The Critical Feature Selection Approach using Ensemble Meta-Based Models to Predict Academic Performances

Muhammad Qasim Memon
Department of Information and Computing,
University of Sufism and Modern Sciences, Pakistan
memon_kasim@usms.edu.pk

Yu Lu
Advanced Innovation Center for Future Education,
Beijing Normal University, China
luyu@bnu.edu.cn

Shengquan Yu
Advanced Innovation Center for Future Education,
Beijing Normal University, China
yusq@bnu.edu.cn

Aasma Memon
School of Management and conomics,
Beijing University of Technology, China
kaasma.bjut@gmail.com

Abdul Rehman Memon
Department of Chemical Engineering,
Mehran University of Engineering and Technology,
Pakistan
enxarm@gmail.com

Abstract: *In this work, machine learning techniques are deemed to predict student academic performances in their historical performance of Final Grades (FGs). Acceptance of Technology enabled the teaching-learning processes, as it has become a vital element to perceive the goal of academic quality. Research is improving and growing fast in Educational Data Mining (EDM) due to many students' information. Researchers urge to invent valuable patterns about students' learning behavior using their data that needs to be adequately processed to transform it into helpful information. This paper proposes a prediction model of students' academic performances with new data features, including student's behavioral features, Psychometric, family support, learning logs via e-learning management systems, and demographic information. In this paper, data collection and pre-processing are firstly conducted following the grouping of students with similar patterns of academic scores. Later, we selected the applicable supervised learning algorithms, and then the experimental work was implemented. The performance of the student's predictive model assessment is comprised of three steps: First, the critical Feature selection approach is evaluated. Second, a set of renowned classifiers are trained and tested. Third, ensemble meta-based models are improvised to boost the accuracy of the classifier. Subsequently, the present study is associated with the solutions that help the students evaluate and improve their academic performance with a glimpse of their historical grades. Ultimately, the results were produced and evaluated. The results showed the effectiveness of our proposed framework in predicting students' academic performance.*

Keywords: *Educational data mining, students' prediction, machine learning, ensemble meta-based models, feature selection.*

Received April 11, 2022; accepted April 28, 2022
https://doi.org/10.34028/iajit/19/3A/12

1. Introduction

An overabundance of data relating to students in digital format permits the educational institutes to analyze the patterns for decision making. The acceptance of technology allows us to process and transform the data that will assist educators, administrators, and policymakers in bettering education quality [9]. higher education sectors believe students' academic performance is one of the essential criteria in assessing them. they also focus on producing graduates with superior academic performances and extra-curricular pursuits. Students' academic performance depends on socioeconomic, personal, and other environmental variables.

Meanwhile, understanding such aspects and their influence on the performance hep them to make an early decision. Evaluating students' information to categorize them in making good decisions or developing their performance is essential to research that focuses primarily on interpreting and analyzing educational data. Later, this data can be used for Learning Analytics (LA), which is provided to stakeholders via data mining methods known as Educational Data Mining (EDM). EDM allows that, in turn, reduces the learning overheads and the time and space to mine the student data and predict their academic performances based on the features, including academic, behavioral, and demographic. Surprisingly, educators associated with this domain require an early warning system as a prompt via EDM

[2]. the main key features of this study examine the status of students (i.e., pass or fail), grades, and final exam marks resembling the debilitation in their performances and achievements.

Underpinning the aims of the present study is the notion that the research described sheds light on the classification in conjunction with ensemble methods evaluated on a new collection of features, including demographic, academic, personal, family, psychometric, and learning log. further, we devised attribute selection which is based upon both a filter and a wrapper-based method. the main contribution of this study is the exploration of the best impactful features that play a vital role in predicting students' performance. Second, analysis of attribute selection based on improvised filter and wrapper based method. Third, a novel ensemble classification approach extracts invisible and intrinsic connections between the feature of students and their performances. fourth, the prediction of students' performances shows accurate and optimized outcomes for the betterment of education institutes.

In the previous works of EDM, prediction of students' performances emphasizes more with a set of features about academic subjects via marks only. The authors analyzed the typical progression of students using only academic marks in the entire degree program from the first to last semester [6]. however, students, especially reserved personalities, always feel less motivated and could not get higher marks. it entails that most researchers have carried predictions analysis with the academic features such as subjects in degree program/exams (i.e., induced with marks only), which is not validated when viewed from socio-economical and psychological lenses [3]. even though students' attainment doesn't rely upon final marks in their academic performance, adhering to functional performance, social attainment, parental support, psychometric analysis, and online activity via learning logs play a vital role in students' performances. Henceforth, researchers were involved in formulating prediction models that were constrained with academic attainment, where they neglected the attainment of social and functional features in their analysis as reported by [1, 4].

Conversely, researchers focused on academic, demographic, personal, social, and family as reported by [10]. the nature of features used in their experiments was lacking the effectually that could not impact the overall performance. for example, some features are used as parental support, assuming the mother and father should have a role during their nurturing. As a result, researchers emphasized their educational data, which was perceived only with parents' basic information and their occupation, which is insufficient in any educational setting. similarly, results of academic features (retained with marks/scores) in predicting students' performances may not be applicable and viable in the long term. In contrast, researchers improved their analysis by including demographic, personal, and academic attributes, but they could not exploit the analysis regarding real scenarios in educational settings. for instance, the prediction of students through online courses is the main focus in EDM (e.g., via MOOC), wherein features including the number of clicks, login attempts, etc., have remained core features in predicting students with low performances [11]

This study has been guided with such features easily accessible in any educational setting, and they can be associated with a student's performance. As discussed in section 1, EDM and LA play a vital part in students' cooperative learning; thereby, our analysis induced with the dataset that provides one of the features known as learning logs. it enables students to log into the system before final exams for their assessment and perform activities like questioning, clarification, and interpretation [8]. in addition, the teacher construes the learning session that shows the superiority of the contribution is imitated in the students' participation. we used quantitative and qualitative indicators before exams, such as the learning logs and psychometric features, in conjunction with Data Mining (DM) techniques to determine and build substitute representation and models for underlying data. in contrast, we use educational data in predicting the academic performance of students that in turn evaluating the effect of the different features, namely Demographic (DE), Academic (AC), Personal (PE), Psychometric (PS), Family (FA), and Learning Logs (ll). we examined our analysis in the realm of Machine Learning (ML), namely, tree, logistic, function, neural network, rule, and instance-based algorithms with ensemble methods such as boosting and bagging, as they enhance the model's accuracy.

This paper is the extension of already published paper, which include ensemble meta-base prediction model. we improvised a novel Ensemble Classification Approach (ECA), which applies different DM techniques in predicting the current status of students, and it analyses all features that a student ought to be reflected in any general educational setting.

The paper is further organized: section 2 presents the background and related literature. section 3 discusses the data collection and pre-processing steps followed in this study. section 4 describes the methodology of the proposed ensemble classification approach. section 5 presents the discussion of the result analysis. Finally, section 6 presents the conclusions.

2. Data Description

The progression period was analyzed from 2016-to 2018, wherein each year, students were logged into the learning log session once before the final exams. as mentioned in section 2, the dataset is collected from

the online educational system Smart Learning Partner (SLP) at advanced innovation center for future education, beijing normal university [5]. the data gathering consists of 11814 students from 31 local schools in beijing that were categorized into seven subjects. the features are divided into six categories: de, pe, ac, ps, fa, and ll. overall, students cover only three concepts in biology subject. the final grade comprises three partial components scores (i.e., PS, Learning Logs Score (LS), and Final Score (FS)). each score is weighted differently: ps is 25%, ll is 35%, and fs is 40%. This formula applies equally to all students and is a curricular definition fed into the data pre-processing (see Figure 1).

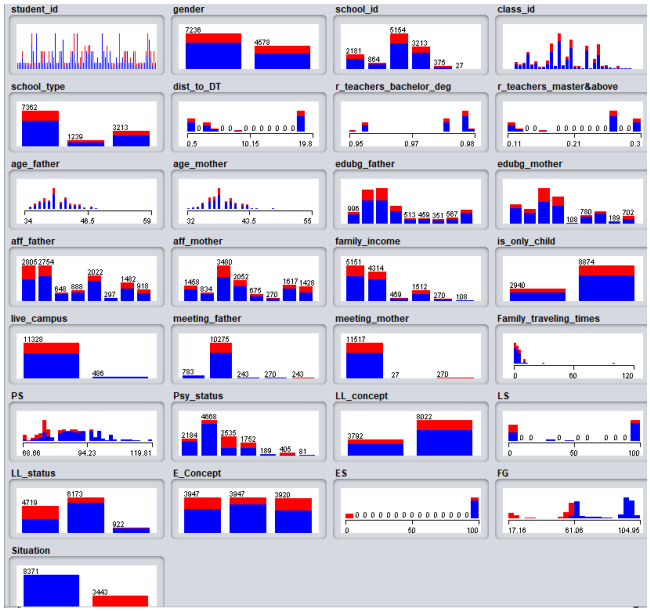

Figure 1. Correlation of all features used in this study.

Further, we eliminated the duplicate and null value records, and the main concern was to execute a procedure to anonymize the data to comply with international data protection standards. thus, this procedure is induced by removing or replacing the personal data fields (identification number, exam id, and course id). the experiments were performed in the microsoft azure studio, rapid miner, weka (https://www.cs.waikato.ac.nz/ml/weka/) (waikato environment for knowledge analysis), as they are well-known ML platforms to carry out a series of experiments. before data is fed into the experiments, it is presented via visual graphs to escalate the correlation between all the columns concerning the final result (pass or fail the course, column "situation"; red = “fail,” blue = “pass”).

3. Methodology

ECA follows the classification models evaluated based on cross-validation and split validation (see Figure 2). in the pre-processing phase, raw data is transformed into an organized format, and it is collected from SLP into a single repository (i.e., a data management system). redundancy is a common problem occurred when integrating data, and this is why we used a centralized database system to fetch students' data uniquely via queries. data consistency is handled by filtering the missing and noisy data, and the dataset occupied by this study does not contain any missing and outliers.

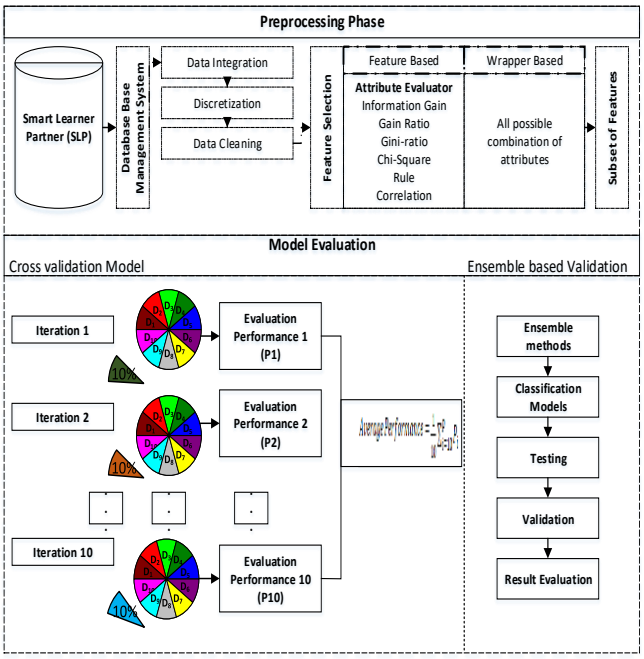

Figure 2. A framework of an ensemble classification approach.

Further, we transformed specified data from numerical values into bi-nominal values, such as attributes including 'live campus', and 'is_only_child' are converted into bi-nominal. The target variable (class label) is known as 'situation', which describes whether a student passes or fails. once data is ready, Attribute Selection (AS) is implemented using filter and wrapped-based methods. filter-based includes 'attribute evaluator' via search methods, namely 'ranker' and 'greedy stepwise'. on the contrary, wrapper methods use combinations of each pair of attributes and evaluate the best features that achieve higher accuracy. a 10-fold cross-validation model is performed in the model evaluation phase to validate the classification results. After that, ensemble-based classification is performed via split validation (i.e., 70% training and 30% testing). Finally, the comparison and results of ensemble meta-based classification against baseline model performances (via cross-validation) are validated.

4. Experiment Results

The proposed framework evaluates the analysis of the classification models with ensembles methods that can achieve better performances in predicting students' performances. we explain three essential issues:

1. Attribute analysis using attribute selection methods,
2. Analysis of state-of-the-art prediction models
3. Prediction model performances via ensemble methods. experiments were performed using Rapid miner studio, microsoft azure studio, and weka. the performances are measured in terms of accuracy, kappa, and F1-score via both 10-cross validation (i.e., baseline performances), and split validation (i.e., ensemble methods). experimental results highlight the RQs as discussed in section 1.

4.1. Filter Based Attribute Selection Via Ranking with Models

Another way to get feature weights in filter-based methods is to use the model-based attribute selection. it provides a resultant weight vector that describes whether an attribute is essential for the learning algorithm. the concrete calculation scheme is different for all learners. after that, we selected those models that can compute the weight score of attributes and evaluate their effect on student performance. weight score is evaluated using the models, including:

1. Decision tree - Gini Index (DT-GI).
2. Decision tree - Information Gain (DT-IG).
3. Decision tree -Gini Ratio (DT-GR).
4. Random Forest (RF).
5. Deep Learning (DL).
6. Logistic Regression (LR).
7. Naïve Bayes (NB).

each attribute can be distinguished in each model through a stacked bar via its unique pattern/color (see Figure 3).

It is worth noting that most selection of attributes remained the same as they found using the rank-based method, adhering to a total of 18 attributes and estimated score of weight ranging between 0.01 and 0.60. the weight score is described as the number of times the model ranks each attribute.

4.2. Wrapper Based Attribute Selection Using A Possible Combination Of All Attributes

Attributes used in this study are devised into six categories, and results performances are measured in terms of accuracy, kappa, and F1-score. first, we choose a single category of attributes (i.e., de, pe, fa, ac, ps, and ll), and the results indicate that demographic attributes gained low accuracy. at the same time, academic and family features achieved higher and constant scores in each model (see Figure 4). at the same time, remaining attributes' scores (e.g., pe, ps, ll) fluctuate from low or high scores due to the worst performances of some classification models. In the combination of two categories (such as fa, ps, pe and ll were given good performances as shown in Figure 5. additionally, fa+ll, ps+ll, and pe+ll gained highest score (.89, 0.9, 0.93), (0.86, 0.60, 0.91), and (0.9, 0.7, 0.93), respectively. In a set of three combination of categories, including family and learning log in combination with demographic, personal, and academic attributes showed better performances. results indicate that de+fa+ll, pe+fa+ll, and ac+fa+ll achieved the better scores (0.90, 0.74, 0.93), (0.90, 0.74, 0.93), and (0.90, 0.74, 0.93), respectively (see Figure 6). In a combination of four categories; de+pe+fa+ll, pe+fa+ac+ll, and ac+fa+ps+ll achieved improved scores (0.90, 0.77, 0.93), (0.90, 0.74, 0.93), and (0.90, 0.74, 0.93), respectively (see Figure 7).

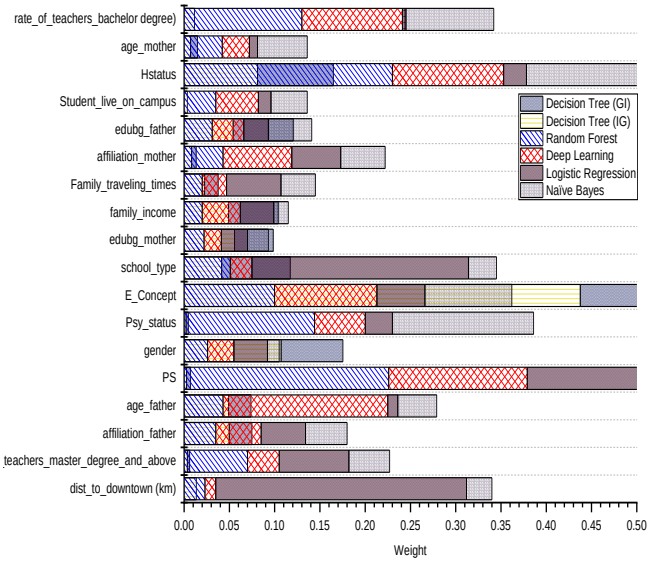

Figure 3. Attribute selection: performances of the model-based method.

4.3. Effect oF Attributes VIA Decision Trees

Figure 8 presents features' effects, including the academic, family, and psychometric categories. These attributes include “distance-to-downtown,” “students living on campus,” and teachers with bachelor's degrees, which impacts the final situation of students where the rate of teachers with bachelor's degrees per school is more significant. similarly, family attributes such as mother and father students impact their final situation as pass or fail. thus, it concludes that students can pass their exams with superior scores, including family, academic, and psychometric attributes. additionally, failed students could emphasize their academic, family, and psychometric scores that may help them pass the exams.

4. Ensemble Meta-Based Model

This experiment integrates classification models with ensemble meta based models (i.e., bagging and boosting and adaboost). only seven classification models were selected in this experiment whose performances were better using 10-fold cross-validation as reported in [7]. ensemble meta models

can optimize the performances of the classifiers and evaluate via accuracy, kappa, and F1-score. the results indicate that k-star and DT-IG achieved the highest scores, respectively (see Table 1). overall, the results performances of classification models DT-GI, DT-GR, RF-GI, RF-GR, and k-Star gained superior scores when ensemble with bagging and boosting. results validate that performance of each classification model gained high accuracy when used with all the ensemble meta-based model (i.e., bagging+Boosting+AdaBoost). for the most part, k-star and DT-IG were also found better when integrated with ensemble meta-based predition model. therefore, these results confirm that the ensemble meta-based can boost the prediction model performances.

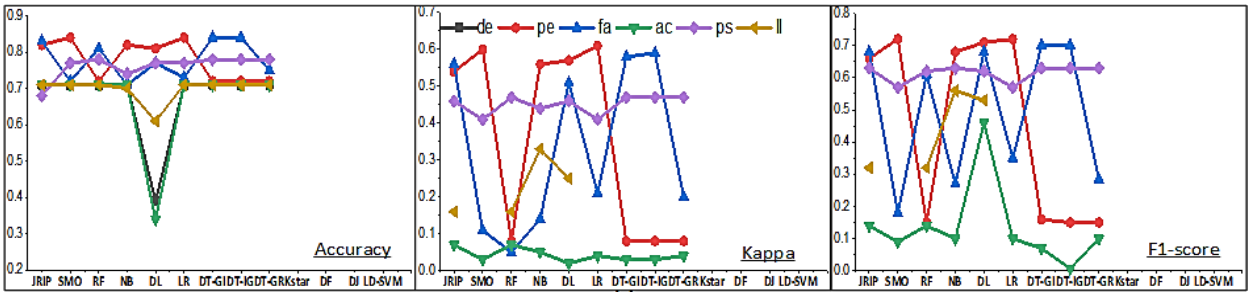

Figure 4. Attribute selection: performance of single category of attributes.

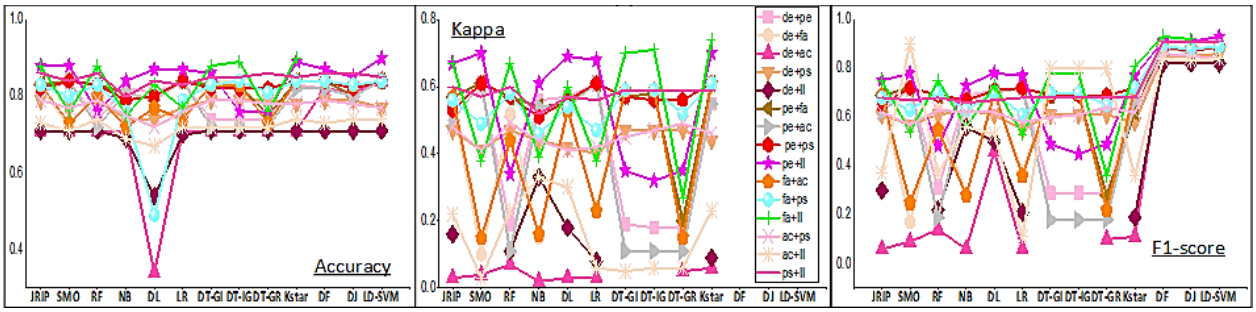

Figure 5. Attribute selection: performance of two category of attributes.

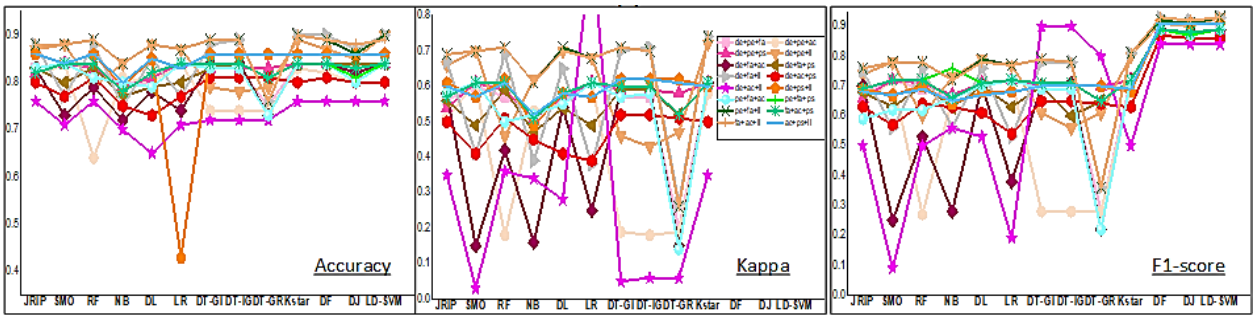

Figure 6. Attribute selection: performance of three category of attributes

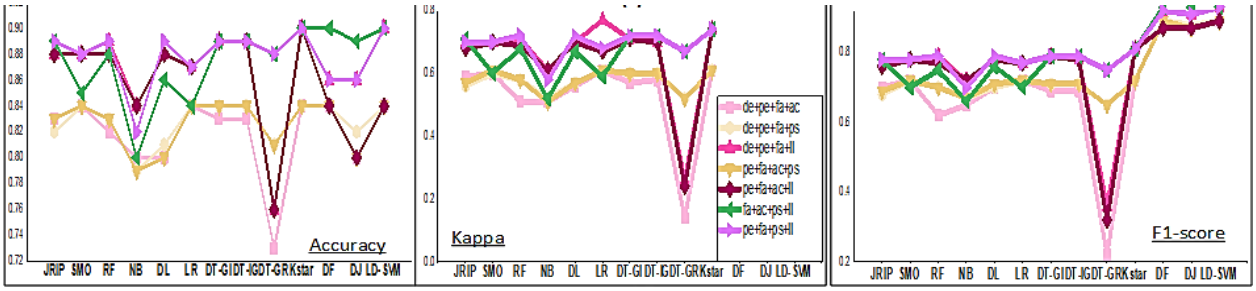

Figure 7. Attribute selection: performance of four category of attributes.

Table 1. Results performance using ensemble meta-based prediction model.

Model	Bagging+ AdaBoost_DT(IG)			Boosting +AdaBoost_DT(IG)			Bagging+Boosting+AdaBoost_DT(IG)		
	Acc	Kappa	FM	Acc	Kappa	FM	Acc	Kappa	FM
DT-IG	.92	.75	.82	.93	.75	.83	.95	.75	.83
DT-GI	.891	.73	.8	.891	.73	.8	.891	.73	.8
DT-GR	.91	.72	.81	.91	.72	.81	.91	.72	.81
RF-IG	.9	.73	.8	.9	.73	.8	.9	.73	.8
RF-GI	.9	.73	.8	.9	.73	.8	.9	.73	.8
RF-GR	.9	.73	.8	.9	.73	.8	.9	.73	.8
K-Star	.92	.745	.81	.93	.745	.82	.94	.745	.82

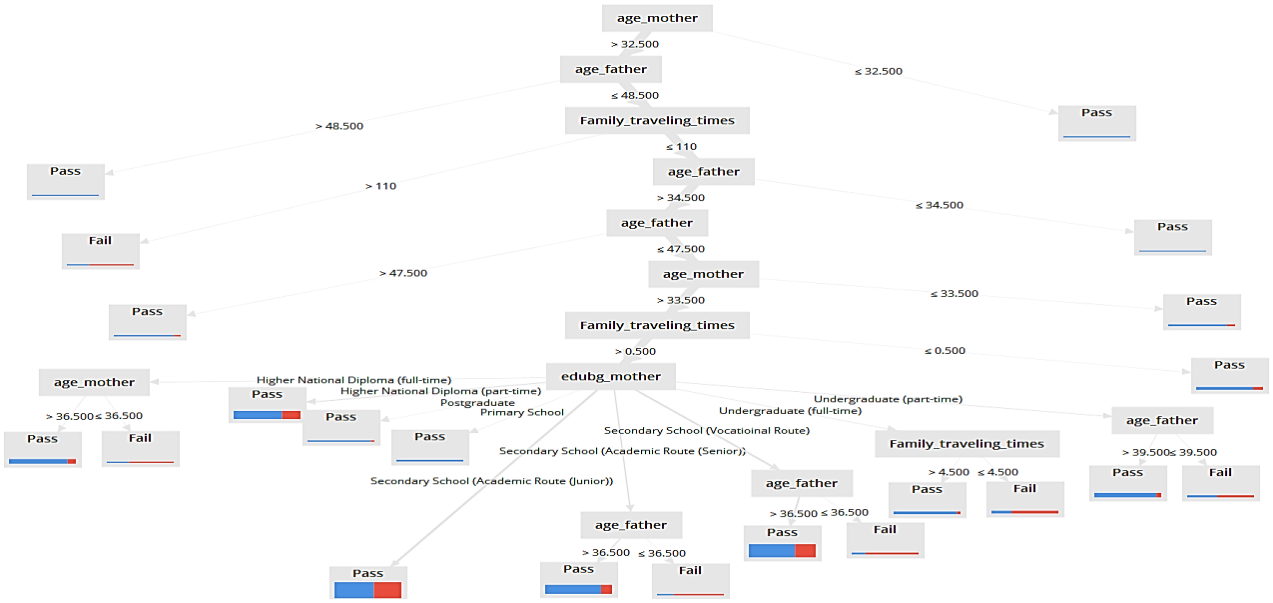

Figure 8. Impact of family attributes.

5. Conclusions

We present a novel classification approach that performs feature selection via an improvised filter and wrapper-based method. We testify our experiments on a large dataset composed of new features in different domains. the proposed framework validates the results via both 10-cross validation and split validation. moreover, ensemble meta-based model identify students' performances, whether they pass or fail in their final exams. it resembles an indiscernible and intrinsic connection between the feature of students and in their performances. our findings indicate that ensemble meta-based methods are devised with a classification model that can help predict students' performances. the results also indicate that features including family, psychometric, and learning logs are more impactful and can be given much attention in any educational setting in predicting student performance.

Acknowledgment

This research is partially supported by the fundamental research funds for the central universities, and the open project for the state key laboratory of cognitive intelligence (No.iED2021-M007)

References

[1] Ahmad Z., Memon M., Memon A., Munshi P., and Memon M., "A New Hybrid Approach of Gravitational Search Algorithm with Spiral-Shaped Mechanism-based RBF Neural Network," *In Proceeding of the 22nd International Arab Conference on Information Technology*, Jordan, pp. 1-6, 2021.

[2] Altrabsheh N., Cocea M., and Fallahkhair S., "Predicting Students' Emotions Using Machine Learning Techniques," *In Proceeding of the International Conference on Artificial Intelligence in Education*, Madrid, pp. 537-540, 2015.

[3] Asif R., Merceron A., Ali S., and Haider N., "Analyzing Undergraduate Students' Performance Using Educational Data Mining," *Computers and Education*,vol. 113, Jordan, pp. 177-194, 2017.

[4] Buenaño-Fernández D., Gil D., and Luján-Mora S., "Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study," *Sustainability*, vol. 11, no. 10, pp. 2833, 2019.

[5] Chen P., Lu Y., Zheng V., and Pian Y., "Prerequisite-Driven Deep Knowledge Tracing," *in Proceeding of the IEEE International Conference on Data Mining*, Singapore, pp. 39-48, 2018

[6] Hirokawa S., "Key Attribute for Predicting Student Academic Performance," *in Proceeding of the 10th International Conference on Education Technology and Computers*, Japan, pp. 308-313. 2018.

[7] Memon M., Qu S., Lu Y., Memon A., and Memon A., "An Ensemble Classification Approach Using Improvised Attribute Selection," i*n Proceeding of the 22nd International Arab Conference on Information Technology*, Jordan, pp. 1-5, 2021.

[8] Memon M., Lu Y., Chen P., Memon A., Pathan M., and Zardari Z., "An ensemble Clustering Approach for Topic Discovery Using Implicit Text Segmentation," *Journal of Information Science*, vol. 47, no. 4, pp. 431-457, 2021.

[9] Paiva R., Bittencourt I., Lemos W., Vinicius A., and Dermeval D,, "Visualizing Learning Analytics and Educational Data Mining Outputs," *in Proceeding of the International Conference on Artificial Intelligence in Education*, Cham, pp. 251-256, 2018.

[10] Shilbayeh S. and Abonamah A, "Predicting Student Enrolments and Attrition Patterns in Higher Educational Institutions Using Machine Learning," *International Arab Journal of Information Technology*, vol. 18, no. 4, pp. 562-567, 2021.

[11] Wan H., Ding J., Gao X., and Pritchard D., "Dropout Prediction in MOOCs using Learners' Study Habits Features," *in Proceeding of the 10th International Conference on Educational Data Mining*, pp. 408-409. 2017.

Muhammad Qasim Memon is currently working as an Assistant Professor in the Department of Computer Science, University of Sufism and Modern Sciences, Bhitshah. Dr. Memon is also a Post-doctorate fellow at the Advanced Innovation Center for Future Education (AICFE), Faculty of Education, Beijing Normal University, China. He received his Ph.D. degree from the School of Software Engineering at Beijing University of Technology, China, in 2018. He received his Bachelor of Engineering and Master of Engineering from Mehran University of Engineering & Technology Jamshoro (MUET) in 2009 and 2014. Dr. Qasim has published several papers in international conferences and research journals indexed in SCI., EI., and Scopus. Dr.Memon's research interests include Educational Data Mining, Text Analytics, Information Extraction, and Technology Education.

Yu Lu received his Ph.D. degree from the National University of Singapore. He is currently an Associate Professor with the Faculty of Education, Beijing Normal University, where he also serves as the director of the artificial intelligence (AI) lab and leads the research team for AI in education. Dr. Lu's research interests include educational data mining, learning analytics, pervasive computing and educational robotics.

Shengquan Yu is the executive director of AICFE at Beijing Normal University, and director of the joint laboratory for Mobile Learning, Ministry of Education-China Mobile Communication corporation. He also serves as the deputy dean of the faculty of education, Beijing Normal University. Professor Yu's research fields include mobile and ubiquitous learning, ICT and curriculum integration, network learning technology, and education information policy.

Aasma Memon completed her Ph.D. from the School of Economics and Management at Beijing University of Technology, China. She received her Bachelor in Arts and Masters in Public Administration from the University of Sindh, Jamshoro, Pakistan, in 2008 and 2012. Her research interests include firm performance and corporate sustainability, human resource management, and data mining.

Abdul Rehman Memon is currently a professor in the department of Chemical Engineering, Mehran University of Engineering & Technology, Jamshoro. Dr. Memon received his Ph.D. in Environmental Engg. from University of Nottingham in 2011. He completed his master and bachelor from MUET Jamshoro in 2004 and 1991. Dr. Memon's research interests include algal biofuels, Waste water Bioremediations, Bioenergy engineering, and pollution control engineering.