Nitte Meenakshi Institute of Technology
Department of Computer Science and Engineering
18CS54 Data Mining
Project Proposal

STUDENT PERFORMANCE ANALYSIS
1NT19CS103, 1NT19CS113, 1NT19CS181, 1NT19CS200
Madhumitha R, Meghana Reddy, Shreya Shettar, Tejashree Krishna Murthy

Introduction :

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD).

Data Mining Task :

Performance of the student in the final exam based on the test scores and various student attributes.Building a Decision Tree model for the dataset. Classify the students into three categories, "good", "fair", and "poor", according to their final exam performance using KNN Algorithm.

Data Set :

This data approaches student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under classification and regression tasks.

The target attribute G3 (predicted) has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades.

Attribute Information:

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

`school`	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
`sex`	student's sex (binary: 'F' - female or 'M' - male)

age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal: 'mother', 'father' or 'other')
traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)

goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese:

G1 - first period grade (numeric: from 0 to 20)
G2 - second period grade (numeric: from 0 to 20)
G3 - final grade (numeric: from 0 to 20, output target)

Methods and Models :

A classification problem has a discrete value as its output. A decision tree model will be implemented to generate rules and identify the important attributes in the decision making process. To execute the decision tree model, attributes are converted into Pandas Series and passed onto functions from in-built modules (sklearn.tree).

We then apply the KNN algorithm to our model. The KNN algorithm assumes that similar things exist in close proximity and thus analyses the nearest neighbors to predict the class label of the test case.

Assessment :

We evaluate the models- Decision tree and KNN model to ensure our project is as accurate as possible.We use Accuracy, Recall and Precision for assessing Classification techniques.

1.Accuracy

The accuracy of a classifier is given as the percentage of total correct predictions divided by the total number of instances.If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known.

2.Recall

Recall also known as True Positivity Rate [TPR] is one of the most used evaluation metrics for an unbalanced dataset. It calculates how many of the actual positives our model predicted as positives (True Positive).

3. Precision

Precision describes how accurate or precise our data mining model is. Out of those cases predicted positive, how many of them are actually positive.

For regression problems, root mean square error (RMSE), sum of squared errors (SSE), mean average error (MAE), etc., evaluation measures are used. These measures are most suitable with continuous values output, unlike classification or clustering, where we deal with discrete output values.

Presentation and Visualization :

We will build a minimalistic website to visualize our model using Front-end technologies like CSS and Javascript frameworks. We are going to display the results of the Python Source Code in a graphical manner.

Roles :

Madhumitha R	-	Working on Data Integration and Visualization
Meghana Reddy	-	Working on Data Analysis
Shreya Shettar	-	Working on Data Integration and Visualization
Tejashree Krishna Murthy	-	Working on Data Analysis

Schedule :

Date	Tasks
05/01/22	Data pre-processing
09/01/22	Loading Dataset and performing Data Mining tasks
12/01/22	Data Visualization
15/01/22	Model Evaluation
17/01/22	Project Report

Bibliography :

- http://archive.ics.uci.edu/ml/datasets/Student+Performance#
- https://www.slideshare.net/HafsaHabib2/student-performance-data-mining-project-report
- https://www.sas.com/en_us/insights/analytics/data-mining.html#:~:text=Data%20mining%20is%20the%20process,relationships%2C%20reduce%20risks%20and%20more
- https://www.mdpi.com/2306-5729/6/11/110/htm
- https://www.pluralsight.com/guides/evaluating-a-data-mining-model