# The Exploration of Modelling for The Student Achievement Predictor

Chengyang Gai

School of Automatic Control, Nanjing University of Information Science and Technology, Nanjing, China

* Corresponding Author Email: 201010013513@stu.swmu.edu.cn

**Abstract.** The project is rooted in the analysis of historical student data. It encompasses a comprehensive approach involving data observation, meticulous analysis, thorough comparison, systematic processing, and efficient coding techniques. The ultimate goal is to harness the power of machine learning to predict students' academic achievements and discern the key features that exert the most significant influence on their learning outcomes. In terms of the machine learning model, this study explored and assessed several machine learning algorithms, including but not limited to the K-Neighbors Classifier, Logistic Regression, and Decision Tree Classifier. These models are scrutinized and fine-tuned to ensure their suitability for the task at hand. Furthermore, a pivotal aspect of this project is identifying the features that wield the greatest impact on students' learning achievements. By employing feature selection techniques, this study aims to uncover the critical factors that can make a difference in educational outcomes. This information can guide educational institutions in designing targeted interventions to support student success. The experimental results obtained from this study demonstrated the effectiveness of the employed machine learning methods.

**Keywords:** Visual Analysis, Machine Learning, Student's Achievement Prediction.

## 1. Introduction

Based on the historical data set of students in a school [1], this paper constructs a student achievement prediction model through python language programming, viewing the data set, visualizing the features of the data set and the relationship between the features and student achievement, data cleaning and data coding, data preprocessing, machine learning model selection and comparison, and data mining process of adjusting and optimizing model parameters. To obtain the features that have a greater impact on student achievement to improve school education level, and to explore the supervised learning type of data mining and modeling ideas suitable for similar research objects.

## 2. Dataset description

Dataset [1] consists of 480 student records and 17 features. The student records in the dataset were collected from two semesters: 245 student records in the first semester and 235 student records in the second semester; The students in the data set were divided into three grades based on their overall grades: high, medium, and low. The fields in the dataset are described in Table 1.

**Table 1.** Description of 17 columns in the dataset [1]

| Features | Instructions |
|---|---|
| gender | Student gender: 'M' (Male), 'F' (Female) |
| NationalITy | Nationality of students: 'Egypt', 'Iran', 'Iraq', 'Jordan', 'KW', 'Lybia', 'Morocco', 'Palestine', 'SaudiArabia', 'Syria', 'Tunis', 'USA', 'lebanon', 'venzuela' |
| PlaceofBirth | Place of birth of students: 'Egypt', 'Iran', 'Iraq', 'Jordan', 'KuwaIT', 'Lybia', 'Morocco', 'Palestine', 'SaudiArabia', 'Syria', 'Tunis', 'USA', 'lebanon', 'venzuela' |
| StageID | Level of education for students: 'HighSchool', 'MiddleSchool', 'lowerlevel' |
| GradeID | Grade: 'G-02', 'G-04', 'G-05', 'G-06', 'G-07', 'G-08', 'G-09', 'G-10', 'G-11', 'G-12' |
| SectionID | Affiliated Classrooms: 'A', 'B', 'C' |
| Topic | Course names: 'Arabic', 'Biology', 'Chemistry', 'English', 'French', 'Geology', 'History', 'IT', 'Math', 'Quran', 'Science', 'Spanish' |
| Semester | Semester: 'F'(First), 'S'(Second) |
| Relation | Parent in charge of student: 'Father', 'Mum' |
| raisedhands | Number of hands raised by students in class: 0-100 |
| VisITedResources | Number of times students access online course content: 0-100 |
| AnnouncementsView | Number of times students check new announcements: 0-100 |
| Discussion | Number of times students participate in discussion groups: 0-100 |
| ParentAnsweringSurvey | Have parents answered the questionnaire provided by the school: 'Yes', 'No' |
| ParentschoolSatisfaction | Parent satisfaction with school: 'Good', 'Bad' |
| StudentAbsenceDays | Absencedays per student: 'Above-7', 'Under-7' |
| Class | A student's grade grade; Three grades based on the student's total grades: 'H'(90 to 100), 'M'(70 to 89), 'L'(0 to 69) |

## 3. Dataset viewing

Data viewing mainly understands the dimension of the data set, the number of rows, column labels, statistical analysis information, the existence and distribution of null values, duplicate values, outliers, etc. of the column values in the data set.

### 3.1. View the data set information

#### 3.1.1 View the overview information

In python, there are corresponding functions for different data structures to realize the query statistics of the data set, the code is implemented as follows: *stu_data.info ()*. The dataset information is shown in Figure 1.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 480 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   gender                    480 non-null    object
 1   NationalITy               480 non-null    object
 2   PlaceofBirth              480 non-null    object
 3   StageID                   480 non-null    object
 4   GradeID                   480 non-null    object
 5   SectionID                 480 non-null    object
 6   Topic                     480 non-null    object
 7   Semester                  480 non-null    object
 8   Relation                  480 non-null    object
 9   raisedhands               480 non-null    int64
 10  VisITedResources          480 non-null    int64
 11  AnnouncementsView         480 non-null    int64
 12  Discussion                480 non-null    int64
 13  ParentAnsweringSurvey     480 non-null    object
 14  ParentschoolSatisfaction  480 non-null    object
 15  StudentAbsenceDays        480 non-null    object
 16  Class                     480 non-null    object
dtypes: int64(4), object(13)
memory usage: 63.9+ KB
```

**Figure 1.** Data set information overview

### 3.1.2 Looking at the first N records raw information

Seeing if the contents of the raw data file are consistent with the data set after importing python. Look at the 5 records starting at the head of the data set and the code is implemented as follows: *stu_data.head(5).T* .

### 3.1.3 View statistics

Using this in python to view statistics for a dataset.

1. Numerical type column characteristics statistics, code implementation: *stu_data.describe()* .

2. Non-numeric type column feature statistics, code implementation: *stu_data.describe(include=[object]).T* .

### 3.1.4 Modify the column label names

Modifying the column label names in the data set.

## 3.2. View abnormal values

Looking at null values, duplicate values, whether outliers are present, and how they are distributed in the data set.

### 3.2.1 Check for null values

Python uses *stu_data.isnull().any()* to check for null values; *stu_data.isnull().T.any ()* checks for line nulls.

### 3.2.2 Check for duplicate values

Duplicate values are found as shown in Figure 2.

| | 326 | 327 |
|---|---|---|
| **Gender** | M | M |
| **Nationality** | Jordan | Jordan |
| **PlaceOfBirth** | Jordan | Jordan |
| **StageID** | lowerlevel | lowerlevel |
| **GradeID** | G-02 | G-02 |
| **SectionID** | A | A |
| **Topic** | French | French |
| **Semester** | F | S |
| **Relation** | Father | Father |
| **RaisedHands** | 10 | 30 |
| **VisitedResources** | 15 | 10 |
| **AnnouncementsView** | 10 | 20 |
| **Discussion** | 21 | 5 |
| **ParentAnsweringSurvey** | No | No |
| **ParentSchoolSatisfaction** | Bad | Bad |
| **StudentAbsenceDays** | Above-7 | Above-7 |
| **Class** | L | L |

**Figure 2.** Duplicate records for data sets

### 3.2.3 Check for outliers

Using Z-score or 0-1 standardization to identify outliers.

### (1) Z-score standardization

Z-score, also known as standard number [2], is used to obtain dimensionless standardized data and realize data comparison of different orders of magnitude.

Z-score is calculated by the formula [3]: $z = (X - \mu) / \sigma$

$X$ represents a single raw data value; $\mu$ represents the population mean; $\sigma$ represents the population standard deviation

Typically, this study considered observations with absolute values of Z-score greater than 3 to be outliers, while for some highly skewed data sets, observations with absolute values of Z-score greater than 2 May be outliers.

### (2) 0-1 standardization

0-1 standardization [2], also known as maximum-minimum standardization, is a process used to scale the original indicator into a range between 0 and 1.

The formula [3] is as follows: $x' = (x - min(x))/(max(x) - min(x))$ .

$x$ is the original value, *min(x)* is the minimum value, *max(x)* is the maximum value, and x' is the normalized value.

## 4. Visual analysis of column characteristics

### 4.1. Visual analysis of frequency chart

Frequency plots use bars to display the observed counts in each box divider to compare counting differences between classes. The following frequency function graph is used to show and analyze the relationships among *ParentAnsweringSurvey*, *Topic*, *Relation*, *gender*, *SectionID*, *StudentAbsenceDays* and *Class*.

### 4.1.1 Correlation analysis

Parents answered whether the survey and the correlation analysis diagram is provided in Figure 3.

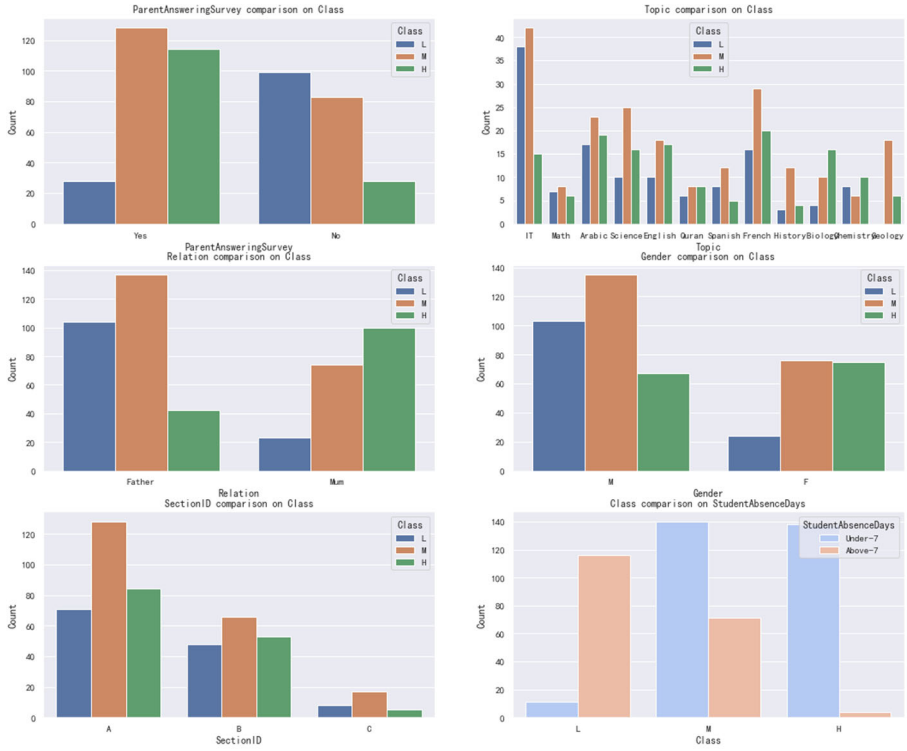


**Figure 3.** Parents answered whether the survey and the correlation analysis diagram of grades were provided by the school.

### 4.1.2 Analysis

'*ParentAnsweringSurvey* comparison on *Class'* in the chart

- Parents of students with average and excellent grades are more willing to participate in school activities, while parents who do not participate in school activities are mostly parents of students with low grades. There was a strong correlation between the two.

'*Topic* comparison on *Class*' in the graph

- There are 12 subjects in total, and only geology has all students achieving average or above.
- The IT course has the largest number of students, and the number of excellent students is relatively small, and the number of students who can master this course is very small.
- In the '*Relation* comparison on *Class*' graph
- Students whose mothers were in charge of the class had higher grades, lower grades and less grades, while those whose fathers were in charge had the opposite.
- Relation column features have a strong relation with Class.

'*gender* comparison on *Class'* in the graph

- The number of boys who got average grades was the largest, followed by the number of boys who failed, and the number of boys who got excellent grades was relatively small. While most of the girls got medium or above, only a small number of them failed, and the overall level was good.

In the '*SectionID* comparison on *Class*' chart

- The distribution of classes and grades conforms to the objective law, and each class has the largest number of people with average grades.

'*Class* comparison on *StudentAbsenceDays*'

- Most of the students at the top of the grade had very low absencedays, and most of the students at the middle of the grade had very low absencedays, so the better the grade, the less absencedays.
- *StudentAbsenceDays* have a strong correlation with *Class*.

### 4.2. Visual analysis of clustering scatter plot

The cluster scatter diagram [4] is a two-dimensional coordinate system classification scatter diagram with non-overlapping data points. It can be used to find the relationships and trends between VisitedResources, RaisedHands, AnnouncementsView, Discussion and Class, respectively. The corresponding results are shown in Figure 4.

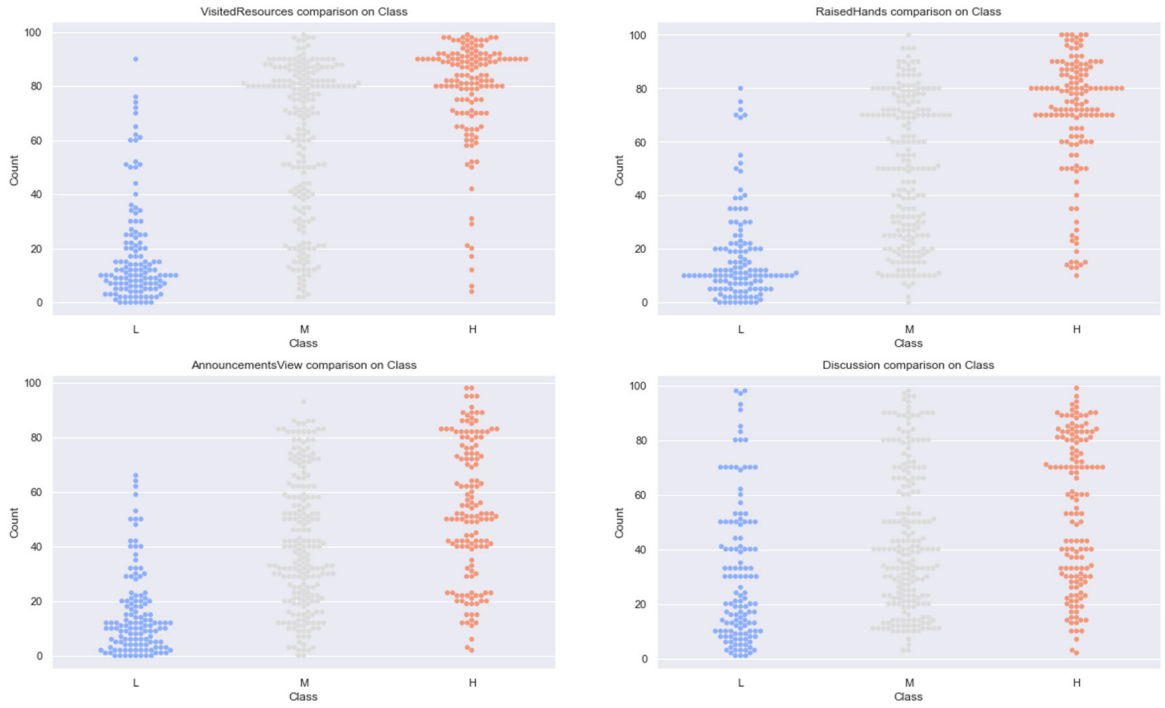


**Figure 4.** *VisitedResources*, *RaisedHands*, *AnnouncementsView*, *Discussion* and *Class* analysis diagram

#### 4.2.1 Analysis

The "*VisitedResources* comparison on *Class*" chart shows that students with higher grades visit online course content more often, while most students with lower grades visit online course content less often. This trend reflects a strong correlation between the two.

The graph of "*RaisedHands* comparison on *Class*" shows that students with higher grades have higher participation in class and raise their hands more often; The students with average grades raise their hands more often than the students with low grades, while the students with low grades overwhelmingly raise their hands less, which reflects the positive and strong correlation between the two variables.

The graph "*AnnouncementsView* comparison on *Class*" showed that most students with low achievement grades had few students check new announcements; The number of students with high and intermediate grades checked more frequently than the number of students with low grades.

The chart of "*Discussion* comparison on *Class*" shows that students with higher grades participate in discussion groups more often, while most students with lower grades participate in discussion groups less frequently. This trend reflects a strong correlation between the two.

## 5. Data preprocessing

After the preliminary review and analysis process, the data set selected in this study is relatively perfect, but in order to make the data set meet the requirements of the model, data preprocessing is required. Data preprocessing methods include data cleaning, conversion, specification, aggregation, sampling and so on.

### 5.1. Data cleaning

Data cleaning is the process of discarding, filling, replacing and removing the missing values, outliers and duplicate values in the data set to remove anomalies, make up for missing values, and reduce the interference of data "impurities" on the model.

#### 5.1.1 Missing value processing

In the process of missing value processing, it is necessary to first carry out the basic analysis of the missing value in the data set, and then process it.

#### 5.1.2 Handling of outliers

Use the Z-score standardized method to determine and view outliers shown in Figure 5.

Displays the outlier record:

| | RaisedHands | VisitedResources | AnnouncementsView | Discussion |
|---|---|---|---|---|
| 19 | 0.755355 | -0.145191 | 0.078291 | 2.018067 |
| 135 | 0.755355 | 0.762647 | 2.147243 | 0.967682 |
| 136 | 0.755355 | 0.762647 | 2.147243 | 0.967682 |
| 138 | 1.731055 | 0.762647 | 2.147243 | 1.692085 |
| 152 | 1.731055 | 1.095521 | 2.260095 | -0.118923 |
| 163 | 0.430121 | 0.823170 | 2.072008 | -0.010262 |
| 240 | 0.820401 | 0.762647 | 2.260095 | 1.909406 |

**Figure 5.** Normalize the values with a Z-score and look at outliers.

The "*Discussion*" recorded in line 19 has values greater than the threshold of 2; "*AnnouncementsView*" recorded in lines 135, 136, 138, 152, 163 and 240 has a value greater than threshold 2.

These values are all "false anomalies", normal states in which the student is particularly active in learning.

#### 5.1.3 Duplicate value processing

Typically, duplicate values in a data set come in two forms: multiple records of exactly the same data value; The body of the record is the same, but a single or small number of characteristic values differ. Remove duplicate values directly, code example: *stu_data.drop_duplicates()* .

### 5.2. Data encoding classification

The algorithm and model studied in this paper are built on the basis of mathematics and numbers and cannot handle categorical variables. Therefore, 13 classification features need to be encoded in the data set adopted in this study, and there are two commonly used ones: One-Hot Encoding and Label Encoding.

#### 5.2.1 One-Hot Encoding processing

Single-hot encoding is implemented by the python code *get_dummies()* [5]. .

Analysis: After the unique thermal coding processing, the 11 classification features of the data set were converted into a data set with 65 binary variables.

#### 5.2.2 Label Encoding processing

Label Encoding is the process of assigning a uniquely identified integer to different class values within a class data variable.

### 5.3. Separate target data

Student grades in the data set are the target of the forecast, so student grades need to be separated from the data set.

### 5.4. Training set and test set processing

Using data from the same data set to train and test the model can improve the training accuracy of the model. Therefore, the data is split into the training set and the test set according to the required distribution ratio.

## 6. Train and test model

Using the pre-processed data to train and test the algorithm model and select the model with the best prediction effect according to the training results.

### 6.1. K-nearest neighbor, logistic regression, linear support vector machine and support vector machine model

Compare test K-nearest neighbor, logistic regression, linear support vector machine and support vector machine model [6-10] and output a comparison table of classification accuracy of the test set.

| | model | accuracy score |
|---|---|---|
| 0 | K-Nearest Neighbor Classifier | 0.683333 |
| 1 | Logistic Regression | 0.683333 |
| 2 | Linear Support Vector Classification | 0.591667 |
| 3 | Support Vector Classification | 0.616667 |

**Figure 6.** Nearest neighbor, logistic regression, linear support vector machine and support vector machine model training results

As can be seen from the above Figure 6, K-nearest neighbor and logistic regression model have better prediction effect.

### 6.2. Decision tree and three kinds (Gauss, Bernoulli, polynomial) Bayesian models

Train and test decision tree and three kinds (Gauss, Bernoulli, polynomial) Bayesian models, and output training and testing accuracy table, select the appropriate model.

| | model | accuracy score |
|---|---|---|
| 0 | DecisionTreeClassifier | 0.641667 |
| 1 | GaussianNB | 0.675000 |
| 2 | BernoulliNB | 0.658333 |
| 3 | MultinomialNB | 0.633333 |

**Figure 7.** Decision tree and three kinds of (Gauss, Bernoulli, polynomial) Bayesian model training results

As can be seen from the above Figure 7, decision tree model and Gauss Bayes model has suitable and better effect.

## 7. Model parameter tuning

Through the comparison of two modeling effects, three suitable models with good effects are selected: logistic regression, K-nearest neighbor, decision tree; Then the parameters were adjusted to further improve the prediction effect of the model.

### 7.1. Logistic regression

*C*, *penalty* and *class_weight* parameters of model are combined to test the model accuracy and output the results; then selected the most appropriate combination of parameters.

| | parameter_list | accuracy |
|---|---|---|
| **0** | (1.6, l1, balanced) | 0.766667 |
| **1** | (1.4, l1, balanced) | 0.766667 |
| **2** | (0.6, l1, None) | 0.766667 |
| **3** | (0.8, l1, None) | 0.766667 |
| **4** | (1.2, l1, balanced) | 0.766667 |
| **5** | (1.8, l2, balanced) | 0.764583 |
| **6** | (1.8, l1, balanced) | 0.764583 |
| **7** | (0.8, l1, balanced) | 0.764583 |
| **8** | (1, l1, balanced) | 0.762500 |
| **9** | (1.6, l2, balanced) | 0.762500 |
| **10** | (1.4, l2, balanced) | 0.762500 |

**Figure 8.** Verification of logistic regression model parameter tuning

When l1 regularization term is adopted, regularization intensity is 1.2, and sample classification weights are balanced, the model effect is improved most obviously in Figure 8.

### 7.2. K-nearest neighbor

K-nearest neighbor model has two key parameters, *K* and *weights*. Different combinations of two parameters are used to verify the precision of the model, and output the results; then select the most accurate and suitable combination of parameters.

| | parameter_list | accuracy |
|---|---|---|
| **0** | (1, uniform) | 1.000000 |
| **1** | (1, distance) | 1.000000 |
| **2** | (2, distance) | 1.000000 |
| **3** | (3, distance) | 1.000000 |
| **4** | (4, distance) | 1.000000 |
| **5** | (5, distance) | 1.000000 |
| **6** | (6, distance) | 1.000000 |
| **7** | (7, distance) | 1.000000 |
| **8** | (8, distance) | 1.000000 |
| **9** | (2, uniform) | 0.808333 |
| **10** | (3, uniform) | 0.804167 |

**Figure 9.** K-Nearest neighbor model parameter tuning verification.

It can be seen that when the number of neighbor samples is 5 in Figure 9, the effect of K-nearest neighbor and weighted k-nearest neighbor model is the same, and the same as before.

### 7.3. Decision Tree

Decision tree model is used to verify the output of the model under different *criterion*, *max_depth* and *class_weight* parameters assignment, and according to the output results, an optimal parameter combination value is selected in the parameter combination.

| | parameter_list | accuracy |
|---|---|---|
| 0 | (gini, 5, balanced) | 0.733333 |
| 1 | (gini, 4, balanced) | 0.733333 |
| 2 | (entropy, None, None) | 0.725000 |
| 3 | (gini, 6, None) | 0.716667 |
| 4 | (gini, 5, None) | 0.716667 |
| 5 | (entropy, 6, balanced) | 0.708333 |
| 6 | (gini, 3, balanced) | 0.708333 |
| 7 | (entropy, 8, None) | 0.700000 |

**Figure 10.** Verification of decision tree model parameter tuning

When Gini coefficient is used to calculate the impurity, the maximum tree depth is set to 5 and the sample classification weight is balanced, the model effect is improved most obviously, and it is also the best model at present, with the correct rate of 73% shown in Figure 10.

## 8. Feature importance ranking of decision tree model

```
StudentAbsenceDays          0.371006
VisitedResources            0.209106
RaisedHands                 0.105386
Relation                    0.084998
AnnouncementsView           0.079035
ParentAnsweringSurvey       0.044179
ParentSchoolSatisfaction    0.023712
Discussion                  0.023688
GradeID                     0.022171
Topic                       0.021788
PlaceOfBirth                0.008952
SectionID                   0.005980
Gender                      0.000000
Nationality                 0.000000
StageID                     0.000000
Semester                    0.000000
dtype: float64
```

**Figure 11.** Weight of the influence of features on the model

Based on Figure 11, It can be seen that *StudentAbsenceDays*, *VisitedResources*, *RaisedHands, Relation* and *AnnouncementsView* have the most weight and are the most important.

So far, through exploratory data analysis, data preprocessing, model selection and parameter tuning, this study finally built a decision tree model to predict students' grades.

## 9. Conclusion

This study finds that there are many factors affecting students' achievement, but the most influential features are mainly concentrated in two categories. First, students' active learning ability, such as *StudengtAbsenceDays*, *VisitedResources*, *RaisdHands* and *AnnouncementsView*; Second, the

influence of family, such as *Relation* and *ParentAnsweringSurvey*. This situation is different from the traditional "cognitive hypothesis" that middle school students' nationality, birthplace, gender and other characteristics have a great impact on student achievement. The result can be explained from the above visualization analysis of column features and the importance ranking of features in the decision tree model.

Finally, the results of this study are consistent with reality, and the model constructed can accurately predict students' achievement and influencing factors and provide reference and direction for schools to improve students' achievement. However, the data set used in this study did not consider the facilities and conditions of the school itself, such as the professional level of teachers, the knowledge reserve of the school, the equipment of teaching facilities, etc. Further research and analysis can be done in these aspects in the future.

## References

[1] Amrieh E A, Thair H, and Ibrahim A. Mining educational data to predict student's academic performance using ensemble methods. International journal of database theory and application 9.8, 2016, 119-136.

[2] Ali P J M, Faraj R H, Koya E, et al. Data normalization and standardization: a technical report. Mach Learn Tech Rep, 2014, 1(1): 1-6.

[3] Gareth J et al. An Introduction to Statistical Learning with Applications in Python,First Printing:July 5 , 2023

[4] Wang Y, Han F, Zhu L, et al. Line graph or scatter plot? automatic selection of methods for visualizing trends in time series. IEEE transactions on visualization and computer graphics, 2017, 24(2): 1141-1154.

[5] Duan X X, Deep Migration Out of Python Machine Learning, Tsinghua University Press, 2018 ISBN 978-7-302-50323-1,69&135ex=14

[6] Othman M F B, Abdullah N B, Kamal N F B. MRI brain classification using support vector machine, 2011 fourth international conference on modeling, simulation and applied optimization. IEEE, 2011: 1-4.

[7] Qiu Y, Chen P, Lin Z, et al. Clustering Analysis for Silent Telecom Customers Based on K-means++, 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2020, 1: 1023-1027.

[8] Bansal M, Goyal A, Choudhary A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short-term memory algorithms in machine learning. Decision Analytics Journal, 2022, 3: 100071.

[9] Delen D, Kuzey C, Uyar A. Measuring firm performance using financial ratios: A decision tree approach. Expert systems with applications, 2013, 40(10): 3970-3983.

[10] Kim S Y, Upneja A. Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Economic Modelling, 2014, 36: 354-362.