Prediction Five Feature Importance for Intention to Enroll High School Student using Random Forest and Decision Tree

Hendra Achmadi

Universitas Pelita Harapan, Tangerang, Indonesia

hendra.achmadi@uph.edu

ABSTRACT

The number of prospective students enrolling in higher education, especially in private universities, is a serious problem. The decline in the number of prospective students that occurred during the COVID-19 pandemic from 2019 to 2022 has also become a serious problem for private universities in Indonesia. Therefore, this research focuses on finding the main characteristics of high school students in choosing private universities in Jakarta and its surroundings. The research method used is data mining, using primary data obtained from questionnaires distributed to high school students in grades 11 and 12 in the area, with a total of 438 respondents, which then went through a data cleaning process, producing 295 respondents. Using the Random Forest method in determining 5 important features and the Learning Supervisor maps out what important features should be taken into consideration in decision making for high school students. By using the random forest algorithm, an accuracy of 67 percent is obtained. Then by using the decision tree algorithm, machine learning will map the decisions of high school students. And the results illustrate that the first thing that is the main consideration is the father's education, and the second is which school he comes from, and the third is the mother's education and then how much transport money is given, and the last is what department he is from.

Keywords: Mapping the decision of the high school student; Data Mining; Random forest; Decision tree

1. INTRODUCTION

In the last three years, the education sector in Indonesia has experienced significant development. This is proven by the increase in the number of educational institutions by 42.55% from 2017 to 2018. In Banten Province, there was an increase of 38.65% in the same period. This increase occurred due to additional types of educational institutions, such as polytechnics, thus providing more options for high school graduates. Therefore, higher education institutions must compete in attracting new students from high school graduates. This is expected to last until 2022.

Competition between higher education institutions will encourage the development of more effective marketing strategies to attract the interest of high school students and convince them to choose universities or polytechnics as their choice after graduation. In Indonesia, there are two types of higher education institutions, namely universities and polytechnics, so high school graduates have many options for continuing their education.

For universities, the number of new students is very important to maintain their existence in the higher education industry. Prospective students in higher education are mostly high school students aged 17 and 18 years, who belong to Generation Z. According to research by

Kusumaningtyas et al. (2020), Generation Z has good technological literacy skills. Apart from that, Generation Z's lifestyle has unique characteristics in decision making, one of which is doing online window shopping, as mentioned in Santoso Geovani and Anna's research (Santoso & Triwijayati, 2018).Therefore, one way is needed to describe the decision consideration patterns of high school students, namely by using random forests and decision trees.

2. LITERATURE REVIEW

2.1 Data Mining Process

The Data Mining process is carried out by preparing data and continuing with data processing or data cleaning, here data preparation begins for further processing, for example whether the data has the type of number or factor or date, and then the data in the data cleaning is also carried out by removing special characters , then after that a transformation is carried out, namely changing the data from cleansing data to target data, namely the next process is carrying out data mining or data modeling based on methods that are suitable for the data, and the last is the process of interpreting the knowledge obtained from data processing. (Jiawei, 2012)

2.2. Random Forest and Decision Tree

According to (Charbuty & Abdulazeez, 2021) Decision tree algorithms are supervised machine learning algorithms used for classification and regression tasks. It creates a tree-like model of decisions and their possible consequences based on a set of input features. The algorithm iteratively partitions the data into smaller and smaller subsets based on feature values, creating a tree structure where each internal node represents a decision rule based on feature values and each leaf node represents a class label or predicted value.

As stated by Duarte et al. (2019) The decision tree algorithm works by recursively partitioning a dataset into subsets based on input feature values. At each step of the partitioning process, the algorithm selects the features that best separate the dataset into target variable classes or minimize the variance of the target variable. These features are then used to create nodes in the decision tree. The algorithm continues to partition the dataset at each node until a stopping criterion is met, such as when all instances in a subset belong to the same class of target variable.

According to Gkikas et al. (2022) one of the advantages of the decision tree algorithm is its interpretability, because the decision path can be easily visualized and understood. However, decision trees are prone to overfitting, especially when the tree is deep and complex. To avoid overfitting, techniques such as pruning, regularization, and ensemble methods such as Random Forest and gradient boosting can be applied.

2.3 Supervisor Learning

According to (Duarte et al., 2019) classification techniques are usually computer programs that learn from given input data, and use this training data with the aim of learning to classify based on observed patterns in the data. On the other hand supervised learning for regression is a set of algorithms used to predict continuous values.

According to (Charbuty & Abdulazeez, 2021), the DT algorithm is part of the family of supervised learning algorithms, and its main goal is to build a training model that can be used to predict the class or value of a target variable through inferred learning decision rules. from training data.

According to (Müller & Guido, 2017) Supervised learning is a type of machine learning where an algorithm learns from a set of labeled data to make predictions or decisions about new, unseen data. In supervised learning, the algorithm is trained on a set of input data and corresponding output data, also known as labels. The algorithm learns to map input data to output data by generalizing patterns in the training data. Supervised learning is becoming an area for much research activity in machine learning. Many of the supervised learning techniques have found application in the processing and analysis of various data

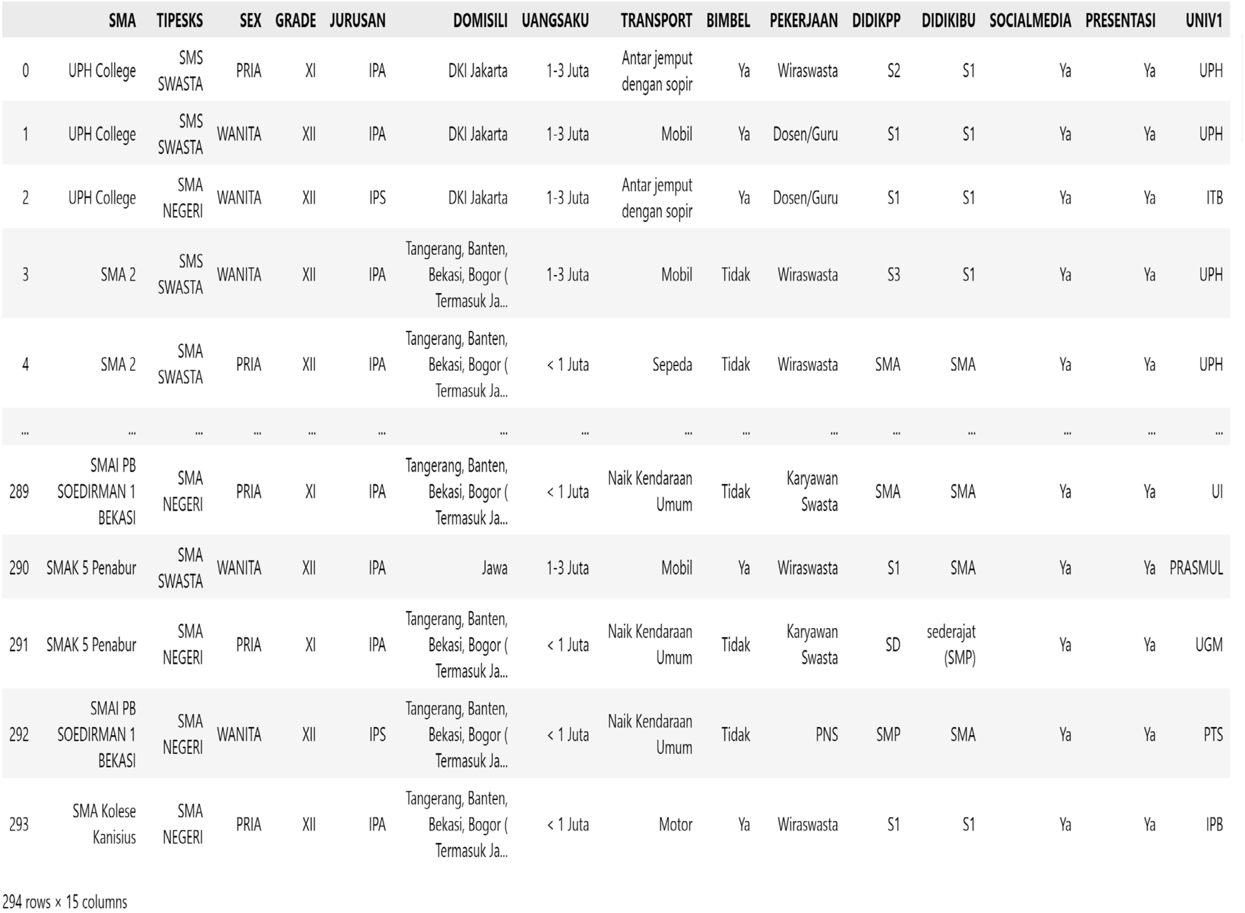
3. METHODOLOGY

This research uses a data mining method, the first is to find out a picture of each pending customer profile taken through a questionnaire to 202 respondents using Google forms, then data processing and data cleaning are carried out using data mining methods, so that it can be known from fifteen characteristics or features, where these features are important for determining decisions, using the random forest algorithm, will use the decision tree algorithm. To make random forest algorithm calculations using the Python program, and to make a decision tree algorithm using Python too.

4. RESULT

Data Preparation

Primary data was taken from questionnaires distributed to high school students in grades 11 and 12 in the Jakarta area and surrounding areas using Google Form, and 438 respondents were obtained, and then data cleaning was carried out and 295 respondents remained.



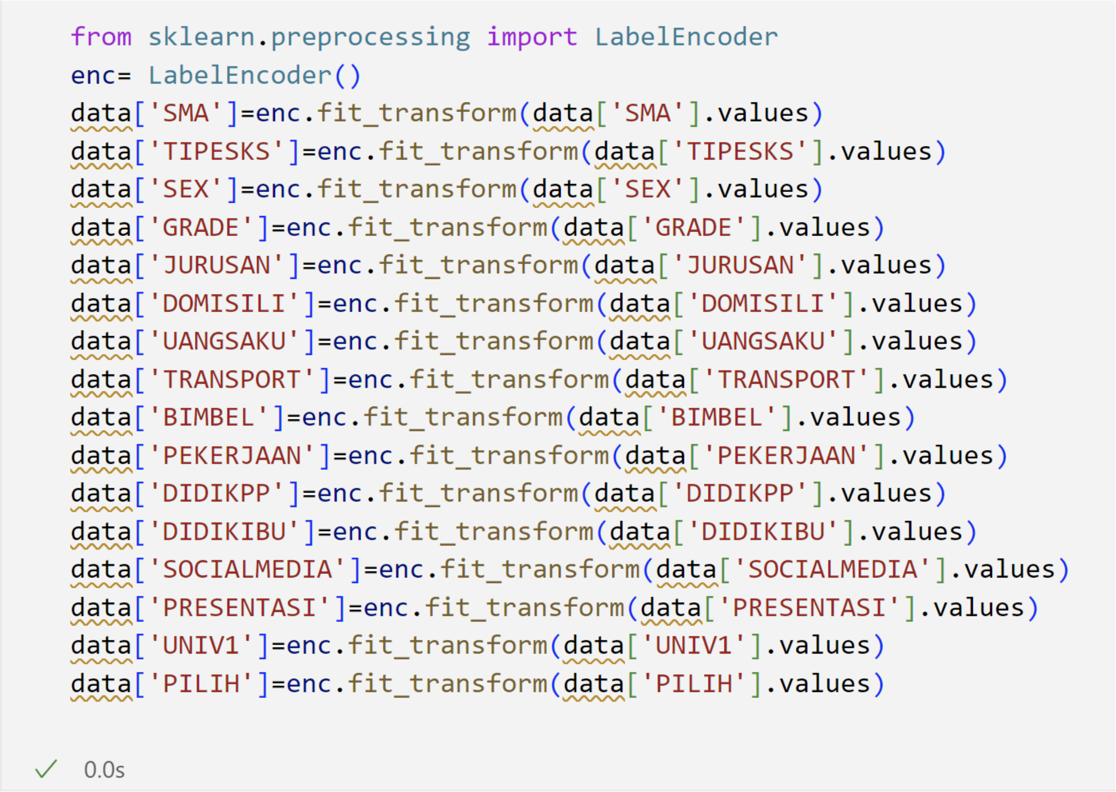
	SMA	TIPESKS	SEX	GRADE	JURUSAN	DOMISILI	UANGSAKU	TRANSPORT	BIMBEL	PEKERJAAN	DIDIKPP	DIDIKIBU	SOCIALMEDIA	PRESENTASI	UNIV1
0	UPH College	SMS SWASTA	PRIA	XI	IPA	DKI Jakarta	1-3 Juta	Antar jemput dengan sopir	Ya	Wiraswasta	S2	S1	Ya	Ya	UPH
1	UPH College	SMS SWASTA	WANITA	XII	IPA	DKI Jakarta	1-3 Juta	Mobil	Ya	Dosen/Guru	S1	S1	Ya	Ya	UPH
2	UPH College	SMA NEGERI	WANITA	XII	IPS	DKI Jakarta	1-3 Juta	Antar jemput dengan sopir	Ya	Dosen/Guru	S1	S1	Ya	Ya	ITB
3	SMA 2	SMS SWASTA	WANITA	XII	IPA	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	1-3 Juta	Mobil	Tidak	Wiraswasta	S3	S1	Ya	Ya	UPH
4	SMA 2	SMA SWASTA	PRIA	XII	IPA	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	< 1 Juta	Sepeda	Tidak	Wiraswasta	SMA	SMA	Ya	Ya	UPH
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
289	SMAI PB SOEDIRMAN 1 BEKASI	SMA NEGERI	PRIA	XI	IPA	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	< 1 Juta	Naik Kendaraan Umum	Tidak	Karyawan Swasta	SMA	SMA	Ya	Ya	UI
290	SMAK 5 Penabur	SMA SWASTA	WANITA	XII	IPA	Jawa	1-3 Juta	Mobil	Ya	Wiraswasta	S1	SMA	Ya	Ya	PRASMUL
291	SMAK 5 Penabur	SMA NEGERI	PRIA	XI	IPA	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	< 1 Juta	Naik Kendaraan Umum	Tidak	Karyawan Swasta	SD	sederajat (SMP)	Ya	Ya	UGM
292	SMAI PB SOEDIRMAN 1 BEKASI	SMA NEGERI	WANITA	XII	IPS	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	< 1 Juta	Naik Kendaraan Umum	Tidak	PNS	SMP	SMA	Ya	Ya	PTS
293	SMA Kolese Kanisius	SMA NEGERI	PRIA	XII	IPA	Tangerang, Banten, Bekasi, Bogor (Termasuk Ja...	< 1 Juta	Motor	Ya	Wiraswasta	S1	S1	Ya	Ya	IPB

294 rows × 15 columns

Figure 1. Data Primer

Encoding

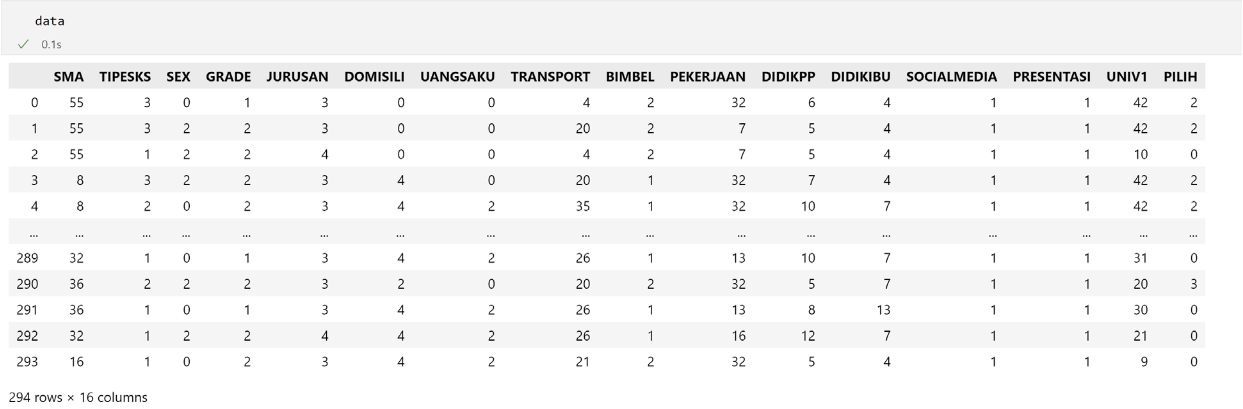
The next process is, the data will be encoded.



```
from sklearn.preprocessing import LabelEncoder
enc= LabelEncoder()
data['SMA']=enc.fit_transform(data['SMA'].values)
data['TIPESKS']=enc.fit_transform(data['TIPESKS'].values)
data['SEX']=enc.fit_transform(data['SEX'].values)
data['GRADE']=enc.fit_transform(data['GRADE'].values)
data['JURUSAN']=enc.fit_transform(data['JURUSAN'].values)
data['DOMISILI']=enc.fit_transform(data['DOMISILI'].values)
data['UANGSAKU']=enc.fit_transform(data['UANGSAKU'].values)
data['TRANSPORT']=enc.fit_transform(data['TRANSPORT'].values)
data['BIMBEL']=enc.fit_transform(data['BIMBEL'].values)
data['PEKERJAAN']=enc.fit_transform(data['PEKERJAAN'].values)
data['DIDIKPP']=enc.fit_transform(data['DIDIKPP'].values)
data['DIDIKIBU']=enc.fit_transform(data['DIDIKIBU'].values)
data['SOCIALMEDIA']=enc.fit_transform(data['SOCIALMEDIA'].values)
data['PRESENTASI']=enc.fit_transform(data['PRESENTASI'].values)
data['UNIV1']=enc.fit_transform(data['UNIV1'].values)
data['PILIH']=enc.fit_transform(data['PILIH'].values)

✓ 0.0s
```

Figure 2. Encoding Process

The Result of encoding process can be see at figure 3.

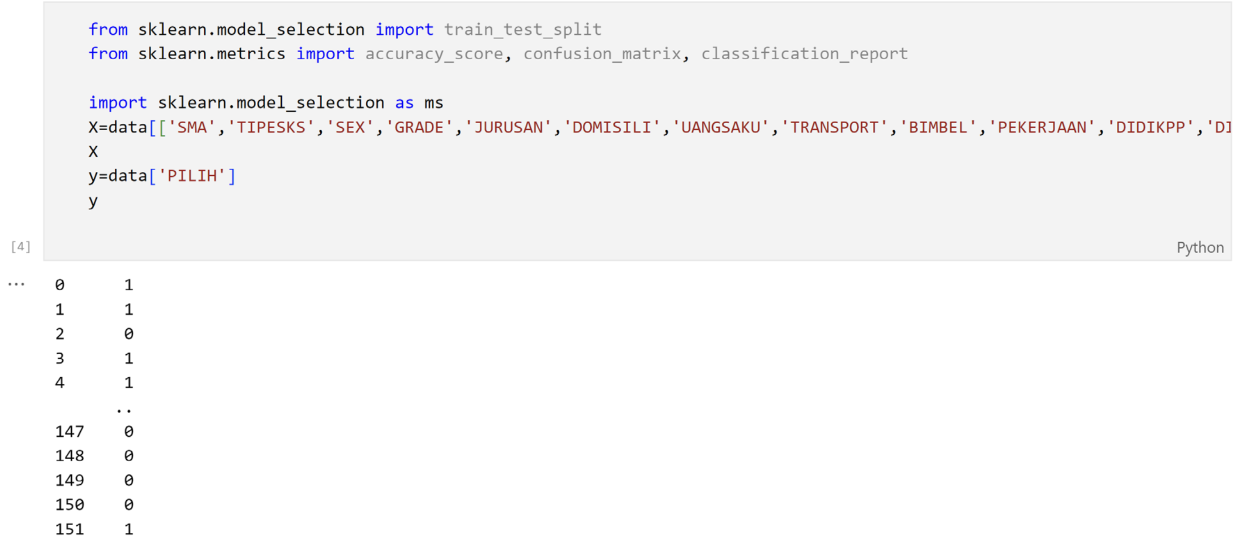


```
data
✓ 0.1s
```

	SMA	TIPESKS	SEX	GRADE	JURUSAN	DOMISILI	UANGSAKU	TRANSPORT	BIMBEL	PEKERJAAN	DIDIKPP	DIDIKIBU	SOCIALMEDIA	PRESENTASI	UNIV1	PILIH
0	55	3	0	1	3	0	0	4	2	32	6	4	1	1	42	2
1	55	3	2	2	3	0	0	20	2	7	5	4	1	1	42	2
2	55	1	2	2	4	0	0	4	2	7	5	4	1	1	10	0
3	8	3	2	2	3	4	0	20	1	32	7	4	1	1	42	2
4	8	2	0	2	3	4	2	35	1	32	10	7	1	1	42	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
289	32	1	0	1	3	4	2	26	1	13	10	7	1	1	31	0
290	36	2	2	2	3	2	0	20	2	32	5	7	1	1	20	3
291	36	1	0	1	3	4	2	26	1	13	8	13	1	1	30	0
292	32	1	2	2	4	4	2	26	1	16	12	7	1	1	21	0
293	16	1	0	2	3	4	2	21	2	32	5	4	1	1	9	0

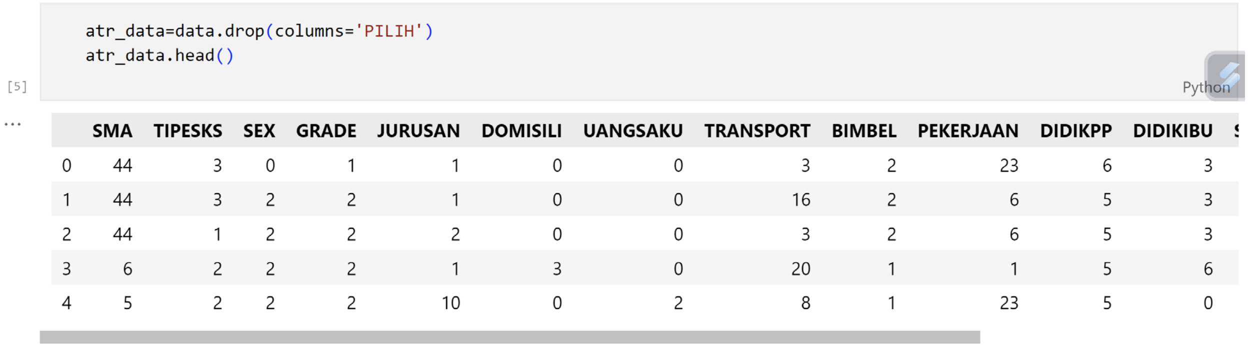
294 rows × 16 columns

Figure 3. The Result of Encoding Process



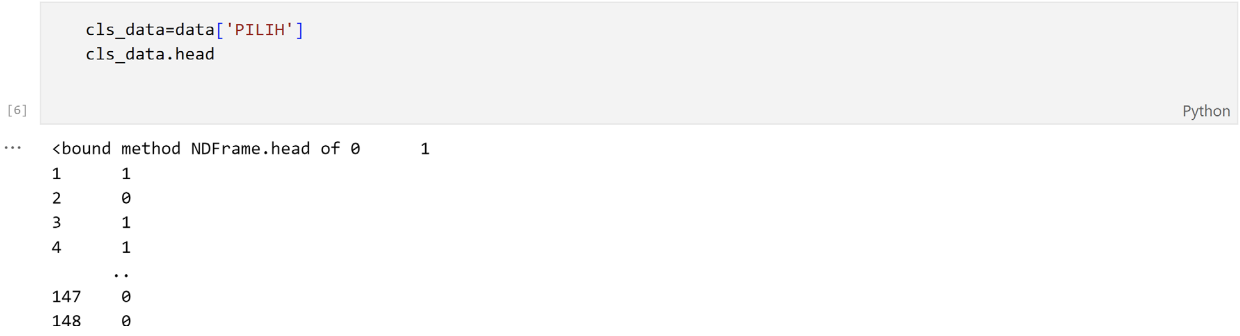
```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

import sklearn.model_selection as ms
X=data[['SMA','TIPESKS','SEX','GRADE','JURUSAN','DOMISILI','UANGSAKU','TRANSPORT','BIMBEL','PEKERJAAN','DIDIKPP','DI
X
y=data['PILIH']
y
```

```
0      1
1      1
2      0
3      1
4      1
      ..
147    0
148    0
149    0
150    0
151    1
```



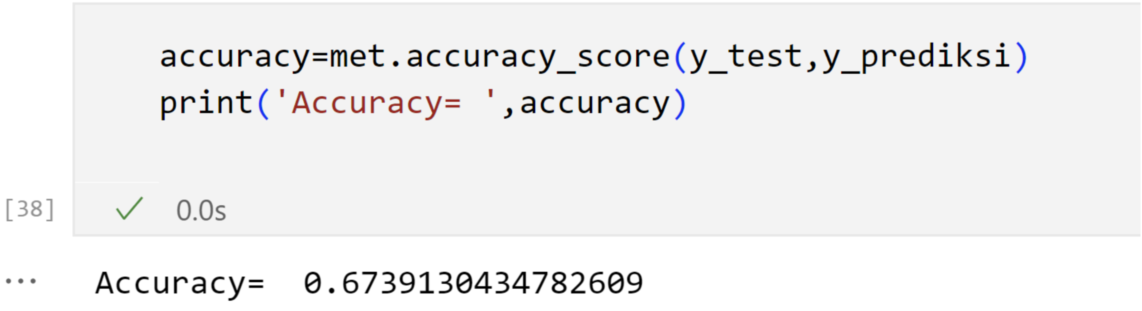
```
atr_data=data.drop(columns='PILIH')
atr_data.head()
```

	SMA	TIPESKS	SEX	GRADE	JURUSAN	DOMISILI	UANGSAKU	TRANSPORT	BIMBEL	PEKERJAAN	DIDIKPP	DIDIKIBU	S
0	44	3	0	1	1	0	0	3	2	23	6	3	
1	44	3	2	2	1	0	0	16	2	6	5	3	
2	44	1	2	2	2	0	0	3	2	6	5	3	
3	6	2	2	2	1	3	0	20	1	1	5	6	
4	5	2	2	2	10	0	2	8	1	23	5	0	



```
cls_data=data['PILIH']
cls_data.head
```

```
<bound method NDFrame.head of 0      1
1      1
2      0
3      1
4      1
      ..
147    0
148    0
```

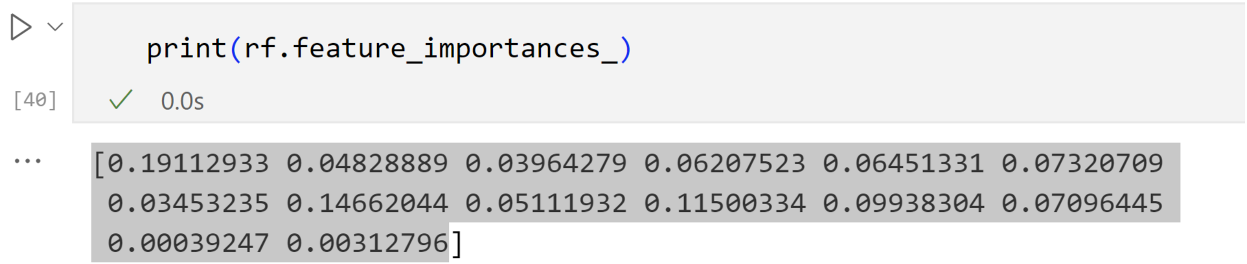


```
accuracy=met.accuracy_score(y_test,y_prediksi)
print('Accuracy= ',accuracy)
```

```
Accuracy=  0.6739130434782609
```

Figure 4. The Supervised Learning Process

Proses selanjutnya adalah dengan menggunakan proses Supervisory learning, di mana data akan dibagi menjadi 2 yaitu file training dan file testing, kemudian baru dijalankan algoritma logistic regression.



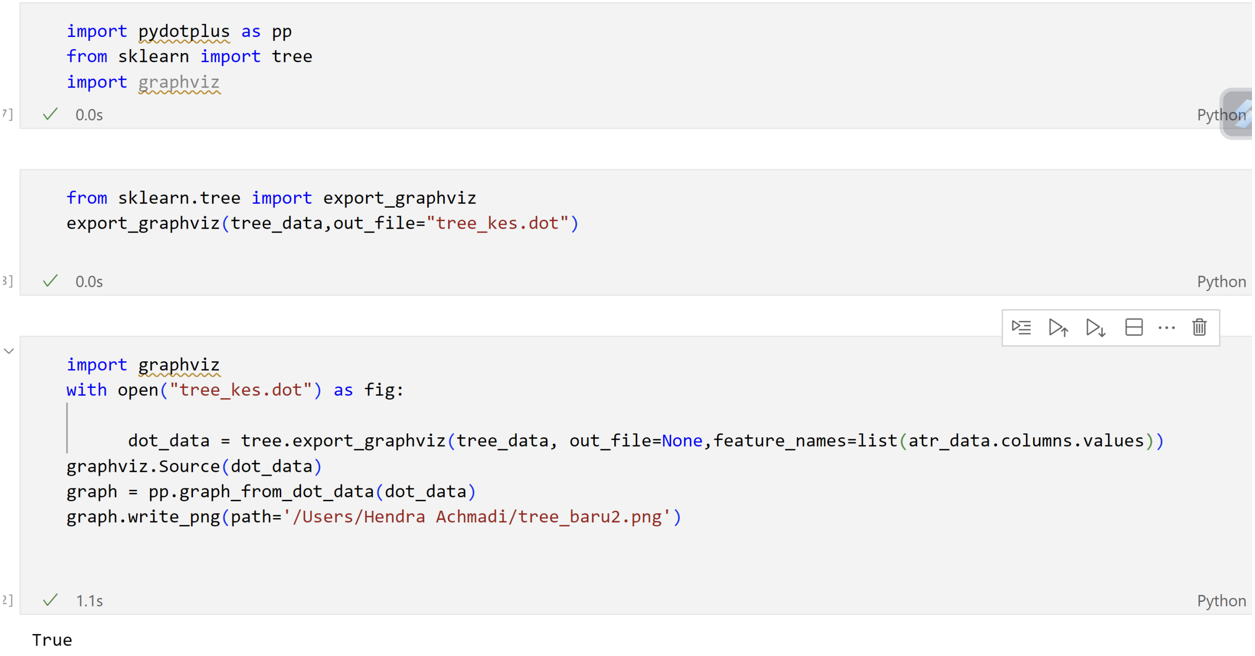
```
print(rf.feature_importances_)
```

[40] ✓ 0.0s

```
[0.19112933 0.04828889 0.03964279 0.06207523 0.06451331 0.07320709
 0.03453235 0.14662044 0.05111932 0.11500334 0.09938304 0.07096445
 0.00039247 0.00312796]
```

After that the process is continuing with determine five feature importance from the largest to smallest value

No	Feature Importance	Value
1	SMA	0,1911
8	TRANSPORT	0,1466
10	PEKERJAAN	0,115
11	DIDIKPP	0,0993
6	DOMISILI	0,0732
12	DIDIKIBU	0,0709
5	JURUSAN	0,0645
4	GRADE	0,062
9	BIMBEL	0,0511
2	TIPESKS	0,0482
3	SEX	0,0396
7	UANGSAKU	0,0345
14	PRESENTASI	0,0031
13	SOCIALMEDIA	0,0003



```
import pydotplus as pp
from sklearn import tree
import graphviz
```

```
7]  ✓  0.0s                                                    Python
```

```
from sklearn.tree import export_graphviz
export_graphviz(tree_data,out_file="tree_kes.dot")
```

```
3]  ✓  0.0s                                                    Python
```

```
import graphviz
with open("tree_kes.dot") as fig:

    dot_data = tree.export_graphviz(tree_data, out_file=None,feature_names=list(atr_data.columns.values))
graphviz.Source(dot_data)
graph = pp.graph_from_dot_data(dot_data)
graph.write_png(path='/Users/Hendra Achmadi/tree_baru2.png')
```

```
2]  ✓  1.1s                                                    Python
```

```
True
```

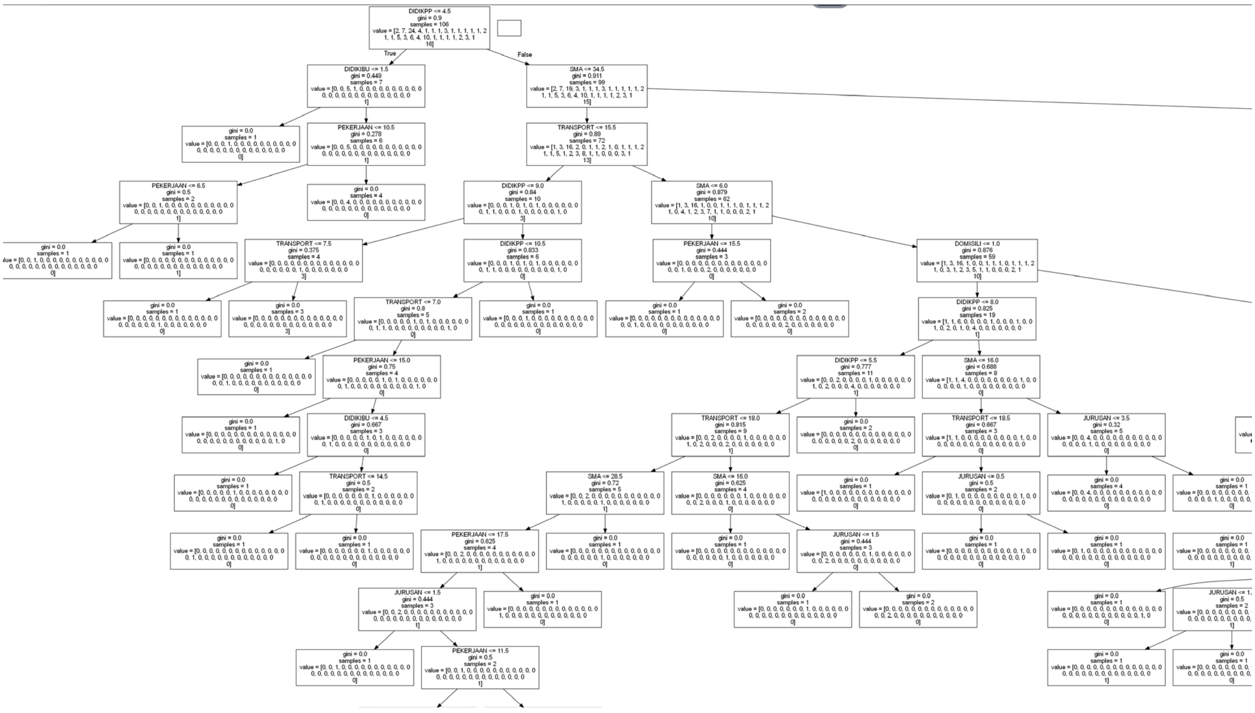


Figure 5. Decision Mapping with Decision Tree

5. CONCLUSION

By using the random forest algorithm, an accuracy of 67 percent is obtained. Then by using the decision tree algorithm, machine learning will map the decisions of high school students. And the results illustrate that the first thing that is the main consideration is the father's education, and the second is which school he comes from, and the third is the mother's education and then how much transport money is given, and the last is what department he is from.

REFERENCES

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Duarte, V., Zuñiga-Jara, S., & Contreras, S. (2019). *Machine Learning and Marketing: a literature review*. https://ssrn.com/abstract=4006436

Jiawei. (2012). *Data Mining Third Edition*.

Mansoor, F. (2022). *Increasing Generalizability: Naïve Bayes Vs K-Nearest Neighbors*. https://doi.org/10.21203/rs.3.rs-1578985/v1

Müller, A. C., & Guido, S. (2017). *Introduction to Machine Learning with Python A GUIDE FOR DATA SCIENTISTS Introduction to Machine Learning with Python*.

Muzumdar, P., Prasad Basyal, G., & Vyas, P. (2022). An Empirical Comparison of Machine Learning Models for Student's Mental Health Illness Assessment. In *Asian Journal of Computer and Information Systems* (Vol. 10, Issue 1). www.ajouronline.com

Rish, I. (2000). *An empirical study of the naive Bayes classifier*.