of binary classification in the case of an unbalanced data set. The applied task is to assess the chances of the entrant's admission to HEI.

2. Materials and methods

This study provides performance evaluation of the different ML-based classifiers for solving the task of prediction the possibility graduate admission. Modeling was conducted using Orange software [9]. Orange software is an online tool for data visualization, machine learning and open-source data analysis. It is equipped with a visual programming interface for fast high-quality data analysis and interactive data visualization.

2.1.1. Dataset descriptions

The task of binary classification was investigated in this work. The task was to predict whether the entrant will enter the university or not. The evaluation of the effectiveness of different classifiers was performed based on a real set of data [10] on admission to the US HEI. The authors collected information on admission to graduate school at 29 United States(US) universities. The data set contains 1653 records, nine attributes [10]. Author considered the below features for dataset: English test score,

Graduate Record Examination Score, Quantitative Reasoning sections (Gre Score Quat), Verbal Reasoning (Gre Score Verbal), Paper Published, Ranking, Undergraduation Score, Work experience.

The task is to determine whether the candidate will enter the Computer Science program at the chosen university or not (binary classification). The data set contains 574 successful entry cases and 1079 unsuccessful ones.

Since the authors of this set chose the most important attributes for solving the task, the set was cleared of omissions and anomalies [10], its previous processing in this work was only to normalize the data, which were then processed by the studied classifiers.

2.1.2. Modeling of the ML-based classifiers

In this work, we investigated the accuracy of solving the classification problem using a number of existing methods of machine learning, in particular:

- SVM;
- Naïve Bayes;
- Logistic regression;
- k-nearest neighbors (kNN);
- Neural Network;
- Tree;
- Stochastic gradient descent (SGD).

The simulation was performed using Orange software. The block diagram of this process is presented in Fig. 2.

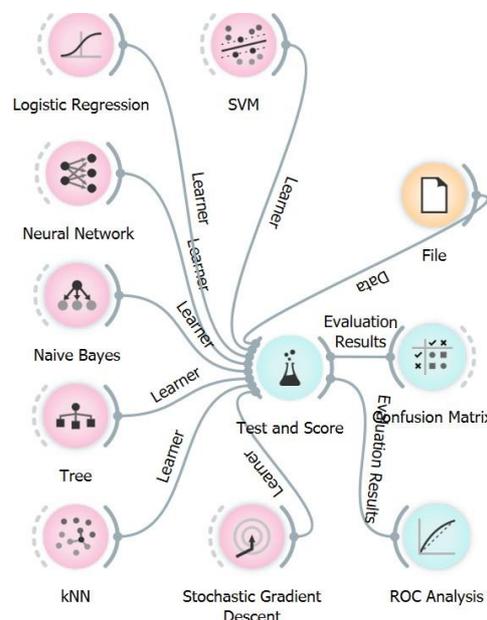


Figure 2: Modeling of the effectiveness of the ML-based classifiers using Orange software

According to the research methodology, the authors selected the optimal operating parameters of all studied classifiers. To ensure the reliability of the obtained result, the procedure of forming training and test samples in the ratio of 80% to 20% was random. In addition, each of the studied classifiers was run 100 times. Then the results were averaged and displayed on the screen.

Performance evaluation was conducted using such indicators: total accuracy, F-measure, Precision and Recall measures [11]. Visual analysis was performed using ROC-curves for each of the two classes separately.

3. Results and discussion

The results of all studied classifiers are summarized in Fig. 3.

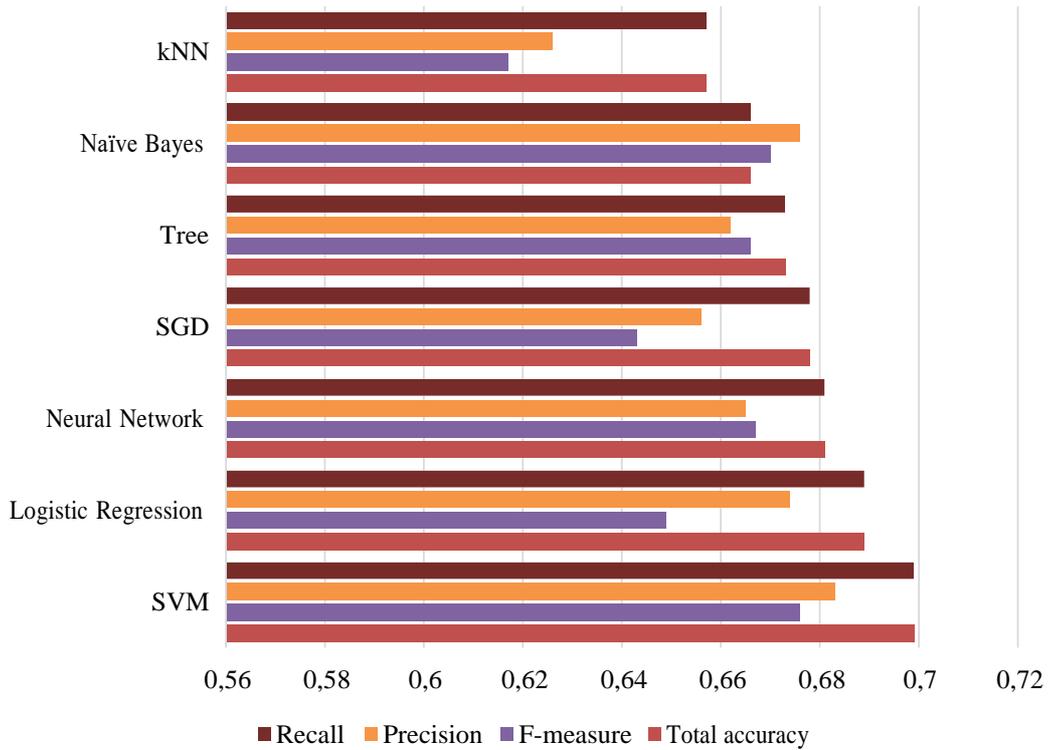


Figure 3: Results of the performance evaluation of all investigated classifier based on the Total Accuracy, F-measure, Precision and Recall indicators.

As can be seen from Fig.3 the kNN method demonstrates the lowest accuracy of work on all performance indicators. This was to be expected given that this simple non-parametric method has a very small number of parameters that need to be used to customize it for a specific task. A number of methods (Naïve Bayes, Tree, SGD and Neural Network) show approximately the same results for Total accuracy. However, SGD, despite a number of advantages in particular in terms of performance, shows very low performance for the F-measure. However, since this measure ignores true negative results, it should be neglected in case of solving an unbalanced problem. It should be noted that the data set processed in this paper is unbalanced (about 65% of one class and 35% of another).

Logistic regression classifier shows significantly better results on three indicators besides F-measure. However, this optimization algorithm is inferior in the fast-learning procedure to many of those studied in this work.

The highest accuracy on all four performance indicators was obtained when using SVM. The application of this fast and efficient algorithm for solving binary classification task in our case has fully justified itself. Despite this, the total accuracy reaches only 70%, which is quite a bit to solve the task.

To visualize the results of the study ROC-curve was used (Fig. 4). This is one of the most commonly used methods of demonstrating the results of binary classification.

The ROC-curve shows the dependence of the number of *True Positive* values on the number of *False Positive* values for each separate class. Accordingly, the studied classifier, the ROC-curve of which is above and to the left of the graph, demonstrates greater accuracy.

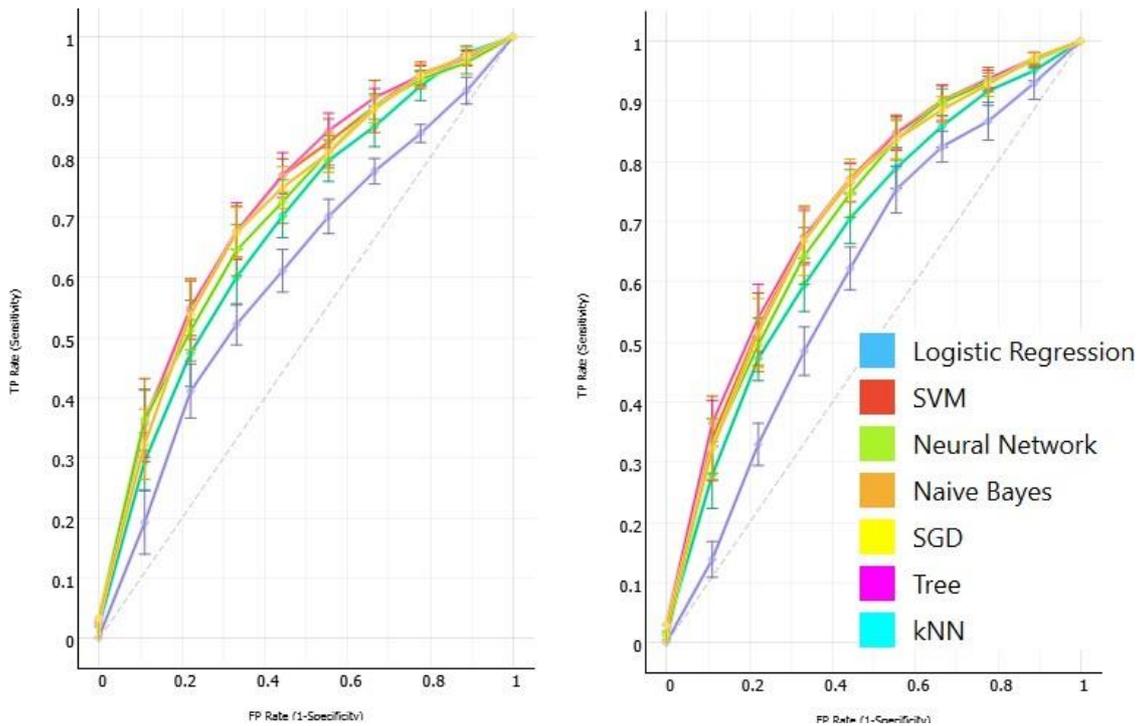


Figure 4: ROC-curves based on all classifiers: a) for first class; b) for second class.

As can be seen from both graphs of Figs. 4, ROC-curves several algorithms almost overlap. This indicates that they are approximately equally effective. This is confirmed by the results presented in Fig. 3. However, SVM shows slightly better results. This is also confirmed by numerical estimates of all four performance indicators from Figs. 3. It should also be noted the high speed of its work on samples of medium size [12]. All this provides the possibility of applying this method when building a real system for predicting the success of the entrant's entry to HEI.

4. Conclusions

The current state of development of decision support systems for entrants in the choice of HEI requires the use of data mining approaches. One of the typical tasks that can be the basis of such systems is classification. Today there are many different ML algorithms for its solution. The work aimed to evaluate their effectiveness in solving the problem.

The work of the studied classifiers was simulated using a sample of 1653 surveys on the results of admission to the specialty of Computer Science in 29 universities in the United States. The sample was normalized and randomly divided into two parts - training and test. To ensure the reliability of the obtained result, the work of each of the studied classifiers was performed ten times, then the results were averaged, and the final result was formed.

It has been experimentally established that the highest accuracy is obtained when using a classifier based on SVM with rbf-core.

Although the SVM-based classification method showed the highest accuracy, this value is still not high enough to use this classifier to develop real systems. Therefore, further research will be conducted in the direction of developing ensembles based on this method, particularly using a stacking approach, to improve the accuracy of the classifier. In addition, this approach will increase the reliability of classification subsystems based on it by using four or more models to obtain the final solution of the system.

The need for the entrant to making a decision about their choice of HEI and specialty to entry arises every year during the admission campaign. The task of supporting applicants remains relevant every year for all educational institutions. From the research, we could see the effectiveness of the application of methods and techniques of ML. However, in addition to defining the most effective method, there

are other tasks that require further research. A preliminary study of a variety of independent traits that may affect the outcome of prediction and recommendation in each individual case is critical.

5. References

- [1] Chakrabarty N., Chowdhury S., Rana S. (2020) A Statistical Approach to Graduate Admissions' Chance Prediction. In: Saini H., Sayal R., Buyya R., Aliseri G. (eds) *Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems*, vol 103. Springer, Singapore. https://doi.org/10.1007/978-981-15-2043-3_38
- [2] Saurabh Singhal, Ashish Sharma Prediction of Admission Process for Gradational Studies using AI Algorithm, *European Journal of Molecular & Clinical Medicine*, 2020, Volume 7, Issue 4, Pages 116-120
- [3] Devarapalli D.J. (2021) Classification Method to Predict Chances of Students' Admission in a Particular College. In: Gunjan V.K., Zurada J.M. (eds) *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications. Advances in Intelligent Systems and Computing*, vol. 1245. Springer, Singapore. https://doi.org/10.1007/978-981-15-7234-0_19
- [4] Amal AlGhamdi, Amal Barsheed, Hanadi AlMshjary, Hanan AlGhamdi A Machine Learning Approach for Graduate Admission Prediction IVSP '20: Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing March 2020 Pages 155–158 <https://doi.org/10.1145/3388818.3393716>
- [5] P. Zhezhnych, O. Berezko, K. Zub, and I. Demydov, 'Analysis of Features and Abilities of Online Systems and Tools Meeting Information Needs of HEIs' Entrants', *CEUR-WS.org*, vol. 2616, pp. 76–85, 2020.
- [6] P. Zhezhnych, I. Demydov, O. Berezko, and A. Shilinh, 'Corporate Culture Influence on the HEI's Information Image on the Internet', *Proceedings of the 1st International Workshop on Control, Optimisation and Analytical Processing of Social Networks (COAPSN-2019)*, vol. 2392, pp. 286–296, 2019.
- [7] Fedushko S., Ustyianovych T. (2020) Predicting Pupil's Successfulness Factors Using Machine Learning Algorithms and Mathematical Modelling Methods. In: Hu Z., Petoukhov S., Dychka I., He M. (eds) *Advances in Computer Science for Engineering and Education II. ICCSEEA 2019. Advances in Intelligent Systems and Computing*, vol 938. Springer, Cham. https://doi.org/10.1007/978-3-030-16621-2_58
- [8] Shilinh A., Zhezhnych P., Shakhovska N., Algorithm for forming the offer of educational services by higher education institutions to improve the technology of processing educational content by potential entrants, *Proceedings of the 1st Symposium on Information Technologies and Applied Sciences, IT and AS Volume 2824*, Pages 110 – 119, 2021
- [9] J. Demšar et al., 'Orange: Data Mining Toolbox in Python', *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.
- [10] A. Singh, A. Dhar, N. Jami, and S. Kashyap, 'Data Science Engineering Methods and Tools', pp. 1–14, 2017.
- [11] H. Dalianis, 'Evaluation Metrics and Evaluation', in *Clinical Text Mining*, Cham: Springer International Publishing, 2018, pp. 45–53. doi: 10.1007/978-3-319-78503-5_6.
- [12] I. Izonin, A. Trostianchyn, Z. Duriagina, R. Tkachenko, T. Tepla, and N. Lotoshynska, 'The Combined Use of the Wiener Polynomial and SVM for Material Classification Task in Medical Implants Production', *International Journal of Intelligent Systems and Applications*, vol. 10, no. 9, pp. 40–47, Sep. 2018, doi: 10.5815/ijisa.2018.09.05.