Predict Students' Dropout and Academic Success using Classification Models

LECTURER - ASST. PROF. DR. PRAPAPORN TECHAANGKOON

Linn Htet Aung

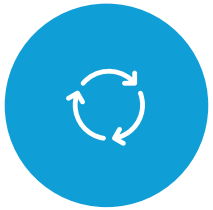
671615515

Linnhtetaung_1@cmu.ac.th

Motivation

COUNTRY GDP IS HIGHLY INFLUENCED BY THE LABOR FORCE

ESPECIALLY SKILLED LABOR FORCE



COUNTRY ECONOMIC CYCLE IS CAN BE DESCRIBED BY AD-AS DIAGRAM.

AGGREGATE DEMAND AND AGGREGATE SUPPLY

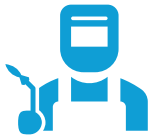
AGGREGATE SUPPLY IS THE TOTAL AMOUNT OF GOODS (INCLUDING SERVICES) SUPPLIED BY BUSINESSES WITHIN A COUNTRY AT A GIVEN PRICE LEVEL. (HTTPS://WWW.ECONLIB.ORG/)

LAND , LABOR, CAPITAL, TECHNOLOGY

Motivation

Changes in Education and Skills

Impact: A more educated and skilled workforce can produce more output and adapt to new technologies, enhancing long-run AS.



Examples:

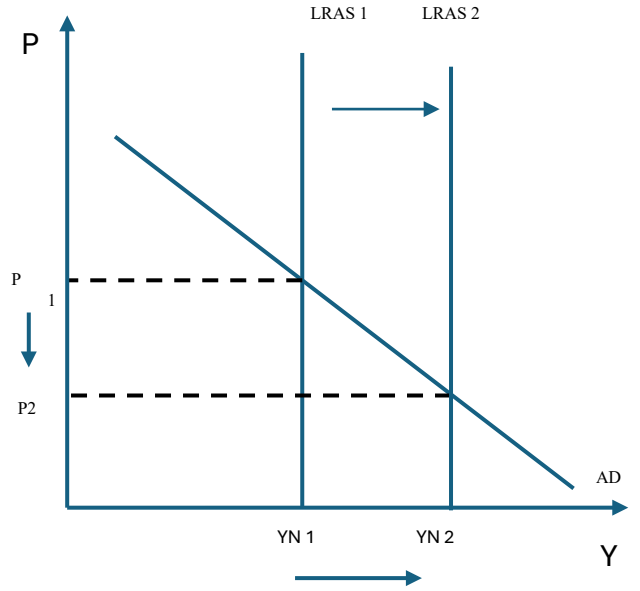
Investments in STEM (Science, Technology, Engineering, and Mathematics) education leading to a more innovative workforce in countries like South Korea.

Vocational training programs in Germany that equip workers with specialized skills, increasing productivity.

University and colleges are places where skilled labor are produced.

How?

When skilled labor increases in the labor market, output in the market will rise, pushing the long-run aggregate supply (LRAS) to the right. This causes the long-term output level to increase, leading to higher GDP and a fall in the price level, which is a positive sign for the economy.



Dataset

Predict Students' Dropout and Academic Success

Donated on 12/12/2021

A dataset created from a higher education institution (acquired from several disjoint databases) related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and...

Dataset Characteristics
Tabular

Subject Area
Social Science

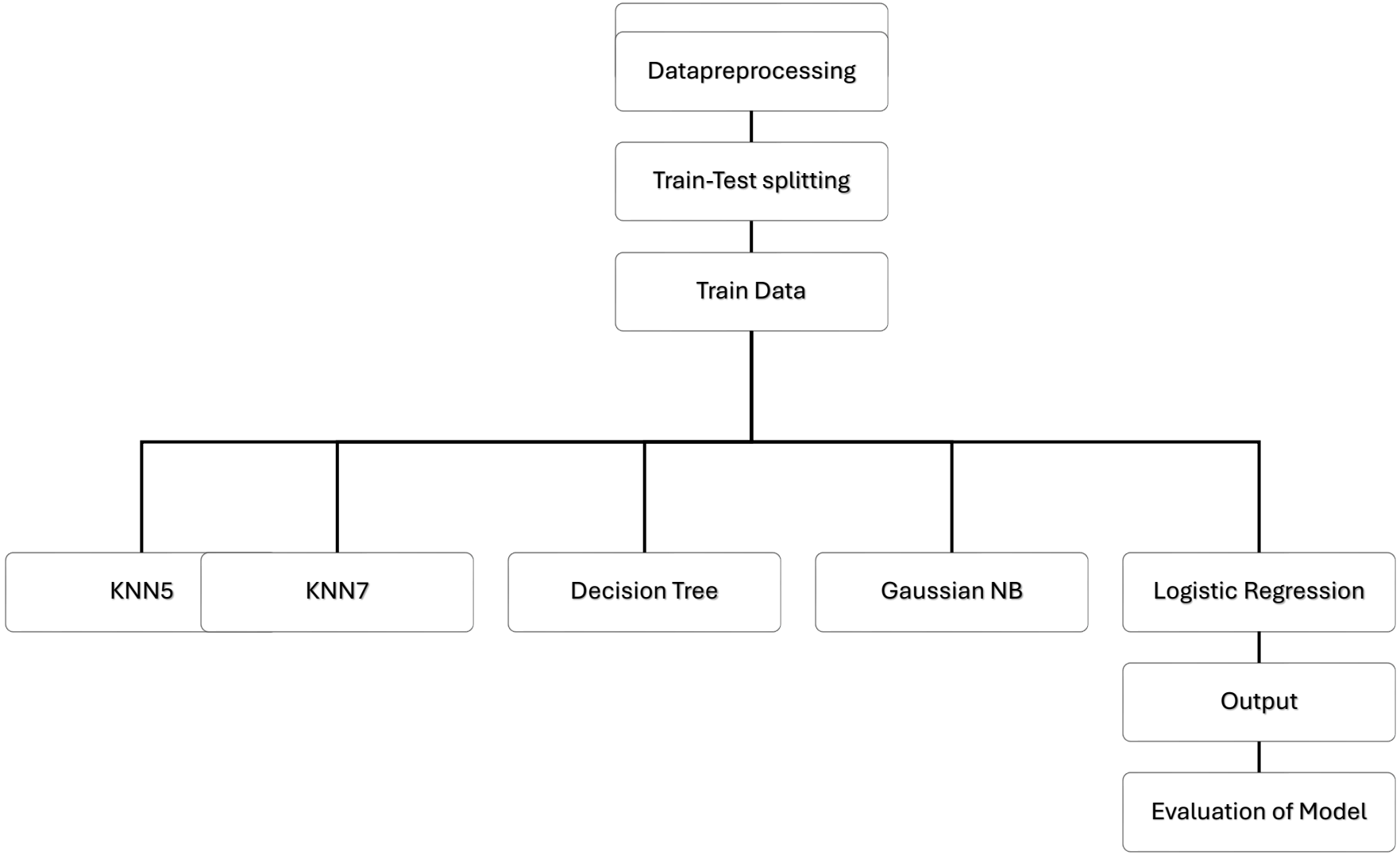
Associated Tasks
Classification

Feature Type
Real, Categorical, Integer

 # Instances
4424

 # Features
36

Method

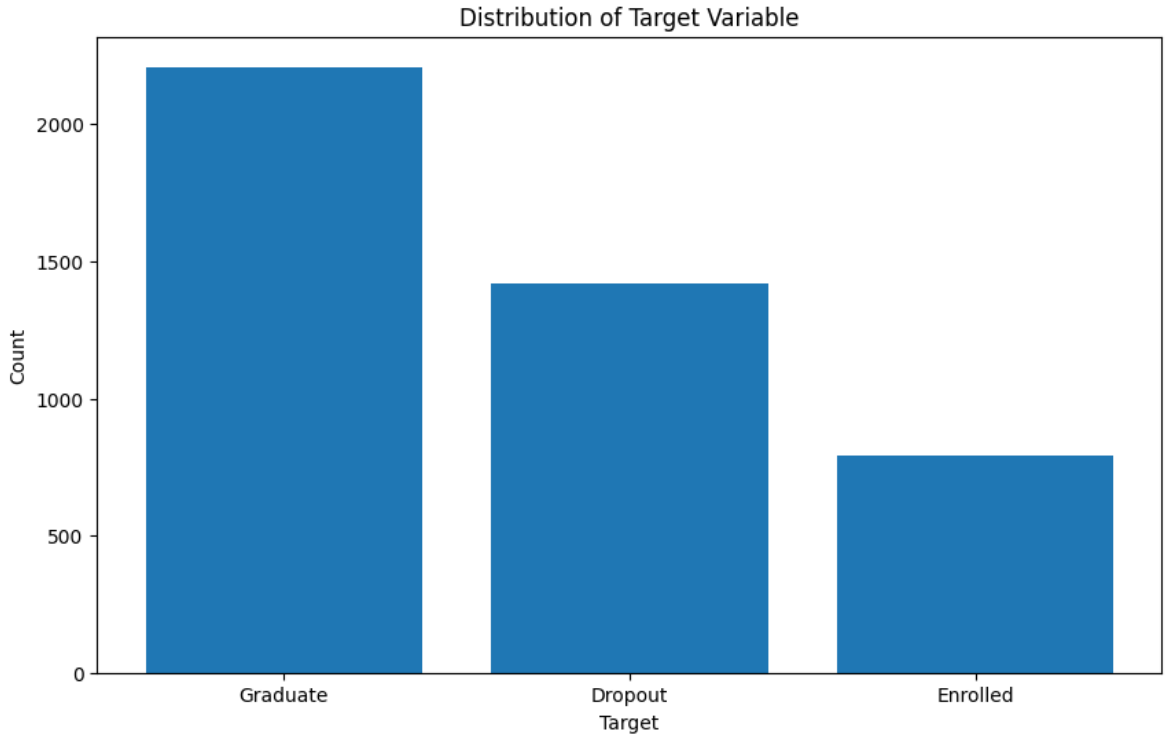

Datapreprocessing
Train-Test splitting
Train Data
KNN5
KNN7
Decision Tree
Gaussian NB
Logistic Regression
Output
Evaluation of Model

Target value

```
[84] data['Target'].value_counts()
```

Target	count
Graduate	2209
Dropout	1421
Enrolled	794

dtype: int64

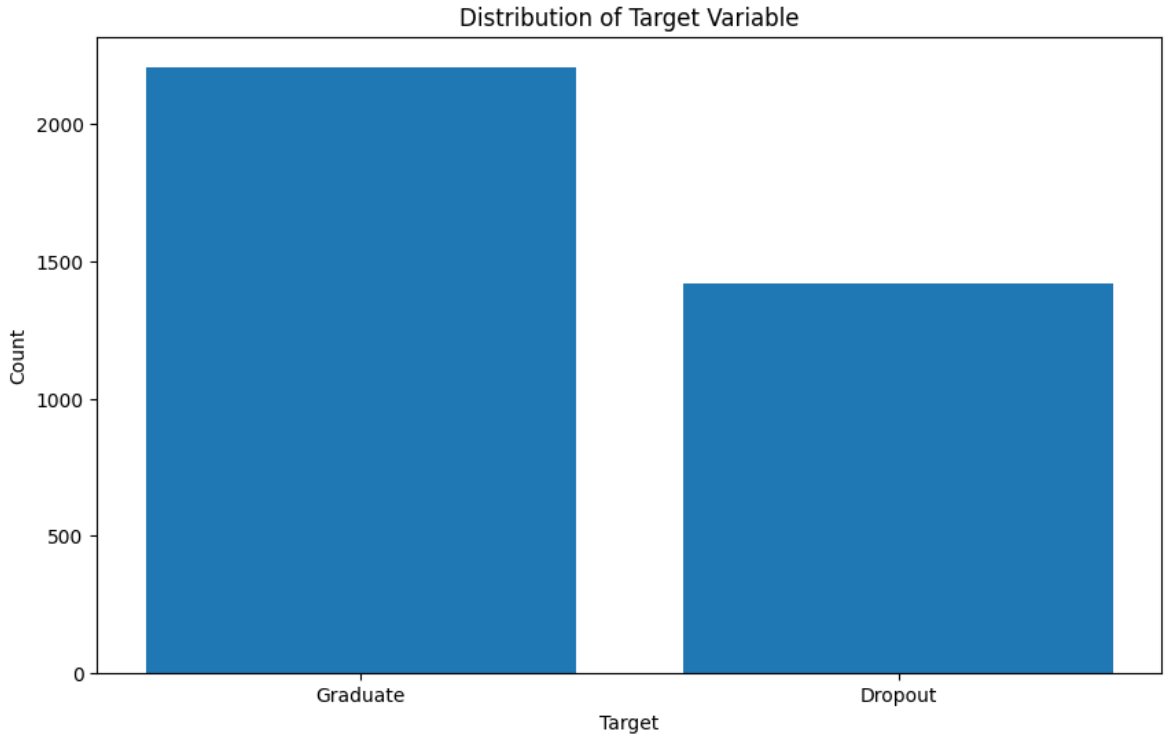


Drop the “enrolled” data

```
data = data[data.Target != 'Enrolled']
data['Target'].value_counts()
```

Target	count
Graduate	2209
Dropout	1421

dtype: int64



```
#encodee the target
data['Target1'] = LabelEncoder().fit_transform(data['Target'])
data['Target1'].value_counts()
```

Target1	count
1	2209
0	1421

dtype: int64

```
[ ] data['Target1'] =data['Target1'].apply(lambda x: 1 if x == 0 else 0 )
data['Target1'].value_counts()
```

Target1	count
0	2209
1	1421

dtype: int64

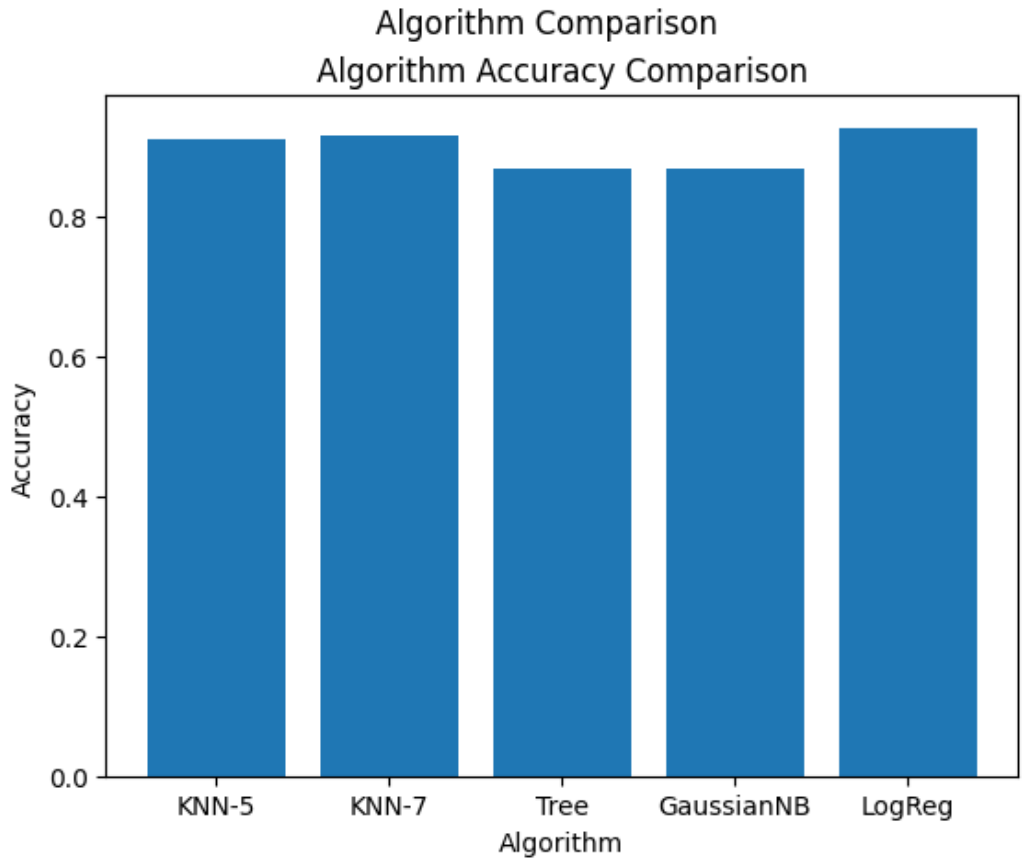
Encoding the data

Correlation

	Target
Target	1.000000
Curricular units 2nd sem (approved)	0.624157
Curricular units 2nd sem (grade)	0.566827
Curricular units 1st sem (approved)	0.529123
Curricular units 1st sem (grade)	0.485207
Tuition fees up to date	0.409827
Scholarship holder	0.297595
Curricular units 2nd sem (enrolled)	0.175847
Curricular units 1st sem (enrolled)	0.155974
Admission grade	0.120889
Displaced	0.113986
Previous qualification (grade)	0.103764

```
X = data[["Previous qualification (grade)",
          "Displaced",
          "Admission grade",
          "Curricular units 1st sem (enrolled)",
          "Curricular units 2nd sem (enrolled)",
          "Scholarship holder",
          "Tuition fees up to date",
          "Curricular units 1st sem (approved)",
          "Curricular units 1st sem (grade)",
          "Curricular units 2nd sem (approved)",
          "Curricular units 2nd sem (grade)"]].values
```

Results and Evaluation



Model	Accuracy	Precision	Recall	F1-Score
KNN-5	0.910468	0.926070	0.838028	0.879852
KNN-7	0.915978	0.930502	0.848592	0.887661
Tree	0.867769	0.815436	0.855634	0.835052
GaussianNB	0.867769	0.894958	0.750000	0.816092
LogReg	0.926997	0.913978	0.897887	0.905861

Question ?