**International Journal of System Dynamics Applications**
Volume 10 • Issue 3 • July-September 2021

# Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques

Deepti Aggarwal, JSS Academy of Technical Education, Noida, India

Sonu Mittal, Jaipur National University, Jaipur, India

https://orcid.org/0000-0002-0452-9652

Vikram Bali, JSS Academy of Technical Education, Noida, India

https://orcid.org/0000-0002-2809-8455

## ABSTRACT

The academic institutions are focusing more on improving the performance of students using various data mining techniques. Prediction models are designed to predict the performance of students at a very early stage so that preventive measures can be taken beforehand. Various parameters (academic as well as non-academic) are considered to predict the student performance using different classifiers. Normally, academic parameters are given more weightage in predicting the academic performance of a student. This paper compares the two models: one built using academic parameters only and another using both academic and non-academic (demographic) parameters. The primary data set of students has been taken from a technical college in India, which consists of data of 6,807 students containing attributes. Synthetic minority oversampling technique filter is applied to deal with the skewed data set. The models are built using eight classification algorithms that are then compared to find the parameters that help to give the most appropriate model to classify a student based on his performance.

## KEYWORDS

Academic Parameters, Classification, Educational Data Mining, Ensemble Learning, Multi-Layer Perceptron, Non-Academic Parameters, Prediction, Synthetic Minority Oversampling Technique (SMOTE)

## INTRODUCTION

In today's competitive world, students are taking admission in different courses. But, all of the students are not able to complete their course because of some reason. Due to which, many students withdraw their admission from the course in between. Various factors, academic as well as non-academic, are associated with a student, which can help in predicting the performance of a student at a very early stage, so that some actions may be taken to improve student's performance.

While predicting the academic performance of a student, normally academic parameters like class X %age, class XII %age, Gap Year, etc. are given more weightage and considered vital as compared

DOI: 10.4018/IJSDA.2021070103

to non-academic (demographic) parameters like Gender, Address, Parent Income, Age, Category, etc. In this paper, prediction models are built using academic parameters only and using combination of academic and non-academic parameters also. The models are created using eight algorithms: Logistic Regression, Support Vector Machine(SVM), Multi-Layer Perceptron(MLP), J48, Random Forest, AdaBoost, Bagging and Voting, which are then compared to show that a model built using combination of academic and non-academic parameters is better as compared to a model built using academic parameters only. The demographic details of a student are equally important in predicting the performance of a student.

The data set is taken from three programs running in a technical institute in Uttar Pradesh, India. The data set contains the non-academic (demographic) details as well as the academic details of a student. The data set consists of 6807 samples with 20 attributes. The class variable taken is ‘Admission Status’ that tells if a student has completed his course or withdrawn his/her admission before completion. The models are built in a simulated environment using Waikato Environment for Knowledge Analysis (WEKA).

The paper contains the description of the dataset along with the academic and non-academic parameters, which is then followed by the experiments performed for showing the significance of non-academic parameters for building a model for predicting student’s performance. The results using academic parameters only and all parameters are shown in a tabular form and compared graphically. The paper concludes that the non-academic parameters are highly significant in predicting the student’s performance at an early stage.

## RELATED WORK

Bhardwaj (2020) used Artificial Neural Network to develop a prediction model for predicting annual medical claims and found that recurrent neural network outperformed feedforward neural network in terms of accuracy.

Yu *et al*. (2019) suggested an effective solution “active online-weighted ELM (AOW-ELM)”, based on “extreme learning machine (ELM) classification model”. Aggarwal *et al*. (2019) compared the studies on different machine learning techniques along with the feature selection techniques. The author focused on the correlation thresholds and variance thresholds for performing feature selection. Aggarwal *et al*. (2019) performed experiment on student data containing academic and non-academic attributes using six classifiers and found that MLP and Random Forest are the most promising classification algorithms to predict students’ performance. Panda (2019) introduced a hybrid classification method by combining distribution base balance-based instance selection and radial basis function neural network classifier to obtain a software defect prediction model. The software metrics with publicly available historical software defect datasets collected from several projects were used to build the prediction model. Abdollahi & Ebrahimi (2019) predicted the behaviour of a theatre complex in Iran for the year 2022 based on the assessment of the complex over the period 2012-2015 and also offered some insights into the problems and suggested practical solutions.

Aggarwal (2018) gave an overview of the machine learning techniques, tools and challenges for doing sentiment analysis. The author discussed the rule-based approach, lexicon-based approach, machine learning approach and hybrid approach for performing sentiment analysis. Collell *et al*. (2018) studied the combination of a bagging ensemble and threshold-moving and demonstrated its competitiveness on multiclass data using decision trees and neural networks. Mirza *et al*. (2018) combined the decision tree classification algorithm with SMOTE and achieved high accuracy on the model to predict diabetes prognosis. Elharakany *et al*. (2018) showed that ICT facilities play an important role in higher education. It is also an important factor in choosing the university by a student at the time of admission. Majhi (2018) used feed foreward neural network for breast cancer classification trained by a sine-cosine algorithm. The experiment was performed on the dataset of

Wisconsin hospital and the results showed that the proposed approach is quite robust and effective as compared to other classification alforithms.

Haixiang *et al*. (2017) provide a review of detecting rare event from an imbalanced learning perspective. Rare events, that are not in favour of mankind or society, often require responses from humans' decision-making. As the name suggests, rare events are observed rarely in day-to-day life. The authors provided a comprehensive classification of existing domains of imbalanced learning. Soni *et al*. (2017) computed the quality of higher technical Institute by using various attributes like placements, faculty strength, student's satisfaction, faculty satisfaction, etc. The authors constructed a dynamic model for policy planning to attain optimum quality in higher technical education system.

Krawczyk *et al*. (2016) discussed latest research challenges faced while learning from imbalanced data set related to real-world applications. The authors explained multiple aspects of imbalanced learning like classification, regression, clustering, big data analytics and mining data streams, which provides a thorough guide to evolving issues in various domains. Salunkhe *et al*. (2016) presented a novel approach that reduce the imbalance between the classes by applying pre-processing to the imbalanced dataset taken from KEEL repository. A comparative analysis shows the performance improvement in terms of Area under ROC Curve (AUC). Yijing *et al*. (2016) proposed a "multiple classifier system" to deal with multi class imbalanced learning problem, to distinguish between different kinds of imbalanced data.

Sun *et al*. (2015) proposed an "ensemble learning method, that converts an imbalanced data set into multiple balanced data sets and later builds different classification models on these multiple data sets using a particular classification algorithm". Sarakit *et al*. (2015) used SMOTE to balance the YouTube dataset and tested using the classifiers: Decision Tree, multinomial Naïve Bayes and Support Vector Machines. The results showed that Support Vector Machine gives the highest accuracy with 93.30% on filtering task and 89.44% on classification. Table 1 summarizes some of the related work done in the field of EDM.

After studying the various researches, that have been done for doing predictive analysis through different educational data mining techniques, the authors found that the parameters considered for building the prediction model are of two types: Academic and Non-academic and found that the following eight classification algorithms: Logistic regression, Support Vector Machine, J48 Decision Tree, Multilayer Perceptron, Random Forest, Voting, AdaBoost and Bagging are the most promising classifiers to build a prediction model.

## MATERIALS AND METHODS

The data set comprises of the students' details taken from three different programs of a technical institutes in the state of Uttar Pradesh, India. The students' details consist of their demographic details as well as academic details. The data set contains 6807 instances and 20 attributes. The academic and non- academic attributes contained in the data set are shown in Table 2 and Table 3.

Figure 1 depicts the sample view of the attribute set along with the values of attributes. The various attributes like Year of admission, Category, Gender, Year of birth, Month of birth, Age at the time of admission, Quota, Permanent State, Class X %age, Class XII %age, etc. are shown in the screenshot along with their values of different instances.

## MODELLING

Since the data set used in this research has imbalanced data, accuracy cannot be considered the appropriate measure to evaluate the classification algorithms. Hence, in this research, the evaluation metric used is F1-Score through which the performance of different classifiers is compared. The classifiers are evaluated using the following metrics:

**Table 1. Comparison of work done in the field of Educational Data Mining**

| S. No. | Paper Title | Author | Technique | Dataset | Accuracy |
|---|---|---|---|---|---|
| 1 | Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods | Berens *et al*. (2018) | ANN, Regression, AdaBoost, DT | Higher Educational Statistical Agency | 93% |
| 2 | Data mining for modelling students' performance: A tutoring action plan to prevent academic dropout | Burgos *et al*. (2018) | Logistic Regression | UDIMA | 97.13% |
| 3 | Early detection of university students with potential difficulties | Hoffait *et al.* (2017) | RF, LR, ANN | University of Li`ege (Belgium) | 70.6% for Logistic Regression |
| 4 | Predicting Student Performance using Advanced Learning Analytics | Daud *et al*. (2017) | SVM, C4.5, CART, BN, NB | Universities of Pakistan | 86.7% F1-Score of SVM |
| 5 | Towards the integration of multiple classifier pertaining to the Student's performance prediction | Pandey and Taruna (2016) | DT, K-NN and Aggregating One-Dependence Estimators (AODE) | Engineering college in India (Source not identified) | 98.96% for K-NN |
| 6 | Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis | Asri *et al*. (2016) | SVM, DT, NB, K-NNs | Wisconsin Breast Cancer | 97.13% |
| 7 | Modeling and Predicting Students' Academic Performance Using Data Mining Techniques | Mueen *et al*. (2016) | NB, MLP, C4.5 | SILO | 85.7% for Naïve Bayes |
| 8 | Machine Learning Application in MOOCs: Dropout Prediction | Liang *et al*. (2016) | SVM, LR, RF, GBDT | XuetangX platform | 88% for GBDT |
| 9 | A Review on Predicting Student's Performance using Data Mining Techniques | Shahiri *et al*. (2015) | NN, NB, K-NN, SVM and DT | Malaysia University | 98% for Neural network |
| 10 | Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory | Xing *et al*. (2015) | DT, LR, ANN, NB | Unidentified Source | 77.7% for Naïve Bayes |
| 11 | Exploring Machine Learning Methods to Automatically Identify Students in Need of Assistance | Ahadi *et al*. (2015) | BN, NB, DT, Conjunctive Rule, PART, AD Tree, J48, RF | University of Helsinki | 93% for J48 |
| 12 | Predicting Students Performance in Educational Data Mining | Guo *et al*. (2015) | NB, MLP, SVM, SPPN | Junior high schools in Hubei province. | 77.2% for SPPN |
| 13 | Classification and prediction-based data mining algorithms to predict slow learners in education sector | Kaur *et al*. (2015) | MLP, NB, SMO, J48 and REPTree | High School (Unidentified) | 93% for Decision tree |
| 14 | Predicting Student Performance by Using Data Mining Methods for Classification | Kabakchieva (2013) | DT, NB, K-NN | Bulgarian university | 66.59% for Decision Tree |
| 15 | Data Mining Approach for Predicting Student Performance | Osmanbegovic *et al*. (2012) | NB, MLP, C4.5 | University of Tuzla | 76.65% for Naïve Bayes |
| 16 | A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education | Kotsiantis *et al*. (2010) | NB, NN and the WINNOW algorithms | Students' Registry of the HOU | 78.95% for ensemble method |

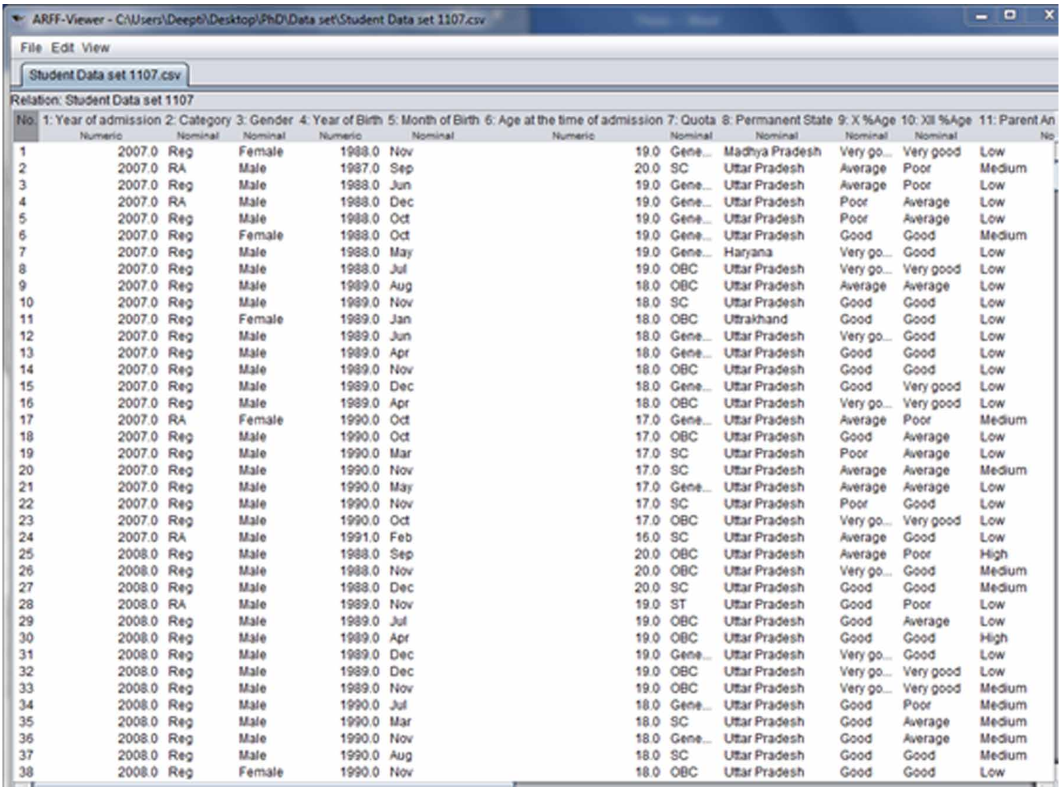
**Table 2. Academic parameters used in Data Set**

| Attribute Label | Values |
|---|---|
| **Year of Admission** | Numeric |
| **X %Age** | (Poor, Average, Good, Very Good, Excellent)<br>If X %age < 60%, then Poor<br>If 60%<= X %age < 70%, then Average<br>If 70%<= X %age < 80%, then Good<br>If 80%<= X %age < 90%, then Very Good<br>If 90%<= X %age <= 100%, then Excellent |
| **XII %Age** | (Poor, Average, Good, Very Good, Excellent)<br>If XII %age < 60%, then Poor<br>If 60%<= XII %age < 70%, then Average<br>If 70%<= XII %age < 80%, then Good<br>If 80%<= XII %age < 90%, then Very Good<br>If 90%<= XII %age <= 100%, then Excellent |
| **X Pass Year** | Numeric |
| **XII Pass Year** | Numeric |
| **Gap Year (Gap after class XII)** | Numeric |
| **Program** | (B.Tech., M.C.A., M.B.A.) |
| **Branch** | (IT, EE, EC, ME, MT, CS, CE, IC, EEE, M.C.A., M.B.A.) |
| **Admission Through** | (Counselling, Vacant Seat, Direct)<br>Counselling – State counselling<br>Vacant Seat – Counselling at college<br>Direct – Management quota |
| **Entrance Test Year** | Numeric |
| **Course completed in stipulated time** | (Yes, No) |
| **Admission status** | (Alumni, Admission Withdrawn) |

**Table 3. Non- Academic parameters used in Data Set**

| Attribute Label | Values |
|---|---|
| **Category** | (Reg, LE, RA)<br>Reg - Regular<br>LE - Lateral Entry<br>RA - Re-Admission |
| **Gender** | (Male, Female) |
| **Year of Birth** | Numeric |
| **Month of Birth** | (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec) |
| **Age at the Time of Admission** | Numeric |
| **Quota** | (General, SC, ST, OBC) |
| **Permanent State** | 21 Indian states |
| **Parent Annual Income (F)** | (Low, Medium, High)<br>If PAI<300000, then Low<br>If PAI>=300000 but less than 500000, then Medium<br>If PAI>=500000, then High |

**Figure 1. Screenshot of dataset with attributes**



ARFF-Viewer - C:\Users\Deepti\Desktop\PhD\Data set\Student Data set 1107.csv

File Edit View

Student Data set 1107.csv

Relation: Student Data set 1107

| No. | 1: Year of admission | 2: Category | 3: Gender | 4: Year of Birth | 5: Month of Birth | 6: Age at the time of admission | 7: Quota | 8: Permanent State | 9: X %Age | 10: XII %Age | 11: Parent An |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Numeric | Nominal | Nominal | Numeric | Nominal | Numeric | Nominal | Nominal | Nominal | Nominal | No |
| 1 | 2007.0 | Reg | Female | 1988.0 | Nov | 19.0 | Gene... | Madhya Pradesh | Very go... | Very good | Low |
| 2 | 2007.0 | RA | Male | 1987.0 | Sep | 20.0 | SC | Uttar Pradesh | Average | Poor | Medium |
| 3 | 2007.0 | Reg | Male | 1988.0 | Jun | 19.0 | Gene... | Uttar Pradesh | Average | Poor | Low |
| 4 | 2007.0 | RA | Male | 1988.0 | Dec | 19.0 | Gene... | Uttar Pradesh | Poor | Average | Low |
| 5 | 2007.0 | Reg | Male | 1988.0 | Oct | 19.0 | Gene... | Uttar Pradesh | Poor | Average | Low |
| 6 | 2007.0 | Reg | Female | 1988.0 | Oct | 19.0 | Gene... | Uttar Pradesh | Good | Good | Medium |
| 7 | 2007.0 | Reg | Male | 1988.0 | May | 19.0 | Gene... | Haryana | Very go... | Good | Low |
| 8 | 2007.0 | Reg | Male | 1988.0 | Jul | 19.0 | OBC | Uttar Pradesh | Very go... | Very good | Low |
| 9 | 2007.0 | Reg | Male | 1989.0 | Aug | 18.0 | OBC | Uttar Pradesh | Average | Average | Low |
| 10 | 2007.0 | Reg | Male | 1989.0 | Nov | 18.0 | SC | Uttar Pradesh | Good | Good | Low |
| 11 | 2007.0 | Reg | Female | 1989.0 | Jan | 18.0 | OBC | Uttrakhand | Good | Good | Low |
| 12 | 2007.0 | Reg | Male | 1989.0 | Jun | 18.0 | Gene... | Uttar Pradesh | Very go... | Good | Low |
| 13 | 2007.0 | Reg | Male | 1989.0 | Apr | 18.0 | Gene... | Uttar Pradesh | Good | Good | Low |
| 14 | 2007.0 | Reg | Male | 1989.0 | Nov | 18.0 | OBC | Uttar Pradesh | Good | Good | Low |
| 15 | 2007.0 | Reg | Male | 1989.0 | Dec | 18.0 | Gene... | Uttar Pradesh | Good | Very good | Low |
| 16 | 2007.0 | Reg | Male | 1989.0 | Apr | 18.0 | OBC | Uttar Pradesh | Very go... | Very good | Low |
| 17 | 2007.0 | RA | Female | 1990.0 | Oct | 17.0 | Gene... | Uttar Pradesh | Average | Poor | Medium |
| 18 | 2007.0 | Reg | Male | 1990.0 | Oct | 17.0 | OBC | Uttar Pradesh | Good | Average | Low |
| 19 | 2007.0 | Reg | Male | 1990.0 | Mar | 17.0 | SC | Uttar Pradesh | Poor | Average | Low |
| 20 | 2007.0 | Reg | Male | 1990.0 | Nov | 17.0 | SC | Uttar Pradesh | Average | Average | Medium |
| 21 | 2007.0 | Reg | Male | 1990.0 | May | 17.0 | Gene... | Uttar Pradesh | Average | Average | Low |
| 22 | 2007.0 | Reg | Male | 1990.0 | Nov | 17.0 | SC | Uttar Pradesh | Poor | Good | Low |
| 23 | 2007.0 | Reg | Male | 1990.0 | Oct | 17.0 | OBC | Uttar Pradesh | Very go... | Very good | Low |
| 24 | 2007.0 | RA | Male | 1991.0 | Feb | 16.0 | SC | Uttar Pradesh | Average | Good | Low |
| 25 | 2008.0 | Reg | Male | 1988.0 | Sep | 20.0 | OBC | Uttar Pradesh | Average | Poor | High |
| 26 | 2008.0 | Reg | Male | 1988.0 | Nov | 20.0 | OBC | Uttar Pradesh | Very go... | Good | Medium |
| 27 | 2008.0 | Reg | Male | 1988.0 | Dec | 20.0 | SC | Uttar Pradesh | Good | Good | Medium |
| 28 | 2008.0 | RA | Male | 1989.0 | Nov | 19.0 | ST | Uttar Pradesh | Good | Poor | Low |
| 29 | 2008.0 | Reg | Male | 1989.0 | Jul | 19.0 | OBC | Uttar Pradesh | Good | Average | Low |
| 30 | 2008.0 | Reg | Male | 1989.0 | Apr | 19.0 | OBC | Uttar Pradesh | Good | Good | High |
| 31 | 2008.0 | Reg | Male | 1989.0 | Dec | 19.0 | Gene... | Uttar Pradesh | Very go... | Good | Low |
| 32 | 2008.0 | Reg | Male | 1989.0 | Dec | 19.0 | OBC | Uttar Pradesh | Very go... | Very good | Low |
| 33 | 2008.0 | Reg | Male | 1989.0 | Nov | 19.0 | OBC | Uttar Pradesh | Very go... | Very good | Medium |
| 34 | 2008.0 | Reg | Male | 1990.0 | Jul | 18.0 | Gene... | Uttar Pradesh | Good | Poor | Medium |
| 35 | 2008.0 | Reg | Male | 1990.0 | Mar | 18.0 | SC | Uttar Pradesh | Good | Average | Medium |
| 36 | 2008.0 | Reg | Male | 1990.0 | Nov | 18.0 | Gene... | Uttar Pradesh | Good | Average | Medium |
| 37 | 2008.0 | Reg | Male | 1990.0 | Aug | 18.0 | SC | Uttar Pradesh | Good | Good | Medium |
| 38 | 2008.0 | Reg | Female | 1990.0 | Nov | 18.0 | OBC | Uttar Pradesh | Good | Good | Low |

- Precision (P):

$$\text{Precision} = \frac{TP}{TP + FP}$$

i.e number of true positive classifications divided by the sum of true positive classifications and false positive classifications

- Recall (R):

$$\text{Recall} = \frac{TP}{TP + FN}$$

i.e number of true positive classifications divided by the sum of true positive classifications and false negative classifications

- F-measure (F1-Score)

F-measure is the harmonic mean of precision and recall. i.e.:

$$\text{F-measure} = \frac{2 * P * R}{(P + R)}$$

The proposed system has been divided into two parts. The first part (Experiment 1) includes building the prediction models using different classifiers with academic parameters and the second part (Experiment 2) builds the prediction model using the same eight classifiers with both Academic as well as Non-Academic parameters. The experiments first balance the dataset using SMOTE filtering and then predicts whether a student will withdraw his/her admission or complete the course.

### Experiment 1

In the first experiment, the prediction models are build using eight classification algorithms: Logistic regression, SVM, J48 Decision Tree, Multilayer Perceptron, Random Forest, Voting, AdaBoost and Bagging. The parameters considered for making the model are academic parameters only: Year of Admission, X %age, XII % age, X Pass Year, XII Pass Year, Gap Year, Program, Branch, Admission Through, Entrance Test Year and Course completed in stipulated time. The Precision, Recall and F1-Score values for all the models are shown in Table 4. The table shows that the highest F1-Score achieved using classifiers is 79.6% (using Logistic Regression, Multi-Layer Perceptron and Voting meta classifier, where voting classifier is an ensemble learning method using J48 Decision Tree and Multi-Layer Perceptron).

### Experiment 2

In the second experiment, the prediction models are build using eight classification algorithms with SMOTE: Logistic regression, SVM, J48 Decision Tree, Multilayer Perceptron, Random Forest, Voting, AdaBoost and Bagging. All the parameters (academic as well as non-academic) are considered for making the model. The Precision, Recall and F1-Score values for all the models are shown in Table 5. The table shows that the highest F1-Score achieved using classifiers is 93.8% (using Random Forest meta classifier).

## RESULTS AND DISCUSSION

The experiments conducted in this research allows us to compare the models predicted using Academic parameters only and using all (academic & demographic) parameters. The F1- Score is used for comparing the performance of different models. The comparison of F1-Score is shown in Table 6,

**Table 4. Detailed Accuracy of classifiers for class 'Admission Withdrawn' with Academic Parameters only**

| Classifier | Precision | Recall | F1-Score (%age) |
|---|---|---|---|
| **J48 Decision Tree** | 72.9 | 84.4 | 78.2 |
| **Logistic Regression** | 72.4 | 88.3 | 79.6 |
| **Multi-Layer Perceptron** | 70.1 | 92.0 | 79.6 |
| **Support Vector Machine** | 68.9 | 93.3 | 79.3 |
| **AdaBoost** | 64.1 | 98.2 | 77.6 |
| **Bagging** | 71.3 | 86.2 | 78.1 |
| **Random Forest** | 75.6 | 80.7 | 78.0 |
| **Voting** | 72.0 | 89.0 | 79.6 |

**Table 5. Detailed Accuracy of classifiers for class 'Admission Withdrawn' with All Parameters**

| Classifier | Precision | Recall | F1-Score (%age) |
|---|---|---|---|
| **J48 Decision Tree** | 93.5 | 92.9 | 93.2 |
| **Logistic Regression** | 91.0 | 89.6 | 90.3 |
| **Multi-Layer Perceptron** | 92.5 | 90.5 | 91.5 |
| **Support Vector Machine** | 96 | 89 | 92.4 |
| **AdaBoost** | 100 | 85.9 | 92.4 |
| **Bagging** | 96.9 | 87.1 | 91.8 |
| **Random Forest** | 97 | 90.8 | 93.8 |
| **Voting** | 93.1 | 91.4 | 92.3 |

**Table 6. Comparison of F1-Score using Academic Parameters and All Parameters**

| Classifier | F1-Score using Academic Parameters | F1-Score using All Parameters |
|---|---|---|
| **J48 Decision Tree** | 78.2 | 93.2 |
| **Logistic Regression** | 79.6 | 90.3 |
| **Multi-Layer Perceptron** | 79.6 | 91.5 |
| **Support Vector Machine** | 79.3 | 92.4 |
| **AdaBoost** | 77.6 | 92.4 |
| **Bagging** | 78.1 | 91.8 |
| **Random Forest** | 78.0 | 93.8 |
| **Voting** | 79.6 | 92.3 |

where for a particular classifier, F1-Score of the model built using academic parameters only and F1-Score of the model built using all parameters is compared. The values in the Table 6 clearly indicates that the F1-Score is significantly higher if the model is built using all parameters for all the classifiers. The maximum F1-Score obtained using academic parameters only is 79.6% (with Logistic Regression, Multi-Layer Perceptron and Voting meta classifiers) whereas the highest F1-Score obtained using all parameters is 93.8% (using Random Forest classifier), which is significantly higher. The results obtained through the experiments conducted allow us to conclude that non-academic parameters, or demographic parameters, like age, gender, location, family income etc. can't be ignored while predicting the performance of a student. Only the academic parameters are not sufficient to predict whether a student will be able to cope up with the course or not. The best results are obtained only with the combination of both academic and non-academic parameters.

The charts in Figure 2 and Figure 3 clearly shows that the precision and F1-Score is higher for all classifiers when modelling is done using a combination of academic and non-academic parameters, i.e. all parameters.

## CONCLUSION

Predicting a student's performance is very crucial in today's competitive scenario. Normally, it is seen that the student's academic performance can be predicted using his/her previous academic parameters like Class X marks, Class XII marks, and so on. Based on academic parameters, a students' performance

**Figure 2. Comparison chart of Precision value for class 'Admission Withdrawn'**

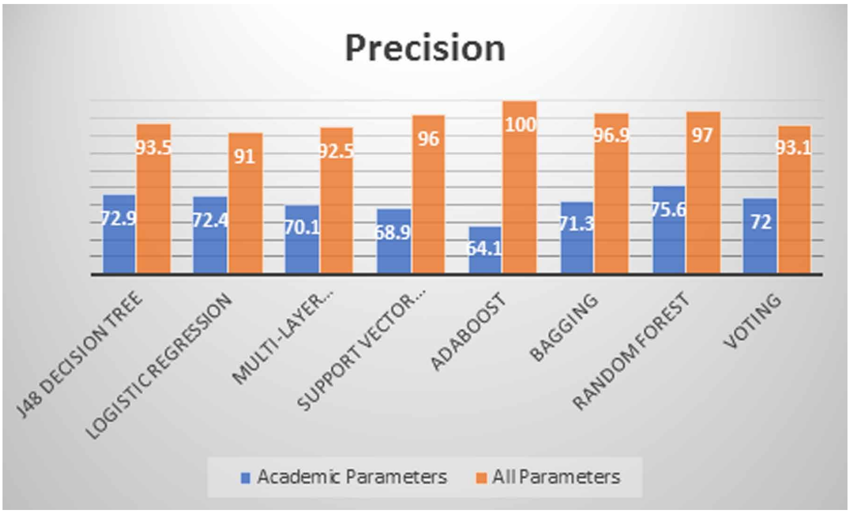
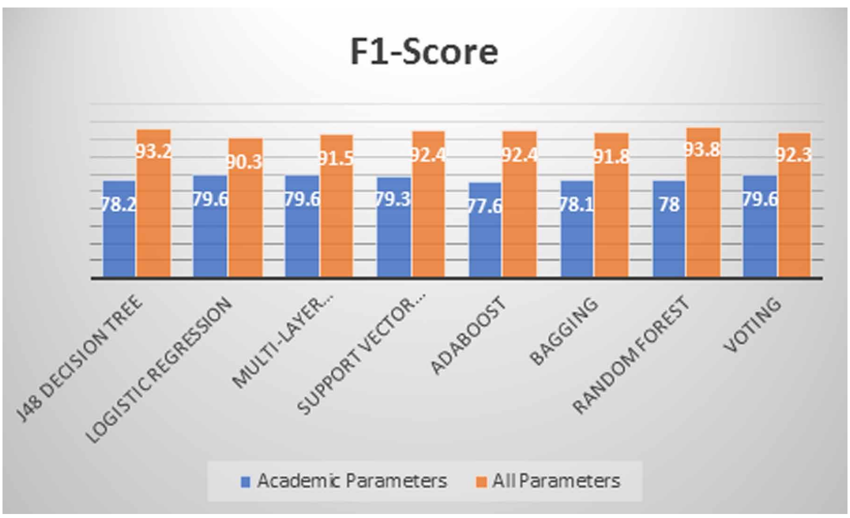



**Figure 3. Comparison chart of F1-Score for class 'Admission Withdrawn'**




can be predicted very well. If a student has performed well in class X or class XII or both, he/she is most likely to perform good in his/her graduation. On the other hand, if the performance in class X or class XII is poor, it's difficult for a student to cope up in his/her graduation also. But, the research in this paper allows us to conclude that a student's performance doesn't depend mainly on academic parameters, but also on the demographic (non-academic) parameters like Gender, Location, Parent's Income, Age, etc. The experiments conducted in this research allow us to conclude that if non-academic parameters are also considered along with academic parameters for predicting students' performance, the resultant models are much more effective. This has been proved by comparing F1-Score, which has improved in almost all the classification models if non-academic parameters are also considered with academic parameters. The results allow us to conclude that only the combination of academic and non-academic parameters can give us the most appropriate prediction model.

## ACKNOWLEDGMENT

The authors are highly grateful to the Principal, Management and Department of Computer Science and Engineering of JSS Academy of Technical Education, Noida, India and Jaipur National University, Jaipur, India to provide complete support in carrying out the research work and writing this research paper. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## REFERENCES

Abdollahi, H., & Ebrahimi, S. B. (2019). Modeling and Investigating the Economy and Production Structure of Iran Public Theater: A System Dynamics Approach. *International Journal of System Dynamics Applications*, *8*(1), 60–78. doi:10.4018/IJSDA.2019010104

Aggarwal, D., Mittal, S., & Bali, V. (2019). Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques. *International Journal of Recent Technology and Engineering, 8*(2S7), 496-503.

Aggarwal, D., Mittal, S., & Bali, V. (2019). An Insight into Machine Learning Techniques for Predictive Analysis and Feature Selection. *International Journal of Innovative Technology and Exploring Engineering*, *8*(9S), 342–349. doi:10.35940/ijitee.I1055.0789S19

Aggarwal, D. G. (2018). Sentiment Analysis: An insight into Techniques, Application and Challenges. *International Journal on Computer Science and Engineering*, *6*(5), 697–703.

Ahadi, A., Lister, R., Haapala, H., & Vihavainen, A. (2015). Exploring Machine Learning Methods to Automatically Identify Students in Need of Assistance. *Proceedings of the eleventh annual International Conference on International Computing Education Research*, 121-130. doi:10.1145/2787622.2787717

Asri, H., Mousannif, H., Moatassime, H. A., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, *83*, 1064–1069. doi:10.1016/j.procs.2016.04.224

BerensJ.SchneiderK.GörtzS.OsterS.BurghoffJ. (2018). *Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods.* CESifo Working Paper No. 7259., CESifo Group Munich. Available at https://ssrn.com/abstract=3275433

Bhardwaj, A. (2020). Health Insurance Claim Prediction Using Artificial Neural Networks. *International Journal of System Dynamics Applications*, *9*(3).

Burgos, C., Campanario, M. L., Peña, D. D. L., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, *66*, 541–556. doi:10.1016/j.compeleceng.2017.03.005

Collell, G., Prelec, D., & Patil, K. R. (2018). A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*, *275*, 330–340. doi:10.1016/j.neucom.2017.08.035 PMID:29398782

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. A., Abbas, F., & Alowibdi, J. S. (2017). Predicting Student Performance using Advanced Learning Analytics. *Proceedings of 26th International Conference on World Wide Web Companion*, 415-421. doi:10.1145/3041021.3054164

Elharakany, R. A., Moscardini, A., Khalifa, N. M., & Elghany, M. M. (2018). Modelling the Effect on Quality of Information and Communications Technology (ICT) facilities in Higher Education: Case Study—Egyptian Universities. *International Journal of System Dynamics Applications*, *7*(3), 1–30. doi:10.4018/IJSDA.2018070101

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015). Predicting Students Performance in Educational Data Mining. *International Symposium on Educational Technology (ISET)*, 125-128. doi:10.1109/ISET.2015.33

Hoffait, A. S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1–11. doi:10.1016/j.dss.2017.05.003

Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetics and Information Technologies*, *13*(1), 61–72. doi:10.2478/cait-2013-0006

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science*, *57*, 500–508. doi:10.1016/j.procs.2015.07.372

Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, *23*(6), 529–535. doi:10.1016/j.knosys.2010.03.010

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. doi:10.1007/s13748-016-0094-0

Liang, J., Li, C., & Zheng, L. (2016). Machine learning application in MOOCs: Dropout prediction. *11th International Conference on Computer Science & Education (ICCSE),* 52-57. doi:10.1109/ICCSE.2016.7581554

Majhi, S. K. (2018). An Efficient Feed Foreword Network Model with Sine Cosine Algorithm for Breast Cancer Classification. *International Journal of System Dynamics Applications*, *7*(2), 1–14. doi:10.4018/IJSDA.2018040101

Mirza, S., Mittal, S., & Zaman, M. (2018). Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree. *International Journal of Applied Engineering Research: IJAER*, *13*(11), 9277–9282.

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modelling and Predicting Students' Academic Performance Using Data Mining Techniques. *International Journal of Modern Education and Computer Science*, *8*(11), 36–42. doi:10.5815/ijmecs.2016.11.05

Osmanbegovic, E., & Suljic, M. (2012). Data Mining Approach for Predicting Student Performance. *Journal of Economics and Business*, *10*(1), 3–12.

Panda, M. (2019). Software Defect Prediction Using Hybrid Distribution Base Balance Instance Selection and Radial Basis Function Classifier. *International Journal of System Dynamics Applications*, *8*(3), 53–75. doi:10.4018/IJSDA.2019070103

Pandey, M., & Taruna, S. (2016). Towards the integration of multiple classifier pertaining to the Student's performance prediction. *Perspectives in Science*, *8*, 364–366. doi:10.1016/j.pisc.2016.04.076

Salunkhe, U. R., & Mali, S. N. (2016). Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach. *Procedia Computer Science*, *85*, 725–732. doi:10.1016/j.procs.2016.05.259

Sarakit, P., Theeramunkong, T., & Haruechaiyasak, C. (2015). Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm. *2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1-5. doi:10.1109/ICAICTA.2015.7335373

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. doi:10.1016/j.procs.2015.12.157

Soni, S., & Chorasia, B. (2017). Policy Planning in Higher Technical Education: A System Dynamic Approach. *International Journal of System Dynamics Applications*, *6*(3), 87–110. doi:10.4018/IJSDA.2017070105

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition*, *48*(5), 1623–1637. doi:10.1016/j.patcog.2014.11.014

Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. *Computers in Human Behavior*, *47*, 168–181. doi:10.1016/j.chb.2014.09.034

Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, *94*, 88–104. doi:10.1016/j.knosys.2015.11.013

Yu, H., Yang, X., Zheng, S., & Sun, C. (2019). Active Learning from Imbalanced Data: A Solution of Online Weighted Extreme Learning Machine. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(4), 1088–1103. doi:10.1109/TNNLS.2018.2855446 PMID:30137013