Data Cleaning:

1. Firstly, the given data was converted manually into test and train csv and data was extracted for further analysis.
2. Removing unwanted information data i.e., cleaning the data.

A screenshot of a computer

Description automatically generated with medium confidence

As we can see there are unwanted columns here ‘Unnamed: 6’, ‘Unnamed:7’ and Attributed information. These need to be removed from the dataset for training data and likewise same has to done on test data based upon the findings.

1. Finding irregularities in the data:

A screenshot of a graph

Description automatically generated with low confidence

As we can see there is irregularity in standard deviation of the data which may result in conflicting the accuracy of the data.

1. Unwanted space.:

Unwanted space is available in one of the column i.e [ ‘UNS’]. The extra spaces are removed from the data set.

1. Data types:

As it is observed that datatype for column is irregular in train and test data.

1. Removing outliers from the data:

• Methodology

Data Exploration:

* Exploring Each Column

STG (Study Time for Goal):

Descriptive Statistics: The mean study time is 0.0767, with a minimum of 0 and a maximum of 0.1.

Histogram: The histogram shows the distribution of study time values, indicating a relatively uniform distribution with most values concentrated around 0.

* SCG (Study Time for Course):

Descriptive Statistics: The mean study time is 0.1033, with a minimum of 0 and a maximum of 0.15.

Histogram: The histogram demonstrates a slightly skewed distribution towards higher study time values.

* STR (Study Time for Related Objects):

Descriptive Statistics: The mean study time is 0.14, with a minimum of 0.05 and a maximum of 0.4.

Histogram: The histogram exhibits a right-skewed distribution, indicating that most study time values are concentrated towards the lower end.

* LPR (Exam Performance for Related Objects):

Descriptive Statistics: The mean exam performance is 0.4067, with a minimum of 0.1 and a maximum of 0.98.

Histogram: The histogram suggests a relatively balanced distribution of exam performance values, with a slight concentration towards higher values.

* PEG (Exam Performance for Goal Object):

Bar Chart: The bar chart presents the count of different exam performance levels (categories) for the goal object. It indicates that the "very\_low" category has the highest count, followed by "Low," "Middle," and "High."

UNS (Knowledge Level):

Bar Chart: The bar chart illustrates the count of different knowledge levels. However, it seems that the dataset provided for this column contains missing values denoted by "None." Further cleaning or imputation of missing values may be required for a complete analysis.

Th

• Results

• Discussion

• Conclusion

• References

ABSTRACT

This study seeks to determine what distinguishes students who are categorised as high-level, middle-level, low-level, and extremely low-level knowledge users by analysing the behaviour of various students. Students' information about study time, repetition, and exam performance was collected, and materials for related objects. We employ the "K nearest neighbours’ classifier with parameter tuning and hill climbing" in this study, which examines several attributes and only chooses those that produce higher predicted results. To analyse user data and more accurately anticipate which user belongs to which group, "decision tree" categorisation is also utilised. Overall, it was clear that a student's "exam performance of goal" determined which category of knowledge consumers they would belong to. With the help of the models mentioned above, educators can identify students who need extra attention and support, as well as provide insights into the factors that contribute to academic success. We are trying to find what makes a user high-level and how can those at the lower level.

improve their performance. The models discussed above can assist instructors identify students who require additional attention and support as well as give them insights into the elements that contribute to academic performance. What makes a high-level user? How can users at lower levels do better? These are the questions we are attempting to answer.

INTRODUCTION

Before the introduction of online educational platforms, it was not clear how various users would interact with their study material. The introduction of online educational platforms enabled educators to observe and track the study patterns, time durations, repetition of goals, and related objects of various users. The goal of a user modelling system is to provide enough or suitable knowledge for students/users. In this process, user models are used for content personalization, accessing the data easily and fast, and streamlining the applications more effectively and efficiently. The user models have been mostly created in web-based applications, especially for online learning environments. Students’ models are composed of static and dynamic data in a web-based adaptive learning environment. Static data represents information about the students such as username, password, and age. Dynamic data represent the knowledge of students about domain-dependent data. Thus, the dynamic data in the user model might be also called the user knowledge model user modelling system (UMS) is used to obtain the behaviour or data of students from their interaction with web applications. This report dwells on understanding patterns in the behaviour of students belonging to different knowledge levels.

METHODOLOGY

The user knowledge was first collected by H. T. Kahraman and Sagiroglu, S., Colak to develop  
intuitive knowledge classifiers and modelling of users' domain-dependent data in web  
Knowledge-Based Systems Our research uses KNN and Decision tree classifications to  
understand and evaluate the classification of students in different knowledge categories. The user data was obtained from the UCI machine learning repository which contained two datasets about user knowledge classification. The first contained training data composed of 258 user records and the second contained testing data composed of 145 user records. The dataset comprised the following columns:

* STG (The degree of study time for goal object materials)
* SCG (The degree of repetition number of a user for goal object materials)
* STR (The degree of study time of a user for related objects with goal object)
* LPR (The exam performance of a user for related objects with goal object)
* PEG (The exam performance of a user for goal objects)
* UNS (The knowledge level of a user)

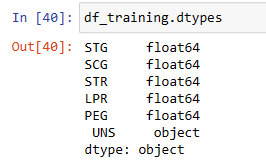
For the **cleaning process**, each column in the dataset went through the following checks.

* **Null value attributes:** I checked the data for possible null value attributes, but none were found.

A screenshot of a computer

Description automatically generated with medium confidence

* **Data type:** I ensured that all the columns had the same datatype.



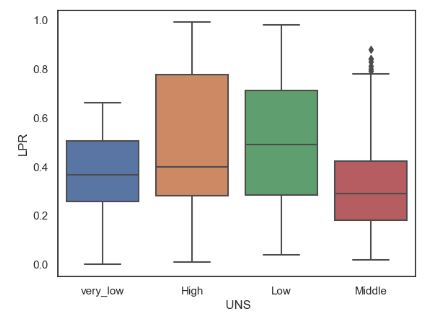
* **Invalid values and white spaces:** I checked the data for invalid values using the describe () method to ensure the maximum and minimum values for each column were well within the specified range. Also stripped the textual columns of any leading or trailing **white** **spaces** and checked if any invalid values are present in the target column.
* **Check for duplicate values:** I have checked duplicate values and there were none.

I now tried to find any patterns in the data given to us. Patterns that will help us classify the students. From the given graph below, it is quite visible that PEG(The exam performance of a user for goal objects) is a clear indicator of a student's knowledge level as it separates the users distinctively. But we know that exam performance is not always a good indicator of a student's knowledge level.

A picture containing text, diagram, screenshot, circle

Description automatically generated

I decided to plot a box plot for each of the columns for each knowledge level group. To see the disparity or similarity in the users.

* UNS VS STG 
* UNS VS STR

A picture containing diagram, rectangle, square, screenshot

Description automatically generated

* UNS vs LPR

A picture containing diagram, rectangle, square

Description automatically generated

* UNS VS SCG

A picture containing diagram, rectangle, square, screenshot

Description automatically generated

Fig 2.3

We use two classification models, K nearest neighbours and decision tree, and the first thing we notice in all of these graphs is that most High-level users are on the higher end of the spectrum in terms of study time, repetition, and the other two aspects. To get the best value of K for the given analysis, I implemented a basic for loop for values of K and accessed the accuracy of the model for each value of K. The K value for the model that provided the h score was the model that required the most study time. The model that was generated gave an accuracy score of **0.86**. To better understand the impact of each feature, I used the parameter tuning and hill climbing technique for K-nearest neighbours. The pseudo-code for the technique that I have used is given below. When conducting this analysis, the data was already divided into testing data and training data. The model was trained on the testing data and checked against the training data.

1. The following actions must be taken for each iteration of the test-train split.

2. Make the initial state the current state by evaluating the initial first characteristic.

3. Continue to improve the current state indefinitely until no more features can be used to do so.   
a) Choose a feature that hasn't yet been applied to the current state and apply it to create a new state.   
b) Carry out these actions to assess the new state.  
 i. If the current state is a goal state, then stop and return to success.  
 ii. If it is an improvement over the situation as it is, make it the situation and go on.  
Continue in the loop until a solution is found if step three is not an improvement over the existing situation.  
Step 4: Print the model accuracy, the test-train split percentage, and the optimum feature selection.

5th step: Exit.

**Modelling:**

For the decision tree algorithm, I used the classifier criterion as ‘Entropy’ as the target value setis not equally distributed. The minimum leaf node sample size has been set to 4 and the Max  
depth has been set to 7 to prevent overfitting the data. Analyzing the generated decision tree has  
the potential of providing insights into what feature differentiates the users. I have re-run the classifiers for different sizes of test/train data sets after combining the initial two testing and training datasets and mapped a graph of their accuracy over different test data set sizes.

RESULTS

On training the KNN model solely based on the different values of K to get the best number of neighbors, we attained the value of K = 3 and an accuracy score of 0.86897 over already split data.To further validate and analyze the accuracy of the model over a large enough dataset, I combined the training and testing data and then split the entire dataset into random training and testing datasets 20 times and gauged the accuracy of the model for the same value of K that was previously generated. The accuracy of the model for different train test splits of the data set was as follows As we can see, the score of the model varied for each train test split, and upon taking the average of those scores, we got an average accuracy of the model for the value of K = 3 as 0.8239.The next thing to analyze was which feature along with PEG had a significant impact on the accuracy of the model. To understand this using KNN, we used the method of parameter tuning and hill climbing function. Now the question comes as to what size of data should be used to train the model and what size should be used to test it. To find the optimal size, the model was run on multiple different sizes of train test splits. PEG and LPR were the two features that were selected in almost all of the test instances indicating that exam results are the prime indicator and classifier when it comes to user knowledge groups. One other feature that appears in many test instances is STG. Indicating the degree of study time is also an important classifier. Model prediction for KNN with parameter tuning

The next classifier we use is the decision tree classifier. For this classifier, we gave the entire

feature set with a minimum leaf node size of 3. The training and testing data were combined to

create one whole user data on which the model was executed. The data was split into training and

testing data similar to the hill climbing technique for KNN. The accuracy of the model for each

testing size was noted. The accuracy is given below

We can see that for the test size of 0.3, we get the maximum accuracy of 0.90. The general split of testing and training data lies around 20-30%. Which is, 70-80% of the data will be used to train the model and the remaining 20-30% will be used to test the model.Upon getting the accuracies of both the models, KNN with parameter tuning and Decision tree over different test train

splits. We plot a line graph to compare their accuracies.Throughout all different sets, KNN seems to be having a better accuracy when compared to decision trees when classifying the data. But for the recommended test train split size(0.3), they seem to have the same accuracy level of

around 0.90.For both models, we first found the parameters that provided the local best result. Then, for these said parameters we used K-fold cross-validation to validate the accuracies of the models. The results for the K-fold cross-validation for KNN with parameter tuning and decision tree are given

below

This confirms the statement that they both have the same accuracy of around 0.90 while

predicting the knowledge level of the students.

DISCUSSION

On the analysis of the graphs and the data models generated. The following results were obtained

1. When we look at the box plot of STG against UNS, it seems as if all students belonging

to all groups study almost for the same duration of time. But we can see that there are

plenty of outliers when it comes to High-level users showcasing that some of those that

are in the high-performance group study for much longer durations than the average

high-performance user to be a part of that group. This showcases that studying for a

longer duration on goal object materials might work for some students.

2. When we look at the box plot of SCG against UNS, it is quite clear that a majority of the

users that belong to the very-low level don't seem to be repeating what they study.

Although a singular outlier might be an exception we will not be considering that during

our analysis. This indicates that perhaps increasing the number of times a user studies

their given study material might result in them improving their scores and promoting

themselves to a higher group.

3. When we look at the box plot of STR against UNS, and the box plot of LPR against

UNS, we can observe that all user groups put in the same amount of effort to study

related objects with the goal object. But there is not much difference in the exam

performance of all the users belonging to the different groups. This says that the exam

performance of users for objects related to the goal object does not define the knowledge

level of the user by itself.

4. If we compare the box plots of all the features for those that belong to the low and middle

knowledge groups, it's visible that those in the low group on average, put less effort in

terms of study time and repetition of study material when compared to those belonging to

the middle group. As they have a relatively high average LPR, if those belonging to the

low group put in more effort, they might get promoted to the medium or even High

intelligence group.

5. If we compare the box plots of all the features for those that belong to the middle and

high knowledge groups. In terms of effort, the middle-knowledge group is on par with the

high-knowledge group. Yet the middle group suffers in the exam performance indicating

that some users might require less effort to perform better than other users.

The graph that the decision tree generates with a test train split of 0.3, is given below

If we take a look at this graph, after the classification based on PEG and LPR is done, those two

are the exam performance for goal objects and related goal objects. The next classifier that the

decision tree uses is SCG. Indicating that the decision tree considers the degree of repetition of

study material as an important classifier. This shows that for a given value of PEG and LPR if

your SCG is lower than a given threshold, the decision tree gets a good idea as to which

knowledge group you belong to.

This is contrary to the results obtained from using KNN where STG(degree of study) is the next

best feature to be considered when classifying the user data.

CONCLUSION

The goal of this research was to understand what makes a user a high-level user and how can

those at the lower level improve their performance. The two training models that we used named

KNN and decision tree showed us that other than a user’s exam performance, The duration for

which a user studies and the number of times he repeats the goal object materials also dictate the

knowledge level of the user. Some users spend a lot of time studying and repeating the materials and are not High-level users. These students showcase a want and need to learn and understand a concept. Giving them guidance and special attention should result in these students performing better in exams and promoting themselves to high-level users. Some users have a high enough STG and yet do not have that high value of LPR and PEG. This could be caused due to factors that are out of the scope of this research. Having these trained models, and live data of the current set of students such as how long they spend time on the online learning platform, how many times they repeat the given materials and their exam results, the teachers can classify the students into the knowledge groups and then based on that, focus particularly more on those students that need more attention. These students are those that have low and middle levels of understanding. The users that belong to the very low level, need to increase their study duration and number of repetitions at first to improve their level of knowledge.

REFERENCES

[1] H. T. Kahraman, Sagiroglu, S., Colak, I.(2013), Developing intuitive knowledge classifier

and modeling of users' domain dependent data in web, Knowledge Based Systems, vol. 37, pp.

283-295,

[2] Kahraman, H. T. (2009). Designing and Application of Web-Based Adaptive Intelligent

Education System. Gazi University Ph. D. Thesis, Turkey, 1-156

[3] V. Tsiriga, M. Virvou, A Framework for the Initialization of Student Models in Web-based

Intelligent Tutoring

[4] Dr. Yongli Ren,(2023) Practical Data Science: Feature Selection (Hill Climbing) Computer

Science & IT School of Science, week6-featureSelection.pdf