Data Science Term Project

Final Report



**Team 13**

소프트웨어학과 201735856 윤혜주

소프트웨어학과 201735881 조소영

소프트웨어학과 201835505 이하영

목차

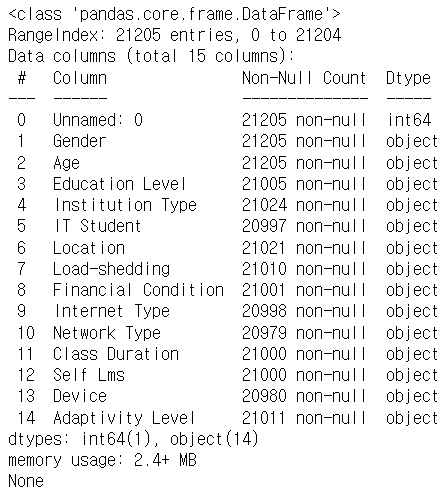
1. **Objective3**
2. **Data description3**
   1. **Data Source3**
   2. **Original Dataset Inspection3**
   3. **Handle ‘Unnamed 0’ columns4**
3. **Data preprocessing6**
   1. **Handle ‘Missing data’6**
   2. **Handle categorical data7**
   3. **Feature selection9**
   4. **Normalization10**
4. **Algorithms11**
   1. **Linear regression11**
   2. **Decision Tree classification11**
   3. **KNN classification13**
5. **Evaluation15**
   1. **K-Fold cross validation15**
   2. **Ensemble learning – Bagging15**
6. **GUI15**
7. **Role16**
8. **Conclusion16**
9. **Reference 17**
10. **Objective**

코로나 상황과 통신 기술의 발달로 온라인, 비대면 학습이 급속도로 발전하였다. 우리는 이 데이터를 분석함으로써 학부모들로 하여금 그들의 아이들에게 효과적인 교육방법을 고를 수 있게 하고 교육 contents 제공자들이 타겟을 잘 설정할 수 있도록 돕는 목표를 설정하였다.

1. **Data description**
   1. **Data Source**

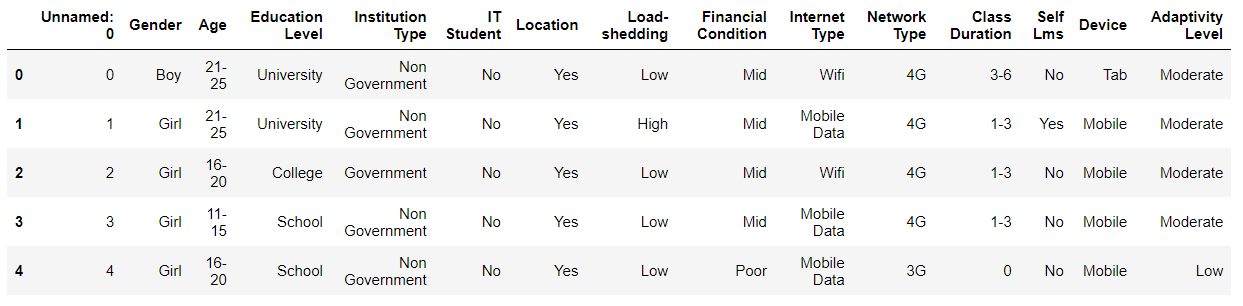
We used 'Students Adaptability Level in Online Education' dataset registered in Kaggle.

The data contains information about student who adapt in online education, for example education level, class duration, age.

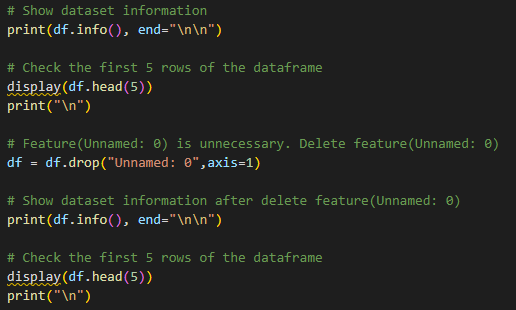
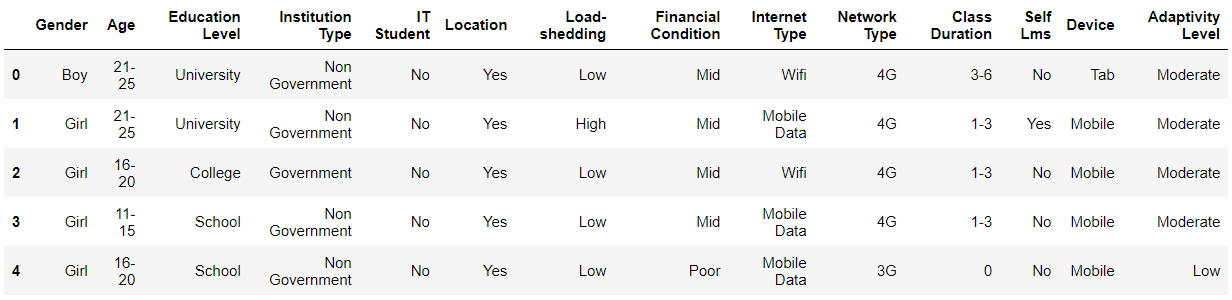
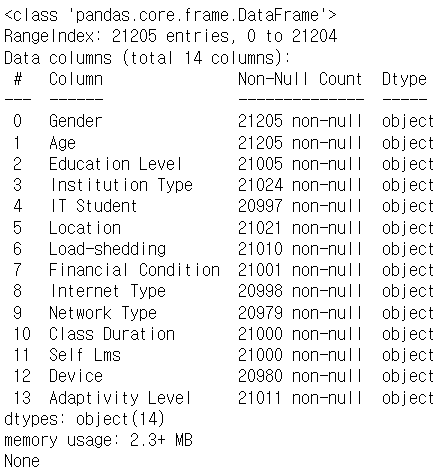
* 1. **Original Dataset Inspection**

**size**: 15 columns X 21205 row

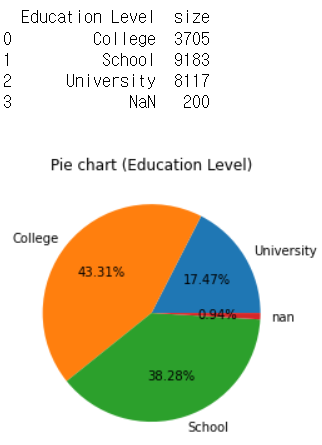
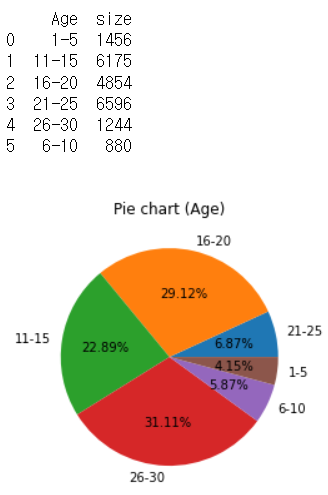
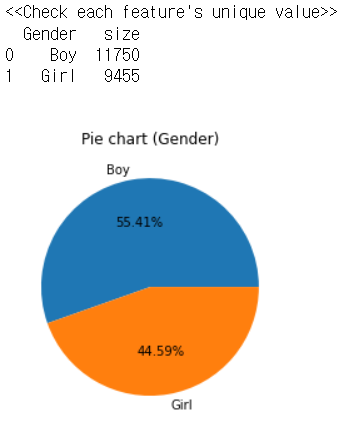
**columns** = [ ‘ ‘, 'Gender', 'Age', 'Education Level', 'Institution Type', 'IT Student', 'Location', 'Load-shedding', 'Financial Condition', 'Internet Type', 'Network Type', 'Class Duration', 'Self Lms', 'Device', 'Adaptivity Level' ]

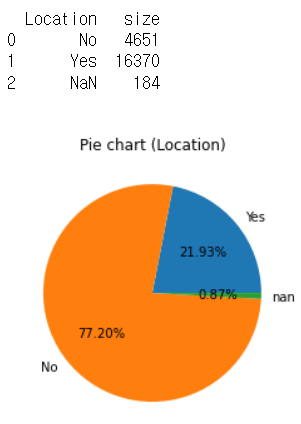
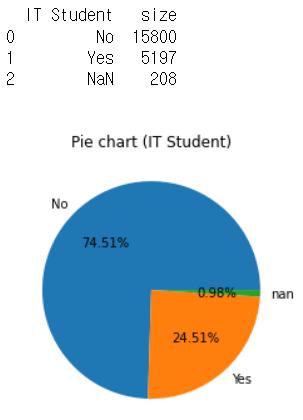
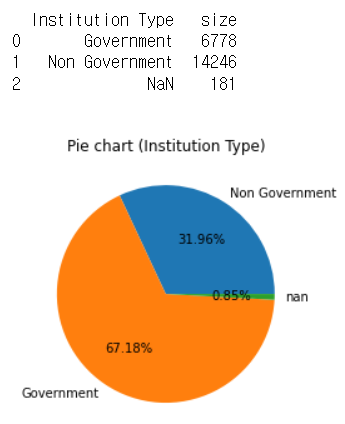


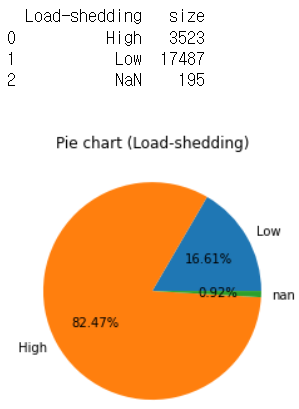
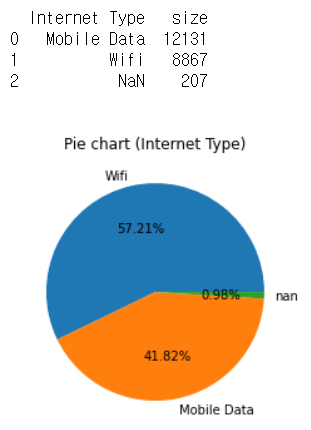
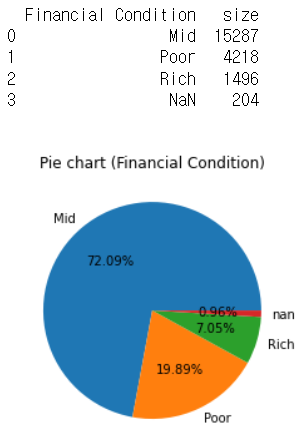
* 1. **Handle ‘Unnamed 0’ columns**

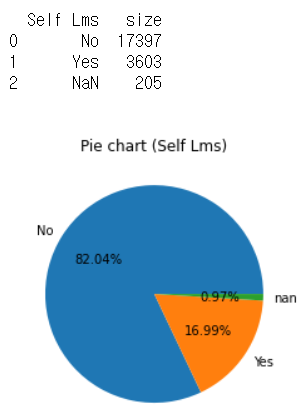
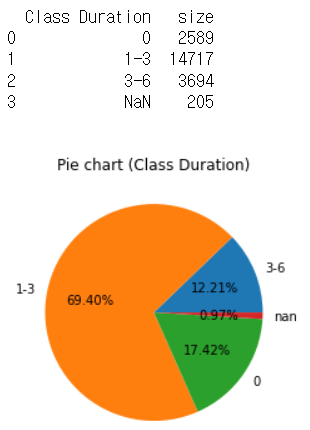
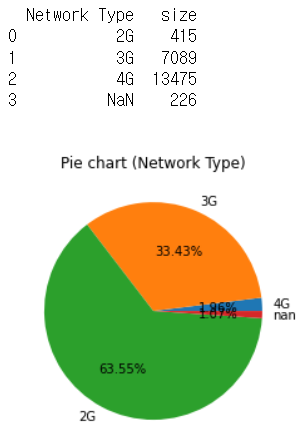
There are ‘Unnamed 0’ columns that just indexing dataset, so we drop the column.  

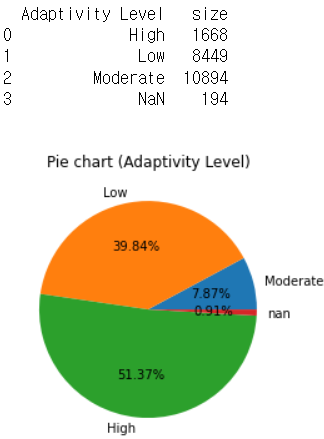
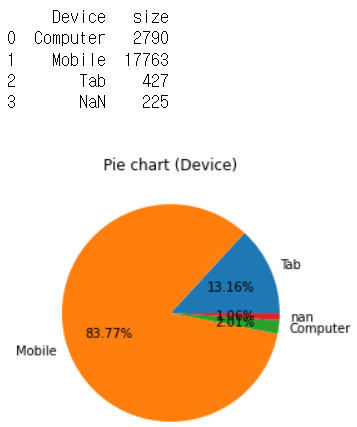
<data visualization>





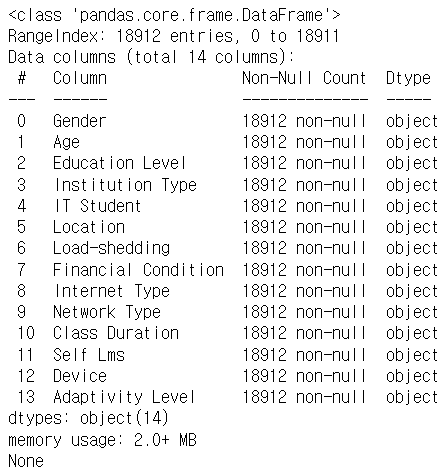
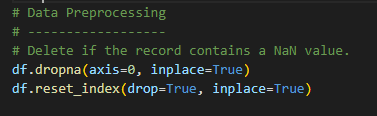
  





1. **Data preprocessing**
   1. **Handle Missing data**

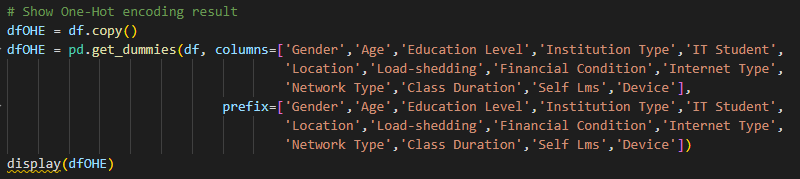
We can find there were less than 1 percent of ‘Missing data’ in several columns. Those percentage is pretty low, so we consider to delete the row which has ‘Missing data’.

* 1. **Handle categorical data**

After processing ‘Missing data’ we have to handle categorical data because machine learning cannot read categorical data.

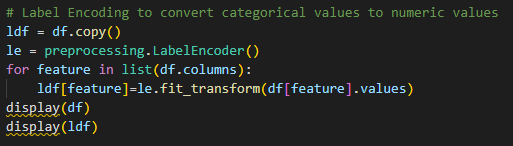
First we try to encode categorical data by ‘One Hot Encoder’. In this case dataset has too many features(36 features) so we don’t use his encoder.

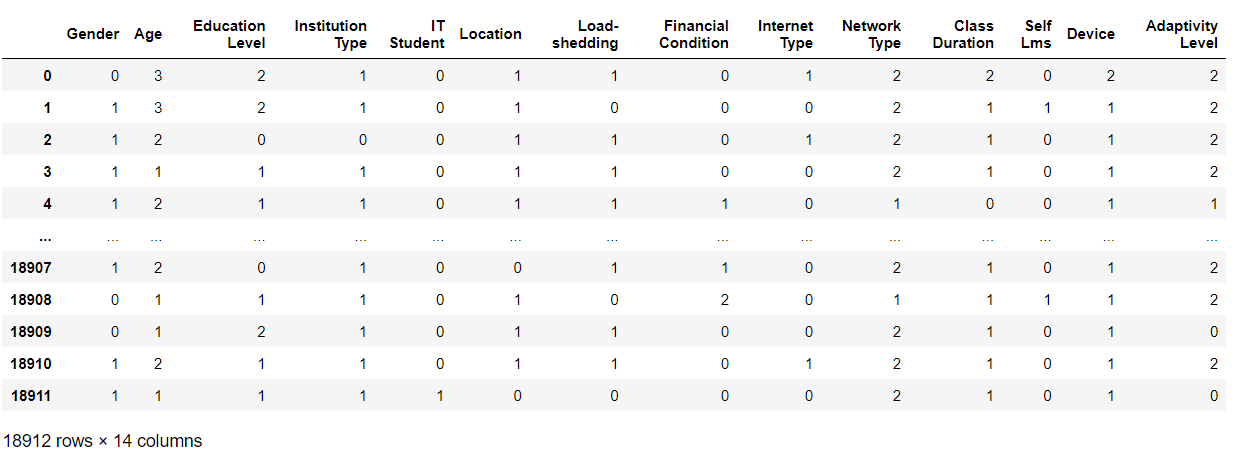






Then, we use ‘Label Encoder’.





* 1. **Feature selection**

Lastly, we have to choose independent features to predict target value. In feature selection we try 2 selection tools, SelectKBest, Extra Tree Classifier. And we choose top 6 features in each tools

* 1. **Normalization**

After processing ‘Categorical data’ we have to Normalized data. We learn 3 kinds of scaler(standard, robust, minmax). We use all those scaler.

First, Standard Scaler

Second, MinMax Scaler

Last, Robust Scaler

1. **Algorithms**
   1. **Linear regression**

We use linear regression

* 1. **Decision Tree Classifier**

We recognize that 'Students Adaptability Level in Online Education' dataset is the data that target value is ‘Categorical data’. So we decided to use ‘Decision Tree classifier’.

* 1. **KNN Classifier**

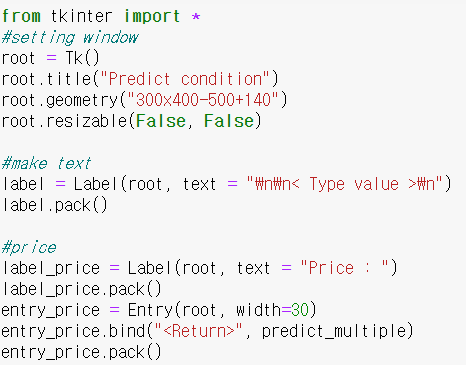
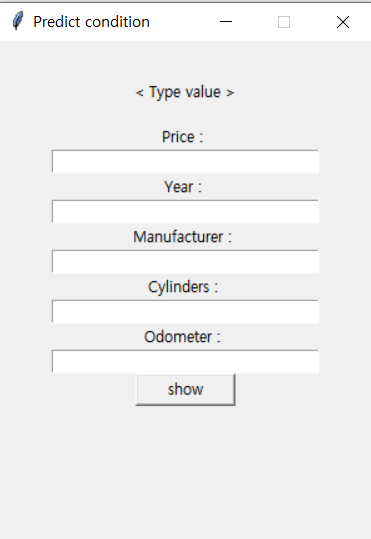
Accuracy of ‘Decision Tree Classifier’ is only about 50%, we decided to use another classifier that ‘KNN Classifier’.

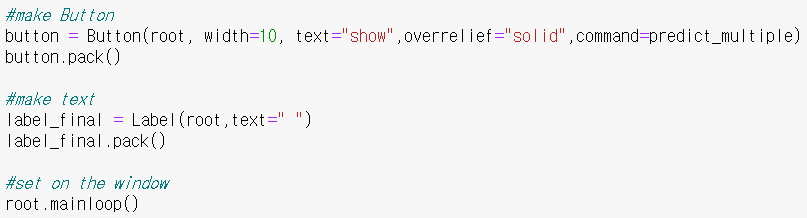
// 여기까지

1. **Evaluation**
   1. **K-Fold cross validation**
   2. **Ensemble learning – Bagging**
2. **GUI**

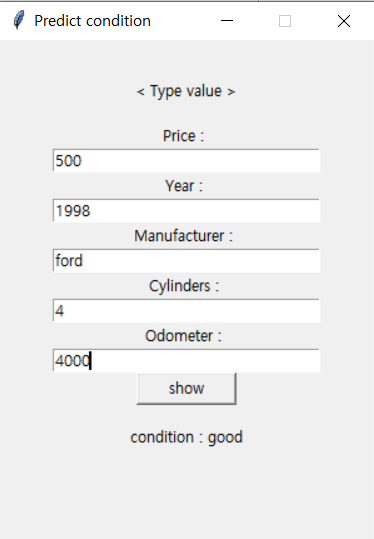
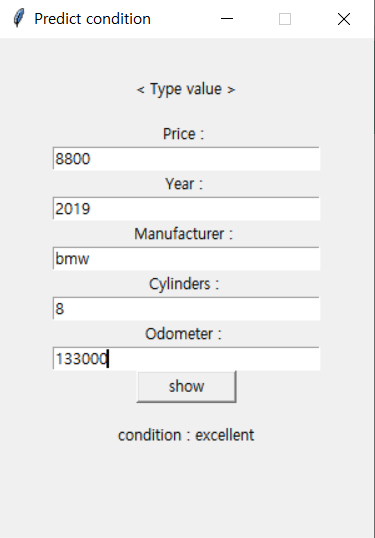
We created a GUI based on the code that we implemented, so that when the user enters it according to the column, we can predict the condition accordingly.

First, set the size and tickle of the window and insert the text, input window, and button to go inside.



Enter a value in the input window entry via the predict\_multiple function and click button to convert it to labelEncoder if it is not a number and show the predicted condition value through the ensemble in gui.

1. **Member role**

|  |  |  |
| --- | --- | --- |
| 윤혜주 | 조소영 | 이하영 |
| Data preprocessing | Data preprocessing | Data preprocessing |
| Data analysis (KNN) | Evaluation | Data analysis  (multiple regression) |
| Final presentation | Ensemble | GUI |
| Final presentation | Proposal PPT | Final PPT |

1. **Conclusion**

Significant difference in prediction accuracy depending on data preprocessing, and realized that preprocessing is very important.

In dealing with quite a lot of data, we experienced an unexpected error, the more systematic and meticulous data pre-processing is necessary.

it is necessary to apply appropriate algorithms according to the data, and all processes are important.

1. **Reference**