The Scientific Journal of Cihan University – Sulaimaniya **PP: 36-59**
Volume (8), Issue (2), December 2024
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Classification and Predicting of Student's Performance using Supervised Machine Learning

Barham A Ahmed[1], Soran A Saeed[2], Hozan K Hamarashid[3]

[1,3] Technical College of Informatics, Sulaimani Polytechnic University, Sulaimani, Iraq

[2]Higher Education and Scientific Affairs Department, Sulaimani Polytechnic University, Sulaimani, Iraq

Email: barham.arif.a@spu.edu.iq[1], soran.Saeed@spu.edu.iq[2], hozan.khalid@spu.edu.iq[3]

Abstract:

Predicting student performance is an issue in educational institutions that researchers frequently discuss as part of improving teaching and learning. If teachers make use of prediction techniques and features, appropriate educational content will be created. The research explores how machine learning can predict student performance using machine learning algorithms such as decision tree, random forest, gradient boosting, and others which are explored. two datasets including student demographic, academic, and behavioral variables have been combined with algorithms such as Decision Trees, Random forests, K-nearest neighbor, Support Vector Machine, Gradient Boosting and Ensemble Voting. The study focused on classification as Machine Learning task by implementing different classifier on two different datasets which are UCI student performance and E-Parwarda. The essential attributes that primarily affect evaluating the student's performance are presented that increase the accuracy of the prediction of the student's performance who wants to start study in university or institutes. The paper concludes that: Two different datasets are utilized to evaluate the models. In addition, various measures are computed such as train time, loss, precision, recall, f-score and accuracy. Consequently, Random Forest achieved highest accuracy (91.67) % based on E-parwarda dataset and GBoost achieved (87.69) % as highest accuracy-based UCI student performance dataset.

Keywords: Student Performance, Academic Achievement, Random Forest Model, Performance Prediction, Supervised Machine Learning, Student demographics.

DOI: http://dx.doi.org/10.25098/8.2.39

الملخص:

التنبؤ بأداء الطلاب هو قضية في المؤسسات التعليمية يناقشها الباحثون بشكل متكرر كجزء من تحسين التدريس والتعلم. إذا استخدم المعلمون تقنيات وميزات التنبؤ، فسيتم إنشاء محتوى تعليمي مناسب. يستكشف البحث كيف يمكن للتعلم الآلي التنبؤ بأداء الطلاب باستخدام خوارزميات التعلم الآلي مثل شجرة القرار والغابة العشوائية وتعزيز التدرج وغيرها والتي تم استكشافها. تم دمج مجموعتين من البيانات بما في ذلك المتغيرات الديموغرافية والأكاديمية والسلوكية للطلاب مع خوارزميات مثل أشجار القرار والغابات العشوائية وأقرب جار K وSVM وتعزيز التدرج والتصويت الجماعي. ركزت الدراسة على التصنيف كمهمة للتعلم الآلي من خلال تنفيذ مصنف مختلف على مجموعتين مختلفتين من البيانات وهما أداء طالب جامعة كاليفورنيا في إيرفين وإي-بارواردا. يتم تقديم السمات الأساسية التي تؤثر بشكل أساسي على تقييم أداء الطالب والتي تزيد من دقة التنبؤ بأداء الطالب الذي يريد بدء الدراسة في الجامعة أو المعاهد. وتخلص الورقة إلى ما يلي: يتم استخدام مجموعتين مختلفتين من البيانات لتقييم النماذج. بالإضافة إلى ذلك، يتم حساب مقاييس مختلفة مثل وقت التدريب والخسارة والدقة والتذكر والنتيجة f والدقة. وبالتالي، حققت الغابة العشوائية أعلى دقة (91.67)٪ بناءً على مجموعة بيانات E-parwarda وحققت GBoost (87.69)٪ كأعلى دقة قائمة على مجموعة بيانات أداء الطلاب في جامعة كاليفورنيا في إيرفين.

الكلمات المفتاحية: أداء الطلاب، التحصيل الأكاديمي، نموذج الغابة العشوائية، التنبؤ بالأداء، التعلم الآلي الخاضع للإشراف، التركيبة السكانية للطلاب.

پوختە:

پێشبینیکردنی ئەدای خوێندکاران پرسێکە لە دامەزراوە پەروەردەییەکان کە توێژەران زۆرجار باسی دەکەن وەک بەشێک لە باشترکردنی فێرکردن و فێربوون. ئەگەر مامۆستایان سوود لە تەکنیک و تایبەتمەندییەکانی پێشبینیکردن وەربگرن، ناوەڕۆکی پەروەردەیی گونجاو دروست دەکرێت. توێژینەوەکە لێکۆڵینەوە لەوە دەکات کە چۆن فێربوونی ئامێر دەتوانێت پێشبینی ئەدای خوێندکار بکات بە بەکارهێنانی ئەلگۆریتمەکانی فێربوونی ئامێر وەک دار بڕیار، دارستانی هەڕەمەکی، بەرزکردنەوەی گرادێنت و ئەوانی تر کە لێکۆڵینەوەیان لەسەر دەکرێت. دوو کۆمەڵە داتا کە گۆڕاوە دیمۆگرافی، ئەکادیمی و ڕەفتارییەکانی خوێندکاران لەخۆدەگرێت لەگەڵ ئەلگۆریتمەکانی وەک دارەکانی بڕیاردان، دارستانە هەڕەمەکییەکان، نزیکترین دراوسێی K، SVM، Gboost و Ensemble Voting. توێژینەوەکە تیشکی خستە سەر پۆلێنکردن وەک ئەرکی فێربوونی ئامێر بە جێبەجێکردنی پۆلێنکەری جیاواز لەسەر دوو کۆمەڵە داتا جیاواز کە بریتین لە ئەدای خوێندکارانی UCI و E-Parwarda. ئەو سیفەتە جەوهەریانەی کە بە پلەی یەکەم کاریگەرییان لەسەر هەڵسەنگاندنی ئەدای خوێندکار هەیە دەخرێنەڕوو کە وردی پێشبینیکردنی ئەدای خوێندکار زیاد دەکەن کە دەیەوێت لە زانکۆ یان پەیمانگاکان دەست بە خوێندن بکات. توێژینەوەکە بەو ئەنجامە دەگا کە: دوو کۆمەڵە داتا جیاوازەکان بەکاردەهێنرێن بۆ هەڵسەنگاندنی مۆدێلەکان. جگە لەوەش پێوەرە جۆراوجۆرەکان حیساب دەکرێن وەک کاتی ڕاهێنان، لەدەستدان، وردبینی، وەبیرهێنانەوە، f-score و وردبینی. لە ئەنجامدا، دارستانی هەڕەمەکی بەرزترین وردبینی (91.67) % بەدەستهێنا لەسەر بنەمای کۆمەڵە داتاکانی E-parwarda و GBoost (87.69) % وەک بەرزترین کۆمەڵە داتاکانی ئەدای خوێندکارانی UCI لەسەر بنەمای وردبینی بەدەستهێنا.

کلیلە وشە: ئەدای خوێندکار، دەستکەوتی ئەکادیمی، مۆدێلی دارستانی هەڕەمەکی، پێشبینی ئەدای کارکردن، فێربوونی ئامێری سەرپەرشتیکراو، دیمۆگرافیای خوێندکار.

1. Introduction

The economic success of every nation is extremely dependent on making higher education more affordable, and this should be one of the government's top priorities. The amount of time students spends studying in order to graduate is a factor contributing to the cost of their education [1]. In real life, predicting student performance is a difficult task [2]. In higher education institutions, student performance is a factor of success. As one of the criteria for a high-quality university, an outstanding record of academic accomplishments boosts the institution's rank. From the student's perspective, maintaining a high level of academic achievement will increase the likelihood of obtaining employment where academic achievement is a primary consideration [1]. The use of information technology (IT) in the education system can help institutions collect a large quantity of student data from various sources [3].

Student performance prediction can be extremely beneficial for educational institutions as well as for educators, allowing them to evaluate student performance more accurately and improve their performance by taking action when necessary. Such a prediction takes into account multiple attributes from offline and online learning settings, such as age, previous academic records, and student family features (size, marital status, etc.).

Educational institutions face a challenging obstacle when attempting to predict student performance. By providing decision-makers, educators, and students with beneficial prediction models, the educational process can be made more effective. In order to improve the accuracy of the model [6]. The study of learning performance prediction offers a basis for instructors to adjust the way they teach for students who may have difficulties by predicting students' performance on future exams, thereby reducing the likelihood of students failing the course and ensuring the quality of learning [7].

In this paper, Machine Learning used for classification as a solution for student performance prediction, the classifiers were (Decision Tree, Random Forest, AdaBoost, GBoost, Voting). Also, two datasets have been used to apply models on them which includes (University of California, Irvine "UCI" student performance and E-parwarda). This study selected features provide crucial insights into the demographics, family dynamics, and academic behaviors of students. This contains the number of siblings, sisters, and children, the student's age, gender, nationality, and place of birth, as well as his or her blood group, parental death, parental separation, and grade. The above-mentioned classifiers tested and applied on both datasets.

The Paper is divided into six sections: the first section introduced an introduction about the subject. Related work is shown in section two and then in third section explained a theoretical background for predicting performance of students. After that, in section four explained the proposed system and how the model is created. In the next section (section five), the experimental results are shown and discussed about the results. In the final section, concluded the summary of the paper.

DOI: http://dx.doi.org/10.25098/8.2.39

2. Literature Review

The literature review section provides an overview of studies focusing on the prediction of student performance using Machine Learning (ML) techniques. These studies investigate the correlation between various features and academic outcomes, evaluate the efficacy of various machine learning algorithms. In this part different datasets were used by the researchers to determine the most effective factors for predicting students' performance. Collectively, the studies provide valuable insights into the application of machine learning in educational settings, thereby guiding efforts to enhance student predictions and interventions.

[20] developed a framework using Machine Learning (ML) for predicting medical students' performance on high-stakes exam as Comprehensive Medical Basic Sciences Examination (CMBSE). To evaluate the models such as (LR, SVM, KNN), ensemble (Voting, stacking, Bagging (BG), RF, Adaptive boosting (ADA), Extreme Gradient boosting (XGB)) are applied to the dataset that consist 32 information about the medical students for five years that contains 1005 records. The article shows that (GPA) and grades has the strongest positive correlation with 38 outcomes.

[21] Tried to find the factors that possibly improve the student's performance by using Pearson correlation for each feature in student G3 Result. It's based on past result that negative impact correlation with grades while Mother's Education will positively impact with grades. The research to improve the factors that actually affects student mark by used Machine Learning (ML) models. The outcome shows that MLP 12-Neuron Model has the best RMSE value (4.32), then Random Forest RMSE (4.52) and Decision Tree RMSE (5.69).

[22] developed a system for student performance prediction using classification in a course. The research discovers the hidden patterns in large amount of available data. UCI Machinery student performance dataset contains 33 attributes and 649 records that utilized with Machine Learning algorithms such as SVM, c4.5, ID3 and Naïve Bayes. The algorithms analyzed on parameters such as accuracy and error rate. The analysis outcome shows that among of the techniques that used Support Vector Machine (SVM) has a most accuracy for classifying the student performance dataset.

[23] purposed a student's performance prediction framework to fix the previous problem utilizing knowledge-based data and learning behavioral data. The proposed system includes two-layer ensemble prediction: several ensemble prediction algorithms. ensemble progressive prediction. The first layer consists of some algorithms such as SVM, KNN and RF. Ensemble progressive prediction is the second layer that integrates that previous and current outcome of the ensemble prediction of the first layer. The aim of the research was to improve the prediction to higher education. Utilized students course data from higher education to graduation that helps to get their learning attributes throw their activity form course-taking.

Another research is by [24] Introduced a student's performance model, that evaluated the effectiveness of student attributes on academic student performance, after data collection and preprocessing applied to Naïve Bayes, KNN, ID3 and SVM. Then evaluated the classifiers and improved performance using ensemble methods (Voting, Bagging and Boosting). The model achieved 92.3% with behavioral features and 88.6% without behavioral features.

DOI: http://dx.doi.org/10.25098/8.2.39

[25] suggested a combination model for student performance prediction based on behavior characteristics of students as a solution to the issue of the imperfect information management platforms and limited data analysis skill in colleges and universities. Basically, the characteristic of GBDT algorithm, KNN, K-mean and ANN are investigated then algorithms are combined to create a prediction model. Then the prediction model combined with management system such as "Campus All in one card". The paper results demonstrate that, compared to a single algorithm, the combined model has faster runtime and greater accuracy and outcomes of the predictions are consistent with the actual situation.

[26] applied Machine learning project to predict student performance. That primary objective of the research to predict and identify student who may fail in semester examinations. This would help the instructors in providing such students with additional assistance. The dataset includes Roll No, Program, Course code, Course Title, Credit Hours, Grade point, Grade, Semester Year and Batch (year of enrolment). The study implemented Machine Learning algorithms such as Neural Network, Naïve Bayes, Support Vector Machine (SVM), and Decision Tree (DT). The result of comparative analysis has been conducted on the accuracy of the utilized algorithms. The outcome of the research demonstrates that Machine Learning is beneficial for forecasting. Although, there is still much work to be done with this technology.

[27] Purposed the use of Automated Machine Learning to improve the student performance prediction accuracy using the available previous to start the academic program. The researcher gathered the information for the study from various academic institutions in the UAE. Relied specially on student records from Admission, Registrar and Student Service agency. The dataset consisted of 1491 student records of 1014 student where in excellent academic standing. The research utilized a 10-fold cross-validation to assess the Ensemble Model accuracy. The model is trained on 90% from the points and evaluated on 10% over the course of separated runs. Using AutoML, the research gained 75.9% overall accuracy in this study with Kapa of 0.5.

[28] shows that the Decision Tree algorithm can precisely predict the academic performance of undergraduate student. As well as first and second year student grades are considered and analyzed with DT algorithm. Also, the study collected data from undergraduate degree colleges of Mumbai university that contains 600 students' data from 2019 to 2020.

[29] investigated multitask Machine Learning student performance prediction framework for traditional classroom-teaching which refers to identification of at-risk students for each course prior to its start. Five real datasets have been collected to student performance prediction. As outcome the research shows that many factors can be affected the process such as health, psychological state and family.

DOI: http://dx.doi.org/10.25098/8.2.39

3. Theoretical Background:

Supervised Learning (SL) attempts algorithms able to reason from externally provided examples in order to generate general hypotheses, which are then used to predict future instances. In other words, the objective of SL is to develop a model of the distribution of class labels based on predictor characteristics [11]. The purpose of training a supervised function during the learning process is to predict the future labels of unobserved data. Regression and classification are the two fundamental problems in supervised learning [12].

A. Classification:

Classification is one of the most widely used methods for predicting the academic performance of students. There are numerous classification methods that have been used for prediction such as Artificial neural network (ANN), decision tree, k-nearest neighbor (KNN), support vector machine (SVM), and Naive Bayes (NB) [3]. Unbalanced datasets necessitate caution when employing classification techniques, as they can produce inaccurate predictive accuracy [11].

1. **Decision Tree:** Decision Tree is also a supervised model for classification, constructed by dividing the dataset into root and node elements. The Decision Tree is used because it can perform well on large datasets with minimal data preparation, unlike Random Forest, a supervised classification algorithm [5]. The tree structure consists of hierarchically organized sets of rules, beginning with root attributes and ending with leaf nodes; every branch of the tree represents one or more outputs from the original dataset. The root node is the node at the summit of the tree without any inbound branches, and all outgoing branches represent each row in the dataset. The internal node of the tree is the node with both incoming and outgoing branches, and it can be used to assess the attribute. The terminal node or leaf is the only descending node with an incoming branch. This node represents the end node in the tree, which may contain multiple leaf nodes that represent the final calculations [12]. In comparison to other models, decision trees provide easier-to-understand categorization principles [4].
2. **Support Vector Machine:** Support Vector Machine (SVM) is one of the most powerful supervised machine learning algorithms that used for classification and regression task. SVM builds hyperplanes in multidimensional space to classify data, thereby separating class levels into distinct cases [10]. Support vector machine is primarily used for classification problems as a classifier. SVMs are extensively used in numerous applications due to their excellent classification performance. In the classification of the binary problem, the instances are differentiated using the hyper plane w Tx + b = 0, where w is a d-dimensional coefficient vector that is normal to the surface's hyperplane. The offset value from the origin is, and x represents the values of the data set. The SVM yields the outcomes w and b. In the linear case, the W can be solved by introducing Lagrangian multipliers. On borders, the points of the data set are known as support vectors [17].
3. **K-Nearest Neighbor (KNN):** KNN is an easy-to-understand machine learning algorithm in which an object is graded based on the majority vote of its neighbors [12]. Based on similarity measures such as a Euclidean distance metric and majority voting of the K nearest training sample class allocated to the test sample, it is computationally straightforward. KNN is an instance-based learner, also known as a lazy learner, because it defers training until a new student (test sample)

DOI: http://dx.doi.org/10.25098/8.2.39

needs to be classified (i.e., there is no training phase) and relies heavily on the matching scheme [18]. Calculating Euclidean distance is depicted in Fig. (1) [19]. The following disadvantages can be enumerated for KNN [18]:

- Every new student must calculate the distance to every training sample, resulting in a very high computational cost.
- The capacity needs are massive compared to the size of the training set.
- The accuracy rate of KNN on multidimensional data sets is minimal.

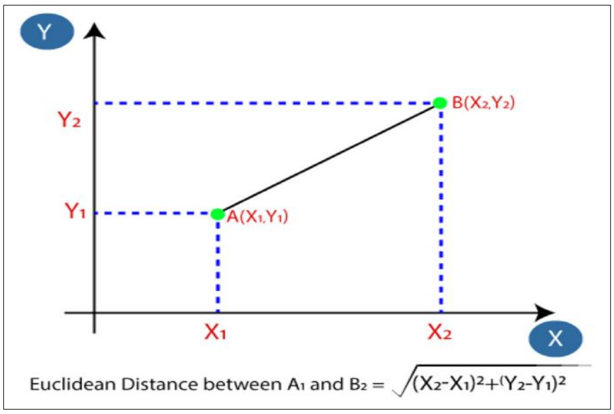

Y
Y_2
Y_1
$B(X_2,Y_2)$
$A(X_1,Y_1)$
X_1
X_2
X
Euclidean Distance between A_1 and B_2 = $\sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

Fig. (1): Calculation of Euclidean Distance b/w two points [19].

4. **Naive Bayes:** NB is a classifier that uses probability and the Bayesian Theorem to generate a prediction based on a set of predefined features [10]. The Naive Bayes (NB) algorithm was the most suitable for predicting both student performance and dropout probability. However, each case study has its own characteristics and nature; consequently, different techniques can be chosen to predict students' behavior [11]. As the most basic variation of the Bayesian network, the Nave-Bayes classification model is commonly considered. This model implies that, given the target attribute state, every feature attribute is independent of the other attributes. Each instance x within the dataset has attribute values a1, a2, ..., ai. Any value from the predefined finite set V= (v1, v2, ..., vj) is equal to the target function f(x). The equation used by the Naive Bayes model is [1]. Equation 2 illustrates Naïve Bayes as follows:

$Vmax = P(vj)\ vj \in V\ \ Max\Pi\ P(ai\ i|vj)$--(1)
Equation 1: Naïve Bayes.

$P\ (ai\ |v_j)$ and $P\ (v_j)$ can be determined by calculating their frequencies in the training dataset when v represents the model's target [1].

5. **Random Forest:** Random Forest is a reliable tagging technique based on the creation of multiple decision tree models. This method emphasizes two aspects of sampling: minimizing the number of training data and variables. Multiple decision trees are trained on arbitrarily selected subsets of training data to prevent overfitting. The final aggregate is determined through a majority vote on the model result. Consequently, there is less correlation between the models, and the ultimate model is more reliable [20].

Random Forest is a well-known algorithm for both classification and regression tasks in machine learning. It is an extension of Decision Trees in which multiple decision trees are trained on various subsets of data and the final prediction is obtained via a voting mechanism for classification or an averaging mechanism for regression.

The term of "Random" refers to two crucial elements of the algorithm:

- Random Subset: A random subset of the training data is chosen for each tree in the forest to generate diverse and independent trees.
- Random Feature Selection: At each node of the decision tree, a random subset of features is selected for splitting, ensuring that distinct trees utilize distinct subsets of features.

Random Forest (RF) has numerous benefits, such as high accuracy, resistance to overfitting, and strength against noise and outliers. It can also manage large datasets and feature spaces with high dimensions. Due to these benefits, Random Forest is utilized in a variety of applications, including regression, classification, feature selection, and identifying anomalies. Fig. (2) shows the sample of Random Forest.

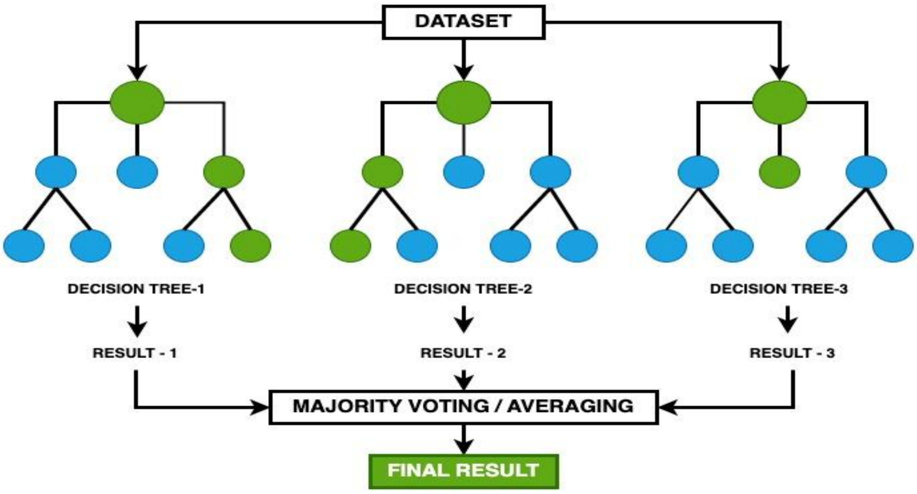


Fig. (2). Random Forest Structure [20].

6. **Gradient Boosting:** Gradient Boosting is an effective boosting algorithm that transforms several poor learners into strong learners by training each new model to minimize the loss function, such as mean squared error or cross-entropy, of the previous model via gradient descent. In each iteration, the algorithm calculates the gradient of the loss function with respect to the current ensemble's predictions and then trains a new weak model reduce this gradient. The new model's predictions are added to the ensemble, and the procedure continues until a stopping criterion is met [20].

DOI: http://dx.doi.org/10.25098/8.2.39

4. The proposed system:

The proposed system is supervised machine learning system that use it to predict student performance. Python programming language is used to implement the system with the Pandas and Scikit-learn libraries to create a model for predicting student performance. Python's adaptability and robust data analysis capabilities, combined with Pandas for data manipulation and preprocessing and Sklearn for machine learning algorithms, make it an ideal tool for data science. The proposed system's architecture begins with the import of the dataset, followed by preprocessing stages that prepare the data for training. The training phase begins once the data have been prepared. The trained model classifies whether the student's performance is high or low during the testing phase. Finally, evaluation metrics are utilized to compute the model's efficacy. The system architecture is shown in Fig. 3

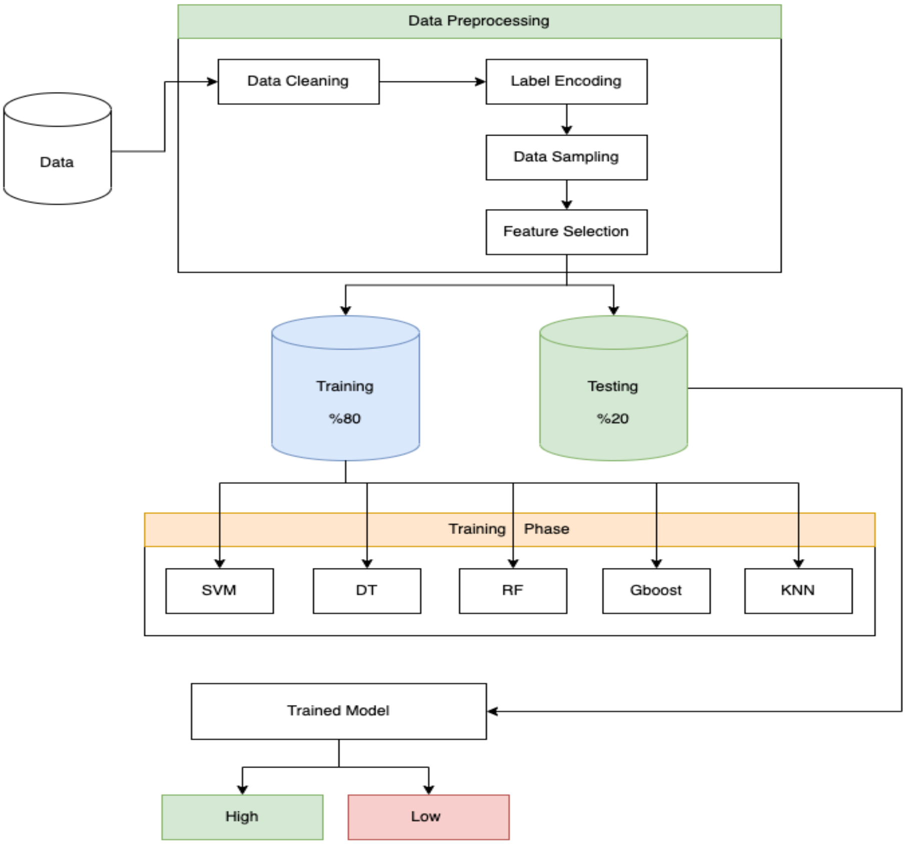


Fig. (3). System Architecture.

DOI: http://dx.doi.org/10.25098/8.2.39

4.1 Dataset: In this study, two different datasets were used to train and test of the model. The first dataset is UCI student performance dataset, and the other dataset is collected from E-Parwarda system of Kurdistan Regional Government. UCI dataset contains 649 instances with 31 features while the E-Parwarda dataset has 426 instance and 20 features.

Table 1. dataset description for UCI and E-Parwarda target variables.			
	Low Performance	High Performance	Total instance
UCI Dataset	546	100	646
E-Parwarda Dataset	328	95	423

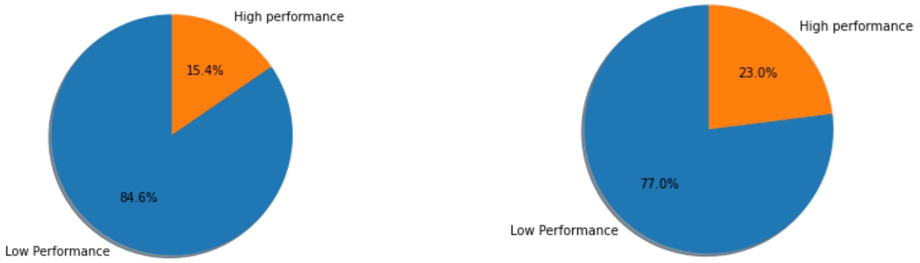


Fig. (4). Description of UCI dataset and E-Parwarda dataset.

The UCI Student Performance dataset includes information regarding the academic performance of students, including gender, age, parental background, study time, previous failures, and internet access. As well as E-Parwarda dataset includes Birthday, Address, Nationality, Blood Group and so on. The hole list of features for datasets shown in table (2), table (2).

Table (2) UCI student performance dataset features.

Feature Synonym	Feature Name	Feature Description	Datatype
F1	School	School name	Text
F2	Sex	Student gender	Boolean
F3	Age	Student age	Integer
F4	Address	Address type	Boolean
F5	famsize	Family size	Text
F6	Pstatus	Parent status	Text
F7	Medu	Mother Education level	Integer
F8	Fedu	Father Education Level	Integer
F9	Mjob	Mother Job	Text
F10	Fjob	Father job	Text
F11	Reason	Reason to choose this school	Text
F12	Guardian	Guardian	Text
F13	Traveltime	Travel time	Integer
F14	Studytime	Study time	Integer
F15	Failures	Number of past class failure	Integer
F16	Schoolsup	School extra support	Boolean
F17	Famsup	Family support	Boolean
F18	Activities	Extra-curricular activities	Boolean
F19	Paid	Extra paid classes within the course subject	Boolean
F20	Nursery	Attended nursery school	Boolean
F21	Higher	Wants to take higher school	Boolean
F22	Internet	Internet access at home	Boolean
F23	Romantic	With a romantic relationship	Boolean
F24	Famrel	Quality family relationship	Integer
F25	Freetime	Free time after school	Integer
F26	Goout	Going out with friends	Integer
F27	Dlac	Workday alcohol consumption	Integer
F28	Walc	Weekend alcohol consumption	Integer
F29	Health	Current health status	Integer
F30	Absences	Number of absences	Integer
F31	Grade	Student Grade	Integer

DOI: http://dx.doi.org/10.25098/8.2.39

Table (3) E-Parwarda dataset features.

Feature Synonym	Feature Name	Feature Description	Datatype
F1	Studentid	Student Identification id	Integer
F2	Nobrother	Number of brothers	Integer
F3	Nosister	Number of sisters	Integer
F4	Childnumber	Child number	Integer
F5	Registerationno	Registration number	Integer
F6	Address	Student address	Text
F7	Birthday	Student birthday	Date
F8	Gender	Student gender	Boolean
F9	Nationality	Student nationality	Text
F10	Nationalitynubmer	Nationality number	Integer
F11	Nation	Student nation	Text
F12	Religion	Student religion	Text
F13	Bloodgroup	Blood group	Text
F14	Status	Study status	Text
F15	Parentdeath	Parental death	Text
F16	Parentseperated	Parents are seperated	Boolean
F17	Studentstatus	Student status	Text
F18	Studytype	Study type	Text
F19	Studylang	Study language	Text
F20	Grade	Student grade	Integer

4.2 Data preprocessing: Data preprocessing is an initial phase in the pipeline for data analysis and machine learning. It is the process of cleaning, transforming, and organizing unprocessed data into a format suitable for analysis or modeling. This process typically involves duties such as handling missing values, standardizing or normalizing characteristics, encoding categorical variables, and removing outliers. Data preprocessing ensures that the data is accurate, consistent, and prepared for further analysis or modeling, thereby enhancing the quality of insights and forecasts that can be derived from the data. Data preprocessing is containing many steps such as Data Cleaning, Label Encoding, Data Sampling, Feature selection.

4.2.1 Data Cleaning: Data cleaning is one of the most important preprocessing steps in Machine Learning, having a clean dataset makes the models more accurate and improve the results. The both datasets examined for missing value and duplicated records. Missing value checked on the dataset missing value record not found on each dataset and no duplicate record founded.

4.2.2 Label Encoding: Label Encoding is another step of the preprocessing processes applied to datasets to convert categorical data to numeric data. In this procedure, a unique integer value is assigned to each unique category or label. Label encoding is designed to facilitate the use of machine learning algorithms that require numeric inputs. For instance, if the dataset contains "Gender" features with the values "Male" and "Female," label encoding would transform them to 0 and 1, respectively.

DOI: http://dx.doi.org/10.25098/8.2.39

By label encoding, we enable predictive models to effectively process and analyze data, resulting in more accurate and accurate predictions.

4.2.3 Data Sampling: Beginning analysis of the dataset demonstrated that "Low performance" 4.34 and 6.5 times lower compared to "High performance". Due to this considerable class imbalance, the model appears to be biased as it learns from a significantly higher proportion of "Low performance" occurrences. For classification problems, a balanced class dataset is important. As the majority of machine learning algorithms used for classification were developed under the assumption that each class contains an equal number of instances, the disparity of types in classification poses challenges for predictive modeling. Consequently, a classification model cannot produce accurate judgments without a balanced classification dataset.

There are a variety of methods for handling an imbalanced dataset. To address this issue, the synthetic minority oversampling technique (SMOTE) was employed. The SMOTE method generates new instances using the KNN algorithm for machine learning. additional instances of the minority class were generated in proportion to the instances of the majority class in order to achieve class parity. To balance the dataset, the minority class must be oversampled unless the number of cases in each category is nearly equal. Following balancing, the minority class was oversampled, causing to increase the data size. The 549 occurrences of each class in UCI and 328 from E-Pawarda finally result in a balanced distribution.

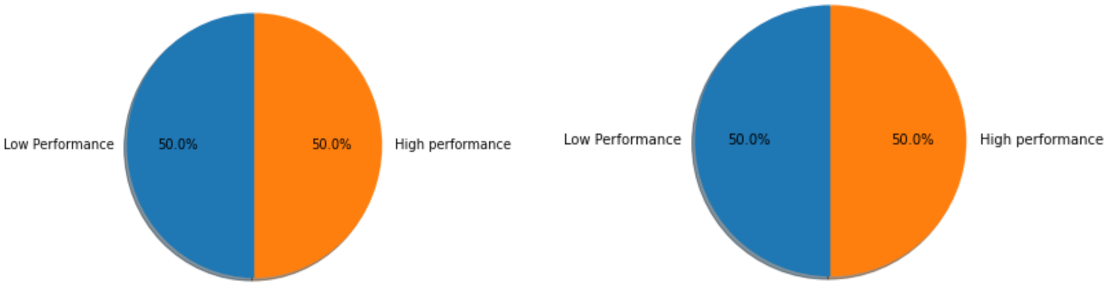


Fig (5). UCI dataset and E-Parwarda dataset after balancing.

4.2.4 Feature Selection: Feature selection is an important step in the machine learning model-building process. It includes finding and choosing the most informative and relevant features from a dataset in order to enhance the performance, efficiency and accuracy of a model. By eliminating irrelevant or redundant features, feature selection reduces the dimensionality of the data, which can result in shorter training durations and prevent overfitting. Filter Strong Correlations feature selection is utilized based on the dataset's characteristics and their respective advantages. The ultimate objective of feature selection is to improve the model's precision, interpretability, and generalizability, allowing for more effective decision-making and valuable data-driven insights. In addition, UCI dataset contains 31 features about students while E-Parwarda which is a second dataset includes 20 features. Several techniques used for features selection includes: Univariate feature selection, Recursive feature elimination, Principal component analysis, independent component analysis and SelectFromModel. Dataset features shown in fig. (6).

DOI: http://dx.doi.org/10.25098/8.2.39

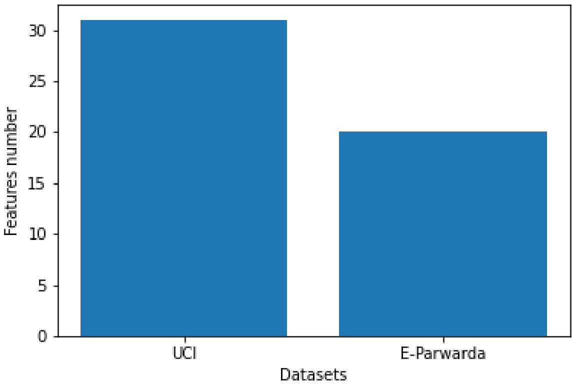


Fig. (6) Dataset features.

4.3 Model Building: Random Forest is a well-known ensemble machine learning algorithm that used for classification and regression tasks. It is an improved version of Decision tree algorithm that involves construction of multiple Decision Tree during the training phase and the combination of their predictions to make a final choice and also known for its robustness, capacity to manage high-dimensional data, and overfitting resistance.

The purposed RF model is to utilize the power of ensemble learning for robust and precise classification. Random Forest avoids the limitations of individual trees and reduces the likelihood of overfitting by constructing multiple decision trees through bootstrapping and randomization of feature values. The model combines the predictions of multiple trees to reach a conclusion, resulting in a more reliable and stable output. The parameter setting of the model shown in table (4).

Table (4). Parameter Setting of Random Forest Model.

Parameter	**value**
N_estimators	200
criterion	gini
max_depth	None
min_sample_split	2
min_sample_leaf	1
max_features	auto
random_state	None

4.4 Model Evaluation: Model evaluation is the assessment of a machine learning model's performance and efficiency. Utilizing different metrics for evaluation to determine how accurately the model predicts outcomes based on unknown data. Common evaluation metrics consist of precision, recall, and the F1 score. By comparing the model's predictions to the actual results, Strengths and weaknesses can be determined, thereby facilitating the identification of potential areas for enhancement or refinement.

Evaluation of the learned model's efficiency is a crucial step in the machine learning process. Models of machine learning are either flexible or non-flexible depending on how well they apply to new input. When a machine learning (ML) model is implemented to new data without being adequately assessed with several kinds of metrics and without depending on accuracy, it is possible

DOI: http://dx.doi.org/10.25098/8.2.39

for the model to generate inaccurate predictions. In addition, the accuracy, precision, recall, and F1 score have been accounted for, when determining the model's classification between High and Low performance and when considering the error factor. Accuracy is the ratio of correct predictions to all other predictions.

Confusion matrix generates a matrix that describes the model's overall efficacy. In this study, for instance, the confusion matrix for binary-classification is a two-by-two matrix.

The confusion matrix displays the number of correct and incorrect classifications for both actual and predicted values, where true positive (TP) demonstrates the number of samples that are correctly classified as positive, true negative (TN) indicates the number of instances that have been correctly recognized as negative, False Positive (FP) indicates the number of samples that are incorrectly classified as positive, and False Negative (FN) shows the number of instances that were mistaken as negative. Table (5) displays the confusion matrix for binary classification.

Table (5). Confusion Matrix.

Actual Values	Predicted Values	
	Negative	Positive
Negative	True Negative (TN)	False Positive (FP)
Positive	False Negative (FN)	True Positive (TP)

According to the confusion matrix, a number of crucial metrics are calculated and considered alongside the model's precision to guarantee the model performs well and has no bias due to factors such as imbalanced dataset. Consequently, precision, recall, and F1 score have been utilized as evaluation metrics for models. Precision represents the accuracy of positive predictions, recall is the proportion of actual positive samples, and F1 is the harmonic mean of precision and recall.

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives}\text{--(2)}$$

Equation (2) Precision calculation formula.

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negatives}\text{---(3)}$$

Equation (3) Recall calculation formula.

$$F1\ Score = 2 \times \frac{Precision\times Recall}{Precision+Recall}\text{---(4)}$$

Equation (4) F1 score calculation formula.

DOI: http://dx.doi.org/10.25098/8.2.39

5.Experimental Result and Discussion:

The selection of these features from the dataset was prompted by their prospective usefulness for comprehending and predicting student performance. The first set of characteristics includes demographic variables such as gender, age, and address type of students, which can provide insight into the influence of sociodemographic characteristics on academic performance. In addition, family-related characteristics such as family size, mother and father's education level, and reason for choosing the school were selected to determine the influence of familial support and socioeconomic background on a student's educational journey. The second set of characteristics examines various academic behaviors and routines of students. Such variables as travel time, study time, and the number of class failures in the past cast light on time management skills and learning difficulties that may be impacting current academic performance. The inclusion of variables such as supplementary paid classes within the course subject and aspirations to pursue higher education indicates a proactive approach to learning, whereas internet access at home can indicate access to online resources and digital literacy. The third set of characteristics investigates personal and lifestyle factors that may affect academic outcomes. Relationships, leisure time, social activities, and alcohol consumption provide insight into the academic and personal balance of students. The number of absences is indicative of student engagement, while the student's grade is the ultimate indicator of academic performance. These diverse features contribute to the development of a comprehensive model that enables educational institutions to support the success of their students. Table (6) shows the selected feature from UCI dataset.

Table (6) UCI student performance dataset features.

Feature ID	Feature Name	Feature ID	Feature Name
F2	Sex	F19	Paid
F3	Age	F21	Higher
F4	Address	F22	Internet
F5	famsize	F23	Romantic
F7	Medu	F25	Freetime
F8	Fedu	F26	Goout
F11	Reason	F27	Dlac
F13	Traveltime	F28	Walc
F4	Studytime	F30	Absences
F15	Failures	F31	Grade

The selected features from the E-Parwarda dataset provide essential information for comprehending various facets of students' lives and prospective factors influencing academic performance. The number of siblings, sisters, and children provides insight into the family structure, while the age and gender of the students contribute to an understanding of their demographics. In addition, the student's nationality and place of birth can provide valuable context regarding cultural and geographical influences. Blood group information could be used to evaluate health-related factors that could affect the health and attendance of students. Parental death and parental separation are significant family-related factors that may affect an individual's emotional health and social support. The final target variable is the student's grade, which serves as the definitive indicator of academic

DOI: http://dx.doi.org/10.25098/8.2.39

performance. By analyzing these characteristics, educators can obtain valuable insights for enhancing support strategies and boosting student success overall. The selected features from E-Parwarda dataset shown in Table (7).

Table (7) E-Parwarda dataset features.

Feature ID	Feature Name	Feature ID	Feature Name
F2	Nobrother	F6	Birthdaycity
F3	Nosister	F13	Bloodgroup
F4	Childnumber	F15	Parentdeath
F7	Birthday	F16	Parentseperated
F8	Gender	F20	Grade
F11	Nation		

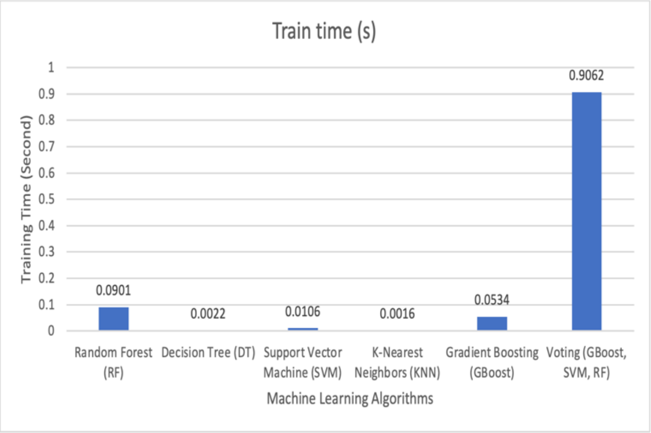
Table (8) shows the properties of the models such as Training time, Testing time, Loss, F-Score, Precision, Recall, Confusion Matrix and Accuracy. KNN has a Lowest training time (0.0016) second and Voting model has a highest training time which is (0.9062) second. As well as the table contains confusion matrix that also used for calculating accuracy of the model. The best accuracy was (87.69%) was achieved by GBoosting while KNN got a lowest accuracy (81.54%).

Table (8) model accuracy, loss, precision and recall based on UCI dataset.

Classifier name	Train time (s)	Test time (s)	loss	F-score	Precision	Recall	Confusion Matrix	Accuracy (%)
Random Forest (RF)	0.0901	0.0098	0.1996	0.8513	0.8409	0.8692	[[3 12] [5 110]]	86.92
Decision Tree (DT)	0.0022	0.0010	0.1996	0.8144	0.8064	0.8231	[[2 13] [10 105]	82.31
Support Vector Machine (SVM)	0.0106	0.0122	1.3975	0.8255	0.8205	0.8308	[[3 12] [10 105]	83.08
K-Nearest Neighbors (KNN)	0.0016	0.0129	3.5937	0.8203	0.8257	0.8154	[[4 11] [13 102]]	81.54
Gradient Boosting (GBoost)	0.0534	0.0025	2.4623	0.8566	0.8478	0.8769	[[3 12] [4 111]]	87.69
Voting (GBoost, SVM, RF)	0.9062	0.0352	0.1997	0.8289	0.8152	0.8462	[[2 13] [7 108]]	84.61

Training time, Loss and accuracy shown in fig. (8), (9), (10) and F-score, Precision, Recall illustrated in fig. (10) that completely displayed the score for above properties for machine learning classifiers such as Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbor, GBoost and Voting.

DOI: http://dx.doi.org/10.25098/8.2.39



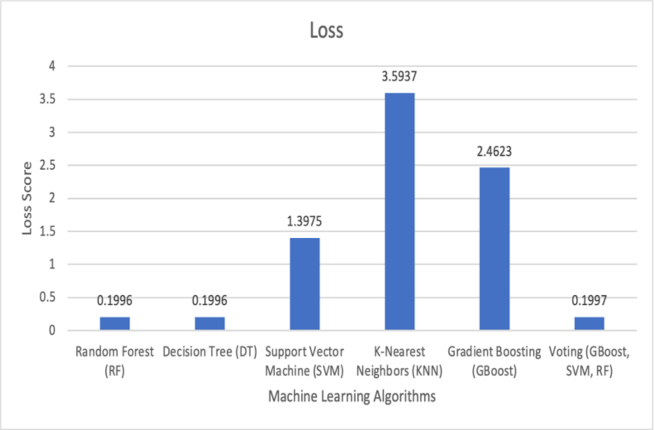
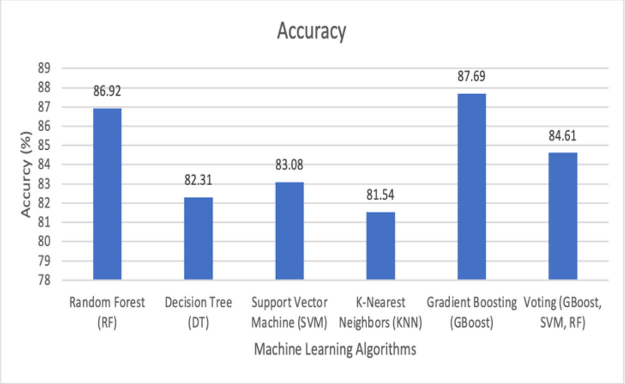


Fig. (7). Training time for UCI Dataset.

Fig. (8). Loss for UCI Dataset.

The Fig. (9) provides a clear visualization of the accuracy achieved by different models, offering insights into their performance on the dataset. Random Forest leads with an accuracy of 86.92%, followed by Decision Tree at 82.31%. Support Vector Machine (SVM) follows closely with an accuracy of 83.08%, while K-Nearest Neighbors (KNN) achieves 81.54%. Gradient Boosting and the Voting ensemble display promising results at 87.69% and 84.61% accuracy.



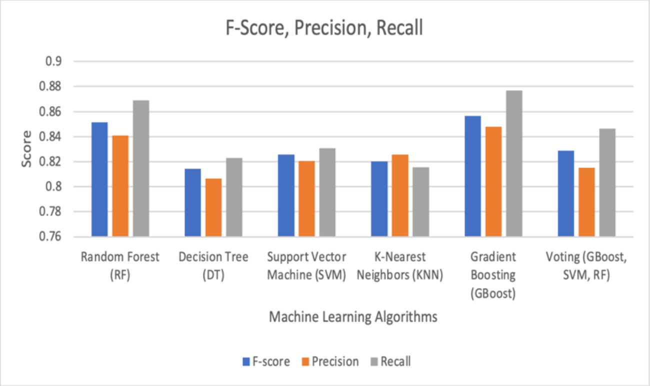


Fig. (9). Accuracy for UCI Dataset.

Fig. (10). F-score, Precision, Recall for UCI Dataset.

The table provides a comprehensive summary of performance metrics for various classifiers on a dataset. Random Forest (RF) demonstrates a training time of 0.1007 seconds and a test time of the same duration. With a loss of 0.3296, RF achieves an F-score of 0.9158, Precision of 0.9281, Recall of 0.9167, and an impressive accuracy of 91.67%. Decision Tree (DT) exhibits efficient training and test times of 0.0025 and 0.0011 seconds respectively. With a loss of 0.3955, DT achieves an F-score of 0.9010, Precision of 0.9057, Recall of 0.9015, and an accuracy of 90.15%.

Support Vector Machine (SVM) records a slightly higher training time of 0.0105 seconds and a test time of 0.0220 seconds. With a higher loss of 9.6236, SVM achieves an F-score of 0.7347, Precision of 0.7288, Recall of 0.7348, and an accuracy of 73.48%. K-Nearest Neighbors (KNN) demonstrates a fast-training time of 0.0020 seconds and a test time of 0.0045 seconds. With a loss of 0.3955, KNN achieves an F-score of 0.8438, Precision of 0.8825, Recall of 0.8485, and an accuracy of 84.84%.

DOI: http://dx.doi.org/10.25098/8.2.39

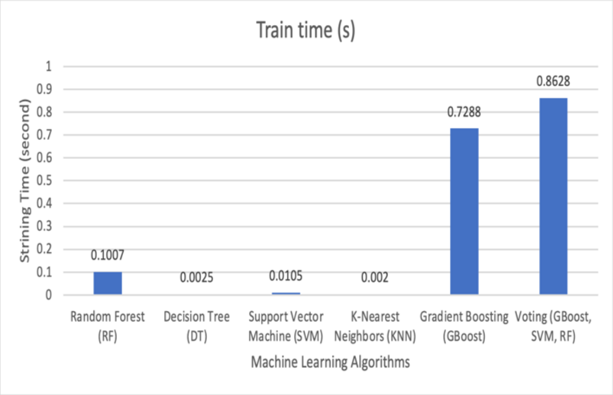
Gradient Boosting (GBoost) showcases a relatively higher training time of 0.7288 seconds and a test time of 0.0113 seconds. With a loss of 0.3955, GBoost achieves an F-score of 0.9087, Precision of 0.9122, Recall of 0.9091, and an accuracy of 90.91%. Similarly, the Voting ensemble also presents a higher training time of 0.8628 seconds and a test time of 0.0439 seconds. With the same loss of 0.3955, the Voting ensemble achieves an F-score of 0.9085, Precision of 0.9148, Recall of 0.9091, and an accuracy of 90.91%.

Table (9) model accuracy, loss, precision and recall based on E-Parwarda dataset.

Classifier name	Train time (s)	Test time (s)	loss	F-score	Precision	Recall	Confusion Matrix	Accuracy (%)
Random Forest (RF)	0.1007	0.1007	0.3296	0.9158	0.9281	0.9167	[[69 0] [11 52]]	91.67
Decision Tree (DT)	0.0025	0.0011	0.3955	0.9010	0.9057	0.9015	[[66 3] [10 53]]	90.15
Support Vector Machine (SVM)	0.0105	0.0220	9.6236	0.7347	0.7388	0.7348	[[48 21] [14 49]]	73.48
K-Nearest Neighbors (KNN)	0.0020	0.0045	0.3955	0.8438	0.8825	0.8485	[[69 0] [20 43]]	84.84
Gradient Boosting (GBoost)	0.7288	0.0113	0.3955	0.9087	0.9122	0.9091	[[66 3] [9 54]]	90.91
Voting (GBoost, SVM, RF)	0.8628	0.0439	0.3955	0.9085	0.9148	0.9091	[[67 2] [10 53]]	90.91

Fig. (11) illustrated training times vary across different algorithms. Random Forest (RF) demonstrates a training time of 0.1007 seconds, while Decision Tree (DT) showcases exceptional efficiency with a remarkably low training time of 0.0025 seconds. Support Vector Machine (SVM) records a slightly higher training time of 0.0105 seconds, and K-Nearest Neighbors (KNN) maintains its agility with a training time of 0.0020 seconds. In contrast, Gradient Boosting (GBoost) presents a higher training time of 0.7288 seconds, and the Voting ensemble (GBoost, SVM, RF) exhibits the longest training time among the classifiers, taking 0.8628 seconds.

DOI: http://dx.doi.org/10.25098/8.2.39



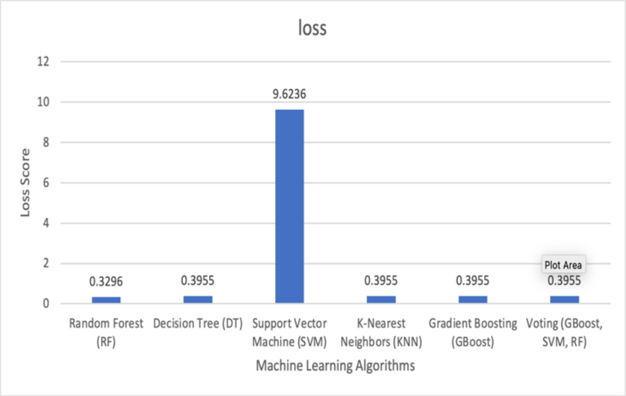
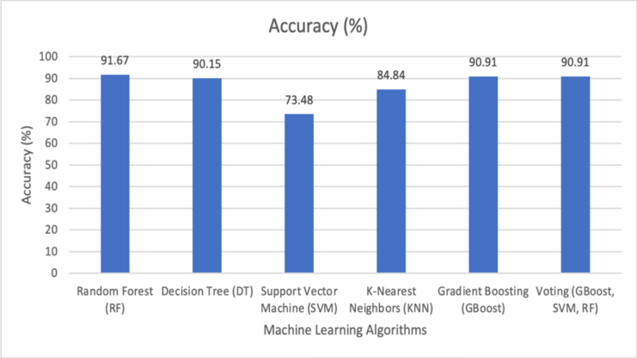


Fig. (11). Training time for E-Parwarda Dataset.

Fig. (12). Loss for E-Parwarda Dataset.

The fig. (13) represents the accuracy of various models, showcasing their respective performance on the dataset. Random Forest leads with an accuracy of 91.67%, followed closely by Decision Tree at 90.15%. Support Vector Machine (SVM) demonstrates an accuracy of 73.48%, while K-Nearest Neighbors (KNN) achieves 84.44%. Both Gradient Boosting and the Voting ensemble share the top spot with an accuracy of 90.91%. This visual representation underlines the models' distinct strengths and highlights their varying degrees of accuracy in predicting outcomes on the given dataset.



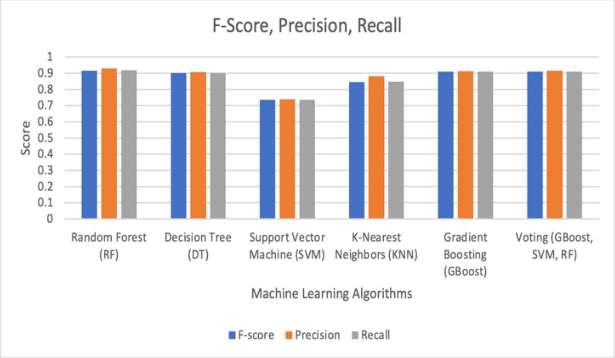


Fig. (13). Accuracy for E-Parwarda Dataset.

Fig (14). F-score, Precision, Recall for E-Parwarda Dataset.

Comparing the efficacy of classifiers on the UCI dataset and the E-Parwarda dataset reveals distinct tendencies. Random Forest demonstrates excellent accuracy on both datasets, obtaining an impressive 86.92% on the UCI dataset and a remarkable 91.67 % on the E-Parwarda dataset. Decision Tree performs well on both datasets, with an accuracy of 82.31 percent on the UCI dataset and a significant increase to 90.15 percent on the E-Parwarda dataset. However, the efficacy of Support Vector Machine (SVM) differs significantly, with an accuracy of 83.08% on the UCI dataset and a notable decrease to 73.15% on the E-Parwarda dataset. K-Nearest Neighbors (KNN) maintains a high degree of accuracy, obtaining 81.54% on the UCI dataset and 84.88% on the E-Parwarda dataset, respectively. Similarly, Gradient Boosting and the Voting ensemble demonstrate robust performance, with 87.69% and 84.61% accuracy on the UCI dataset, respectively, and both classifiers exhibit a slight increase in accuracy on the E-Parwarda dataset, reaching 90.91%. Overall, Random Forest, Decision Tree, Gradient Boosting, and the Voting ensemble exhibit strong adaptability to both datasets. On the other hand, SVM's performance on the E-Parwarda dataset decreases significantly, indicating a potential difficulty in generalizing to diverse data domains. These results highlight the

DOI: http://dx.doi.org/10.25098/8.2.39

significance of selecting classifiers based on the particular characteristics and distribution of the data for optimal prediction accuracy as shown in table (10), table (11).

Table (10). Machine Learning Algorithms accuracy based on UCI dataset.

Classifiers	UCI Dataset
Random Forest (RF)	86.92%
Decision Tree (DT)	82.31%
Support Vector Machine (SVM)	83.08%
K-Nearest Neighbors (KNN)	81.54%
Gradient Boosting (GBoost)	87.69%
Voting (GBoost, SVM, RF)	84.61%

Table (11). Machine Learning Algorithms accuracy based on E-Parwarda dataset.

Classifiers	E-Parwarda Dataset
Random Forest (RF)	91.67%
Decision Tree (DT)	90.15%
Support Vector Machine (SVM)	73.15%
K-Nearest Neighbors (KNN)	84.84%
Gradient Boosting (GBoost)	90.91%
Voting (GBoost, SVM, RF)	90.91%

Conclusion

This paper concluded with a comprehensive analysis of the performance and accuracy of various classifiers on two distinct datasets: the UCI dataset and the E-Parwarda dataset. The results provided valuable insights into the adaptability and efficacy of various classifiers in predicting outcomes across various data domains. High levels of accuracy were achieved by Random Forest and Gradient Boosting, which demonstrated robust and consistent performance across both datasets. Additionally, both Decision Tree and the Voting ensemble demonstrated competitive performance. However, Support Vector Machine (SVM) encountered difficulties when generalizing to the E-Parwarda dataset, resulting in a significant accuracy decrease. These results highlight the significance of selecting appropriate classifiers based on specific data characteristics for optimal prediction accuracy. The results of this study contribute valuable knowledge to the field of machine learning, guiding researchers and practitioners to make informed decisions when applying classifiers to diverse datasets and domains, thereby improving the efficacy of predictive modeling in real-world applications. To maximize classifier performance even further, future research could focus on feature engineering and parameter tuning.

DOI: http://dx.doi.org/10.25098/8.2.39

References

[1] H. Altabrawee, O. Abdul, J. Ali, and Q. Ajmi, "Predicting Students' Performance Using Machine Learning Techniques," 2019.

[2] V. A. Sungar, P. D. Shinde, and M. V Rupnar, "Predicting Student's Performance using Machine Learning," 2017. [Online]. Available: www.caeaccess.org

[3] Y. Baashar, G. Alkawsi, N. Ali, H. Alhussian, and H. T. Bahbouh, "Predicting student's performance using machine learning methods: A systematic literature review," in *Proceedings - International Conference on Computer and Information Sciences: Sustaining Tomorrow with Digital Innovation, ICCOINS 2021*, Institute of Electrical and Electronics Engineers Inc., Jul. 2021, pp. 357–362. doi: 10.1109/ICCOINS49721.2021.9497185.

[4] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, "Classification and prediction of student performance data using various machine learning algorithms," *Mater Today Proc*, Jul. 2021, doi: 10.1016/j.matpr.2021.07.382.

[5] N. R. Beckham, L. J. Akeh, G. N. P. Mitaart, and J. V Moniaga, "Determining factors that affect student performance using various machine learning methods," *Procedia Comput Sci*, vol. 216, pp. 597–603, 2023, doi: 10.1016/j.procs.2022.12.174.

[6] K. Kazi, K. K. Solutions, and M. Solapur, "Implementation of Latest Machine Learning Approaches for Students Grade Prediction," *International Journal of Early Childhood Special Education (INT-JECS)*, vol. Vol 14, no. Issue 03 2022, 2022, doi: 10.9756/INT-JECSE/V14I3.1141.

[7] F. Qiu *et al.*, "Predicting students' performance in e-learning using learning process and behaviour data," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-021-03867-8.

[8] H. Nawang, M. Makhtar, and W. M. A. F. W. Hamzah, "Comparative analysis of classification algorithm evaluations to predict secondary school students' achievement in core and elective subjects," *International Journal of Advanced Technology and Engineering Exploration*, vol. 9, no. 89, pp. 430–445, 2022, doi: 10.19101/IJATEE.2021.875311.

[9] R. Deshmukh, A. Kulkarni, A. Kumthekar, and P. Kottur, "Machine Learning Techniques for Predicting Student Performance," *Mathematical Statistician and Engineering ApplicationsISSN: 2094-0343Mathematical Statistician and Engineering Applications*, vol. 71, no. 4, 2022, [Online]. Available: http://philstat.org.ph

[10] N. Mohammad Suhaimi, S. Abdul-Rahman, S. Mutalib, N. H. Abdul Hamid, and A. Hamid, "Review on Predicting Students' Graduation Time Using Machine Learning Algorithms," *International Journal of Modern Education and Computer Science*, vol. 11, no. 7, pp. 1–13, Jul. 2019, doi: 10.5815/ijmecs.2019.07.01.

[11] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Applied Sciences (Switzerland)*, vol. 10, no. 3. MDPI AG, Feb. 01, 2020. doi: 10.3390/app10031042.

[12] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student Performance Prediction Model based on Supervised Machine Learning Algorithms," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1757-899X/928/3/032019.

DOI: http://dx.doi.org/10.25098/8.2.39

[13] J. Dhilipan, N. Vijayalakshmi, S. Suriya, and A. Christopher, “Prediction of Students Performance using Machine learning,” *IOP Conf Ser Mater Sci Eng*, vol. 1055, no. 1, p. 012122, Feb. 2021, doi: 10.1088/1757-899x/1055/1/012122.

[14] Engr. Sana Bhutto, Dr. Isma Farah Siddiqui, Dr. Qasim Ali Arain, and Maleeha Anwar, “Predicting Students’ Academic Performance Through Supervised Machine Learning,” *2020 International Conference on Information Science and Communication Technology*, 2020.

[15] E. T. Lau, L. Sun, and Q. Yang, “Modelling, prediction and classification of student academic performance using artificial neural networks,” *SN Appl Sci*, vol. 1, no. 9, Sep. 2019, doi: 10.1007/s42452-019-0884-7.

[16] A. J. Khalil, A. M. Barhoom, B. S. Abu-Nasser, M. M. Musleh, and S. S. Abu-Naser, “Energy Efficiency Prediction using Artificial Neural Network,” 2019. [Online]. Available: www.ijeais.org/ijapr

[17] A. U. Haq *et al.*, “Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson’s Disease Using Voice Recordings,” *IEEE Access*, vol. 7, pp. 37718–37734, 2019, doi: 10.1109/ACCESS.2019.2906350.

[18] S. T. Ahmed, R. Al-Hamdani, and M. S. Croock, “Enhancement of student performance prediction using modified K-nearest neighbor,” *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 4, pp. 1777–1783, 2020, doi: 10.12928/TELKOMNIKA.V18I4.13849.

[19] M. Bansal, A. Goyal, and A. Choudhary, “A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning,” *Decision Analytics Journal*, vol. 3, p. 100071, Jun. 2022, doi: 10.1016/j.dajour.2022.100071.

[20] Mastour, Haniye, et al. "Early prediction of medical students' performance in high-stakes examinations using machine learning approaches." Heliyon (2023).

[21] Beckham, Nicholas Robert, et al. "Determining factors that affect student performance using various machine learning methods." Procedia Computer Science 216 (2023): 597-603.

[22] Pallathadka, Harikumar, et al. "Classification and prediction of student performance data using various machine learning algorithms." Materials today: proceedings 80 (2023): 3782-3785.

[23] Priyambada, Satrio Adi, Tsuyoshi Usagawa, and E. R. Mahendrawathi. "Two-layer ensemble prediction of students’ performance using learning behavior and domain knowledge." Computers and Education: Artificial Intelligence (2023): 100149.

[24] Ajibade, Samuel-Soma M., et al. "Utilization of Ensemble Techniques for Prediction of the Academic Performance of Students." Journal of Optoelectronics Laser 41.6 (2022): 48-54.

[25] Chen, Liyan, Lihua Wang, and Yuxin Zhou. "Research on data mining combination model analysis and performance prediction based on students’ behavior characteristics." Mathematical Problems in Engineering 2022 (2022): 1-10.

[26] Sudais, Muhammad, et al. "Students’ academic performance prediction model using machine learning." (2022).

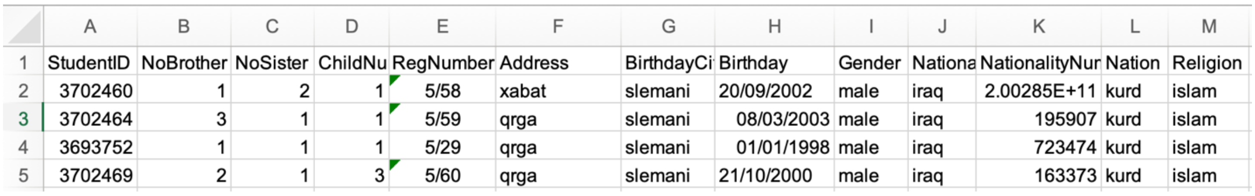
[27] Zeineddine, Hassan, Udo Braendle, and Assaad Farah. "Enhancing prediction of student success: Automated machine learning approach." Computers & Electrical Engineering 89 (2021): 106903.

DOI: http://dx.doi.org/10.25098/8.2.39

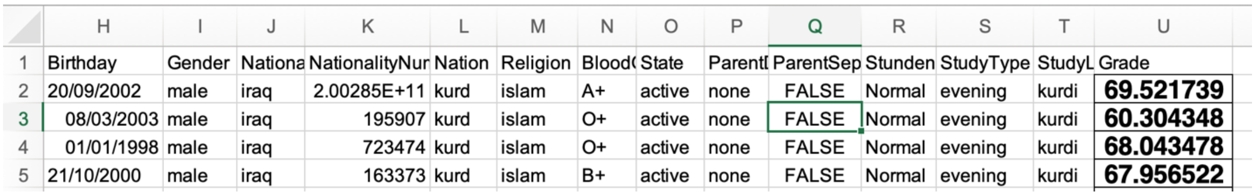
[28] Varade, Rashmi V., and Blessy Thankanchan. "Academic performance prediction of undergraduate students using decision tree algorithm." SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology 13.SUP 1 (2021): 97-100.

[29] Ma, Yuling, et al. "Multi-task MIML learning for pre-course student performance prediction." Frontiers of Computer Science 14 (2020): 1-10.

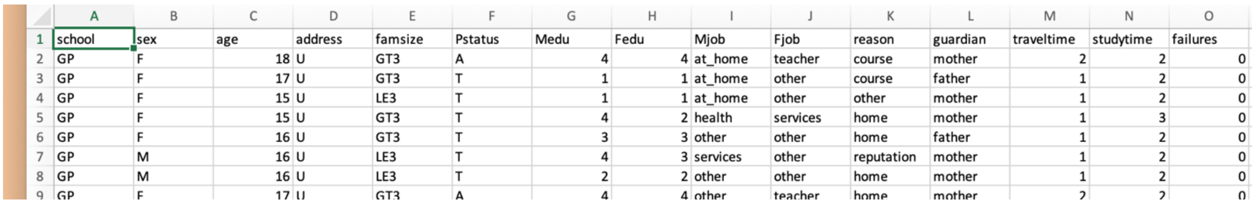
6. Appendix:



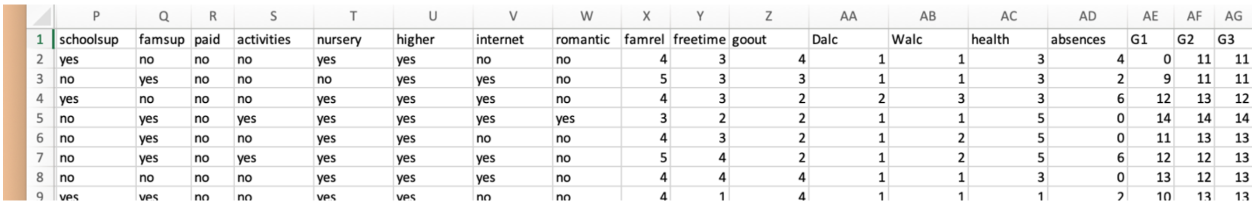
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	StudentID	NoBrother	NoSister	ChildNu	RegNumber	Address	BirthdayCi	Birthday	Gender	Nationa	NationalityNur	Nation	Religion
2	3702460	1	2	1	5/58	xabat	slemani	20/09/2002	male	iraq	2.00285E+11	kurd	islam
3	3702464	3	1	1	5/59	qrga	slemani	08/03/2003	male	iraq	195907	kurd	islam
4	3693752	1	1	1	5/29	qrga	slemani	01/01/1998	male	iraq	723474	kurd	islam
5	3702469	2	1	3	5/60	qrga	slemani	21/10/2000	male	iraq	163373	kurd	islam



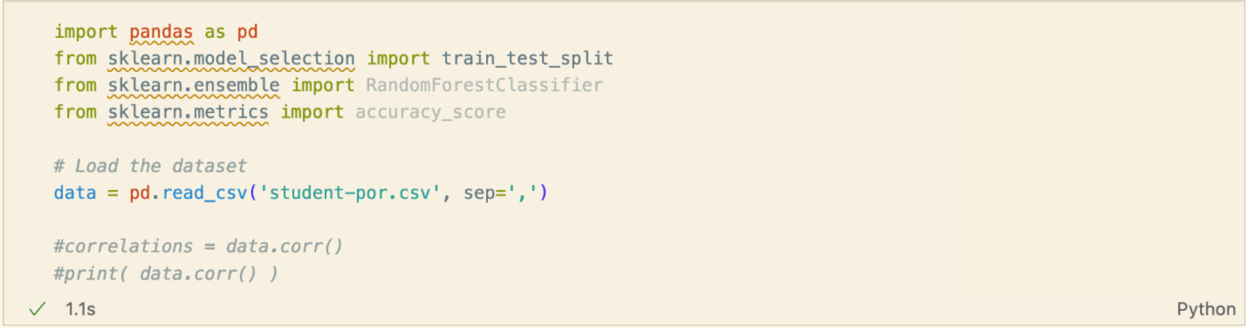
	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Birthday	Gender	Nationa	NationalityNur	Nation	Religion	Blood(	State	ParentI	ParentSep	Stunden	StudyType	StudyL	Grade
2	20/09/2002	male	iraq	2.00285E+11	kurd	islam	A+	active	none	FALSE	Normal	evening	kurdi	**69.521739**
3	08/03/2003	male	iraq	195907	kurd	islam	O+	active	none	FALSE	Normal	evening	kurdi	**60.304348**
4	01/01/1998	male	iraq	723474	kurd	islam	O+	active	none	FALSE	Normal	evening	kurdi	**68.043478**
5	21/10/2000	male	iraq	163373	kurd	islam	B+	active	none	FALSE	Normal	evening	kurdi	**67.956522**



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures
2	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0
3	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0
4	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	0
5	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0
6	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0
7	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0
8	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0
9	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0



	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
2	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	4	0	11	11
3	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	2	9	11	11
4	yes	no	no	no	yes	yes	yes	no	4	3	2	2	3	3	6	12	13	12
5	no	yes	no	yes	yes	yes	yes	yes	3	2	2	1	1	5	0	14	14	14
6	no	yes	no	no	yes	yes	no	no	4	3	2	1	2	5	0	11	13	13
7	no	yes	no	yes	yes	yes	yes	no	5	4	2	1	2	5	6	12	12	13
8	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	13	12	13
9	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	2	10	13	13

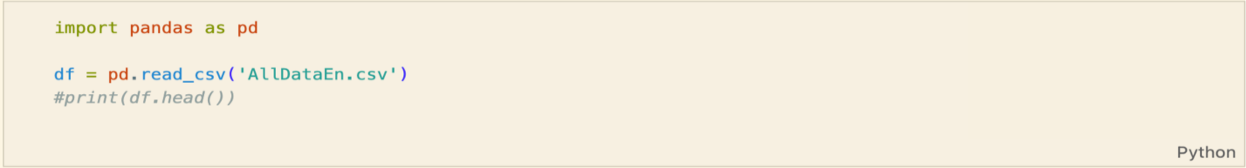


```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

# Load the dataset
data = pd.read_csv('student-por.csv', sep=',')

#correlations = data.corr()
#print( data.corr() )
```

✓ 1.1s Python



```python
import pandas as pd

df = pd.read_csv('AllDataEn.csv')
#print(df.head())
```

Python

DOI: http://dx.doi.org/10.25098/8.2.39