Thiếu Nhi Thánh Thể - Xứ | Machine Learning cơ bản | (238) TRÓT YÊU - TRUN | overleaf - Tìm trên Google | TNHien/Khai thác dữ liệu/L | 19110315_TrinhNgocHien_

localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb

Web Hỗ Trợ | Latex | Library Genesis | TRANG NHÀ - Tủ S... | Moodle HCMUS | Thư - TRỊNH NGỌC... | Drive của tôi - Goo... | Zalo Web | Duolingo - Cách họ... | Sử - web - viethoc.c...

jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Data Mining Lab02

In [1]:

```
# Library
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")

pd.set_option('max.columns',100)
pd.set_option('max.rows',500)
```

Read data and look statistics

In [2]:

```
import pandas as pd

data = pd.read_csv("Titanic.csv")

print("Display all first of 5 rows :")
display(data.head())
print("The shape of data in (nrows,ncols): ",data.shape)

```

```
Display all first of 5 rows :
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S

0:23 ENG 8:34 AM

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

```
The shape of data in (nrows,ncols):  (891, 12)
```

Function find missing percent

In [3]:
```python
def find_missing_percent(data , showresult = True):
    miss_df = pd.DataFrame({'ColumnName':[],'TotalMissingVals':[],'PercentMissing':[]})
    for col in data.columns:
        sum_miss_val = data[col].isnull().sum()
        percent_miss_val = round((sum_miss_val/data.shape[0])*100,2)
        missinginfo = {"ColumnName" : col, "TotalMissingVals" : sum_miss_val, "PercentMissing" : percent_miss_val}
        miss_df = miss_df.append(missinginfo, ignore_index = True)

    miss_df = miss_df[miss_df["PercentMissing"] > 0.0]
    miss_df = miss_df.reset_index(drop = True)
    miss_features = miss_df["ColumnName"].values
    if(showresult):
        print(data.shape)
        display(data.head())
        display(miss_df)
    return miss_df
```

In [4]:
```python
miss_df = find_missing_percent(data)
```

(891, 12)

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

	ColumnName	TotalMissingVals	PercentMissing
0	Age	177.0	19.87
1	Cabin	687.0	77.10
2	Embarked	2.0	0.22

In [5]:
```
drop_cols = list(miss_df[miss_df['PercentMissing'] > 60.0].ColumnName)
print(drop_cols)
data = data.drop(drop_cols,axis=1)
miss_df = find_missing_percent(data)
```

```
['Cabin']
(891, 11)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S

```
['Cabin']
(891, 11)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S

	ColumnName	TotalMissingVals	PercentMissing
0	Age	177.0	19.87
1	Embarked	2.0	0.22

Missing Handling

Function List wise delection

In [6]:
```
def listwise_deletion(data):
    for col in data.columns:
        miss_ind = data[col][data[col].isnull()].index
        data = data.drop(miss_ind, axis = 0)
    return data
```

In [7]:
```
data_lwd = listwise_deletion(data)
miss_df = find_missing_percent(data_lwd)
```

[KHTN] NMMH HK2 21-22 | Machine Learning cơ bản | (238) MIN - ĐỪNG YÊU | overleaf - Tìm trên Google | TNHien/Khai thác dữ liệu/L | 19110315_TrinhNgocHien_

localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb

Web Hỗ Trợ | Latex | Library Genesis | TRANG NHÀ - Tủ S... | Moodle HCMUS | Thư - TRỊNH NGỌC... | Drive của tôi - Goo... | Zalo Web | Duolingo - Cách họ... | Sử - web - viethoc.c...

jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [7]:
```
data_lwd = listwise_deletion(data)
miss_df = find_missing_percent(data_lwd)
```

```
(712, 11)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S

	ColumnName	TotalMissingVals	PercentMissing

In [8]:
```
numeric_cols = data.select_dtypes(['float','int']).columns
categoric_cols = data.select_dtypes('object').columns
print(f"Numeric Columns : {numeric_cols}")
print(f"Categoric Columns : {categoric_cols}")
```

```
Numeric Columns : Index(['Age', 'Fare'], dtype='object')
Categoric Columns : Index(['Name', 'Sex', 'Ticket', 'Embarked'], dtype='object')
```

In [9]:
```
def mean_imputation(data_numeric):
    for col in data_numeric.columns:
        mean = data_numeric[col].mean()
        data_numeric[col] = data_numeric[col].fillna(mean)
    return data_numeric
def mode_imputation(data_categoric):
    for col in data_categoric.columns:
        mode = data_categoric[col].mode().iloc[0]
```

0:14 ENG 8:37 AM

[KHTN] NMMH HK2 21-22 | Machine Learning cơ bản | (238) MIN - ĐỪNG YÊU | overleaf - Tìm trên Google | TNHien/Khai thác dữ liệu/L | 19110315_TrinhNgocHien_

localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb

Web Hỗ Trợ | Latex | Library Genesis | TRANG NHÀ - Tủ S... | Moodle HCMUS | Thư - TRỊNH NGỌC... | Drive của tôi - Goo... | Zalo Web | Duolingo - Cách họ... | Sử - web - viethoc.c...

jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [9]:
```
def mean_imputation(data_numeric):
    for col in data_numeric.columns:
        mean = data_numeric[col].mean()
        data_numeric[col] = data_numeric[col].fillna(mean)
    return data_numeric
def mode_imputation(data_categoric):
    for col in data_categoric.columns:
        mode = data_categoric[col].mode().iloc[0]
        data_categoric[col] = data_categoric[col].fillna(mode)
    return data_categoric
```

In [10]:
```
data_numeric = data[numeric_cols]
data_numeric_mean_imp = mean_imputation(data_numeric)
data_categoric = data[categoric_cols]
data_categoric_mode_imp = mode_imputation(data_categoric)

data_imputed_value = pd.concat([data_numeric_mean_imp, data_categoric_mode_imp], axis = 1)
miss_df = find_missing_percent(data_imputed_value)
```

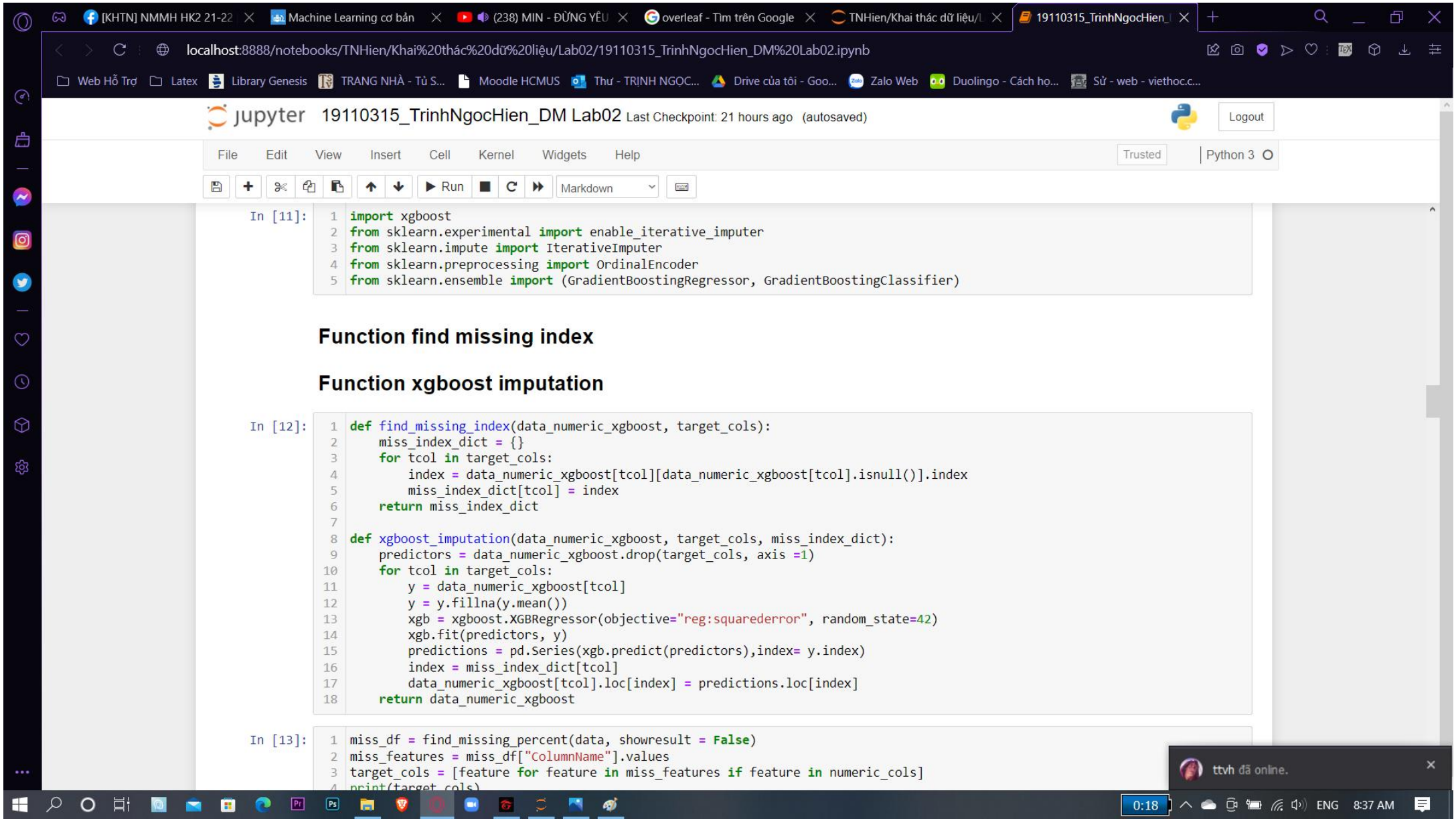
```
(891, 6)
```

	Age	Fare	Name	Sex	Ticket	Embarked
0	22.0	7.2500	Braund, Mr. Owen Harris	male	A/5 21171	S
1	38.0	71.2833	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	PC 17599	C
2	26.0	7.9250	Heikkinen, Miss. Laina	female	STON/O2. 3101282	S
3	35.0	53.1000	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	113803	S
4	35.0	8.0500	Allen, Mr. William Henry	male	373450	S

ColumnName	TotalMissingVals	PercentMissing

In [11]:
```
import xgboost
```

0:18 ENG 8:37 AM


[KHTN] NMMH HK2 21-22
Machine Learning cơ bản
(238) MIN - ĐỪNG YÊU
overleaf - Tìm trên Google
TNHien/Khai thác dữ liệu/L
19110315_TrinhNgocHien_
localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb
Web Hỗ Trợ
Latex
Library Genesis
TRANG NHÀ - Tủ S...
Moodle HCMUS
Thư - TRỊNH NGỌC...
Drive của tôi - Goo...
Zalo Web
Duolingo - Cách họ...
Sử - web - viethoc.c...
jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Markdown
In [11]:
import xgboost
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
from sklearn.preprocessing import OrdinalEncoder
from sklearn.ensemble import (GradientBoostingRegressor, GradientBoostingClassifier)
Function find missing index
Function xgboost imputation
In [12]:
def find_missing_index(data_numeric_xgboost, target_cols):
miss_index_dict = {}
for tcol in target_cols:
index = data_numeric_xgboost[tcol][data_numeric_xgboost[tcol].isnull()].index
miss_index_dict[tcol] = index
return miss_index_dict
def xgboost_imputation(data_numeric_xgboost, target_cols, miss_index_dict):
predictors = data_numeric_xgboost.drop(target_cols, axis =1)
for tcol in target_cols:
y = data_numeric_xgboost[tcol]
y = y.fillna(y.mean())
xgb = xgboost.XGBRegressor(objective="reg:squarederror", random_state=42)
xgb.fit(predictors, y)
predictions = pd.Series(xgb.predict(predictors),index= y.index)
index = miss_index_dict[tcol]
data_numeric_xgboost[tcol].loc[index] = predictions.loc[index]
return data_numeric_xgboost
In [13]:
miss_df = find_missing_percent(data, showresult = False)
miss_features = miss_df["ColumnName"].values
target_cols = [feature for feature in miss_features if feature in numeric_cols]
print(target_cols)
ttvh đã online.
0:18
ENG
8:37 AM

jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [13]:

```
miss_df = find_missing_percent(data, showresult = False)
miss_features = miss_df["ColumnName"].values
target_cols = [feature for feature in miss_features if feature in numeric_cols]
print(target_cols)

data_numeric_xgboost = data[numeric_cols]
miss_index_dict = find_missing_index(data_numeric_xgboost, target_cols)
data_numeric_xgboost = xgboost_imputation(data_numeric_xgboost, target_cols, miss_index_dict)
data_imputed_xgboost = pd.concat([data_numeric_xgboost, data_categoric_mode_imp], axis = 1)

miss_df = find_missing_percent(data_imputed_xgboost)
```

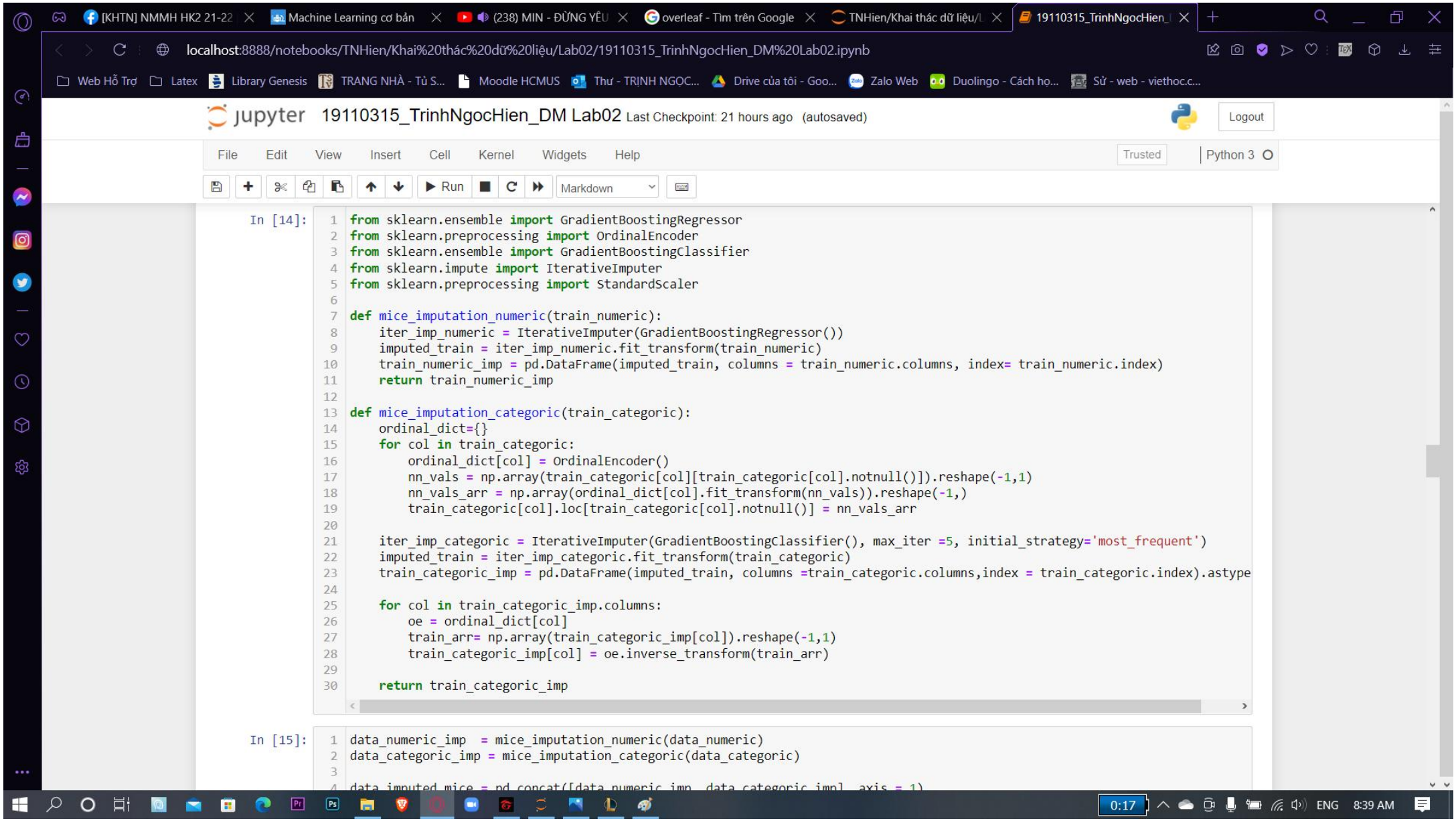
```
['Age']
(891, 6)
```

	Age	Fare	Name	Sex	Ticket	Embarked
0	22.0	7.2500	Braund, Mr. Owen Harris	male	A/5 21171	S
1	38.0	71.2833	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	PC 17599	C
2	26.0	7.9250	Heikkinen, Miss. Laina	female	STON/O2. 3101282	S
3	35.0	53.1000	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	113803	S
4	35.0	8.0500	Allen, Mr. William Henry	male	373450	S

	ColumnName	TotalMissingVals	PercentMissing

Mice Imputation Numberic

Mice imputation Categoric


[KHTN] NMMH HK2 21-22
Machine Learning cơ bản
(238) MIN - ĐỪNG YÊU
overleaf - Tìm trên Google
TNHien/Khai thác dữ liệu/L
19110315_TrinhNgocHien_
localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb
Web Hỗ Trợ
Latex
Library Genesis
TRANG NHÀ - Tủ S...
Moodle HCMUS
Thư - TRỊNH NGỌC...
Drive của tôi - Goo...
Zalo Web
Duolingo - Cách họ...
Sử - web - viethoc.c...
jupyter
19110315_TrinhNgocHien_DM Lab02
Last Checkpoint: 21 hours ago
(autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Markdown
In [14]:
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.preprocessing import OrdinalEncoder
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.impute import IterativeImputer
from sklearn.preprocessing import StandardScaler
def mice_imputation_numeric(train_numeric):
iter_imp_numeric = IterativeImputer(GradientBoostingRegressor())
imputed_train = iter_imp_numeric.fit_transform(train_numeric)
train_numeric_imp = pd.DataFrame(imputed_train, columns = train_numeric.columns, index= train_numeric.index)
return train_numeric_imp
def mice_imputation_categoric(train_categoric):
ordinal_dict={}
for col in train_categoric:
ordinal_dict[col] = OrdinalEncoder()
nn_vals = np.array(train_categoric[col][train_categoric[col].notnull()]).reshape(-1,1)
nn_vals_arr = np.array(ordinal_dict[col].fit_transform(nn_vals)).reshape(-1,)
train_categoric[col].loc[train_categoric[col].notnull()] = nn_vals_arr
iter_imp_categoric = IterativeImputer(GradientBoostingClassifier(), max_iter =5, initial_strategy='most_frequent')
imputed_train = iter_imp_categoric.fit_transform(train_categoric)
train_categoric_imp = pd.DataFrame(imputed_train, columns =train_categoric.columns,index = train_categoric.index).astype
for col in train_categoric_imp.columns:
oe = ordinal_dict[col]
train_arr= np.array(train_categoric_imp[col]).reshape(-1,1)
train_categoric_imp[col] = oe.inverse_transform(train_arr)
return train_categoric_imp
In [15]:
data_numeric_imp = mice_imputation_numeric(data_numeric)
data_categoric_imp = mice_imputation_categoric(data_categoric)
0:17
ENG
8:39 AM

[KHTN] NMMH HK2 21-22 | Machine Learning cơ bản | (238) MIN - ĐỪNG YÊU | overleaf - Tìm trên Google | TNHien/Khai thác dữ liệu/L | 19110315_TrinhNgocHien_

localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb

Web Hỗ Trợ | Latex | Library Genesis | TRANG NHÀ - Tủ S... | Moodle HCMUS | Thư - TRỊNH NGỌC... | Drive của tôi - Goo... | Zalo Web | Duolingo - Cách họ... | Sử - web - viethoc.c...

jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [15]:
```
data_numeric_imp  = mice_imputation_numeric(data_numeric)
data_categoric_imp = mice_imputation_categoric(data_categoric)

data_imputed_mice = pd.concat([data_numeric_imp, data_categoric_imp], axis = 1)
miss_df = find_missing_percent(data_imputed_mice)
```

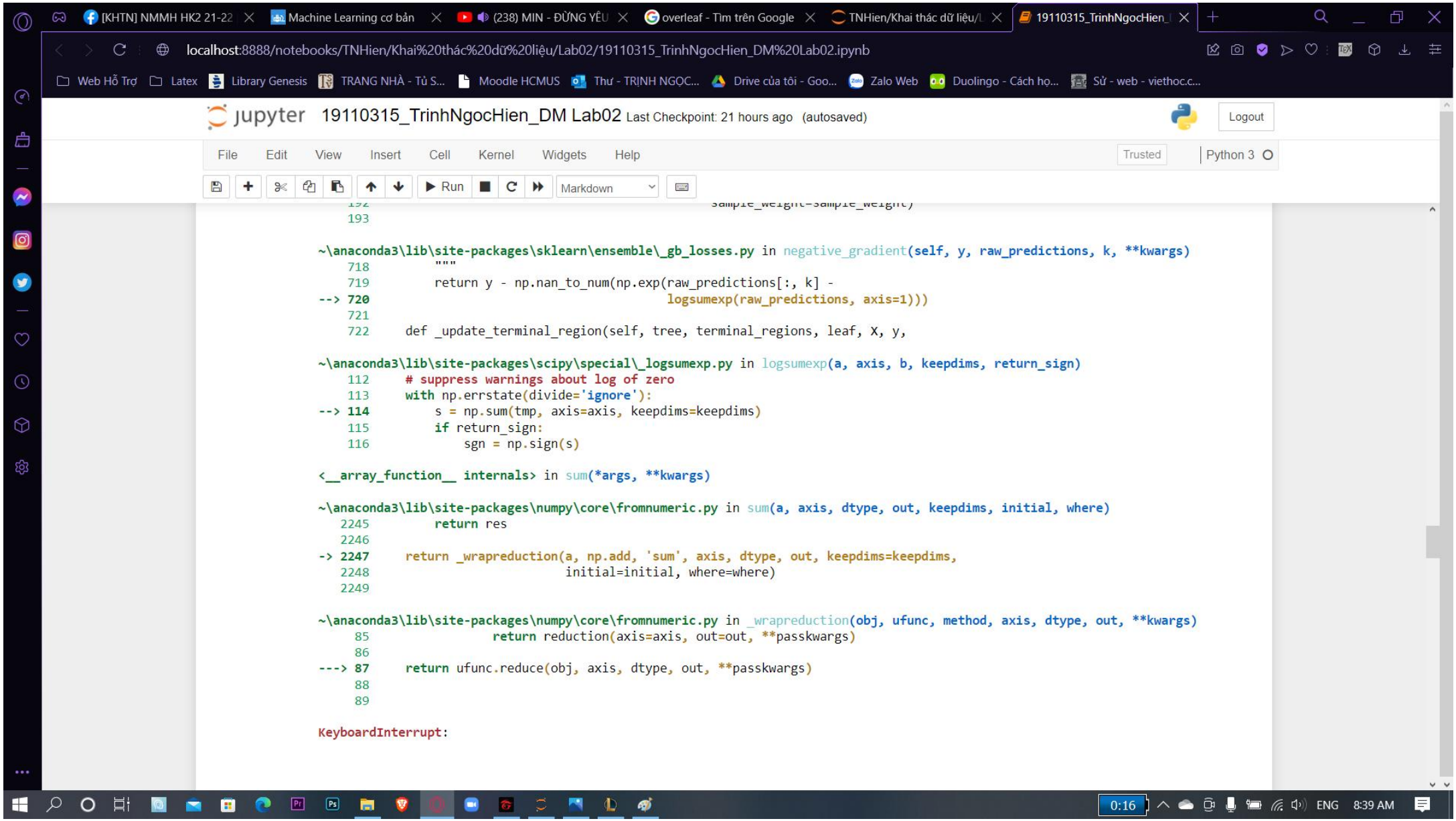
```
---------------------------------------------------------------------------
KeyboardInterrupt                         Traceback (most recent call last)
<ipython-input-15-578255edd542> in <module>
      1 data_numeric_imp  = mice_imputation_numeric(data_numeric)
----> 2 data_categoric_imp = mice_imputation_categoric(data_categoric)
      3
      4 data_imputed_mice = pd.concat([data_numeric_imp, data_categoric_imp], axis = 1)
      5 miss_df = find_missing_percent(data_imputed_mice)

<ipython-input-14-09a8be3e20e0> in mice_imputation_categoric(train_categoric)
     20
     21     iter_imp_categoric = IterativeImputer(GradientBoostingClassifier(), max_iter =5, initial_strategy='most_frequent')
---> 22     imputed_train = iter_imp_categoric.fit_transform(train_categoric)
     23     train_categoric_imp = pd.DataFrame(imputed_train, columns =train_categoric.columns,index = train_categoric.index).a
stype(int)
     24

~\anaconda3\lib\site-packages\sklearn\impute\_iterative.py in fit_transform(self, X, y)
    655                                                         feat_idx,
    656                                                         abs_corr_mat)
--> 657                 Xt, estimator = self._impute_one_feature(
    658                     Xt, mask_missing_values, feat_idx, neighbor_feat_idx,
    659                     estimator=None, fit_mode=True)

~\anaconda3\lib\site-packages\sklearn\impute\_iterative.py in _impute_one_feature(self, X_filled, mask_missing_values, feat_id
x, neighbor_feat_idx, estimator, fit_mode)
    307             y_train = _safe_indexing(X_filled[:, feat_idx],
    308                                      ~missing_row_mask)
--> 309             estimator.fit(X_train, y_train)
    310
```

0:17 ENG 8:39 AM


[KHTN] NMMH HK2 21-22
Machine Learning cơ bản
(238) MIN - ĐỪNG YÊU
overleaf - Tìm trên Google
TNHien/Khai thác dữ liệu/L
19110315_TrinhNgocHien_
localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb
Web Hỗ Trợ
Latex
Library Genesis
TRANG NHÀ - Tủ S...
Moodle HCMUS
Thư - TRỊNH NGỌC...
Drive của tôi - Goo...
Zalo Web
Duolingo - Cách họ...
Sử - web - viethoc.c...
jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Markdown
193
~\anaconda3\lib\site-packages\sklearn\ensemble_gb_losses.py in negative_gradient(self, y, raw_predictions, k, **kwargs)
718 """
719 return y - np.nan_to_num(np.exp(raw_predictions[:, k] -
--> 720 logsumexp(raw_predictions, axis=1)))
721
722 def _update_terminal_region(self, tree, terminal_regions, leaf, X, y,
~\anaconda3\lib\site-packages\scipy\special_logsumexp.py in logsumexp(a, axis, b, keepdims, return_sign)
112 # suppress warnings about log of zero
113 with np.errstate(divide='ignore'):
--> 114 s = np.sum(tmp, axis=axis, keepdims=keepdims)
115 if return_sign:
116 sgn = np.sign(s)
<__array_function__ internals> in sum(*args, **kwargs)
~\anaconda3\lib\site-packages\numpy\core\fromnumeric.py in sum(a, axis, dtype, out, keepdims, initial, where)
2245 return res
2246
-> 2247 return _wrapreduction(a, np.add, 'sum', axis, dtype, out, keepdims=keepdims,
2248 initial=initial, where=where)
2249
~\anaconda3\lib\site-packages\numpy\core\fromnumeric.py in _wrapreduction(obj, ufunc, method, axis, dtype, out, **kwargs)
85 return reduction(axis=axis, out=out, **passkwargs)
86
---> 87 return ufunc.reduce(obj, axis, dtype, out, **passkwargs)
88
89
KeyboardInterrupt:
0:16
ENG
8:39 AM

Data Modelling

In [16]:
```
data_modelling = data_lwd.copy()

```

In [17]:
```
data_modelling
```

Out[17]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S
...	...	...	...	...	...	...	...	...	...	...	...
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	Q

712 rows × 11 columns

In [18]:
```
def FeatureEngineering(data_modelling):
    data_modelling['Total_Age'] = data_modelling['Age'].sum()
    data_modelling['Total_Survived_for_all'] = data_modelling['Survived'].sum()
    data_modelling['Total_Fare_for_all'] = data_modelling['Fare'].sum()
    data_modelling['Total_Age_and_Fare'] = data_modelling['Age'] + data_modelling['Fare']
    return data_modelling
```

[KHTN] NMMH HK2 21-22 | Machine Learning cơ bản | (238) Chi Pu | ANH ƠI | overleaf - Tìm trên Google | TNHien/Khai thác dữ liệu/L | 19110315_TrinhNgocHien_

localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb

Web Hỗ Trợ | Latex | Library Genesis | TRANG NHÀ - Tủ S... | Moodle HCMUS | Thư - TRỊNH NGỌC... | Drive của tôi - Goo... | Zalo Web | Duolingo - Cách họ... | Sử - web - viethoc.c...

jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [18]:

```
def FeatureEngineering(data_modelling):
    data_modelling['Total_Age'] = data_modelling['Age'].sum()
    data_modelling['Total_Survived_for_all'] = data_modelling['Survived'].sum()
    data_modelling['Total_Fare_for_all'] = data_modelling['Fare'].sum()
    data_modelling['Total_Age_and_Fare'] = data_modelling['Age'] + data_modelling['Fare']
    return data_modelling

data_modelling = FeatureEngineering(data_modelling)
display(data_modelling.head())
print(data_modelling.shape)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	Total_Age	Total_Survived_for_all	Total_Fare_for_all	Tc
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S	21105.17	288	24611.883	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C	21105.17	288	24611.883	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S	21105.17	288	24611.883	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S	21105.17	288	24611.883	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S	21105.17	288	24611.883	

(712, 15)

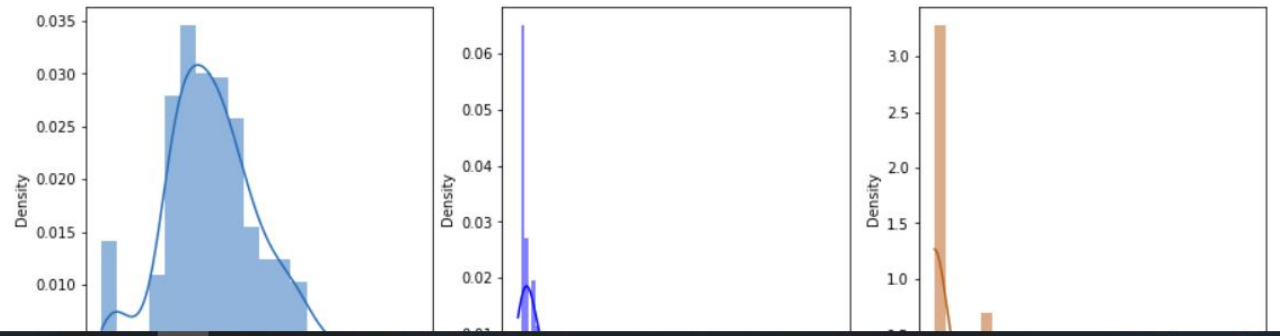
0:15 ENG 8:42 AM

(712, 15)

In [19]:
```
skew_limit = 0.5
skew_vals = data_modelling[numeric_cols].skew()
skew_cols = (skew_vals
             .sort_values(ascending=False)
             .to_frame()
             .rename(columns={0:'Skew'})
             .query('abs(Skew) > {0}'.format(skew_limit)))
display(skew_cols.T)
```

	Fare
Skew	4.667009

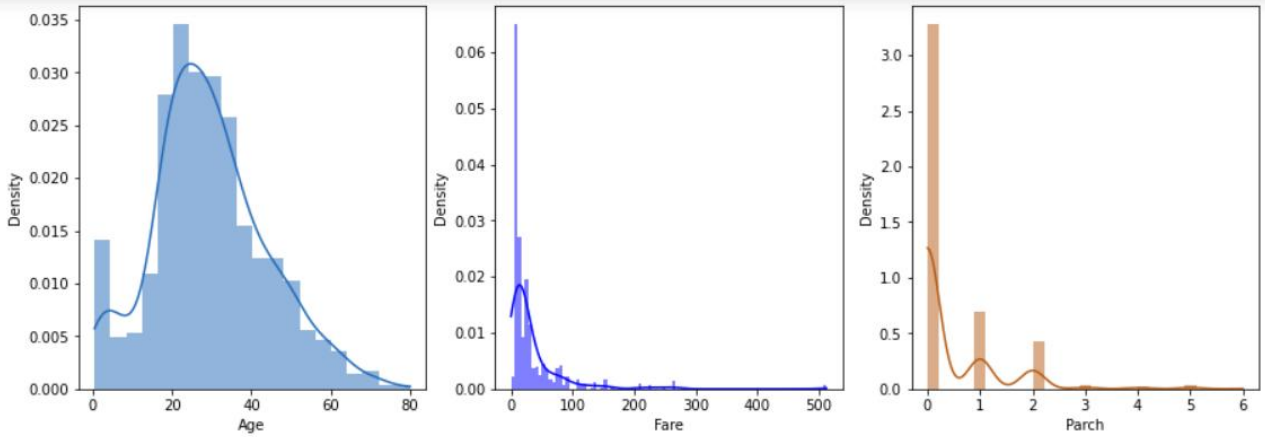
In [20]:
```
fig, (ax_positive, ax_target, ax_negative) = plt.subplots(1, 3, figsize=(15, 5))
sns.histplot(data_modelling['Age'],kde=True, stat='density', linewidth=0, color = '#236AB9', ax=ax_positive)
sns.histplot(data_modelling['Fare'],kde=True, stat='density', linewidth=0, color = 'blue', ax=ax_target)
sns.histplot(data_modelling['Parch'], kde=True, stat='density', linewidth=0,color='#B85B14', ax=ax_negative)
plt.show()
```



jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code



In [21]:
```
def NormalizeSkewedFeatures(data_modelling):
    from scipy.special import boxcox1p
    from scipy.stats import boxcox_normmax
    for col in skew_cols.index:
        if(col != 'Fare'):
            try:
                data_modelling[col] = boxcox1p(data_modelling[col], boxcox_normmax(data_modelling[col] + 1))
            except:
                print(f"column {col} can not apply BoxCox")
                continue
    return data_modelling

data_modelling = NormalizeSkewedFeatures(data_modelling)
data_modelling["Fare"] = np.log1p(data_modelling["Fare"])
```

In [22]:
```
fig, (ax_positive, ax_target, ax_negative) = plt.subplots(1, 3, figsize=(15, 5))
```

[KHTN] NMMH HK2 21-22 | Machine Learning cơ bản | (238) Chi Pu | ANH ƠI | overleaf - Tìm trên Google | TNHien/Khai thác dữ liệu/L | 19110315_TrinhNgocHien_L

localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb

Web Hỗ Trợ | Latex | Library Genesis | TRANG NHÀ - Tủ S... | Moodle HCMUS | Thư - TRỊNH NGỌC... | Drive của tôi - Goo... | Zalo Web | Duolingo - Cách họ... | Sử - web - viethoc.c...

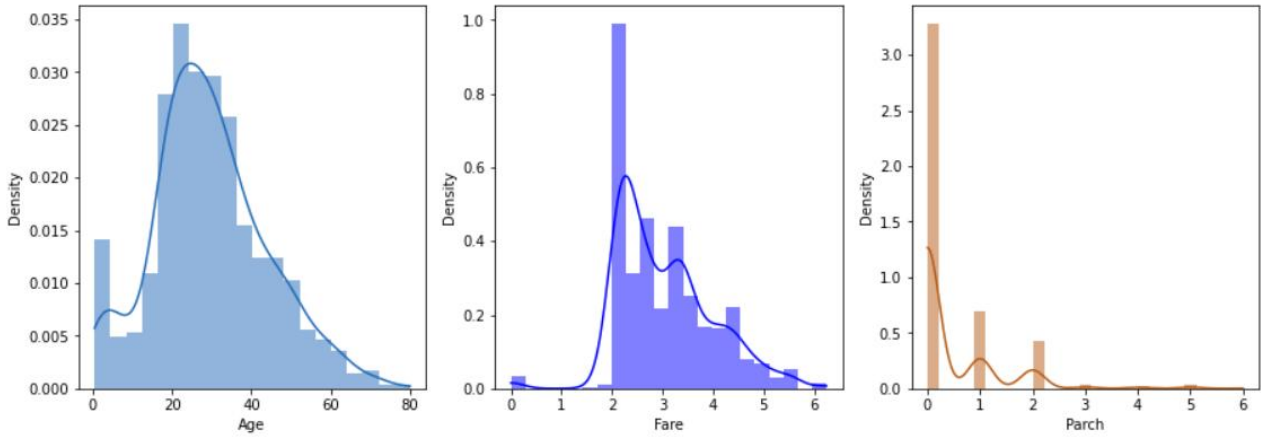
jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [22]:

```
fig, (ax_positive, ax_target, ax_negative) = plt.subplots(1, 3, figsize=(15, 5))
sns.histplot(data_modelling['Age'],kde=True, stat='density', linewidth=0, color = '#236AB9', ax=ax_positive)
sns.histplot(data_modelling['Fare'],kde=True, stat='density', linewidth=0, color = 'blue', ax=ax_target)
sns.histplot(data_modelling['Parch'], kde=True, stat='density', linewidth=0,color='#B85B14', ax=ax_negative)
plt.show()
```



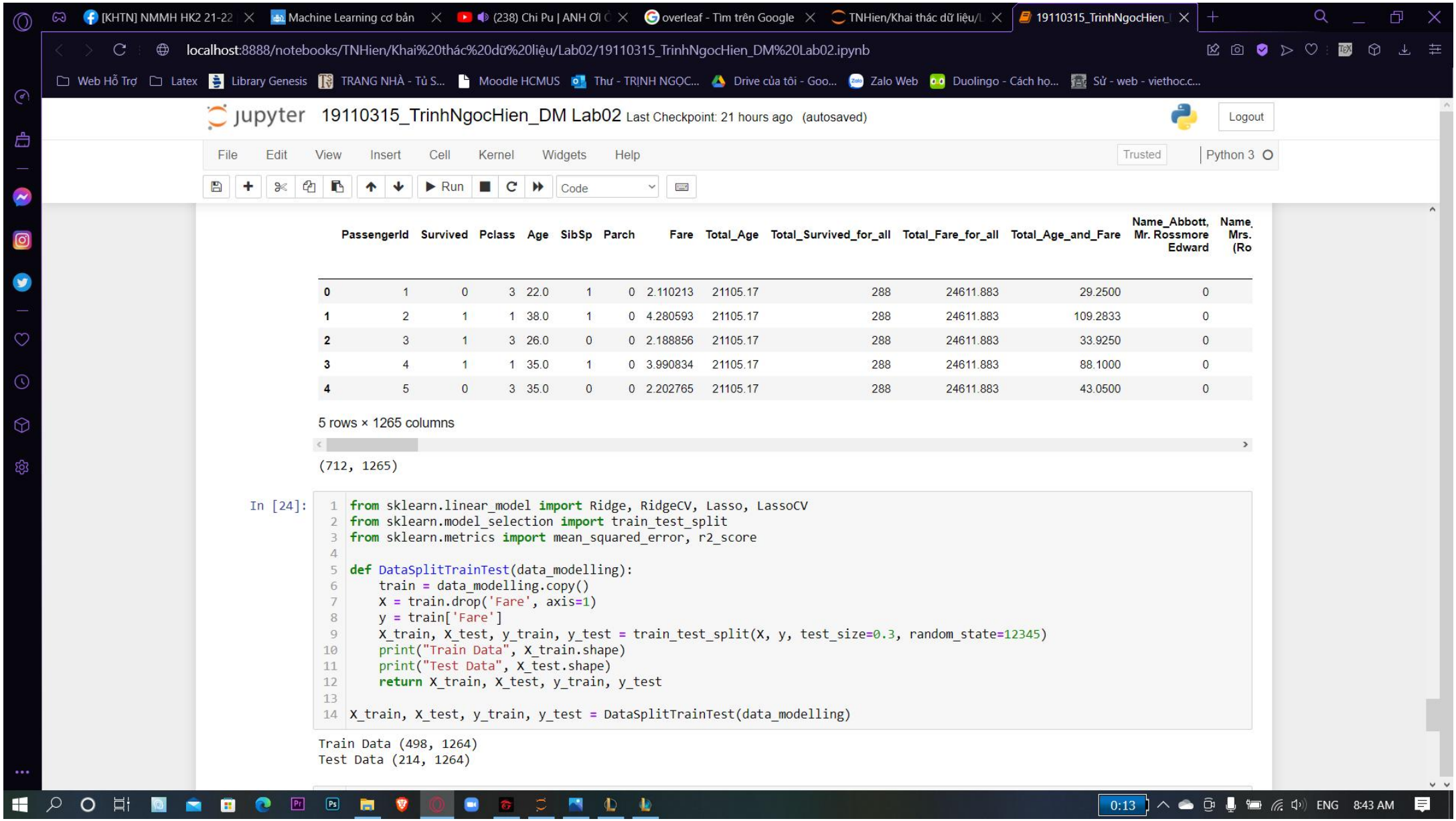
In [23]:

```
def FeatureEncoding(data_modelling):
    data_modelling = pd.get_dummies(data_modelling, columns=categoric_cols, drop_first=True)
    return data_modelling

data_modelling = FeatureEncoding(data_modelling)
display(data_modelling.head())
print(data_modelling.shape)
```

Name_Abbott... Name

0:14 ENG 8:43 AM



[KHTN] NMMH HK2 21-22 | Machine Learning cơ bản | (238) Chi Pu | ANH ƠI | overleaf - Tìm trên Google | TNHien/Khai thác dữ liệu/L | 19110315_TrinhNgocHien_

localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb

Web Hỗ Trợ | Latex | Library Genesis | TRANG NHÀ - Tủ S... | Moodle HCMUS | Thư - TRỊNH NGỌC... | Drive của tôi - Goo... | Zalo Web | Duolingo - Cách họ... | Sử - web - viethoc.c...

jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Total_Age	Total_Survived_for_all	Total_Fare_for_all	Total_Age_and_Fare	Name_Abbott, Mr. Rossmore Edward	Name_ Mrs. (Ro
0	1	0	3	22.0	1	0	2.110213	21105.17	288	24611.883	29.2500	0	
1	2	1	1	38.0	1	0	4.280593	21105.17	288	24611.883	109.2833	0	
2	3	1	3	26.0	0	0	2.188856	21105.17	288	24611.883	33.9250	0	
3	4	1	1	35.0	1	0	3.990834	21105.17	288	24611.883	88.1000	0	
4	5	0	3	35.0	0	0	2.202765	21105.17	288	24611.883	43.0500	0	

5 rows × 1265 columns

```
(712, 1265)
```

In [24]:

```
from sklearn.linear_model import Ridge, RidgeCV, Lasso, LassoCV
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

def DataSplitTrainTest(data_modelling):
    train = data_modelling.copy()
    X = train.drop('Fare', axis=1)
    y = train['Fare']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=12345)
    print("Train Data", X_train.shape)
    print("Test Data", X_test.shape)
    return X_train, X_test, y_train, y_test

X_train, X_test, y_train, y_test = DataSplitTrainTest(data_modelling)
```

```
Train Data (498, 1264)
Test Data (214, 1264)
```

0:13 ENG 8:43 AM

(3) [KHTN] NMMH HK2 21 | Machine Learning cơ bản | (238) Aloha - Cool || Pia | overleaf - Tìm trên Google | TNHien/Khai thác dữ liệu/L | 19110315_TrinhNgocHien_L

localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb

Web Hỗ Trợ | Latex | Library Genesis | TRANG NHÀ - Tủ S... | Moodle HCMUS | Thư - TRỊNH NGỌC... | Drive của tôi - Goo... | Zalo Web | Duolingo - Cách họ... | Sử - web - viethoc.c...

jupyter 19110315_TrinhNgocHien_DM Lab02 Last Checkpoint: 21 hours ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [25]:

```
def BuildLassoModel(X_train, X_test, y_train, y_test):
    lasso = Lasso(max_iter = 100000, normalize = True)
    lassocv = LassoCV(alphas = None, cv = 10, max_iter = 100000, normalize = True)
    lassocv.fit(X_train, y_train)

    lasso.set_params(alpha=lassocv.alpha_)
    lasso.fit(X_train, y_train)

    print('The Lasso:')
    print("Alpha =", lassocv.alpha_)
    print("RMSE =", mean_squared_error(y_test, lasso.predict(X_test), squared=False))
    print("R2 Score = ", r2_score(y_test, lasso.predict(X_test)))
    return lasso

lasso = BuildLassoModel(X_train, X_test, y_train, y_test)
```

```
The Lasso:
Alpha = 3.335189513517674e-05
RMSE = 0.2743078224257173
R2 Score =  0.9089543698393528
```

In [26]:

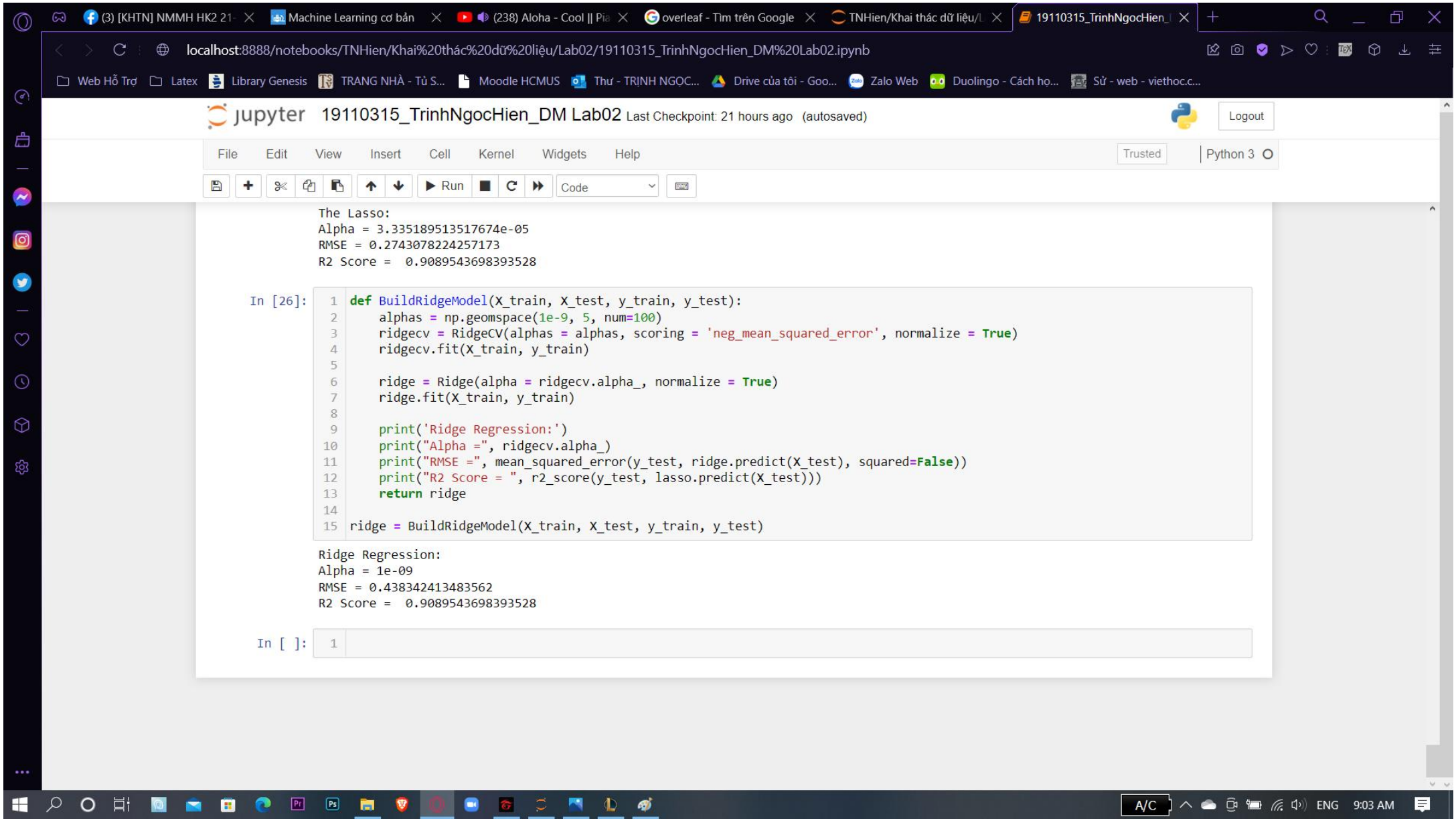
```
def BuildRidgeModel(X_train, X_test, y_train, y_test):
    alphas = np.geomspace(1e-9, 5, num=100)
    ridgecv = RidgeCV(alphas = alphas, scoring = 'neg_mean_squared_error', normalize = True)
    ridgecv.fit(X_train, y_train)

    ridge = Ridge(alpha = ridgecv.alpha_, normalize = True)
    ridge.fit(X_train, y_train)

    print('Ridge Regression:')
    print("Alpha =", ridgecv.alpha_)
    print("RMSE =", mean_squared_error(y_test, ridge.predict(X_test), squared=False))
    print("R2 Score = ", r2_score(y_test, lasso.predict(X_test)))
    return ridge

ridge = BuildRidgeModel(X_train, X_test, y_train, y_test)
```

A/C ENG 9:03 AM


(3) [KHTN] NMMH HK2 21
Machine Learning cơ bản
(238) Aloha - Cool || Pia
overleaf - Tìm trên Google
TNHien/Khai thác dữ liệu/L
19110315_TrinhNgocHien_
localhost:8888/notebooks/TNHien/Khai%20thác%20dữ%20liệu/Lab02/19110315_TrinhNgocHien_DM%20Lab02.ipynb
Web Hỗ Trợ
Latex
Library Genesis
TRANG NHÀ - Tủ S...
Moodle HCMUS
Thư - TRỊNH NGỌC...
Drive của tôi - Goo...
Zalo Web
Duolingo - Cách họ...
Sử - web - viethoc.c...
jupyter
19110315_TrinhNgocHien_DM Lab02
Last Checkpoint: 21 hours ago
(autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
The Lasso:
Alpha = 3.335189513517674e-05
RMSE = 0.2743078224257173
R2 Score = 0.9089543698393528
In [26]:
def BuildRidgeModel(X_train, X_test, y_train, y_test):
alphas = np.geomspace(1e-9, 5, num=100)
ridgecv = RidgeCV(alphas = alphas, scoring = 'neg_mean_squared_error', normalize = True)
ridgecv.fit(X_train, y_train)
ridge = Ridge(alpha = ridgecv.alpha_, normalize = True)
ridge.fit(X_train, y_train)
print('Ridge Regression:')
print("Alpha =", ridgecv.alpha_)
print("RMSE =", mean_squared_error(y_test, ridge.predict(X_test), squared=False))
print("R2 Score = ", r2_score(y_test, lasso.predict(X_test)))
return ridge
ridge = BuildRidgeModel(X_train, X_test, y_train, y_test)
Ridge Regression:
Alpha = 1e-09
RMSE = 0.438342413483562
R2 Score = 0.9089543698393528
In [ ]:
A/C
ENG
9:03 AM