University of Computer Studies Yangon

Student Performance Classification System With Decision Tree Algorithm

Submitted By

Mr.Than Swe Lin

Abstract

- The analysis and evaluation of students’ performance and retaining the standard of education is a very important problem in all the educations situations.
- The most important goal of this project is to analyze and evaluate the school students’ performance by applying data mining classification with decision tree algorithm.
- In this system, data set is considered in prediction of the performance of students.
- Can predict the average of students who passed or failed depend on reading score, writing score and math score.
- Can also analyze read pass, write pass and math pass with all the students of “parental level of education” and “test preparation course”.
- Build decision tree classification with python.

Introduction

- Data mining helps to extract the relevant information from the large and complex databases.
- Data Mining Techniques are useful for data analysis and predictions.
- Classification techniques is an unsupervised learning technique.
- There are various classification techniques such as Decision tree algorithm, Bayesian network and Neural network etc.
- This project propose a classification model particularly decision tree algorithm to predict the performance of students.
- Python Scikit-learn package is used for model construction and evaluation.
- Scikit-learn is a free software machine learning library for the Python programming language.

Applied Language and Tools

Language : Python

Python is a programming language that lets you work quickly and integrate systems more effectively.

Tools : Anaconda Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands

Tools : Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Method

- Decision tree
- It builds classification or regression models in the form of a tree structure.
- It can be used as a decision-making tool, for research analysis, or for planning strategy.
- It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- Decision trees can handle both categorical and numerical data.
- A primary advantage for using a decision tree is that it is easy to follow and understand.

Experiments

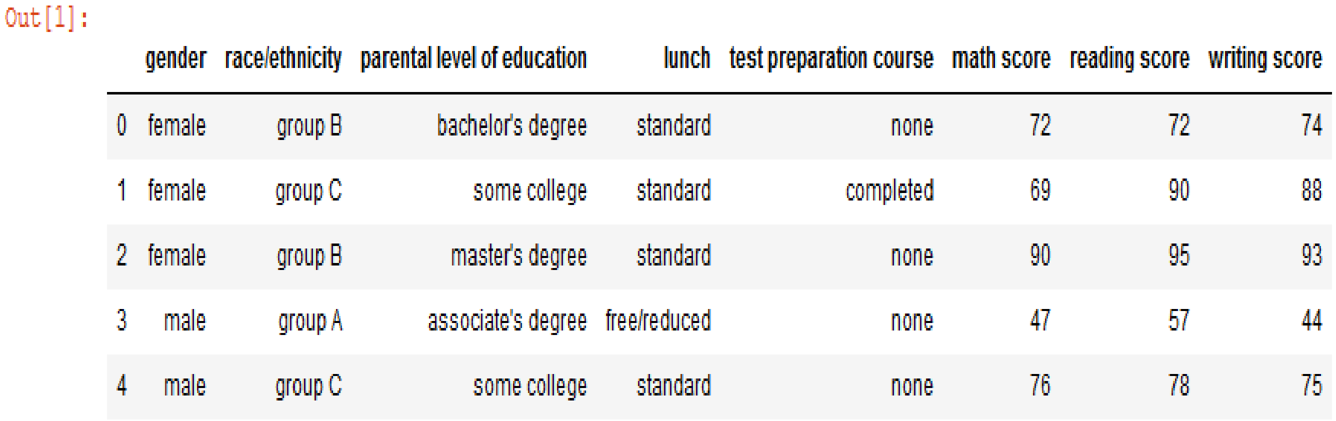
About Dataset

We use the data set from Kaggle (*Kaggle* is the world's largest data science community). There are 1000 instances and 8 attributes of datasets in our project. They are-

Gender	(female, male)
Race/ethnicity	(GroupA, GroupB, GroupC, GroupD, GroupE)
Parental level of education	(bachelor’s degree, some college, master’s degree, associate’s degree, high school, some high school)
Lunch	(Standard, free reduced)
Test preparation course	(none, completed)
Math score	(Minimum, maximum, mean, StdDev)
Reading score	(Minimum, maximum, mean, StdDev)
Writing score	(Minimum, maximum, mean, StdDev)

Data Preprocessing

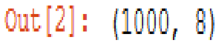
First top 5 row of data set



Out[1]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

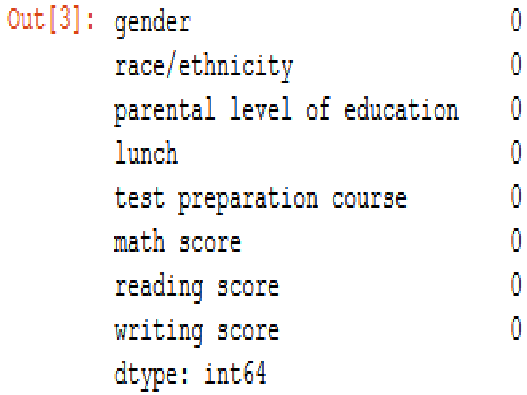
Shape of the data set



```
Out[2]: (1000, 8)
```

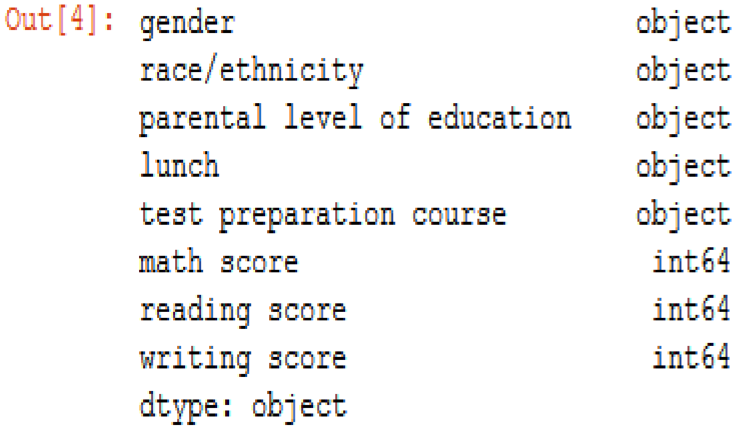
Data Preprocessing

Check the missing values



```
Out[3]: gender                         0
        race/ethnicity                 0
        parental level of education    0
        lunch                          0
        test preparation course        0
        math score                     0
        reading score                  0
        writing score                  0
        dtype: int64
```

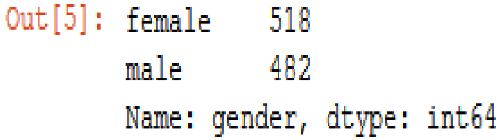
Check the datatype of all the column values



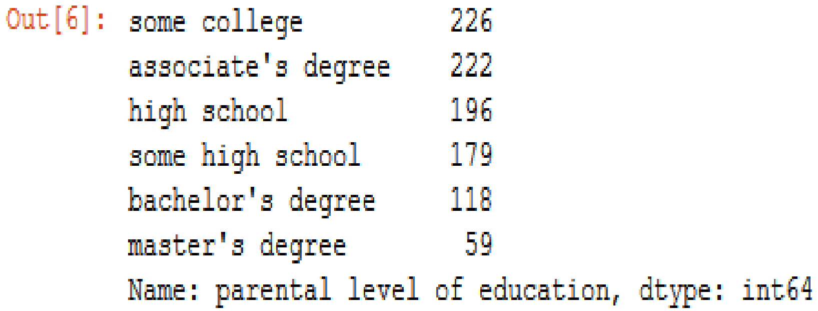
```
Out[4]: gender                         object
        race/ethnicity                 object
        parental level of education    object
        lunch                          object
        test preparation course        object
        math score                      int64
        reading score                   int64
        writing score                   int64
        dtype: object
```

Data Preprocessing

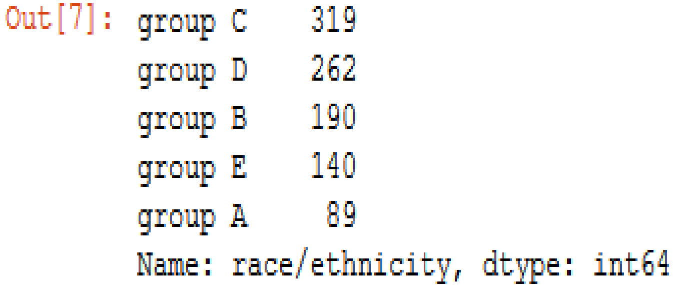
Analyze the values of the columns and check they are numerical or categorical.



```
Out[5]: female    518
        male      482
        Name: gender, dtype: int64
```

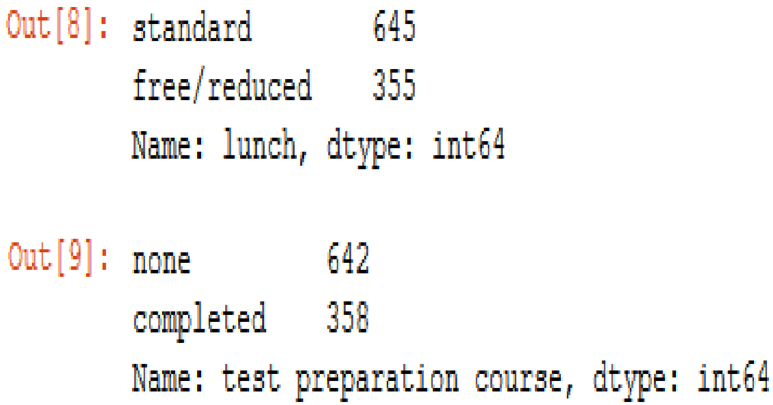


```
Out[6]: some college          226
        associate's degree    222
        high school           196
        some high school      179
        bachelor's degree     118
        master's degree        59
        Name: parental level of education, dtype: int64
```



```
Out[7]: group C    319
        group D    262
        group B    190
        group E    140
        group A     89
        Name: race/ethnicity, dtype: int64
```

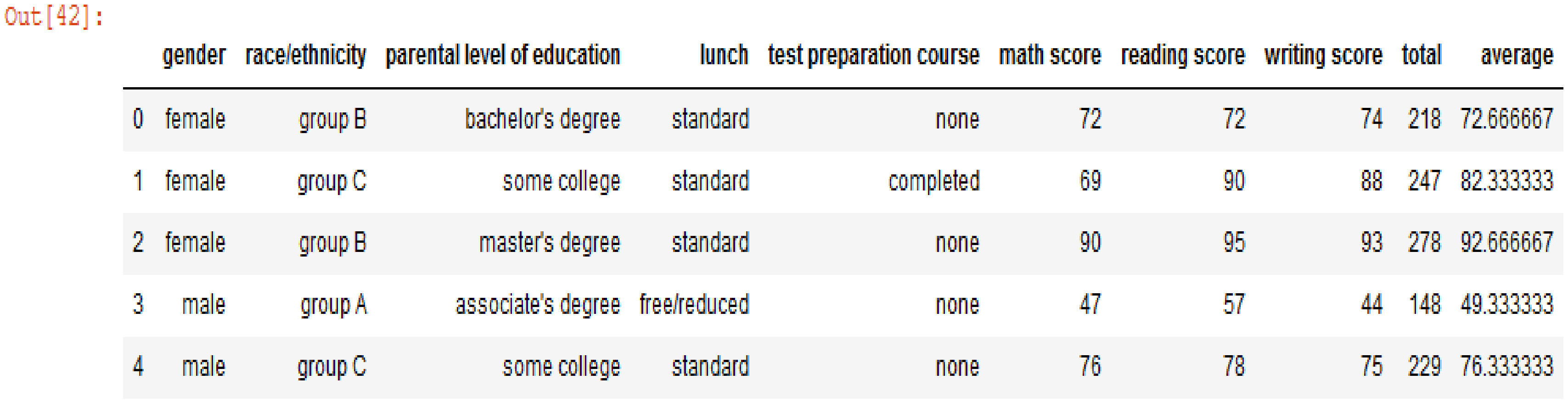
Data Preprocessing



```
Out[8]: standard        645
        free/reduced    355
        Name: lunch, dtype: int64

Out[9]: none         642
        completed    358
        Name: test preparation course, dtype: int64
```

Adding columns "total" and "average" to the dataset

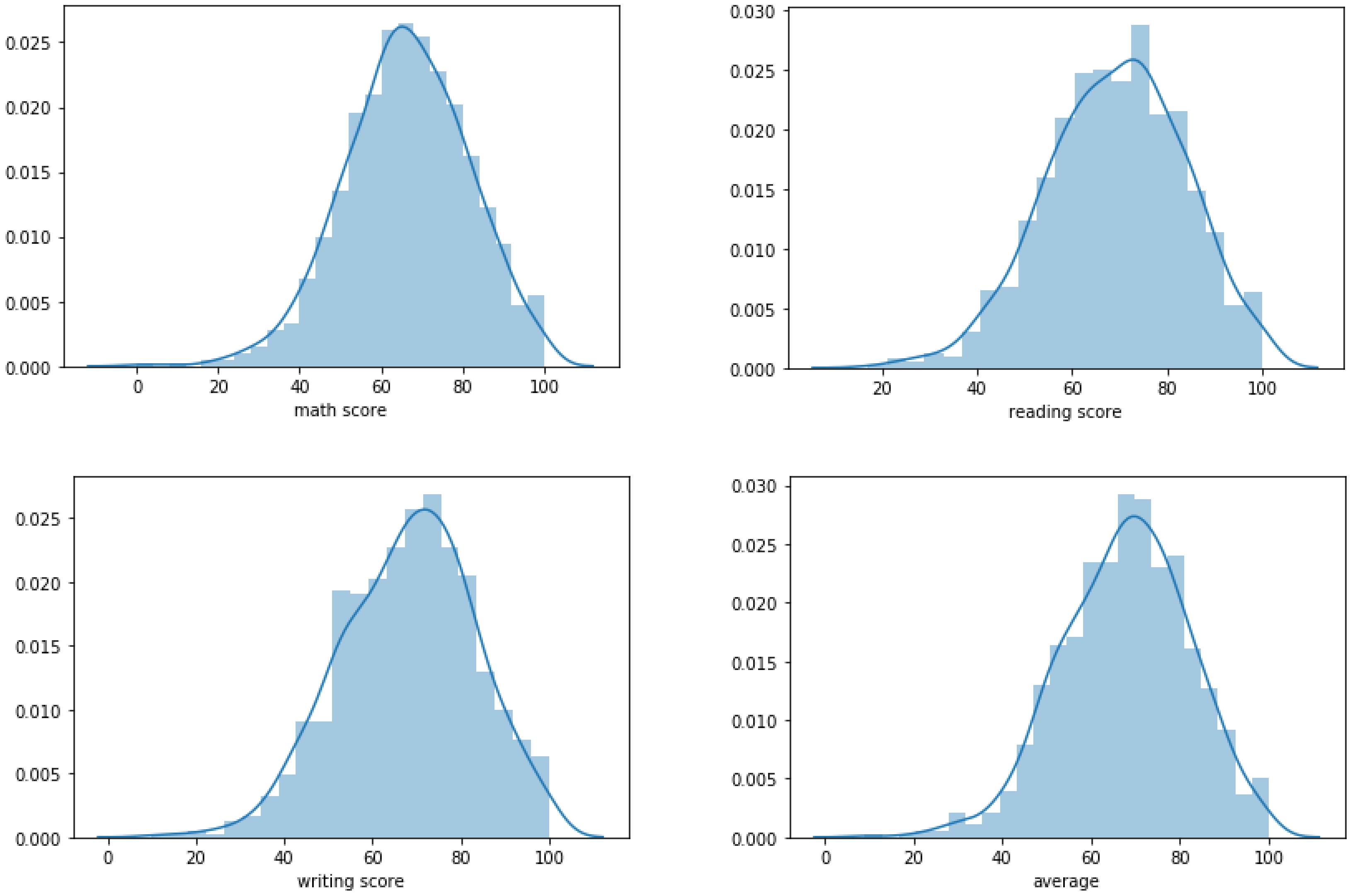


Out[42]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	total	average
0	female	group B	bachelor's degree	standard	none	72	72	74	218	72.666667
1	female	group C	some college	standard	completed	69	90	88	247	82.333333
2	female	group B	master's degree	standard	none	90	95	93	278	92.666667
3	male	group A	associate's degree	free/reduced	none	47	57	44	148	49.333333
4	male	group C	some college	standard	none	76	78	75	229	76.333333

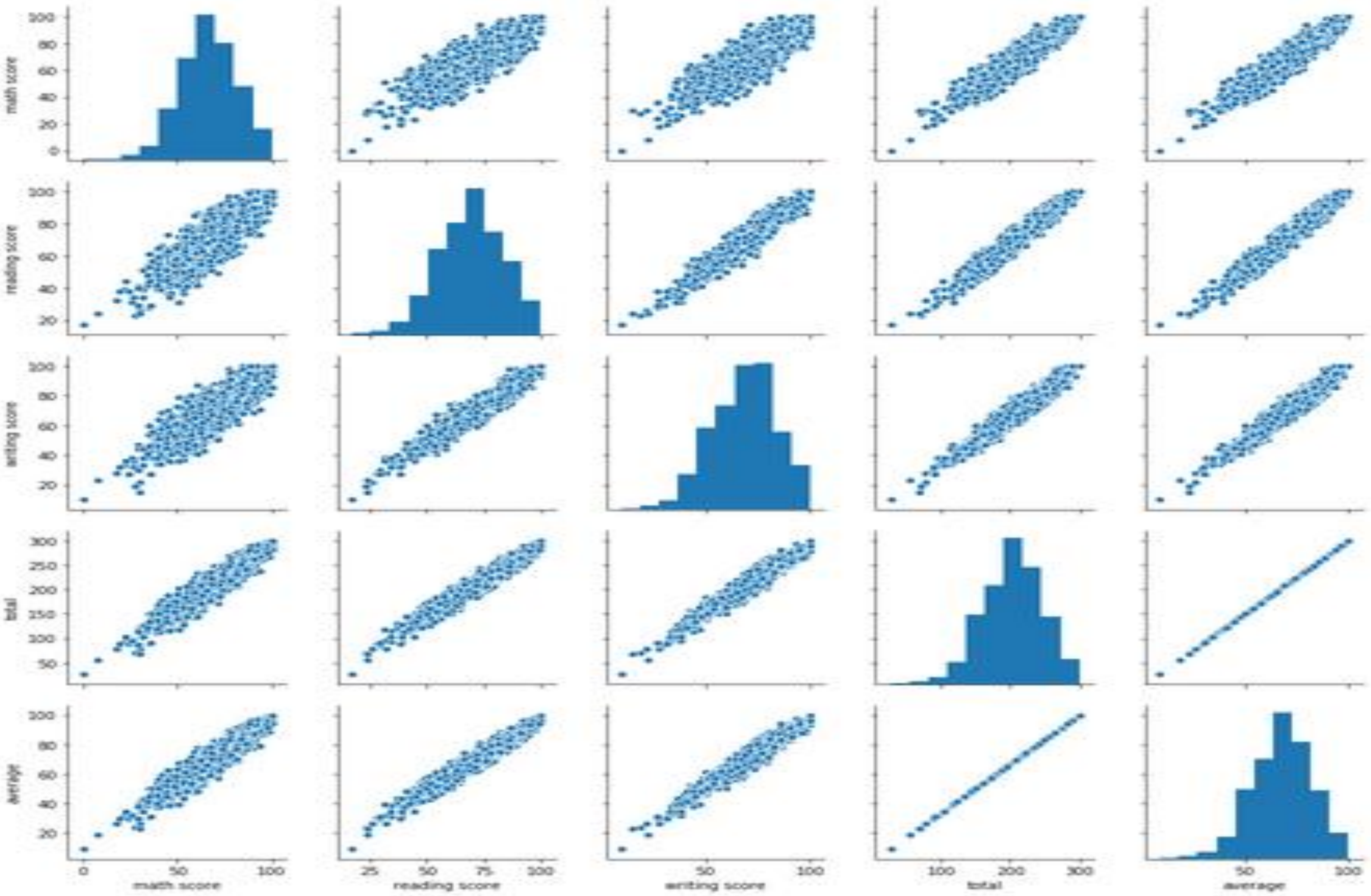
Results

Distribution of the scores



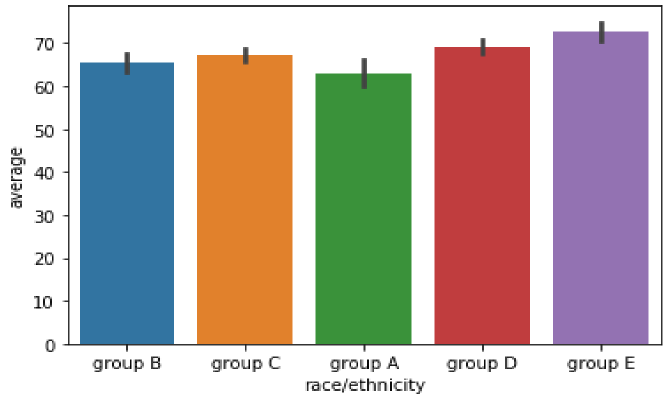
Results

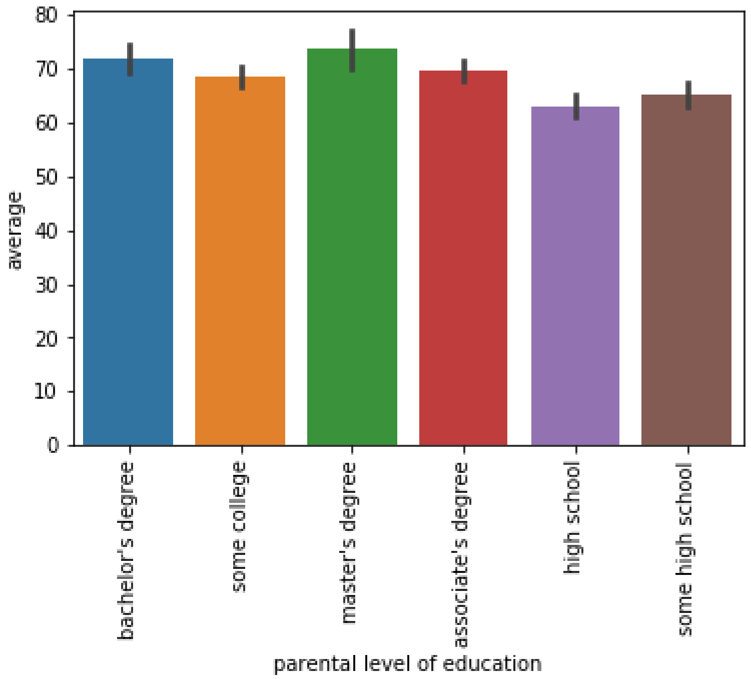
Pairs Plot of the scores

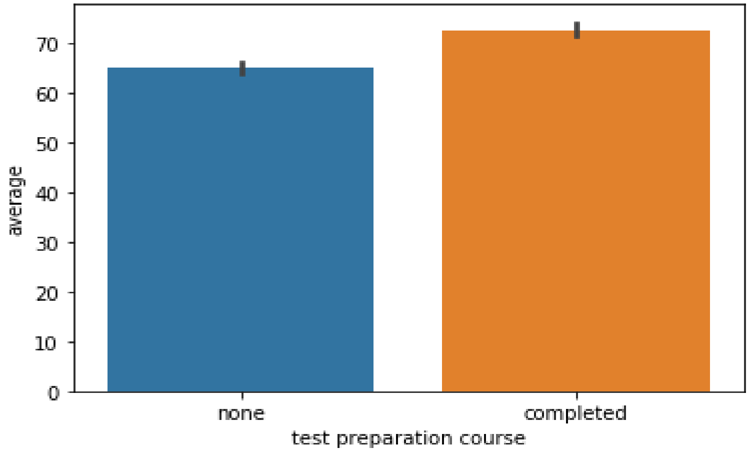


Results

Analyzing the average score of all the students on the basis of “race/ethnicity”, “parental level of education”, “test preparation course”.

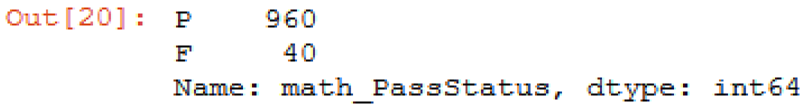




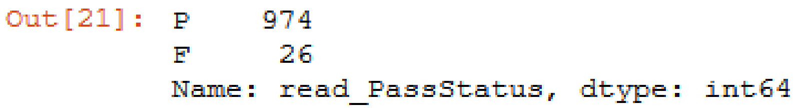


Results

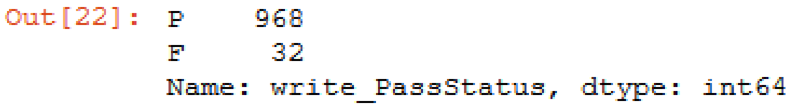
No of student who can pass or fail base on “math score”, “reading score” and “writing score” which is greater than or equal “40” and less than “40”



```
Out[20]: P     960
         F      40
         Name: math_PassStatus, dtype: int64
```



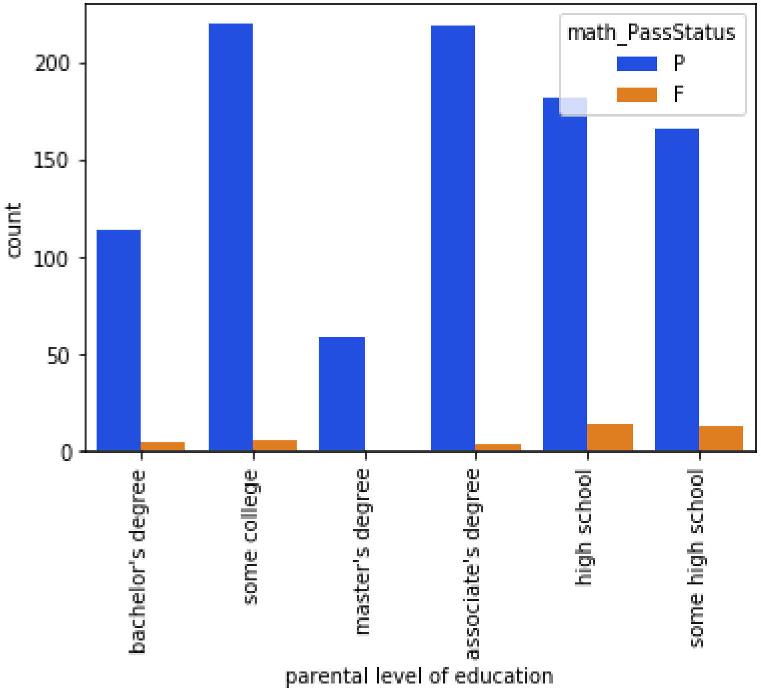
```
Out[21]: P     974
         F      26
         Name: read_PassStatus, dtype: int64
```

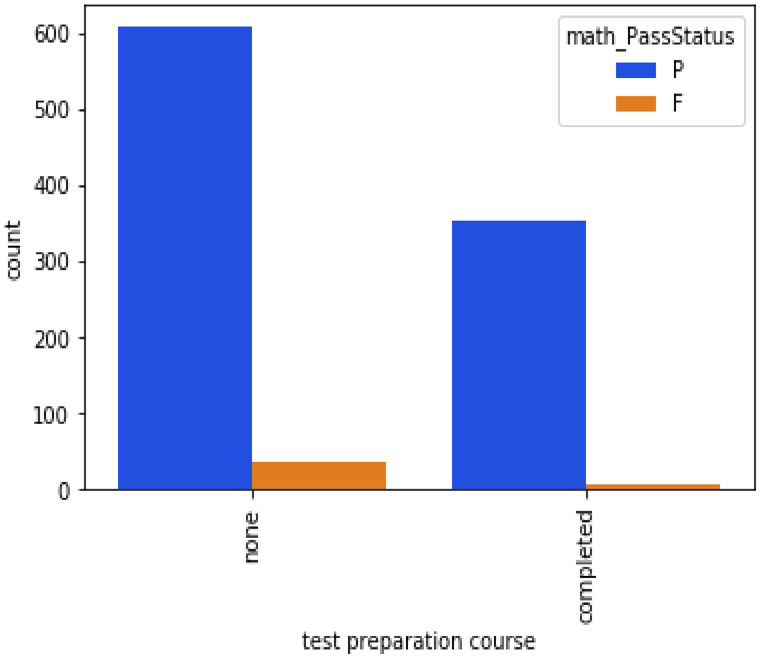


```
Out[22]: P     968
         F      32
         Name: write_PassStatus, dtype: int64
```

Results

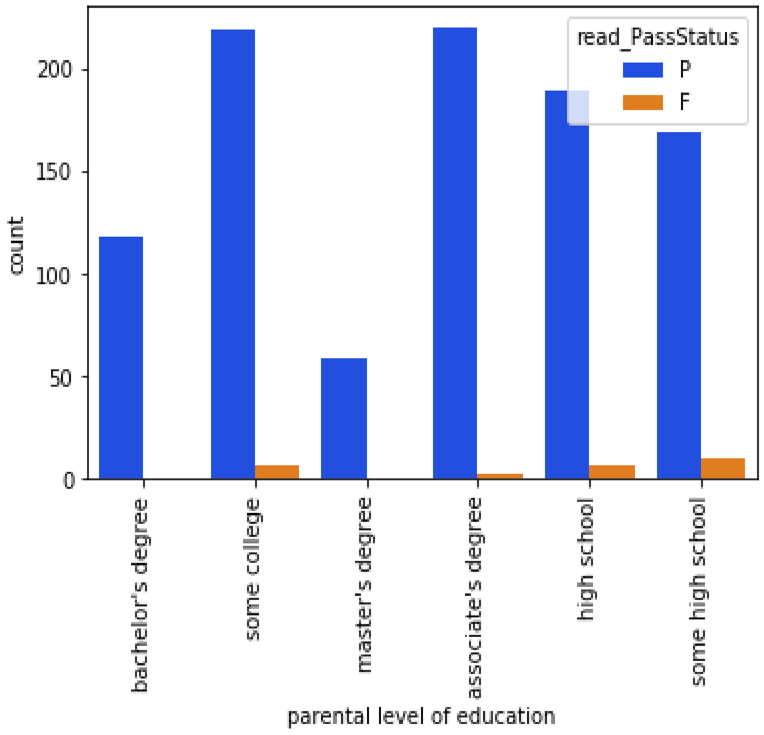
Analyzing the math_passStatus of all the students on the basis of "parental level of education" and "test preparation course".

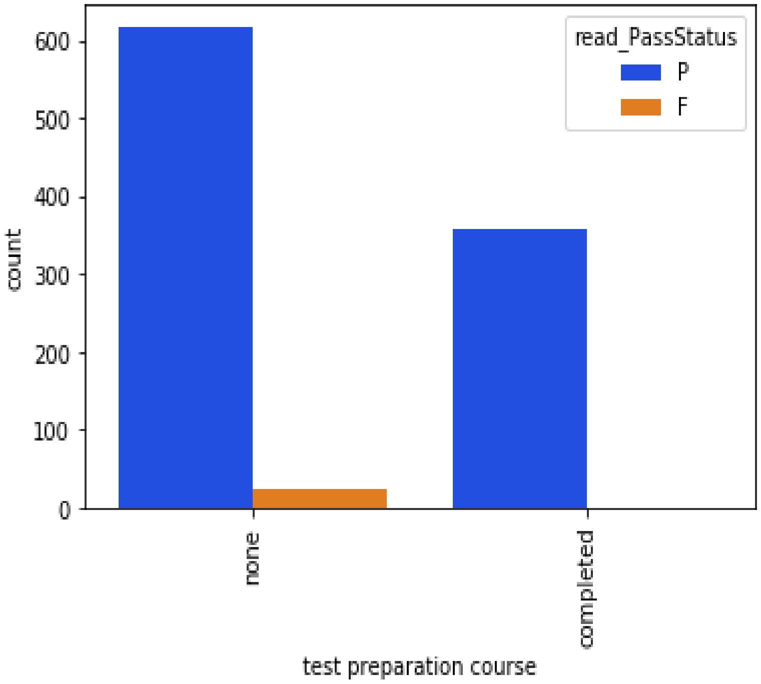




Results

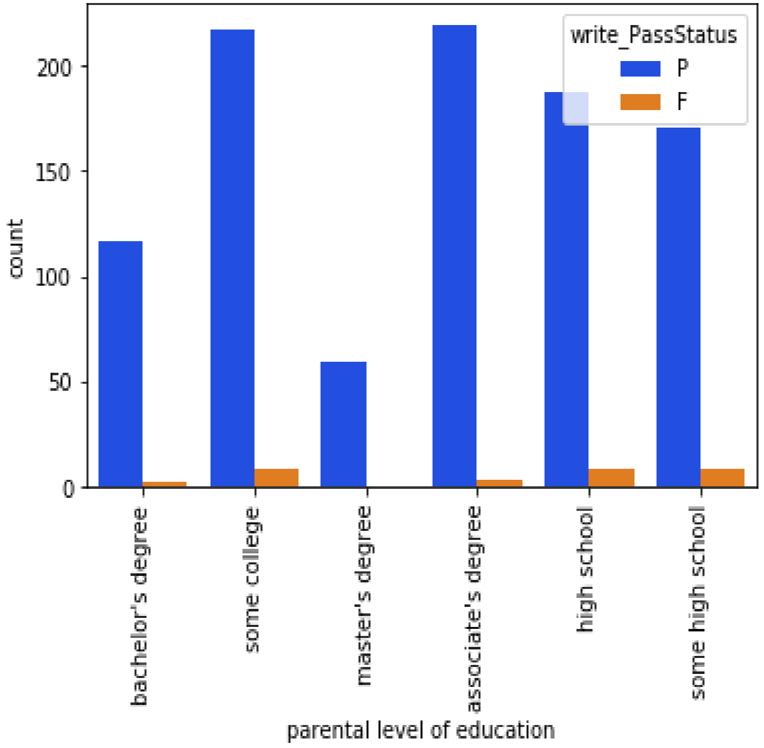
Analyzing the read_passStatus of all the students on the basis of "parental level of education" and "test preparation course".

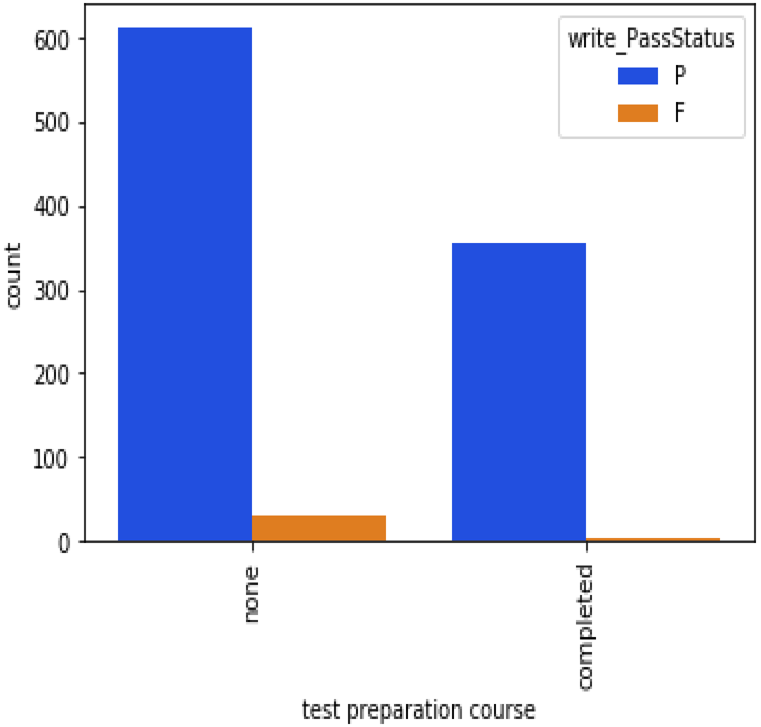




Results

Analyzing the write_passStatus of all the students on the basis of "parental level of education" and "test preparation course".





Related Work

- Data Mining is an emerging methodology used in educational field to enhance the understanding of learning process.
- The application of Data Mining is widely spread in higher education system.
- This paper predicted the performance of students using classification system with decision tree algorithm.
- Different clustering techniques and association rules mining can also use in this project.
- This study helps the teachers to reduce the failing ratio by taking appropriate steps at right time and improve the performance of students.

Conclusion

- In this project, we presented techniques to record student performance base on the United States.
- A classification model has been proposed in this study for predicting students' performance.
- The model obtained accuracy of the classification and it indicates that model is good/bad for forecasting the performance of students.
- Can know easily information of students rating who pass or fail and also predict looking their background (parental level of education and test preparation).
- In future, this project can be useful to support for educational institutions of our country.
- And it is to increase the analysis by using different clustering techniques.

References

- https://medium.com/israjan/students-performance-in-exams-data-analysis-19ca93fccd37
- https://www.datacamp.com/community/tutorials/decision-tree-classification-python
- https://www.kaggle.com/spscientist/students-performance-in-exams

Any Questions?

Thank You