Predicting Early Withdrawal of University Students: A Comparative Study between KNN and Decision Tree

Arham Tariq
Department of Computer Science
University of Central Punjab
Lahore, Pakistan
arhamtariq99@gmail.com

Ahmad Amin
Department of Computer Science
The Superior University Lahore
Pakistan
ahmadaminvhr@gmail.com

Yasir Masood
Department of Computer Science
University of Lahore
Lahore, Pakistan
yasir.masood48@gmail.com

Muhammad Muzaffar
Department of Computer Science
Sharif College of Engineering and Technology
Lahore, Pakistan
m.muzaffar@sharif.edu.pk

Junaid Iqbal
Department of Computer Science
University of Lahore
Lahore, Pakistan
junaid.iqbal7678@gmail.com

2023 4th International Conference on Advancements in Computational Sciences (ICACS) | 978-1-6654-6377-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACS55311.2023.10089706

***Abstract*—**"The rising trend of students dropping out of universities without completing their degrees is becoming a concerning issue for institutions. To address this problem, the reasons behind this phenomenon need to be explored. However, most educational data sets have small sample sizes and varying patterns. Currently, there are few machine learning approaches for Pakistani higher education student performance. This study presents a machine learning-based approach to predict student withdrawals and identify the reasons behind them. The proposed approach compares two supervised ML algorithms, K-NearestNeighbors (KNN) and Decision-Tree (DT). The most important attributes affecting student retention are also determined using the ExtraTreesClassifier ensemble learning algorithm. In our experimental evaluation, the accuracy of KNN was 75%, and 70% for DT."

Index Terms—Student Drop Prediction, Dropping out of university, KNN performance, Decision Tree Comparison, Extra-TreesClassifier.

I. INTRODUCTION

Dealing with student withdrawal has become a major concern for higher authorities and faculty members of universities. There are numerous factors involved that affect student performance in academics. Several social behavior issues, financial instability, and psychological disorders can become a hurdle in student performance. Most of the students in universities are no more teenagers. Many Socio-Economic responsibilities are imposed on them as their age increases, on the other side higher education requires more effort with more consistency. Both of these constraints generate numerous factors that need to be evaluated for the betterment of students' performance in their academic careers. Apart from gaining only profit universities and stakeholders also need to evaluate factors that are important for students to complete their degrees. The main concern of this paper is to predict and evaluate the factors due to which students get a drop out of university. In our proposed technique we used KNN and DT as supervised learning algorithms. To solve this problem, First, we apply both algorithms on the dataset individually and measure the accuracy score for both algorithms. The dataset used in this research was manually collected from the University of South Asia Learning Management System's Database. Data attributes consist of transcript record, students' previous academic results, their employment status, and the distance of their home from university premises. The main contributions of this paper are as follows:

- Every country's student demographics vary in many different aspects. Previously there was not any approach configured for Pakistani private universities comprising Computer Science students. Our approach not only provides a comparative analysis between two different supervised learning-based models. Aside from that, we have extracted information related to major attributes contributing to students dropping out.

Section I: Introduction to implementation guidelines and approach used in the paper. **Section II:** Overview of existing methodology for evaluating students. **Section III:** Problem statement explaining the need for a prediction model. **Section IV:** Description of methodology and related background information. **Sections V, VI, VII, and VIII:** Description of tools and techniques used for implementation, limitations, future work, and conclusion.

978-1-6654-6377-5/23/$31.00 ©2023 IEEE

II. LITERATURE REVIEW

Most universities primarily maintain student data in excel sheets and evaluate students based on traditional factors such as the number of absences or academic performance. In the past, various studies have been conducted to predict student performance using machine learning models. Anwarudin Anwarudin et al. [1] used a dataset from a medical laboratory technician study program and evaluated the performance of their model using 5-fold cross-validation. Tismy Devasia et al. [2] used a dataset with 700 instances and 19 attributes gathered from a university and applied regression, decision trees, neural networks, and Naive Bayesian models. Fei Mi et al. [3] used a temporal dataset from edx and Coursera to predict student dropout using RNN and LSTM-based models. Jiazhen He et al. [4] proposed a prediction model using logistic regression, SVM, random forest, decision tree, naive Bayes, and BayesNet and trained it on a dataset from a course offered at the University of Melbourne. Bani et al. [5] trained a linear regression and support vector regression model on a dataset from a college with 5000 student instances and 16 attributes. Mike Sharkey [6] used a random forest algorithm to predict student dropout using a dataset with 20000 instances and 15 attributes. Z Kovacic et al. [7] used a regression tree to predict student dropout using a dataset from an information systems course with 450 instances. Keith Zvoch [8] used a dataset with 90000 students from over 100 schools in the USA to predict the dropout ratio using a multilevel logistic regression model. Yuda N. Mnyawami et al. [9] trained DT, NB, KNN, and MLP classifiers on a dataset with 85,634 instances and 37 attributes gathered from Tanzanian secondary schools. Alam TM et al. [10] trained decision trees, support vector machines, random forests, rotation forests, and artificial neural networks on a dataset of 10th-grade students in public schools and found that ANNs performed the best with an accuracy of 82.9%. Javed et al. [11] reviewed two types of recommendation systems, content-based and context-based, and discussed their advantages and disadvantages, with realworld examples. The conclusion emphasized the need for further research to improve the effectiveness and efficiency of these systems. Cioffi et al. [12] reviewed the current state and future prospects of using AI and machine learning in smart production, covering recent advances and challenges, applications, and the importance of large, high-quality data sets. The conclusion emphasized the need for continued research and development in this field. Shaukat K et al. [13] presented a new method for extracting data from hyperlinked web pages using a combination of machine learning and web page structure analysis. The results showed improved accuracy and efficiency compared to traditional methods, providing valuable insights into the field of data extraction.

However, the generic data attributes used in these studies may not be suitable for specific organizations and subjects, especially in emerging fields such as Computer Science in Pakistani private universities. There is a need to predict results based on data attributes that are in line with current trends and demands in these universities.

III. PROBLEM STATEMENT

In this section, we will figure out the reasons in order to develop this approach.

A. Need of Prediction Model

Considering the methodologies of conventional programming without machine learning concepts, it was not feasible to judge every student on the same parameters. Every country contains different cultures and demographics. Like this, every higher-level degree contains different complexity levels of courses. Previously there was not any ML-based approach considered for Pakistani-based private universities, especially for Computer Science students. For evaluating core facts and reasons behind students dropping out there was a need for a predictive model that can predict upcoming or existing students using past ones.

IV. METHODOLOGY

A. *Data Description*

The data used in machine learning algorithms is gathered from the University of South Asia Portal Database consisting of attributes such as the total no of F grades, the student's Current incomplete Degree CGPA, previous degree academic results, the distance of the student's house from university, an attribute suggesting student has studied prerequisite or core courses in the previous degree required in current degree and employment status attribute reflecting student is employed or not.

TABLE I
DATA DESCRIPTION TABLE FOR DATA SET

No	Attribute	Description
1	CGPA	CGPA of not more than 4 th semester students
2	Total Number of F Grades	Total Number of F Grades the student has scored till 4th Semester
3	Percentage marks	Percentage of marks student has scored in previous academic degrees
4	Core course status	This attribute reflects student has passed the required core subjects of the degree or not. For example, BSCS students studied Computers or not in intermediate

5	Distance	This attribute suggests how much the university is far away from the house of the student
6	Employment Status	This attribute tells whether students are employed or not.

B. *K-Nearest-Neighbor (KNN)*

KNN, or k-Nearest Neighbors, is a simple and fundamental Supervised learning algorithm that can be used for both classification and regression problems. It classifies instances based on the similarity of the K nearest instances. KNN can handle both categorical and continuous variables.

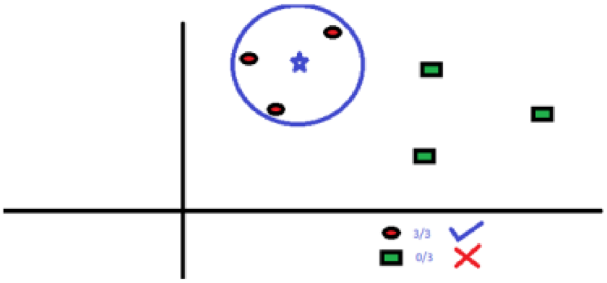


Fig. 1 Working of KNN

C. *Decision Tree(DT)*

A Decision Tree is a tree-based model employed in both regression and classification problems. It is a supervised learning algorithm widely used in machine learning and data mining. The tree is constructed by repeatedly dividing the data into smaller subsets based on the feature that provides the greatest information gain or lowest impurity. The tree's final nodes represent the target variable's predictions. Decision Trees are straightforward to comprehend and interpret, and they can handle both numerical and categorical data. However, they are susceptible to overfitting if the tree becomes too deep and complex. To mitigate this, the tree can be pruned or combined with ensemble methods like Random Forest or Boosting

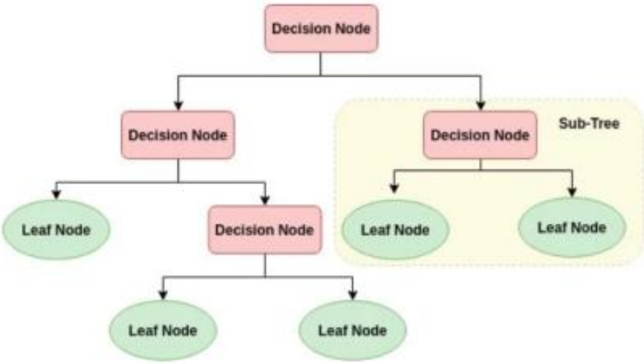


Fig. 2. Working of Decision Tree

D. *ExtraTreeClassifier(ET)*

ExtraTreeClassifier is an implementation of the decision tree algorithm in the scikit-learn library for Python. It uses extra randomization in the tree-building process to make trees more different from each other and to reduce overfitting compared to the standard decision tree algorithm. The ExtraTreeClassifier is typically used for classification problems, where the goal is to predict a categorical outcome based on input features.

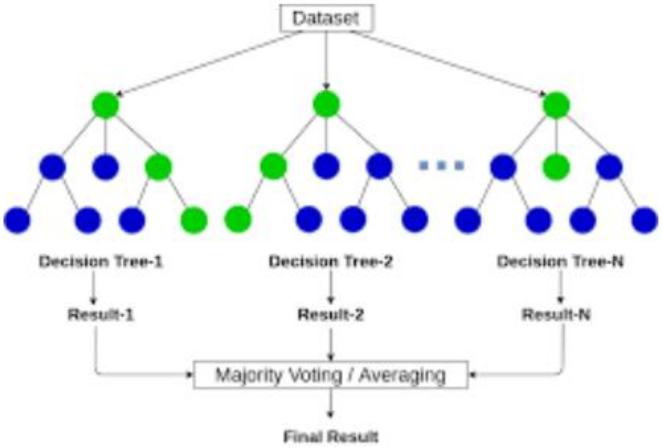


Fig. 3. Working of ExtraTreeClassifier

E. *Results*

Fig.4 Represents the accuracy measure of the KNN model implemented by randomly providing data, divided into a different number of splits.

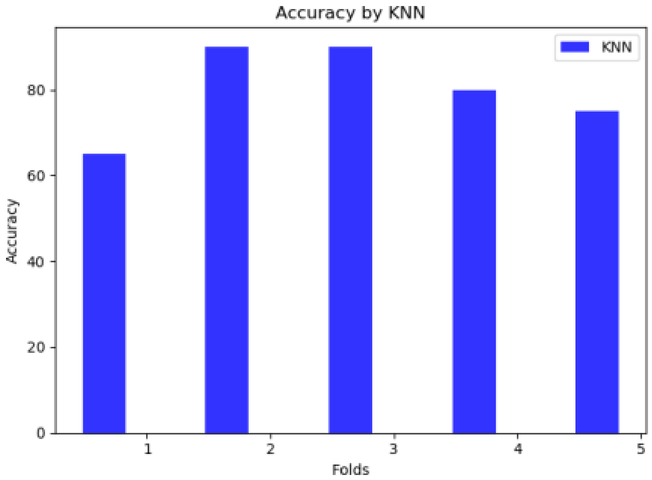


Fig. 4 K-Fold Cross Validation (KNN)

Fig.5 represents a graphical representation of Error Rate compared with different values of K during K –Fold cross validation implemented on the student data model. Fig.6 Illustrates about F1-Score of the algorithm, representing precision, recall, and accuracy of the KNN model prediction two classes as true or false. Information refers to the measure of reduction in Entropy (a measure of disorder). The Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.

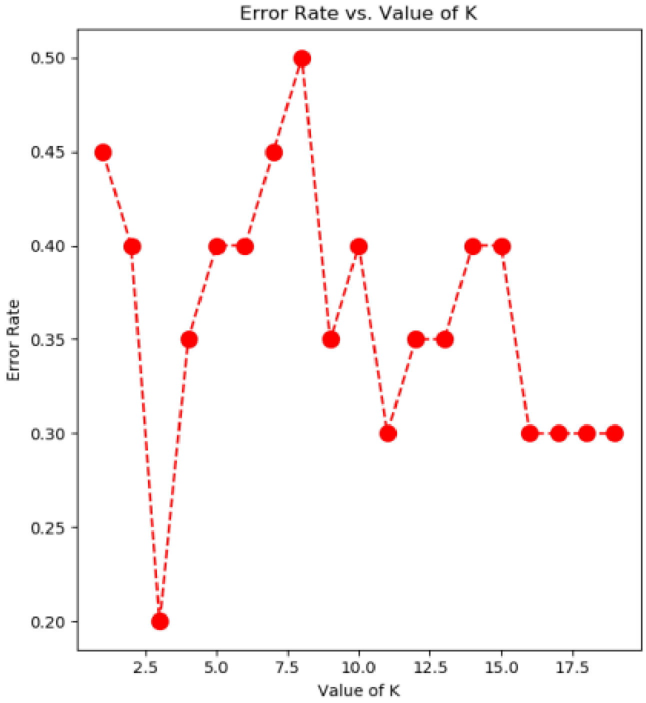
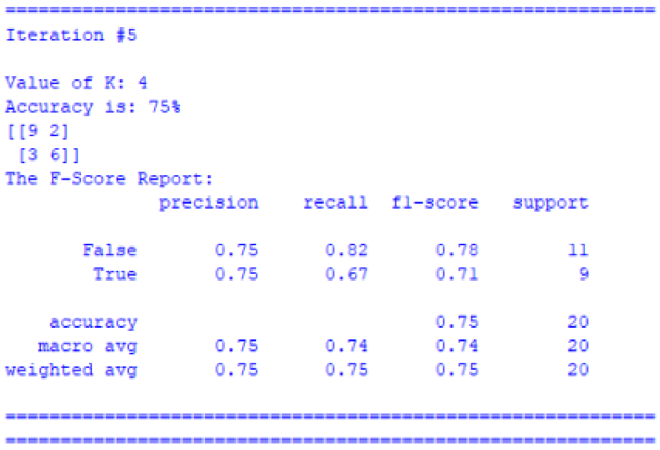


Fig. 5 Error rates for different K values (KNN)



```
============================================================
Iteration #5

Value of K: 4
Accuracy is: 75%
[[9 2]
 [3 6]]
The F-Score Report:
              precision    recall  f1-score   support

       False       0.75      0.82      0.78        11
        True       0.75      0.67      0.71         9

    accuracy                           0.75        20
   macro avg       0.75      0.74      0.74        20
weighted avg       0.75      0.75      0.75        20

============================================================
============================================================
```

Fig. 6 F1-Score Report of KNN Algorithm

Fig.7 represents the accuracy measure of the DT model implemented by randomly providing data, divided into a different number of splits. Fig.8 Illustrates about F1-Score of the algorithm representing precision, recall, and accuracy of the DT.

F. *Selecting Attributes With Highest Predictive Performance*

With the help of an Ensemble Classifier named as ExtraTreesClassifier that uses concepts of Entropy, Information Gain to sort the Attributes on the basis of their contribution to producing results or outcomes.

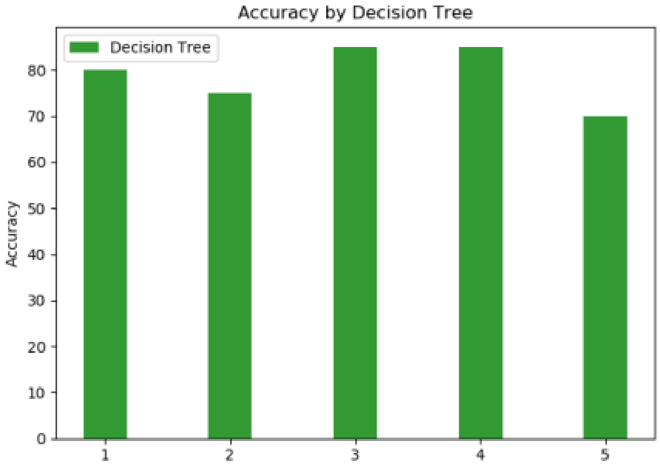
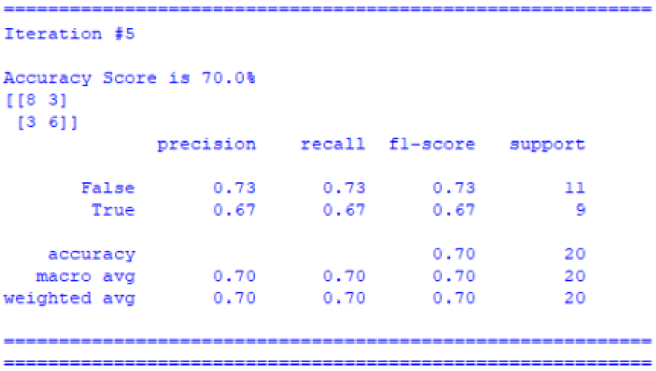


Fig. 7 K –Fold Cross Validation (DT)



```
============================================================
Iteration #5

Accuracy Score is 70.0%
[[8 3]
 [3 6]]
              precision    recall  f1-score   support

       False       0.73      0.73      0.73        11
        True       0.67      0.67      0.67         9

    accuracy                           0.70        20
   macro avg       0.70      0.70      0.70        20
weighted avg       0.70      0.70      0.70        20

============================================================
============================================================
```

Fig. 8 F1-Score Report of Decision Tree Algorithm

TABLE II ATTRIBUTES WITH HIGHEST PREDICTIVE PERFORMANCE

ID	Attribute	Feature Score
1	Percentage marks(previous degree academic results)	0.02049942
2	Total Number of F Grades	0.00627877
3	CGPA (current CGPA)	0.0336012
4	Distance (Total distance of student residence from university)	0.02526638
5	Core course status(Student has studied core courses in previous degree or not)	0.02712336

6	Employment Status(Student is job holder or not currently)	0.88906038

With the help of features important mentioned in Table II Students affair department and faculty members can concentrate on the factors that contribute to the student dropout ratio. According to the importance of the factors one can generate a questionnaire and will be able to provide counseling to the students.

G. *Comparative Analysis of KNN with Decision Tree*

By dividing the data set into multiple segments and testing them on both KNN and Decision Tree we can claim that KNN average accuracy in different data set segments is efficient from Decision Tree in our data model. Fig.9 provides us with a visualization of the comparison between accuracy results generated from both KNN and the Decision Tree.

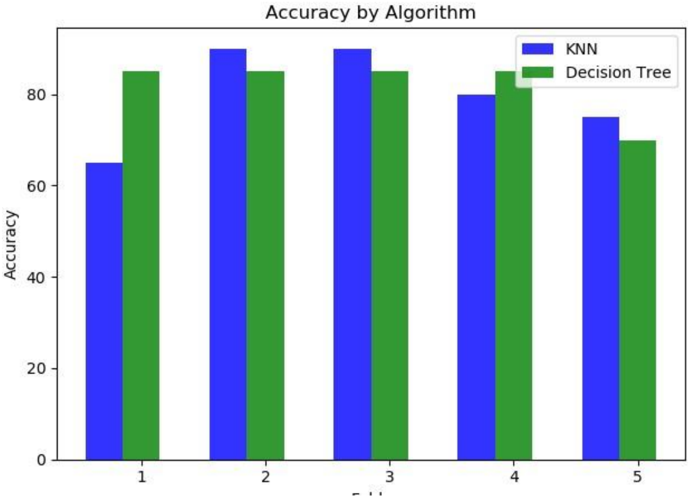


Fig. 9. K –Fold Cross Validation (DT)

The average accuracy of KNN puts 75% accuracy on our data set as compared to 70% of the Decision Tree.

V. TOOLS AND TECHNIQUES

Results were compiled by using different python libraries. The main libraries used in the implementation are: Scipy, Numpy, SciKit-Learn, Beautiful-Soup, Matplotlib and Seaborn.

Two supervised learning algorithms 1 KNN 2 Decision Tree are implemented to predict outcomes and one Ensemble method ExtraTreesClassifier is manipulated for determining the most predictive attributes in the learning process.

VI. Limitations

In this section, limitations in our approach are described. The Limitations of the existing approach can be defined as followings:

- Because the data set used in the approach is a low sample size and imbalance in nature. There are approaches for handling class imbalance and oversampling low-sized data sets. This oversampling if considered then it can generate more optimal results.
- Existing system also extracted Information gain. Considering it, if the classifier was evaluated after Feature selection approaches, then it will help to gain better insights into the classification performance.
- Existing system is only configured for binary-based classification. If multi-label or multi-class classification was adopted then it will add a constraint to the classifier to generate a more complex decision boundary.

VI. Future Work

There are several ways to improve the performance and overcome the limitations of machine learning algorithms. Firstly, additional attributes should be included in the data model, such as the total number of credit hours passed by the student and the major challenges faced in their academic career. Currently, our plan is to integrate the implementation with the university database and schedule regular jobs to collect and analyze each student's data. The resulting list of students at risk of dropping out will be communicated to the relevant authorities for preventive measures and counseling. Finally, incorporating deep learning models such as ANN, LSTM, and RNN with gradient-based optimization is expected to yield even better results.

VII. Conclusion

In our implementation, we have employed three machine learning algorithms to determine the reasons and factors for early withdrawals of currently enrolled students. This was achieved by analyzing attributes such as transcript records, the number of F grades, student employment status, the distance from the university, and the student's previous academic history. We collected data from the university database on both high-performing and low-performing students. The predictive results were integrated with the university web portal to notify stakeholders about the at-risk students and allow for preventive measures to be taken based on the contributing factors. Our next step is to enhance the accuracy of the predictive model by incorporating more attributes, as gathered from various stakeholders.

References

[1] Anwarudin Anwarudin, Widyastuti Andriyani, Bambang Purnomosidi DP, Dommy Kristomo (2022): The Prediction on the students' graduation timeliness using naive bayes classification and k-nearest neighbor, Journal of intelligent software system, Vol. 36.

[2] Devasia, Tismy and Vinushree, TP and Hegde, Vinayak. (2016): Prediction of students performance using Educational Data Mining, 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). Conf., Rostock, Proceeding. pp. 91–95.

[3] Fei, Mi and Yeung, Dit-Yan. (2015): Temporal models for predicting student dropout in massive open online courses, 2015 IEEE International Conference on Data Mining Workshop (ICDMW). Proceeding. pp. 256 – 263.

[4] He, Jiazhen and Bailey, James, Rubinstein, Benjamin IP and Zhang, Rui. (2015). Identifying at-risk students in massive open online courses,2015 Twenty-Ninth AAAI Conference on Artificial Intelligence.

[5] Bani, Mehrdad Haji, Mina. (2017). College Student Retention: When Do We Losing Them?Proceedings of the World Congress on Engineering and Computer Science 2017.

[6] Mike Sharkey and Robert Sanders. (2014).A process for predicting MOOC attrition.Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs.pp. 50—54,2014.

[7] Sampath Z Kovacic and Henrik Bostr om. Predicting student success by mining enrolment data. Research in Higher Education Journal, 2012.

[8] Keith Zvoch. Freshman year dropouts: Interactions between student and school characteristics and student dropout status. Journal of education for students placed at risk, 11(1):97–117, 2006.

[9] Hellen H. Maziku Joseph C. Mushl Yuda N. MnyawamiORCID Icon. Enhanced model for predicting student dropouts in developing countries using automated machine learning approach: A case of tanzanian's secondary schools. Applied Artificial Intelligence, 36(1), 2022.

[10] Alam TM, Mushtaq M, Shaukat K, Hameed IA, Umer Sarwar M, Luo S. A Novel Method for Performance Measurement of Public Educational Institutions Using Machine Learning Models. Applied Sciences. 2021 ; 11(19):9296. https://doi.org/10.3390/app11199296

[11] Javed, U., Shaukat, K., A. Hameed, I., Iqbal, F., Mahboob Alam, T., Luo, S. (2021). A Review of Content-Based and Context-Based Recommendation Systems. International Journal of Emerging Technologies in Learning (iJET), 16(03), pp. 274–306. https://doi.org/10.3991/ijet.v16i03.18851

[12] Cioffi R, Travaglioni M, Piscitelli G, Petrillo A, De Felice F. Artificial Intelligence and Machine Learning Applications in Smart Production: Progress, Trends, and Directions. Sustainability. 2020; 12(2):492. https://doi.org/10.3390/su12020492

[13] Shaukat K, Masood N, Khushi M. A Novel Approach to Data Extraction on Hyperlinked Webpages. Applied Sciences. 2019; 9(23):5102. https://doi.org/10.3390/app9235102