Itroduction

The use of datasets for the study and prediction of student performance has become crucial in today's educational systems. Educational institutions are utilizing machine learning algorithms and advanced analytics to forecast future results and get insights into student performance, thanks to the growing amount of data on student demographics, academic background, and other pertinent criteria. This method assists teachers in pinpointing areas that require development, customizing lessons, and offering focused assistance to pupils who are having difficulty. Additionally, by predicting and analyzing student performance, legislators and school administrators may make better decisions and more efficiently distribute resources. This article will address some of the techniques and resources utilized in this sector as well as the advantages of using datasets for analysis and prediction of student performance.

Comprehending the Problem Statement

The goal of this experiment is to determine how factors including gender, ethnicity, parental education level, test preparation course, and lunch affect students' performance, as measured by test scores.

Giving their students a top-notch education is the main goal of higher education establishments. Knowledge must be found to predict student performance, identify problems with traditional classroom teaching models, detect unfair means used in online examinations, detect abnormal values in student result sheets, and predict student enrollment in specific courses in order to attain the highest level of quality in the educational system. Through the use of data mining tools, this knowledge can be recovered from educational datasets that conceal it.

## Abstract

[Student performance](https://www.sciencedirect.com/topics/social-sciences/student-performance), student progress and student potential are critical for measuring learning results, selecting learning materials and learning activities. However, existing work doesn't provide enough analysis tools to analyze how students performed, which factors would affect their performance, in which way students can make progress, and whether students have potential to perform better. To solve those problems, we have provided multiple analysis tools to analyze student performance, student progress and student potentials in different ways. First, this paper formulates student model with performance related attributes and non-performance related attributes by Student Attribute Matrix (SAM), which quantifies student attributes, so that we can use it to make further analysis. Second, this paper provides a student performance estimation tools using Back Propagation [Neural Network](https://www.sciencedirect.com/topics/social-sciences/neural-network) (BP-NN) based on classification, which can estimate student performance/attributes according to students' [prior knowledge](https://www.sciencedirect.com/topics/social-sciences/prior-knowledge) as well as the performance/attributes of other students who have similar characteristics. Third, this paper proposes student progress indicators and attribute causal relationship predicator based on BP-NN to comprehensively describe student progress on various aspects together with their causal relationships. Those indicators and predicator can tell how much a factor would affect student performance, so that we can train up students on purpose. Finally, this paper proposes a student potential function that evaluates student achievement and development of such attributes. We have illustrated our analysis tools by using real [academic performance](https://www.sciencedirect.com/topics/social-sciences/academic-performance) data collected from 60 high school students. Evaluation results show that the proposed tools can give correct and more accurate results, and also offer a better understanding on student progress.

## Background and related work

Student assessment measures the levels of student achievement in terms of knowledge and abilities. The methods of student assessment contains summative assessment and formative assessment (Oscarson & Apelgren, 2011). Information about student progress is required to be collected before, during and after taking certain learning activities (Feng, Heffernan, Heffernan, & Mani, 2009; Oscarson & Apelgren, 2011). Student progress can be expressed by growth rate (Betebenner, 2009; Stecker, Lembke, &

## Student progress and development

Analyzing student progress is critical. Different subjects (or learning activities (LAs) (Yang, Li, & Lau, 2010)) have different assessment criteria, where some are subject specific, and some are shared among subjects. On the other side, learning styles and learning modes also play significant roles on how a student perform and make progress in different assessment criteria. We have developed student attribute descriptors to provide a more complete picture on student's progress, performance,

## Experiments results and discussions

In order to analyze student learning progress with our BP-NN based student progress indicator by finding out attribute causal relationship, to predict student potential with improved Progress potential function, and to test the student performance estimator based on BP-NN model, we have collected academic data over 6 subjects of 60 high school students from No.83 Xi'an Middle School, China. These data contains their test results in both year 1 and year 2. And also, we ask 6 teachers who taught

**Import Data and Required Packages**

Importing Pandas, Numpy, Matplotlib, Seaborn and Warings Library.

**import** pandas **as** pd

**import** numpy **as** np

**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sns

**import** warnings

warnings.**filterwarnings**("ignore")

Import the CSV Data as Pandas DataFrame

df = pd.read\_csv("data/StudentsPerformance.csv")

df.shape

https://cdn.analyticsvidhya.com/wp-content/uploads/2023/04/df-shape.png

Dataset Information

* gender: sex of students -> (Male/female)
* race/ethnicity: ethnicity of students -> (Group A, B, C, D, E)
* parental level of education: parents’ final education ->(bachelor’s degree, some college, master’s degree, associate’s degree)
* lunch: having lunch before test (standard or free/reduced)
* test preparation course: complete or not complete before test
* math score
* reading score
* writing score

After that, we check the data as the next step. There are a number of categorical features contained in the dataset, including multiple [missing value](https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/) kinds, duplicate values, check data types, and a number of unique value types.

Data Checks to Perform

* Check Missing values
* Check Duplicates
* Check data type
* Check the number of unique values in each column
* Check the statistics of the data set
* Check various categories present in the different categorical column

.

**Check Duplicates**

If checking the our dataset has any duplicated values present or not

df.duplicated().sum()

https://cdn.analyticsvidhya.com/wp-content/uploads/2023/04/df-duplicated-sum-1.png

## Exploring Data (Visualization)

Visualize Average Score Distribution to Make Some Conclusion

* Histogram
* Kernel Distribution Function (KDE)
* Histogram & KDE

In [102]:

**import** pandas **as** pd;

**import** numpy **as** np;

**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sn

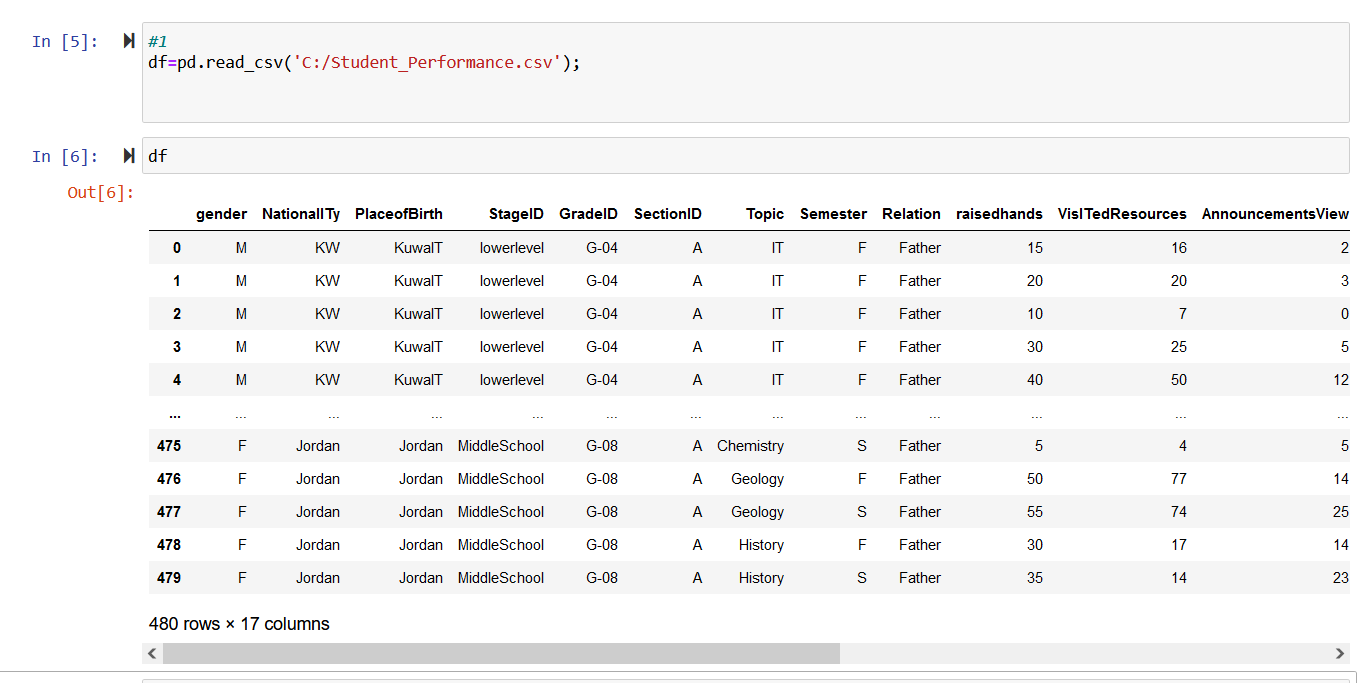
**import** plotly.express **as** pe

**from** sklearn.model\_selection **import** train\_test\_split

**from** sklearn.linear\_model **import** LinearRegression

**import** warnings

warnings.filterwarnings('ignore')



In

*#dataset columns and rows*

In [7]:

Out[7]:

df.columns

Index(['gender', 'NationalITy', 'PlaceofBirth', 'StageID', 'GradeID', 'SectionID', 'Topic', 'Semester', 'Relation', 'raisedhands', 'VisITedResources', 'AnnouncementsView', 'Discussion',

'ParentAnsweringSurvey', 'ParentschoolSatisfaction', 'StudentAbsenceDays', 'Class'],

dtype='object')

# Producing Various descriptive statistics of the analytic dataset.

In [8]:

mean**=**df["Discussion"].mean();

*#Mean: The mean represents the average value of the Discussion*

In [12]:

median**=**df["Discussion"].median();

*#Median: The median represents the middle value of the Discussion when the values are sor*

mode**=**df["Discussion"].mode().tolist();

*#Mode: The mode represents the most frequently occurring Discussion value.*

*#It is useful for identifying the population value that appears with the highest frequenc*

std**=**df["Discussion"].std();

*#Standard Deviation: The standard deviation measures the dispersion or spread of the Disc*

In [13]:

min\_value**=**df["Discussion"].min(); max\_value**=**df["Discussion"].max();

*#Minimum and Maximum: The minimum and maximum values represent the smallest and largest # #the range of Discussion valuesin the dataset. values, respectively.*

In [15]:

quartiles**=**df["Discussion"].quantile([0.25,0.5,0.75]);

*#Quartiles: Quartiles divide the population into four equal parts, each containing 25% of #They help identify the range of values within which a specific percentage of the populat #The quartiles consist of the first quartile (Q1), median (Q2 or 50th percentile), and th*

In [ ]:

*# print(f"mean is {round(mean,2)} \n meadian are {median}\n mode is {mode}\n standard dev*

print(f" manimum value are {min\_value}\n maxximum value are {max\_value}\n Quartiles are {

# comments

## the mean indicates us the average value of the Discussion is 43.28 meadian is shows us the middle value of the Discussion is 39.0 Mode show us the most frequently occurring Discussion value is 70

**The standard deviation measures the dispersion or spread othe Discussion values around the mean is 27.64**

manimum and maximum are show us the lowest value and heightest value are 1 and 99 respectivily

Quartile one showing us 25% are 20.0 Quartile two showing 50% are 39.0

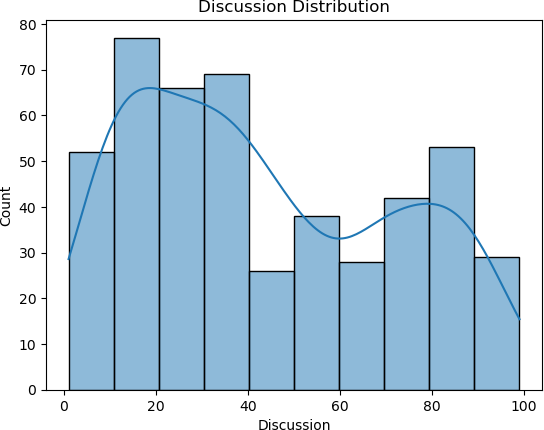
Quartile three 75% are 70.0

# Building possible visualizations

In [84]:

sn.histplot(df["Discussion"],bins**=**10,kde**=True**) plt.title("Discussion Distribution")

plt.show()



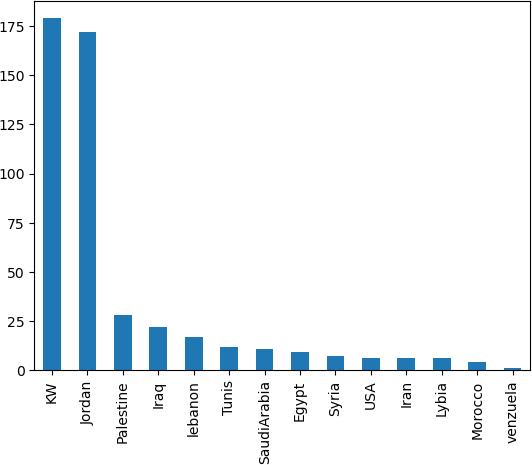
# COMMENTS: this HistGrapsh is showing us The most discussion was between 10 to 20, in general the discussion was 0 to 40.

In [82]:

Out[82]:

df.NationalITy.value\_counts().plot(kind**=**"bar")

<Axes: >



# This grapsh shows the Academic Performance, the most students are from Kuwait = kw and it is more than 175 students, the smallest is from Venezuela and it is only one student.

In [104]:

df.NationalITy.value\_counts()

Out[104]:

KW 179

Jordan 172

Palestine 28

Iraq 22

lebanon 17

Tunis 12

SaudiArabia 11

Egypt 9

Syria 7

USA 6

Iran 6

Lybia 6

Morocco 4

venzuela 1

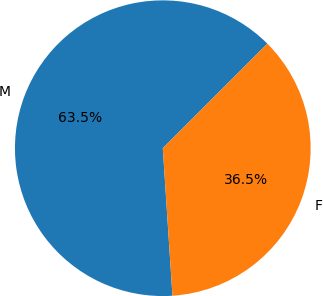
Name: NationalITy, dtype: int64

In [80]:

data**=**df["gender"].value\_counts() label**=**df["gender"].unique()

fig,res**=**plt.subplots()

res.pie(data,labels**=**label,autopct**=**"%1.1f%%",startangle**=**45) plt.show()



# When looking at girls and boys, there are more boys and less girls, boys are more than 60%, girls less than 40%.

1. **checking null values and handling**

In [ ]:

null**=**df.isnull().sum(); *#df.fillna(method="ffill",inplace=True) # or bfill*

|  |  |  |
| --- | --- | --- |
| In [31]: | null |  |
| Out[31]: | gender | 0 |
|  | NationalITy | 0 |
|  | PlaceofBirth | 0 |
|  | StageID | 0 |
|  | GradeID | 0 |
|  | SectionID | 0 |
|  | Topic | 0 |
|  | Semester | 0 |
|  | Relation | 0 |
|  | raisedhands | 0 |
|  | VisITedResources | 0 |
|  | AnnouncementsView | 0 |
|  | Discussion | 0 |
|  | ParentAnsweringSurvey | 0 |
|  | ParentschoolSatisfaction | 0 |
|  | StudentAbsenceDays | 0 |
|  | Class | 0 |
|  | dtype: int64 |  |

## we checked missing value and the result showing us no any missing value

In [60]:

duplicate**=** df.duplicated().sum()

In [46]:

duplicate

Out[46]: 2

## we check duplicated value and we found 2 duplicated value so, we drop duplicate value usnig these steps below

In [113]:

*#Clear duplicated Value*

df.drop\_duplicates(inplace**=True**)

*#df.drop\_duplicates(inplace=True) # xalka waa kan*

In [116]:

*#after clear duplicate value*

df.duplicated().sum()

Out[116]: 0

# investigate the unique values of categories variables

In [118]:

unique\_values**=**df.select\_dtypes(include**=**['object']).nunique() unique\_values

|  |  |  |
| --- | --- | --- |
| Out[118]: | gender | 2 |
|  | NationalITy | 14 |
|  | PlaceofBirth | 14 |
|  | StageID | 3 |
|  | GradeID | 10 |
|  | SectionID | 3 |
|  | Topic | 12 |
|  | Semester | 2 |
|  | Relation | 2 |
|  | ParentAnsweringSurvey | 2 |
|  | ParentschoolSatisfaction | 2 |
|  | StudentAbsenceDays | 2 |
|  | Class | 3 |
|  | dtype: int64 |  |

Gender are categories :Male and female

Nationalyty are: KW, lebanon, Egypt, SaudiArabia, USA, Jordan, venzuela, Iran, Tunis, Morocco, Syria, Palestine,

Iraq, Lybia

PlaceofBirth are 14: KW, lebanon, Egypt, SaudiArabia, USA, Jordan, venzuela, Iran, Tunis, Morocco, Syria, Palestine, Iraq, Lybia

Semester are 2: first and Second

# Building a Linear Regression Mode

In [98]:

*#fig.show()*

In [146]:

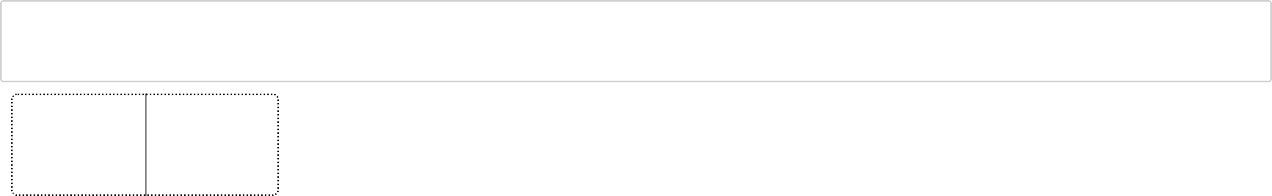
features**=**["raisedhands","VisITedResources"] x**=**df[features]

y**=**df["Discussion"]

In [147]:

xtrain,xtest,ytrain,ytest**=**train\_test\_split(x,y,test\_size**=**0.2)

In [148]:



ml**=**LinearRegression() ml.fit(xtrain,ytrain)

▾ LinearRegression LinearRegression()

Out[148]:

In [149]:

print("Accurancy: ",ml.score(xtest,ytest))

Accurancy: 0.12116356755759583

In [150]:

Discus**=**[[20,40]]

In [152]:

print(ml.predict(Discus))

[34.989666]

# Reporting for summarizing all steps you have already done minimum 150 words but not limited

Title: Exploratory Data Analysis and Linear Regression Modeling for Student Performance Dataset

In this notebook, we conducted a comprehensive analysis of the Student Performance dataset, covering various steps from data

exploration to building a linear regression model. We began by loading the dataset and inspecting its structure, identifying key

features, and understanding the data types.

We then performed exploratory data analysis (EDA) to gain insights into the dataset. This involved generating descriptive statistics,

exploring unique values in categorical variables, and visualizing data distributions using count plots and histograms. We also checked

for missing values and duplicates, employing appropriate handling methods.

The subsequent step involved building a linear regression model to predict the 'raisedhands' variable using selected independent

variables. We interpreted the model parameters, including the intercept and coefficients, to understand their significance in predicting

student engagement. Model evaluation metrics such as Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error

were used to assess the model's accuracy.

The notebook concluded with a scatter plot visualizing the alignment between actual and predicted values, providing a practical

understanding of the model's performance. The entire process aimed to provide a holistic understanding of the dataset and

demonstrate the steps involved in exploratory data analysis and linear regression modeling

In summary

The prediction of the student's performance ends here. Let's go over what we did. We began by outlining our problem statement, researching the techniques we will employ, and setting up the implementation pipeline for regression. After that, we practically implemented several identification and regression methods, including CatBoosting, AdaBoost, Lasso, K-Neighbors, Decision Tree, Random Forest, XGB, and Linear regression. Next, we evaluated these models' performances side by side. Finally, we developed a linear regression model that demonstrated its optimal performance for student performance prediction tasks.

This student performance prediction's main conclusions are:

For many schools, identifying the predictor of student performance is crucial.

In comparison to other regression problems, linear regression yields higher accuracy.

Linear

## References (32)

* H. Alzoubi et al.

### [Fast and accurate global motion estimation algorithm using pixel subsampling](https://www.sciencedirect.com/science/article/pii/S0020025508001436)

### Information Science

(2008)

* C.M. Chen et al.

### [Personalized e-learning system using Item Response Theory](https://www.sciencedirect.com/science/article/pii/S036013150400034X)

### Computers & Education

(April 2005)

* M. Oscarson et al.

### [Mapping language teachers' conceptions of student assessment procedures in relation to grading: A two-stage empirical inquiry](https://www.sciencedirect.com/science/article/pii/S0346251X11000157)

### Journal of System

(2011)

* M.B. Wieling et al.

### [The Impact of online video lecture recordings and automated feedback on student performance](https://www.sciencedirect.com/science/article/pii/S0360131509002784)

### Compertes & Education

(May 2010)

* Y.F. Yang et al.

### [Conceptions of and approaches to learning through online peer assessment](https://www.sciencedirect.com/science/article/pii/S095947520900005X)

### Learning and Instruction

(2010)

* R. Alkhasawneh et al.

### ‘Modeling student retention in science and engineering disciplines using neural networks

* R. Baker et al.

### v26.pdf">Contextual slip and prediction of student performance after use of an intelligent tutor

### UMAP

(2010)

* G. Barata et al.

### Early predication of student profile based on performance and gaming preferences

### IEEE Transaction on Learning Technologies

(2016)

* D. Betebenner

### Norm- and criterion-referenced student growth

### Educational Measurement: Issues and Practice

(2009)

* J.C. Bezdek

### Pattern recognition with fuzzy objective functional algorithms

(1981)