Data Science Internship

BY MOHAMMAD ARSHADULLA NOOR
AMINA MAHNOOR

02

Contents

Problem Statement:

The aim is to visualize data and get insights from the dataset. To analyze how different categories play a role in income prediction.

Introduction:

The adult income dataset that involves predicting personal income levels as above or below 50,000 per year based on personal details such as relationship and education level. This means that techniques for imbalanced classification can be used whilst model performance can still be reported using classification accuracy, as is used with balanced classification problems.

Data Preprocessing

In this, you will discover how to develop and evaluate a model for the imbalanced adult income classification dataset.

We evaluate:

- How to load and explore the dataset and generate ideas for data preparation and model selection.

In this project, we use a standard imbalanced machine learning dataset referred to as the "Adult Income" dataset.

Out[6]:

	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32555	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32556	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32557	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32558	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
32559	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

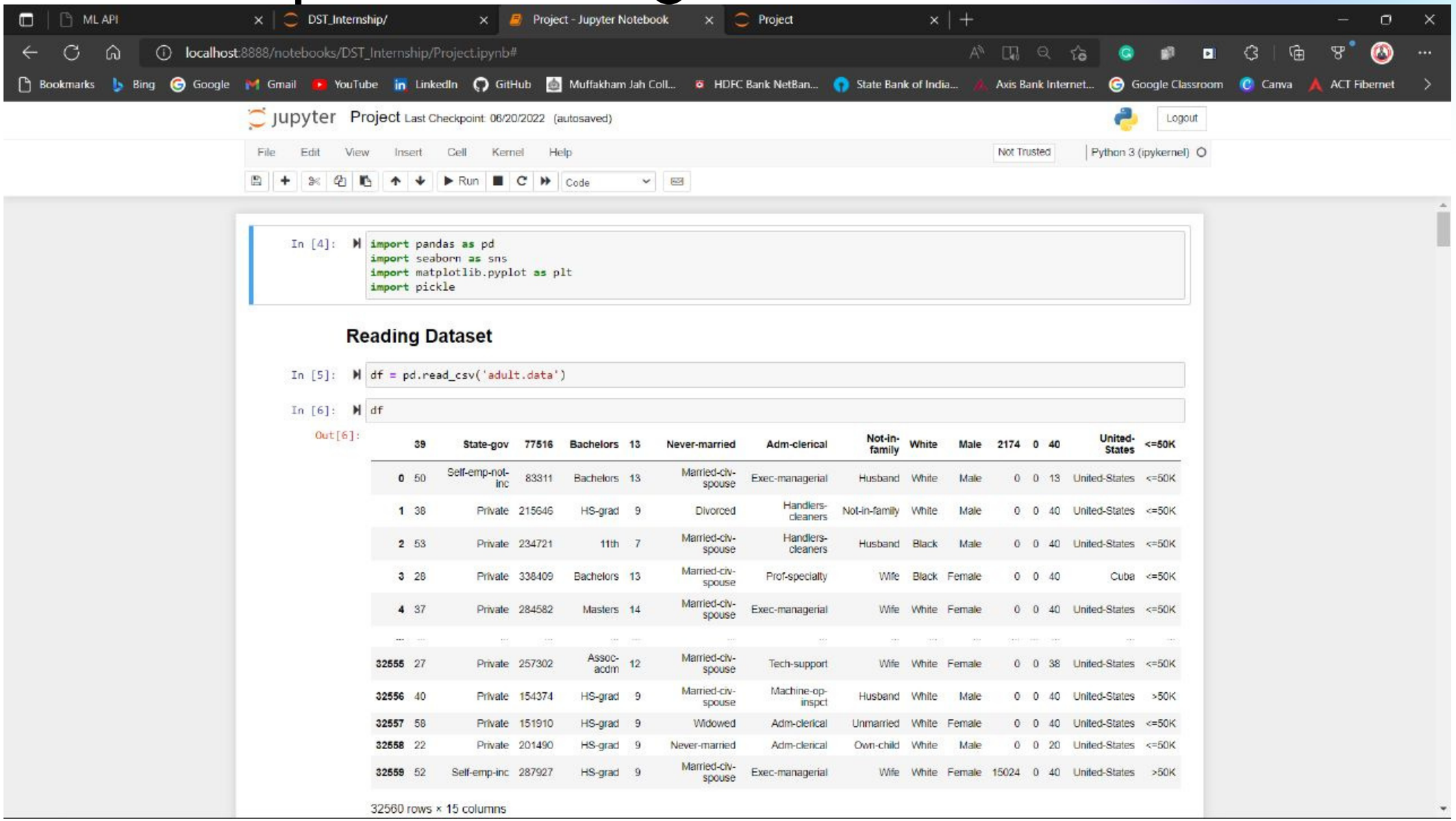
Data Preprocessing

The dataset provides 14 input variables that are a mixture of categorical, ordinal, and numerical data types. The complete list of variables is as follows:

In [7]:
```
col_labels = ['age', 'workclass', 'fnlwgt', 'education', 'education-num', 'marital_status', 'occupation', 'relationship',
              'race', 'sex', 'capital-gain', 'capital-loss', 'hours-per-week', 'native-country', 'income']
df.columns = col_labels
```

The dataset contains missing values that are marked with a question mark character (?).
There are a total of 32,560 rows of data, and 15 with columns of data.

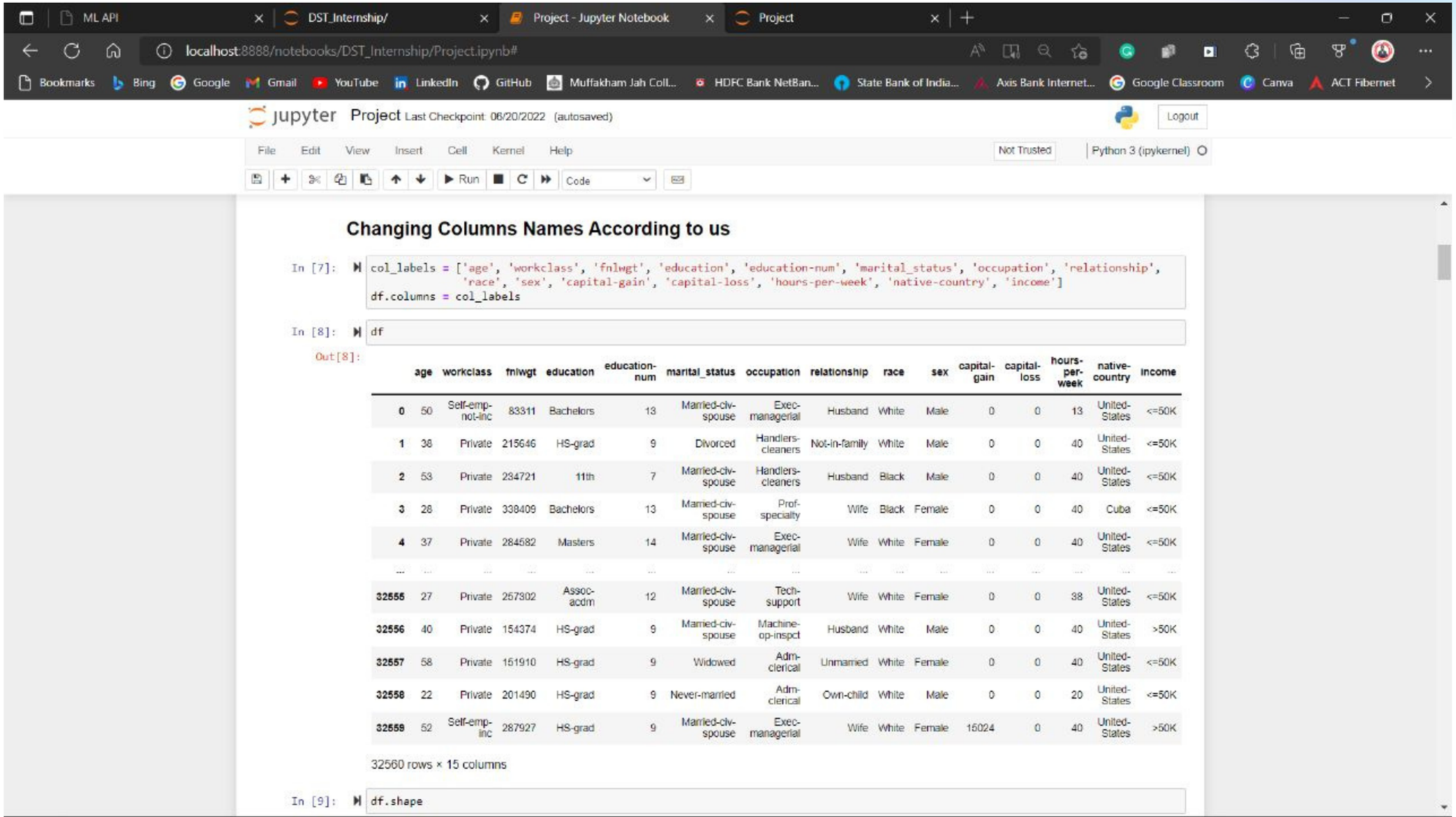
Data Preprocessing



	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32555	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32556	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32557	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32558	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
32559	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

32560 rows × 15 columns

Data Preprocessing



Changing Columns Names According to us

In [7]:
```
col_labels = ['age', 'workclass', 'fnlwgt', 'education', 'education-num', 'marital_status', 'occupation', 'relationship',
              'race', 'sex', 'capital-gain', 'capital-loss', 'hours-per-week', 'native-country', 'income']
df.columns = col_labels
```

In [8]:
```
df
```

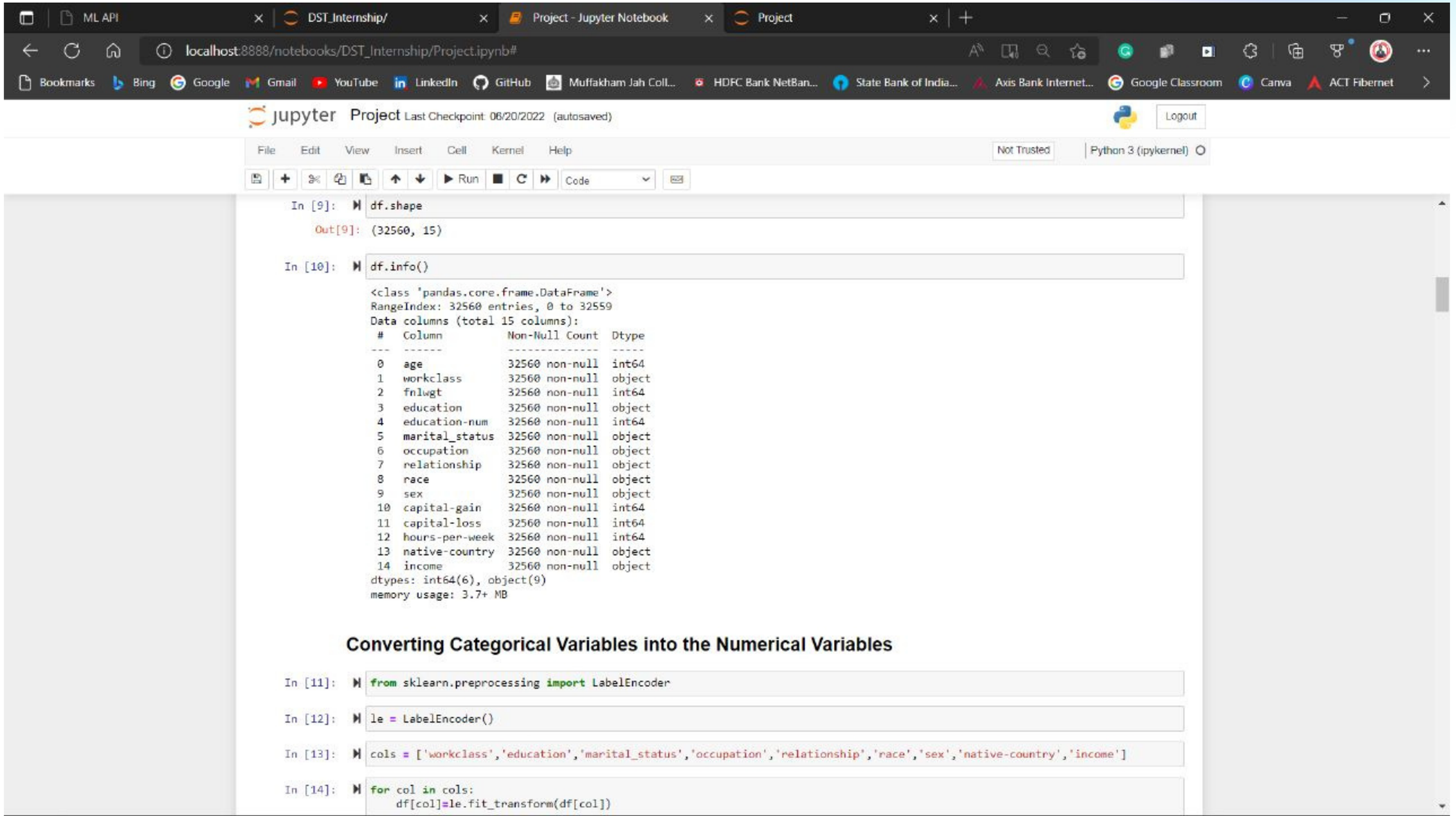
Out[8]:

	age	workclass	fnlwgt	education	education-num	marital_status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32555	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32556	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32557	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32558	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
32559	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

32560 rows × 15 columns

In [9]:
```
df.shape
```

Data Preprocessing



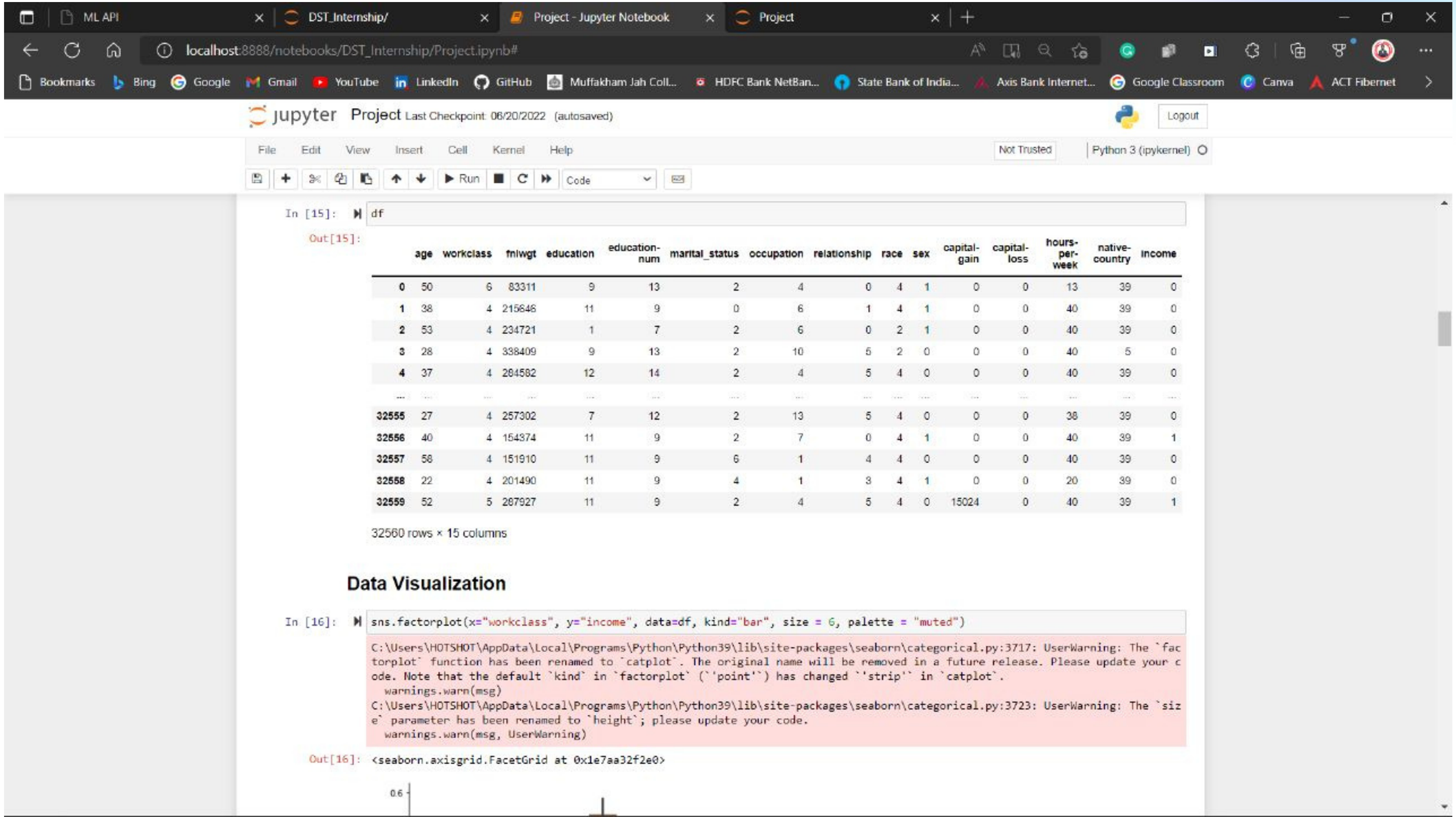
Data Visualization:

After data preprocessing, data visualization is done using the given dataset to check to analyze the income for different categories.

There are two class values '>50K' and '<=50K', meaning it is a binary classification task. The classes are imbalanced, with a skew toward the '<=50K' class label.
''>50K': majority class, approximately 25%.
''<=50K': minority class, approximately 75%.

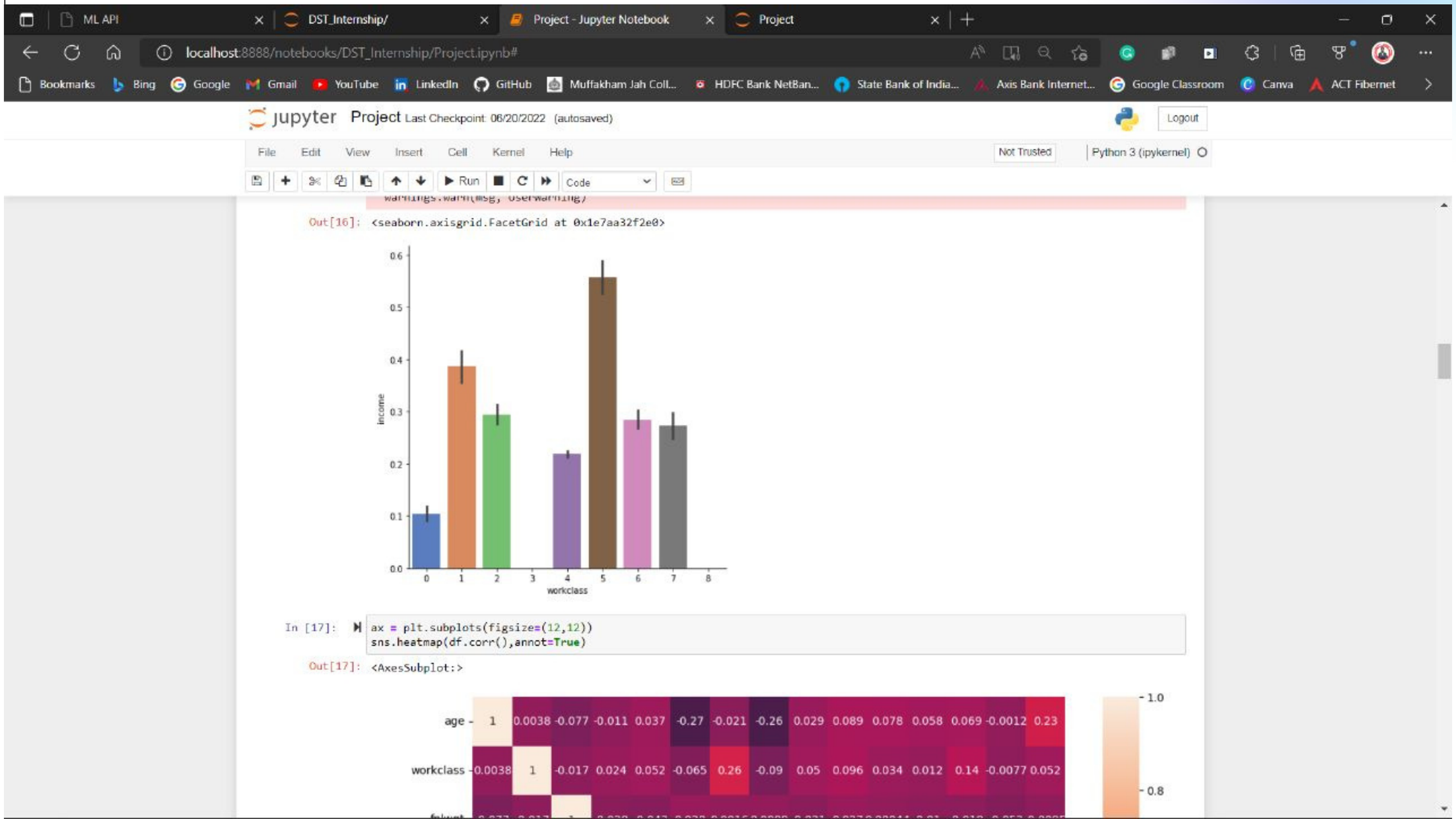
Different data visualizations are applied to the dataset like bargraph, headmap,countplot to analyze the income for different categories.

Data Visualization:

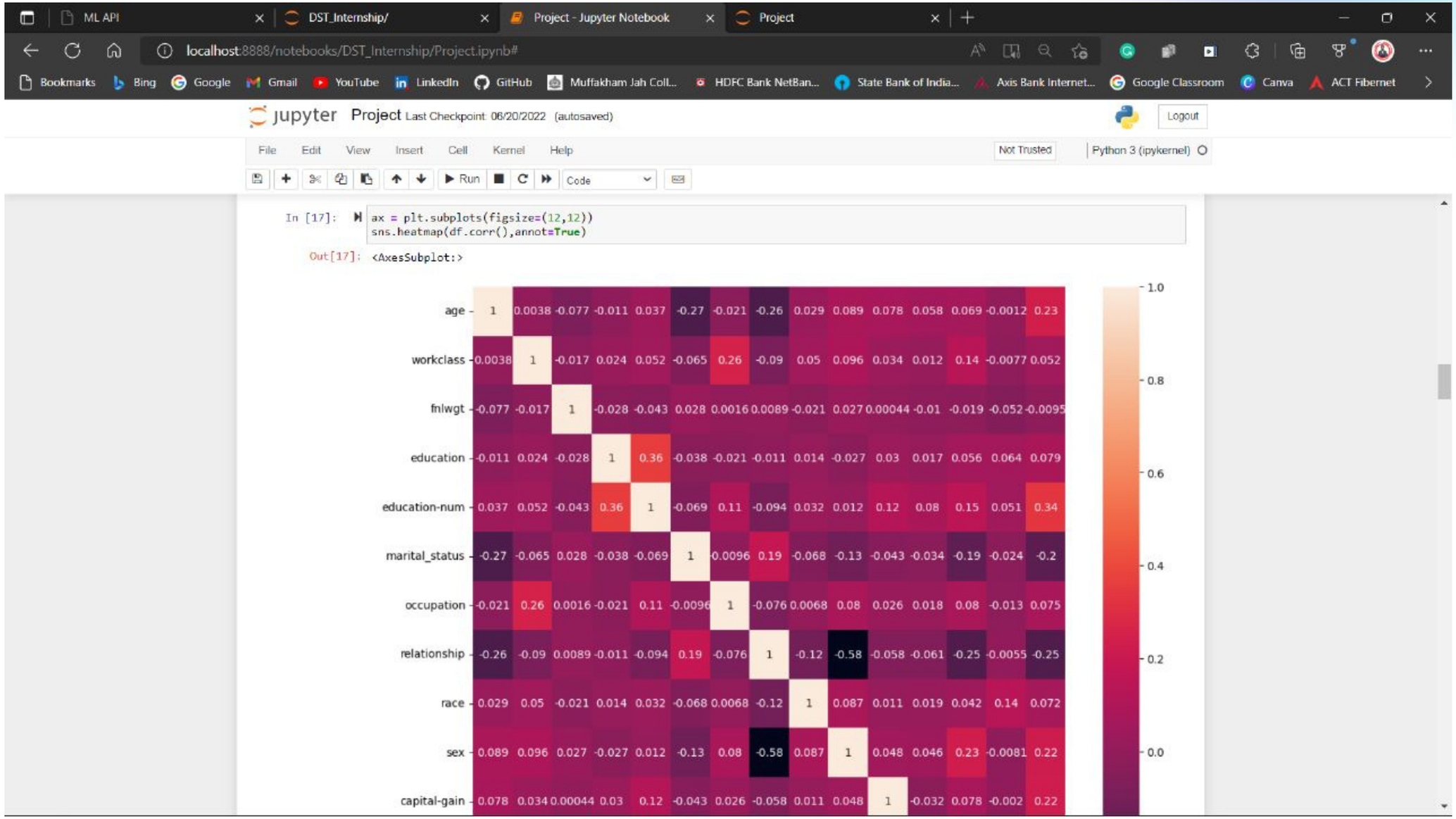


	age	workclass	fnlwgt	education	education-num	marital_status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	50	6	83311	9	13	2	4	0	4	1	0	0	13	39	0
1	38	4	215646	11	9	0	6	1	4	1	0	0	40	39	0
2	53	4	234721	1	7	2	6	0	2	1	0	0	40	39	0
3	28	4	338409	9	13	2	10	5	2	0	0	0	40	5	0
4	37	4	284582	12	14	2	4	5	4	0	0	0	40	39	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32555	27	4	257302	7	12	2	13	5	4	0	0	0	38	39	0
32556	40	4	154374	11	9	2	7	0	4	1	0	0	40	39	1
32557	58	4	151910	11	9	6	1	4	4	0	0	0	40	39	0
32558	22	4	201490	11	9	4	1	3	4	1	0	0	20	39	0
32559	52	5	287927	11	9	2	4	5	4	0	15024	0	40	39	1

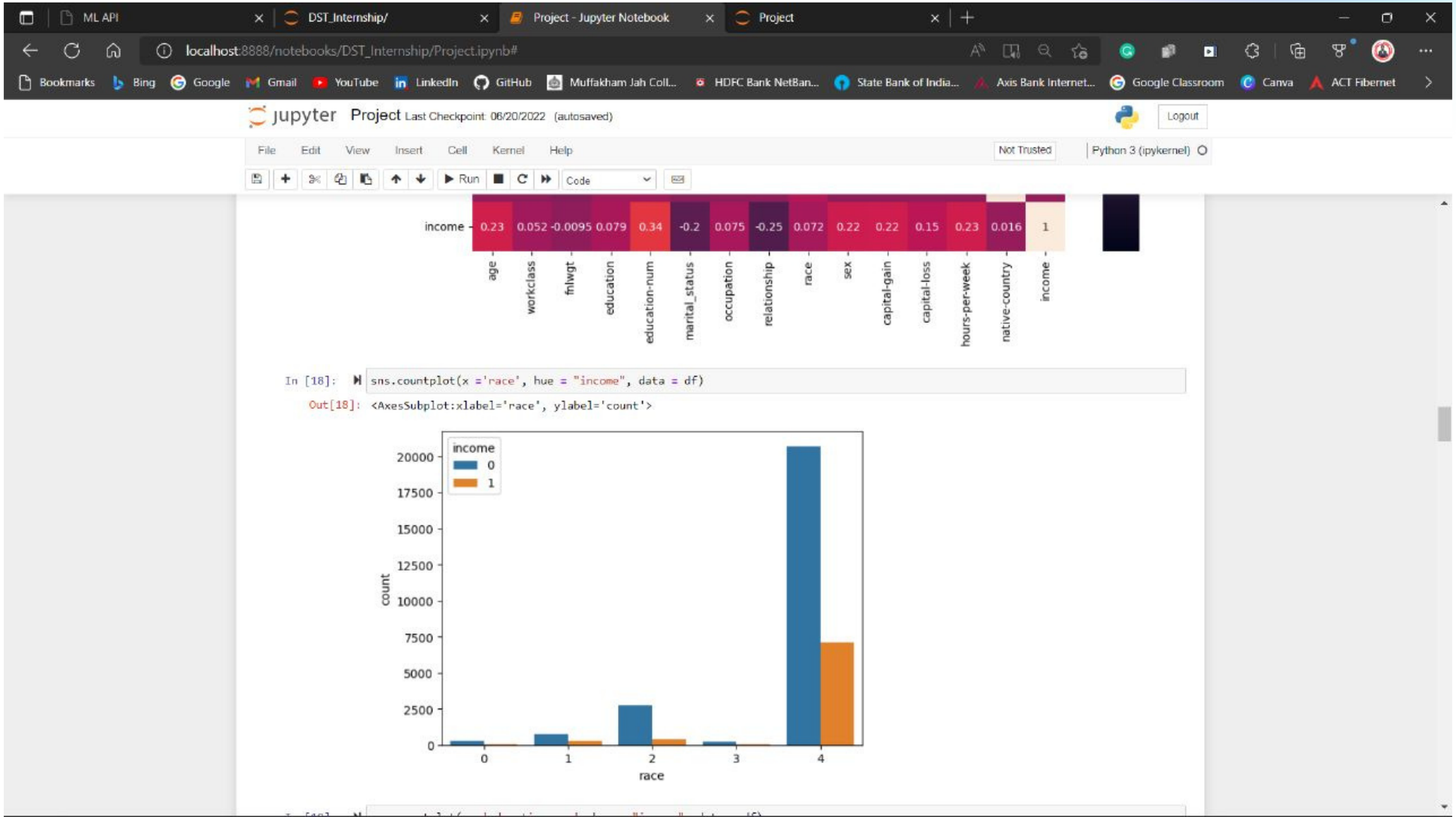
Data Visualization:



Data Visualization:



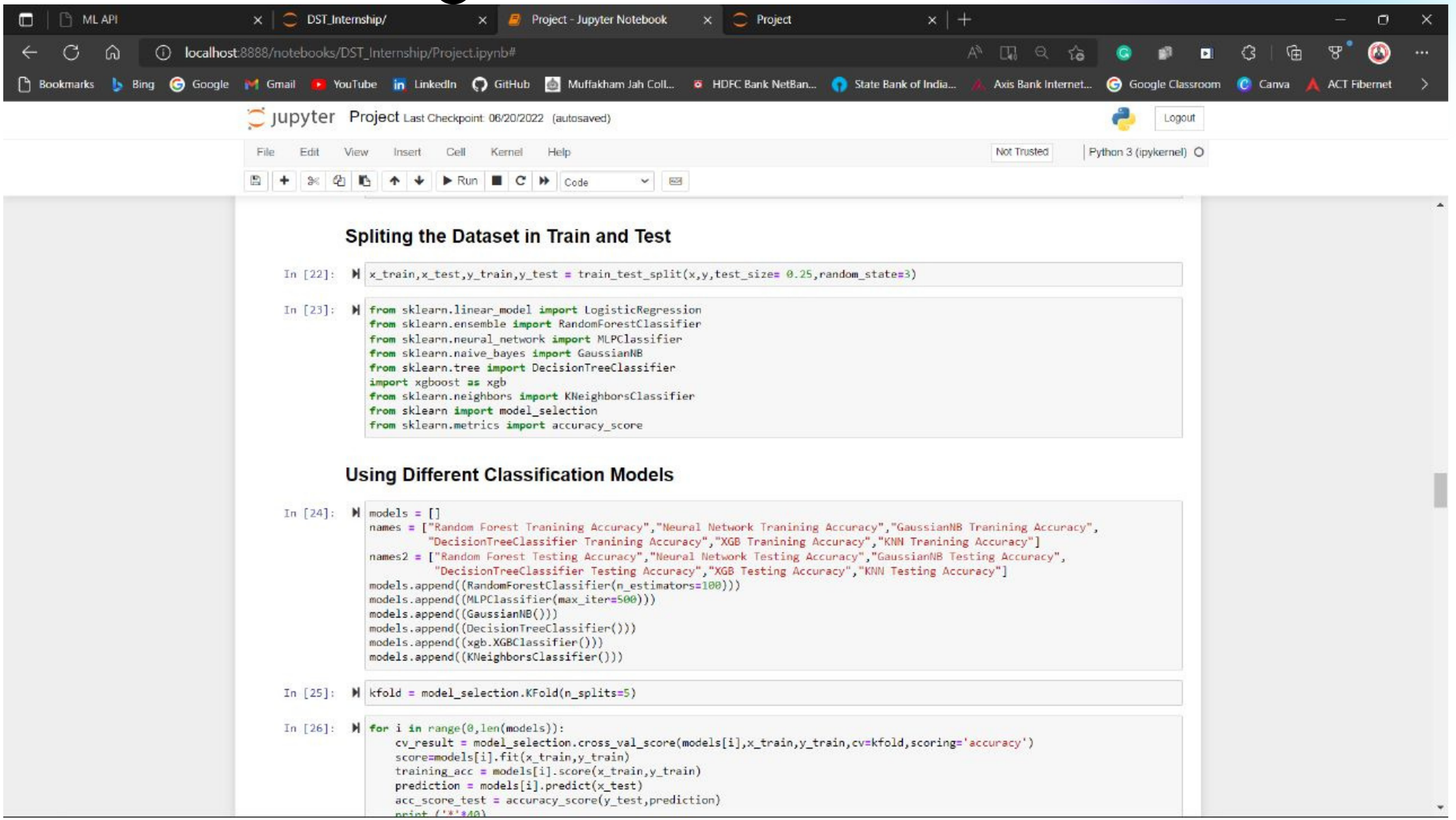
Data Visualization:



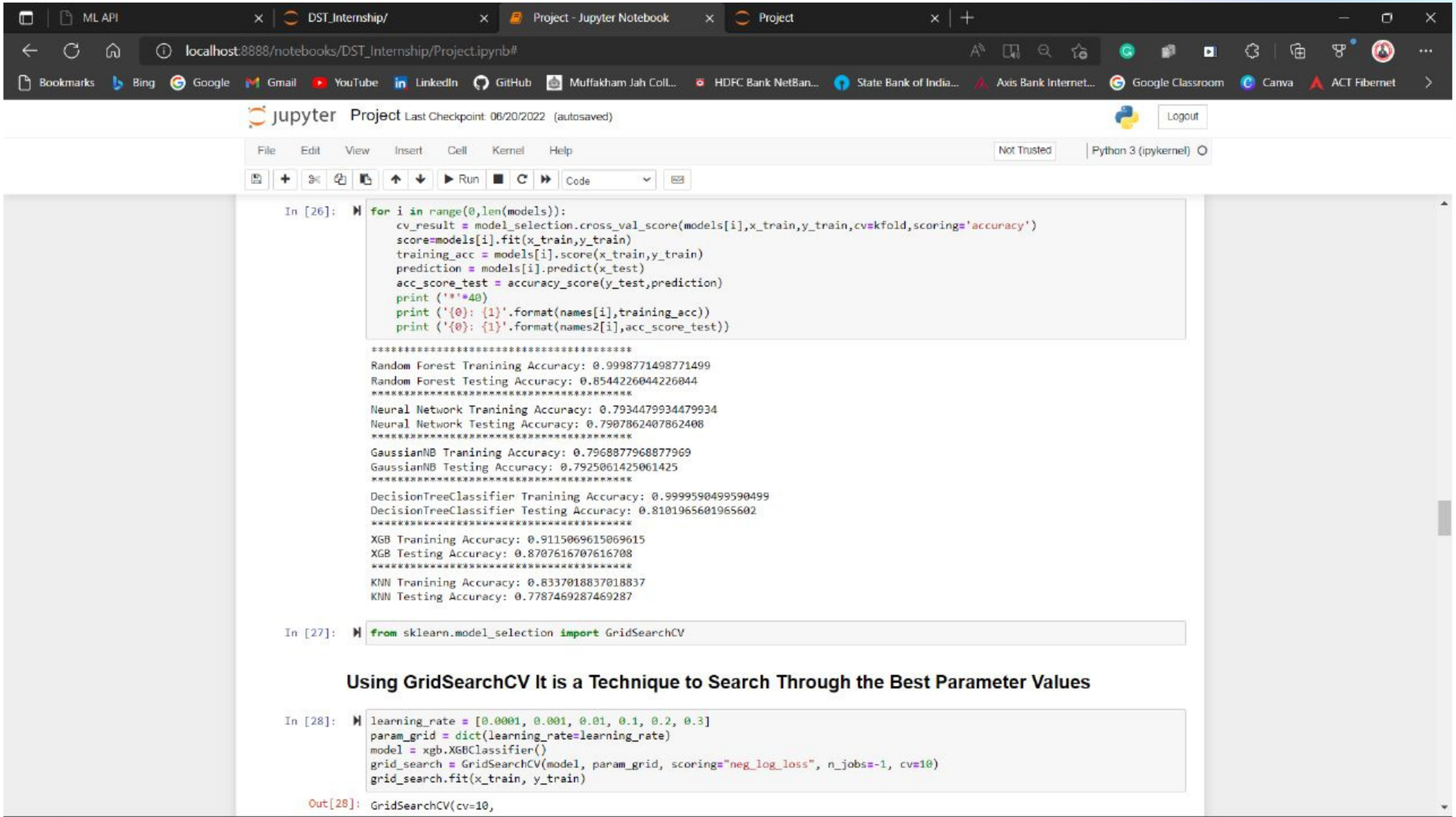
Model Building:

Model Building is done using different types of classification models.Using predefined train and test sets, reported good testing classification accuracy of about 87.30 percent by using XGB Classifier.
This might provide a target to aim for when working on this dataset.

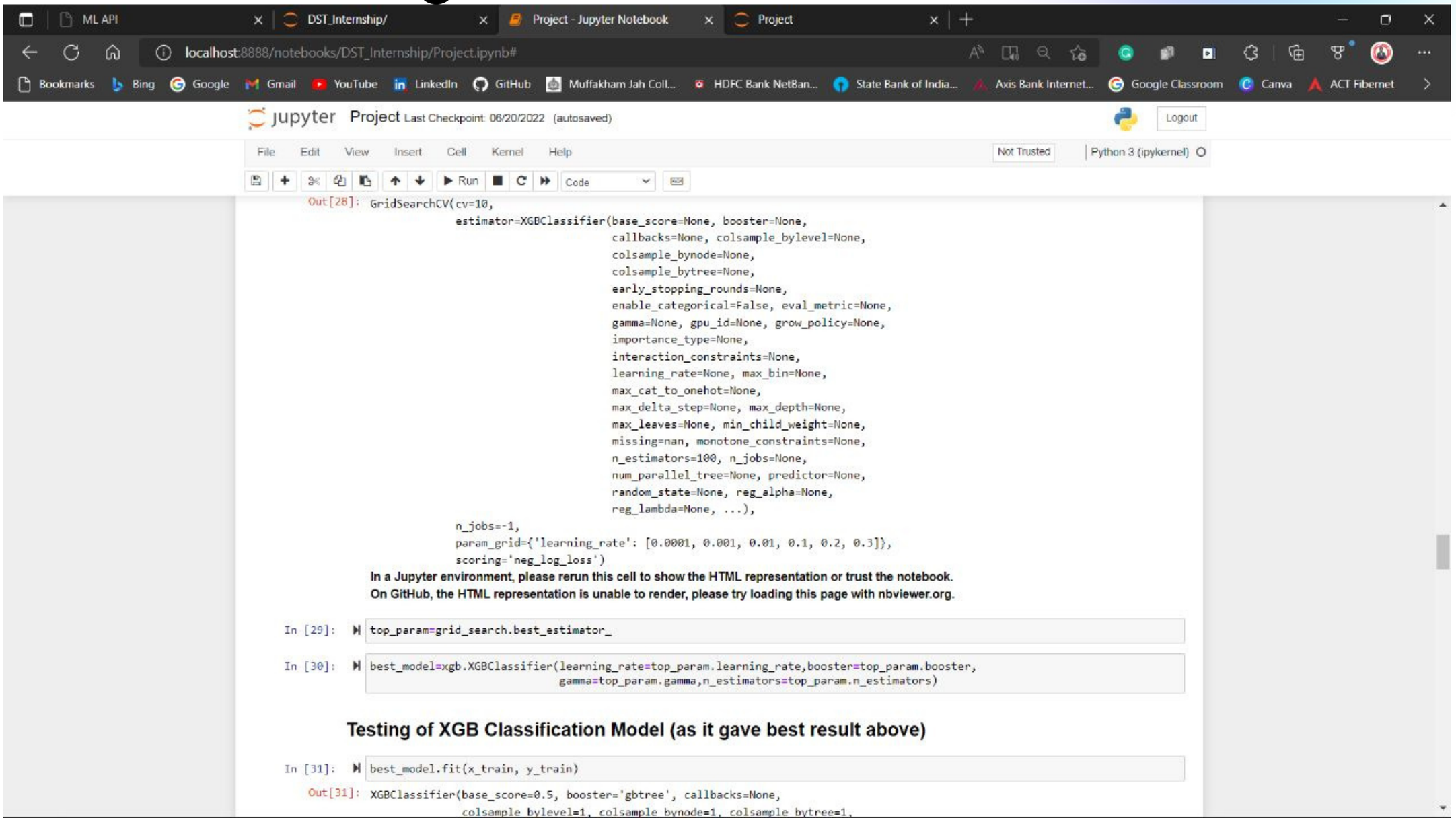
Model Building:



Model Building:



Model Building:

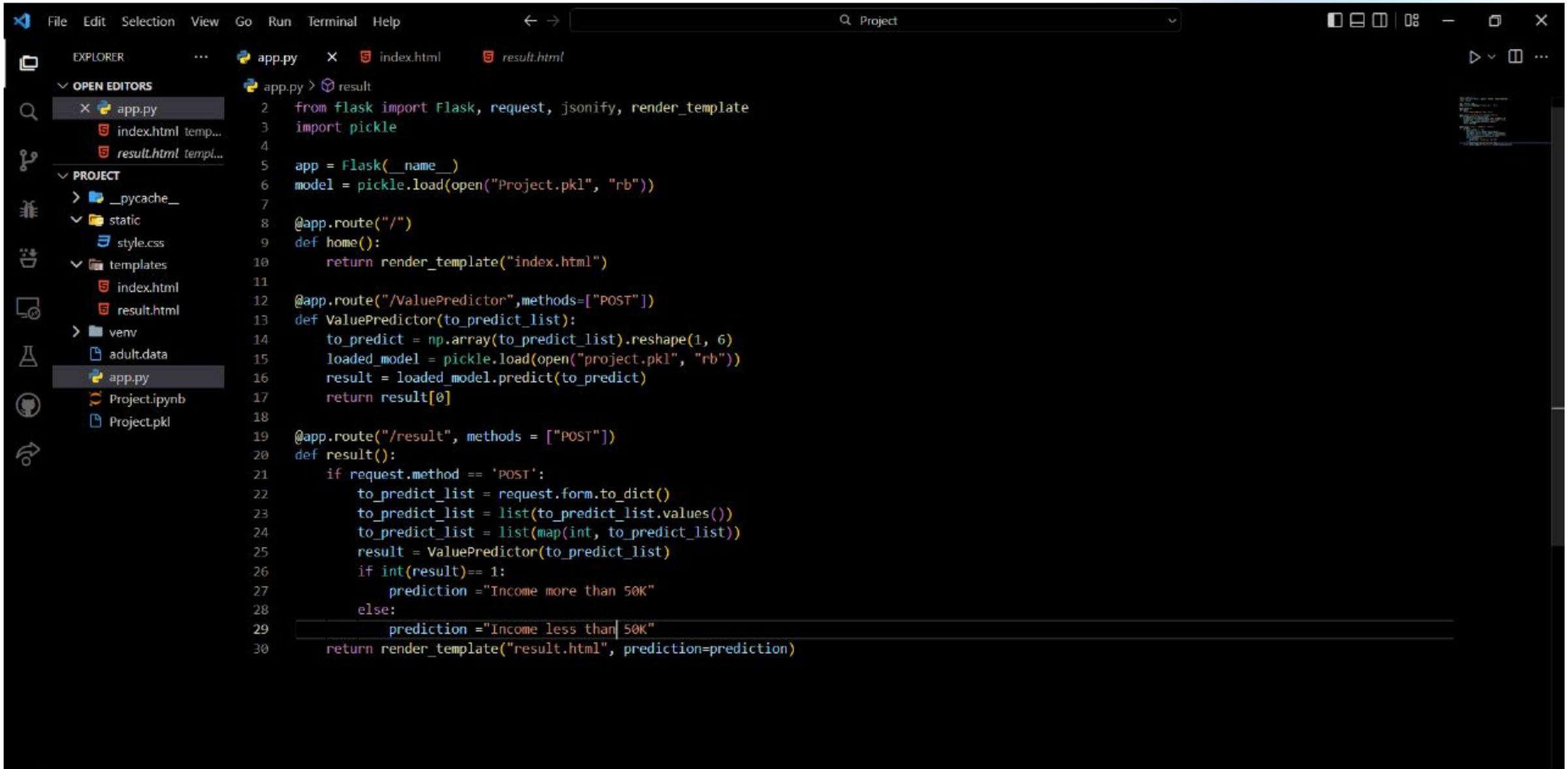


Conclusion:

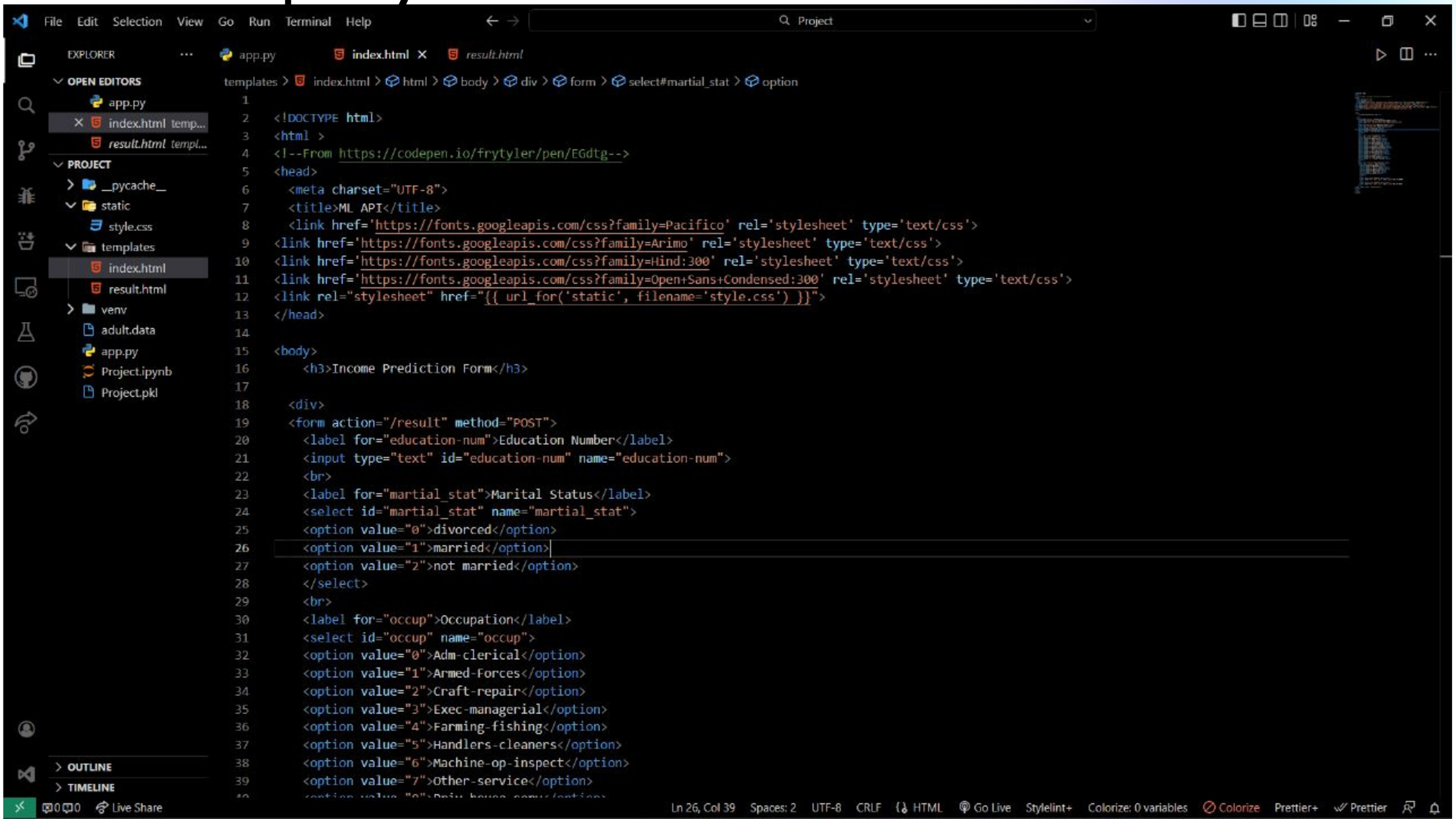
By looking above values we can say that education-num, marital status, relationship and capital gain are the features that are important for income prediction.

Model Deployment:

Model Deployment:

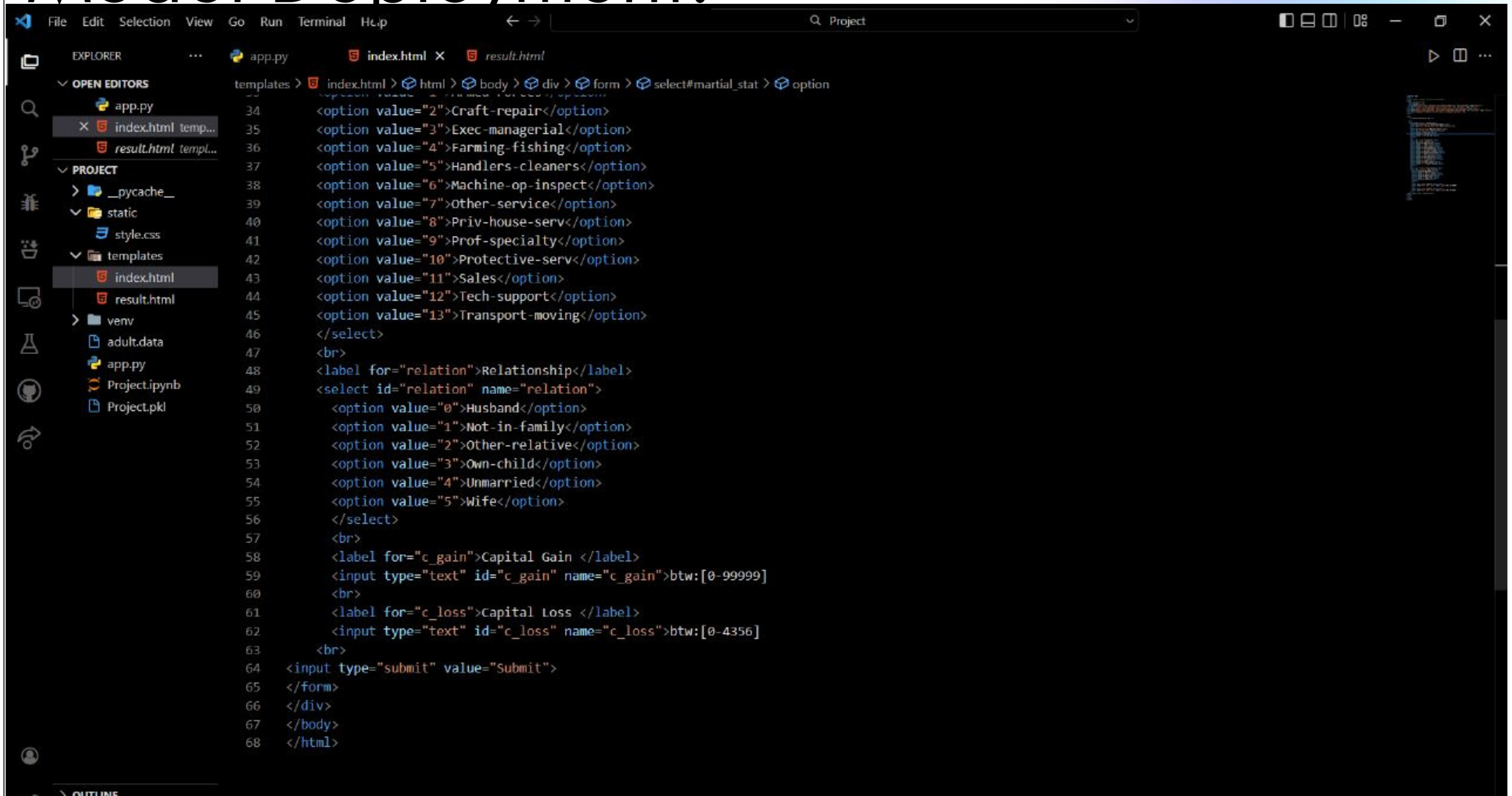


Model Deployment:

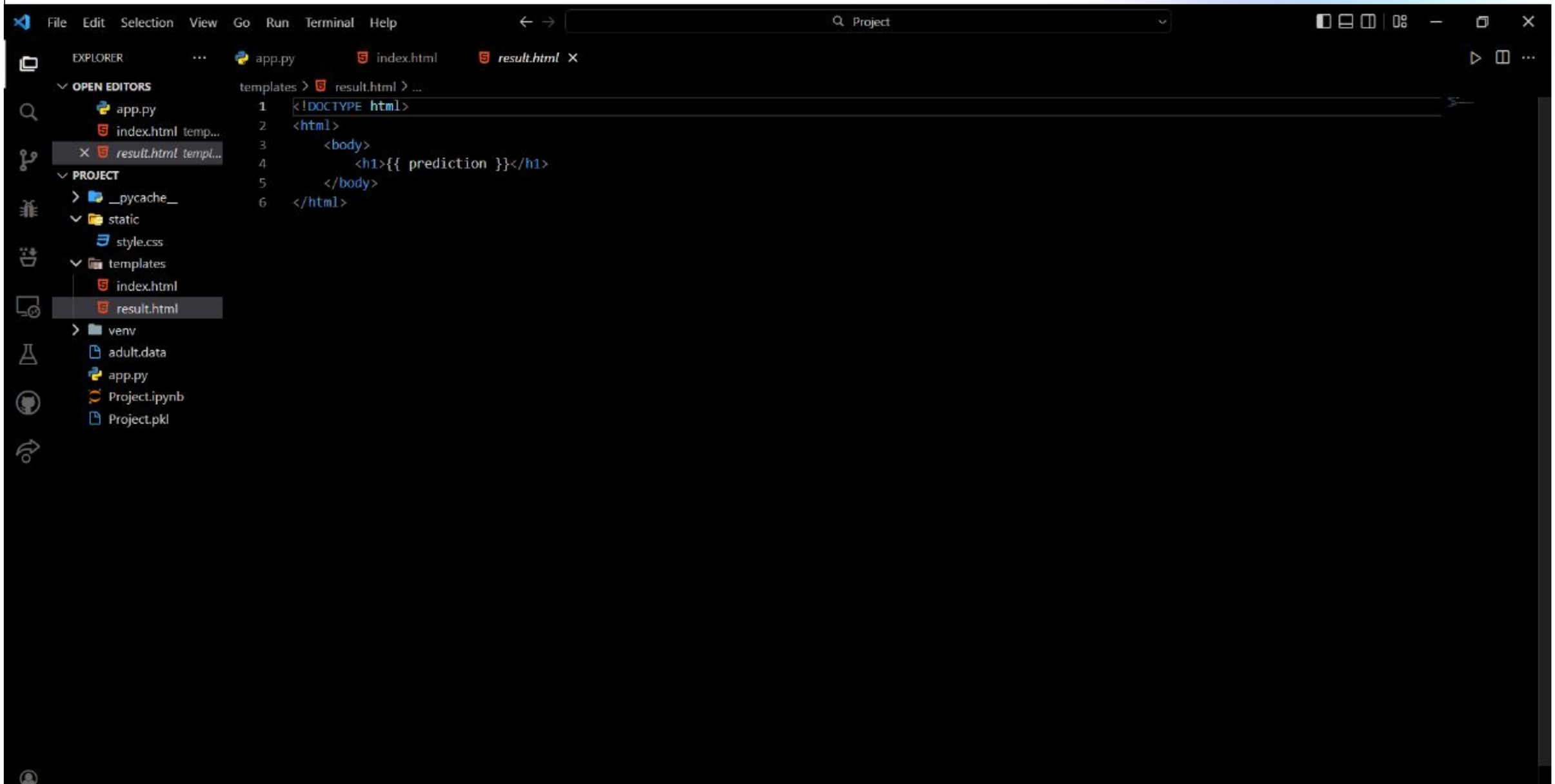



```
<!DOCTYPE html>
<html >
<!--From https://codepen.io/frytyler/pen/EGdtg-->
<head>
  <meta charset="UTF-8">
  <title>ML API</title>
  <link href='https://fonts.googleapis.com/css?family=Pacifico' rel='stylesheet' type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Arimo' rel='stylesheet' type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Hind:300' rel='stylesheet' type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Open+Sans+Condensed:300' rel='stylesheet' type='text/css'>
<link rel="stylesheet" href="{{ url_for('static', filename='style.css') }}">
</head>

<body>
    <h3>Income Prediction Form</h3>

  <div>
  <form action="/result" method="POST">
    <label for="education-num">Education Number</label>
    <input type="text" id="education-num" name="education-num">
    <br>
    <label for="martial_stat">Marital Status</label>
    <select id="martial_stat" name="martial_stat">
    <option value="0">divorced</option>
    <option value="1">married</option>
    <option value="2">not married</option>
    </select>
    <br>
    <label for="occup">Occupation</label>
    <select id="occup" name="occup">
    <option value="0">Adm-clerical</option>
    <option value="1">Armed-Forces</option>
    <option value="2">Craft-repair</option>
    <option value="3">Exec-managerial</option>
    <option value="4">Farming-fishing</option>
    <option value="5">Handlers-cleaners</option>
    <option value="6">Machine-op-inspect</option>
    <option value="7">Other-service</option>
```

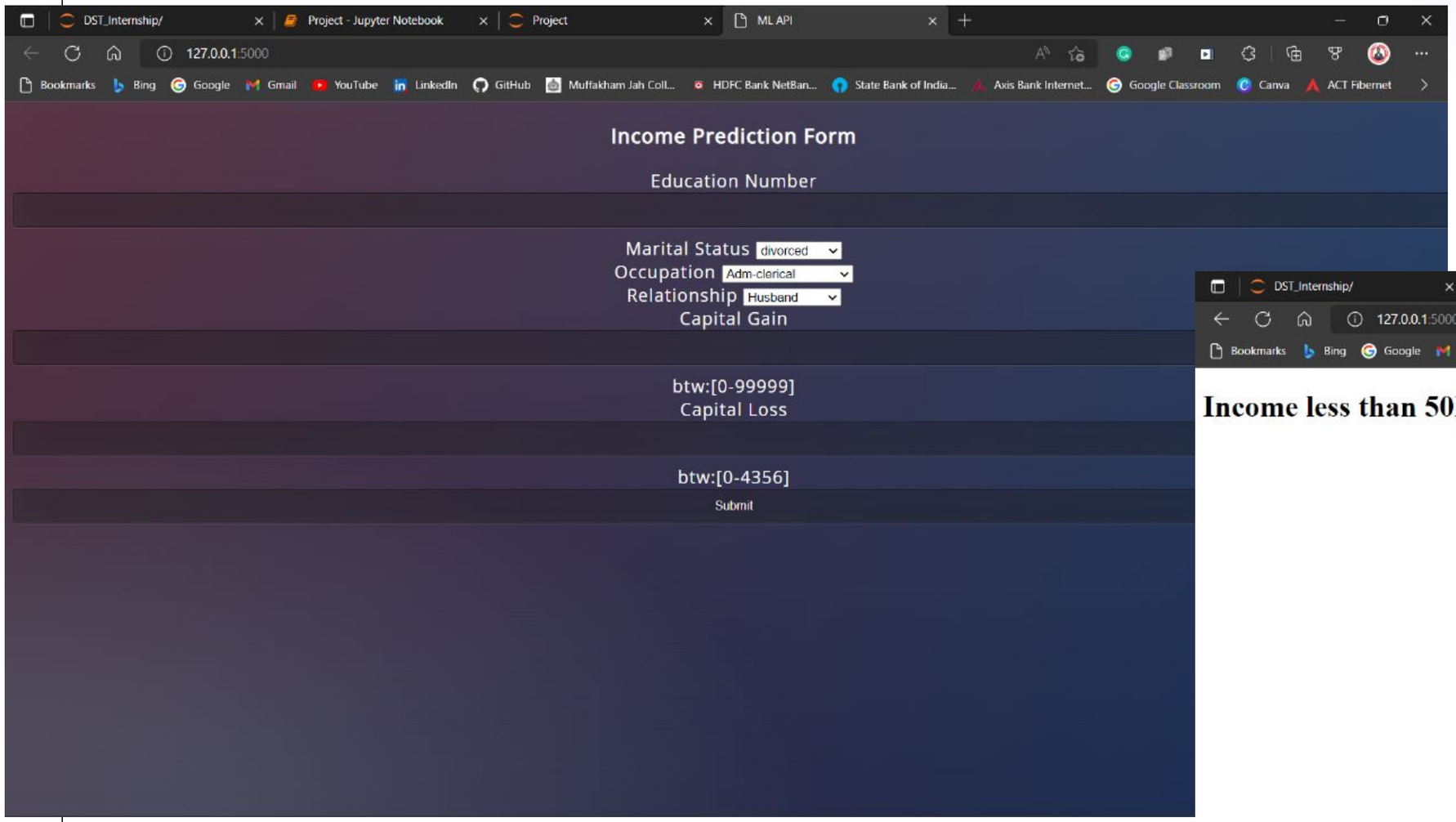
Model Deployment:

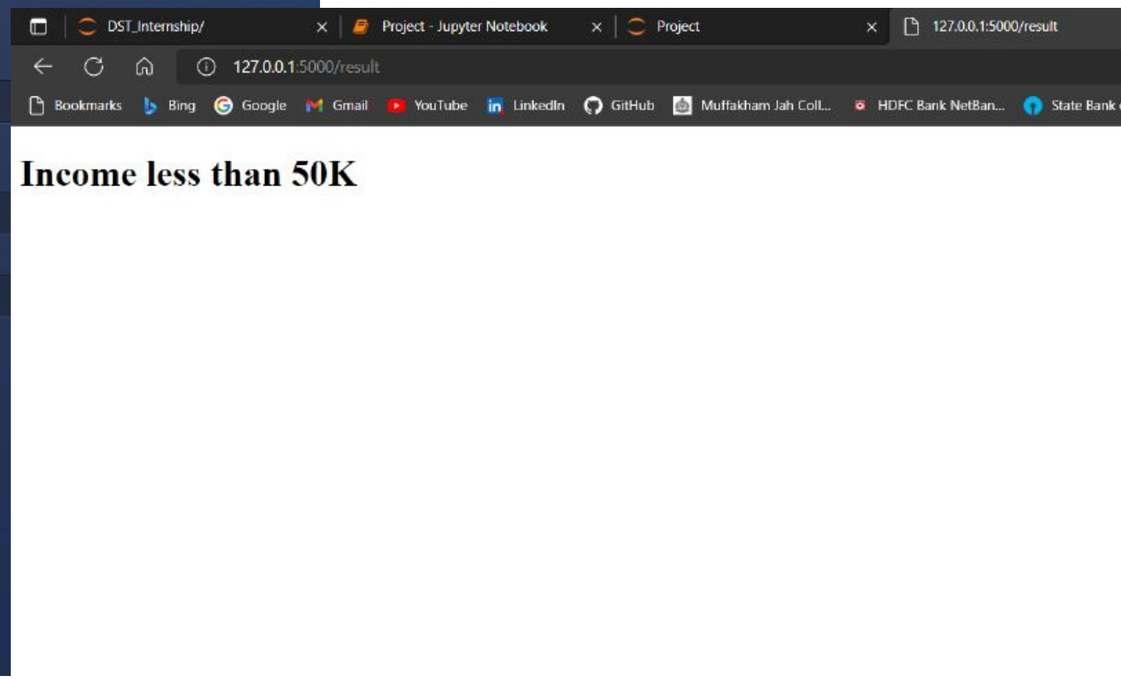


RESULT:


File Edit Selection View Go Run Terminal Help
Project
EXPLORER
app.py
index.html
result.html
OPEN EDITORS
app.py
index.html temp...
result.html templ...
PROJECT
__pycache__
static
style.css
templates
index.html
result.html
venv
adult.data
app.py
Project.ipynb
Project.pkl
templates > result.html > ...
1 <!DOCTYPE html>
2 <html>
3 <body>
4 <h1>{{ prediction }}</h1>
5 </body>
6 </html>

RESULT:


DST_Internship/
Project - Jupyter Notebook
Project
ML API
127.0.0.1:5000
Bookmarks
Bing
Google
Gmail
YouTube
LinkedIn
GitHub
Muffakham Jah Coll...
HDFC Bank NetBan...
State Bank of India...
Axis Bank Internet...
Google Classroom
Canva
ACT Fibernet
Income Prediction Form
Education Number
Marital Status
divorced
Occupation
Adm-clerical
Relationship
Husband
Capital Gain
btw:[0-99999]
Capital Loss
btw:[0-4356]
Submit


DST_Internship/
Project - Jupyter Notebook
Project
127.0.0.1:5000/result
127.0.0.1:5000/result
Bookmarks
Bing
Google
Gmail
YouTube
LinkedIn
GitHub
Muffakham Jah Coll...
HDFC Bank NetBan...
Income less than 50K