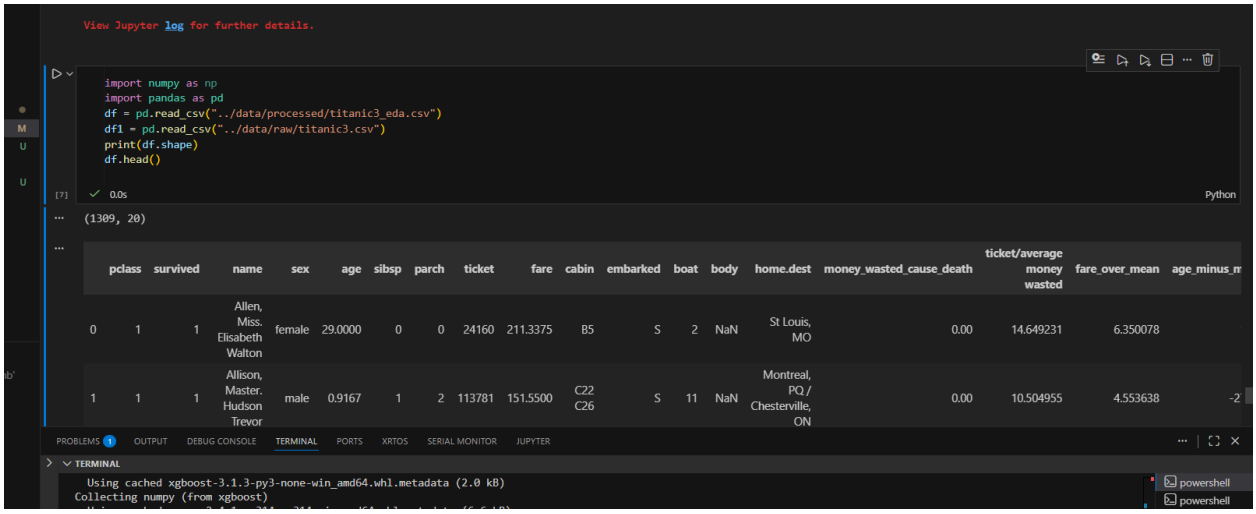
Week 1 Progress Report – Machine Learning Foundations

Days 1–2: NumPy & Pandas

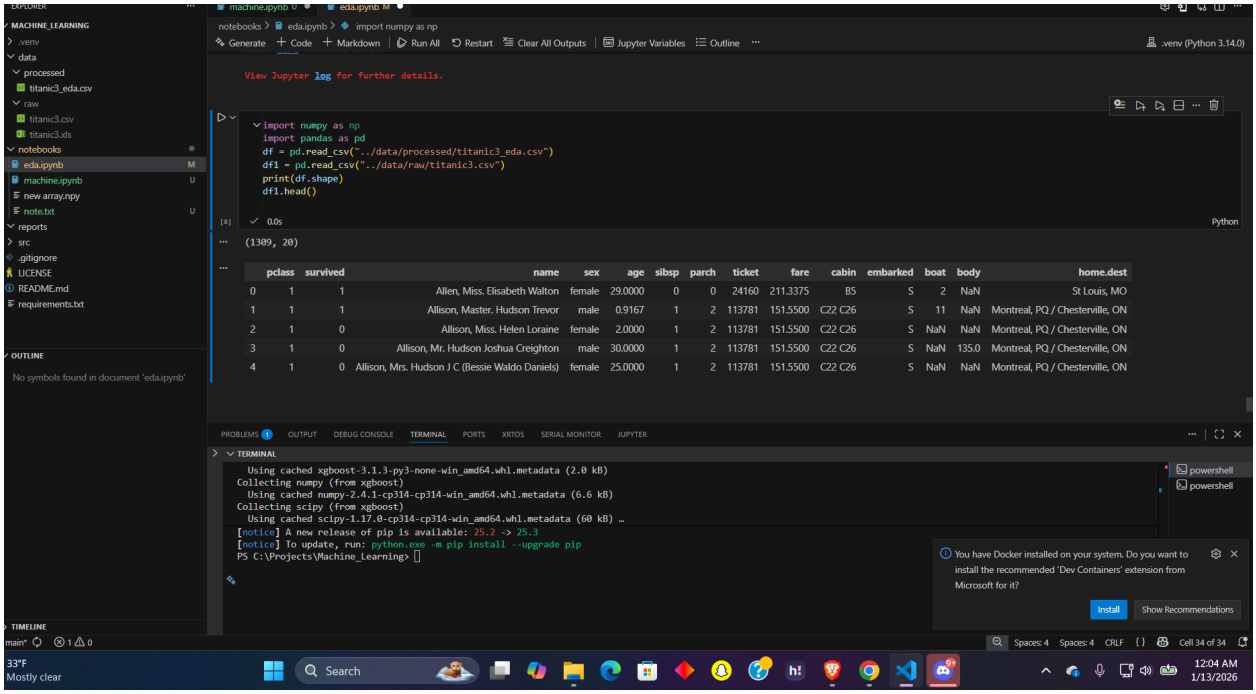
Loaded and explored the raw Titanic dataset using Pandas.
Added new column features, including a binary `is_child` feature derived from age.
Practiced vectorized operations, boolean masking, and conditional column creation without loops.
Partitioned between, processed, or raw, made the attributes non-null,
Was successfully able to broadcast, DataFrame manipulation, handling missing values.

Learning outcome: using numPy and Pandas to manipulate data, create some new columns

processed:



Raw:

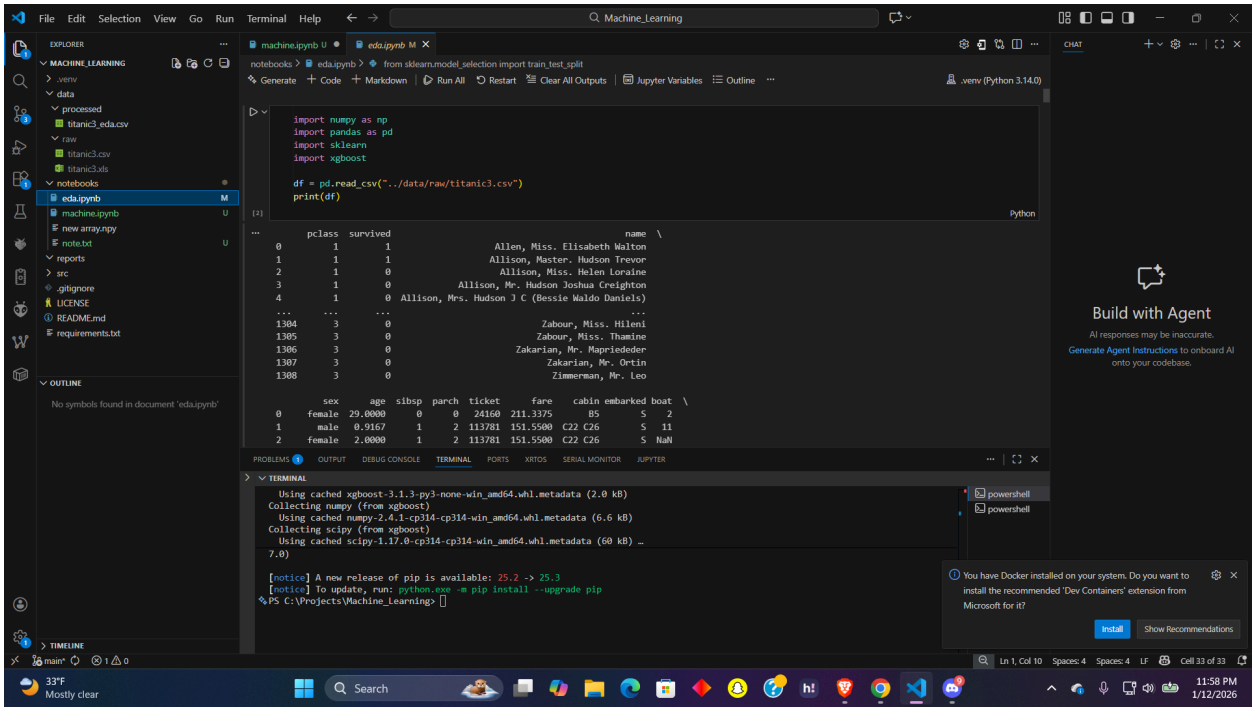


```
import numpy as np
import pandas as pd
df = pd.read_csv("../data/processed/titanic3_eda.csv")
df1 = pd.read_csv("../data/raw/titanic3.csv")
print(df.shape)
df1.head()
```

(1309, 20)

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1	1	Allen, Miss. Elisabeth Walton	female	29.0000	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1	0	Allison, Miss. Helen Loraine	female	2.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON

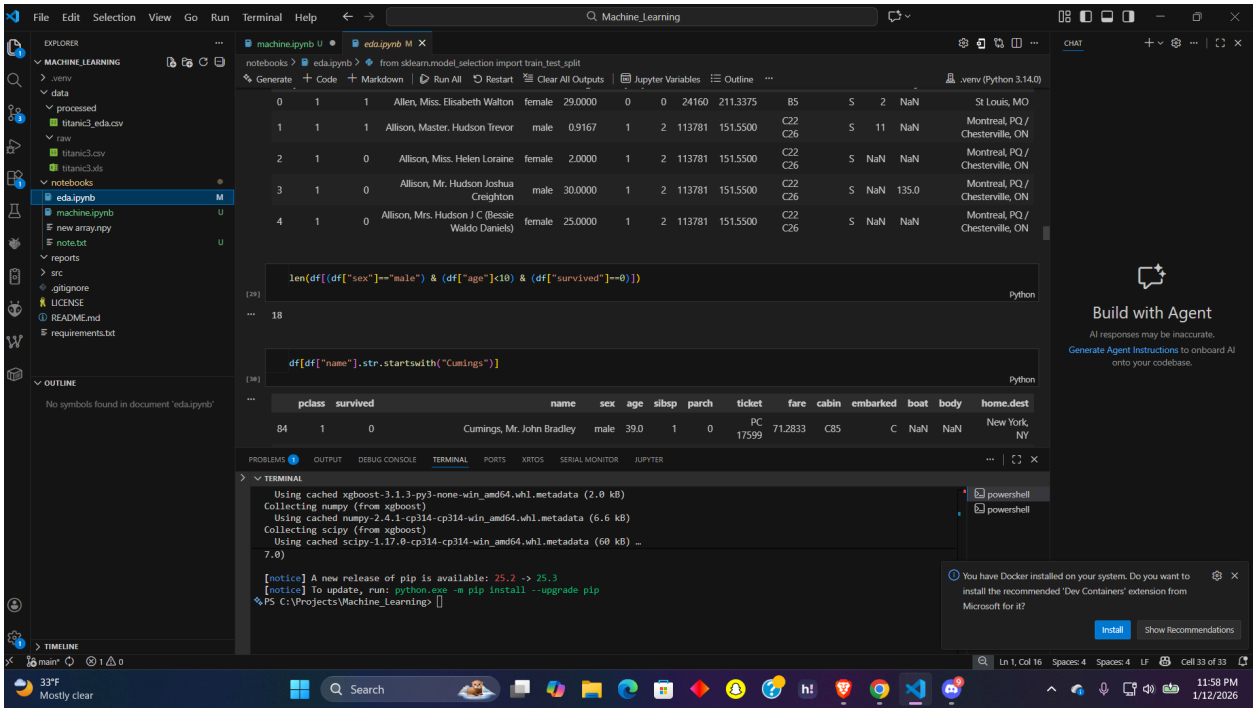
```
Using cached xgboost-3.1.3-py3-none-win_amd64.whl.metadata (2.0 kB)
Collecting numpy (from xgboost)
  Using cached numpy-2.4.1-cp314-cp314-win_amd64.whl.metadata (6.6 kB)
Collecting scipy (from xgboost)
  Using cached scipy-1.17.0-cp314-cp314-win_amd64.whl.metadata (60 kB) …
[notice] A new release of pip is available: 25.2 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
PS C:\Projects\Machine_Learning>
```



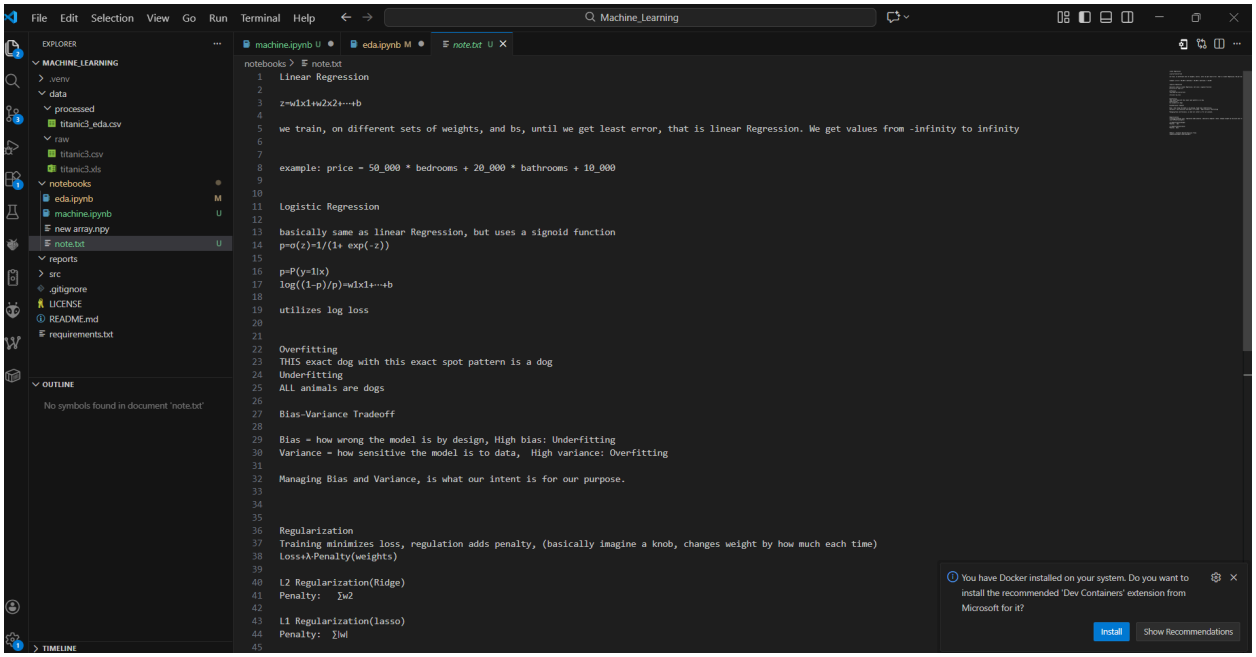
```
import numpy as np
import pandas as pd
import sklearn
import xgboost

df = pd.read_csv("../data/raw/titanic3.csv")
print(df)
```

```
      pclass  survived                                             name \
0          1         1                    Allen, Miss. Elisabeth Walton
1          1         1                   Allison, Master. Hudson Trevor
2          1         0                     Allison, Miss. Helen Loraine
3          1         0             Allison, Mr. Hudson Joshua Creighton
4          1         0  Allison, Mrs. Hudson J C (Bessie Waldo Daniels)
...      ...       ...                                              ...
1304       3         0                               Zabour, Miss. Hileni
1305       3         0                              Zabour, Miss. Thamine
1306       3         0                          Zakarian, Mr. Mapriededer
1307       3         0                                Zakarian, Mr. Ortin
1308       3         0                              Zimmerman, Mr. Leo

          sex      age  sibsp  parch  ticket      fare    cabin embarked boat  \
0      female  29.0000      0      0   24160  211.3375       B5        S    2
1        male   0.9167      1      2  113781  151.5500  C22 C26        S   11
2      female   2.0000      1      2  113781  151.5500  C22 C26        S  NaN
```

```
Using cached xgboost-3.1.3-py3-none-win_amd64.whl.metadata (2.0 kB)
Collecting numpy (from xgboost)
  Using cached numpy-2.4.1-cp314-cp314-win_amd64.whl.metadata (6.6 kB)
Collecting scipy (from xgboost)
  Using cached scipy-1.17.0-cp314-cp314-win_amd64.whl.metadata (60 kB) …
7.0)

[notice] A new release of pip is available: 25.2 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
PS C:\Projects\Machine_Learning>
```



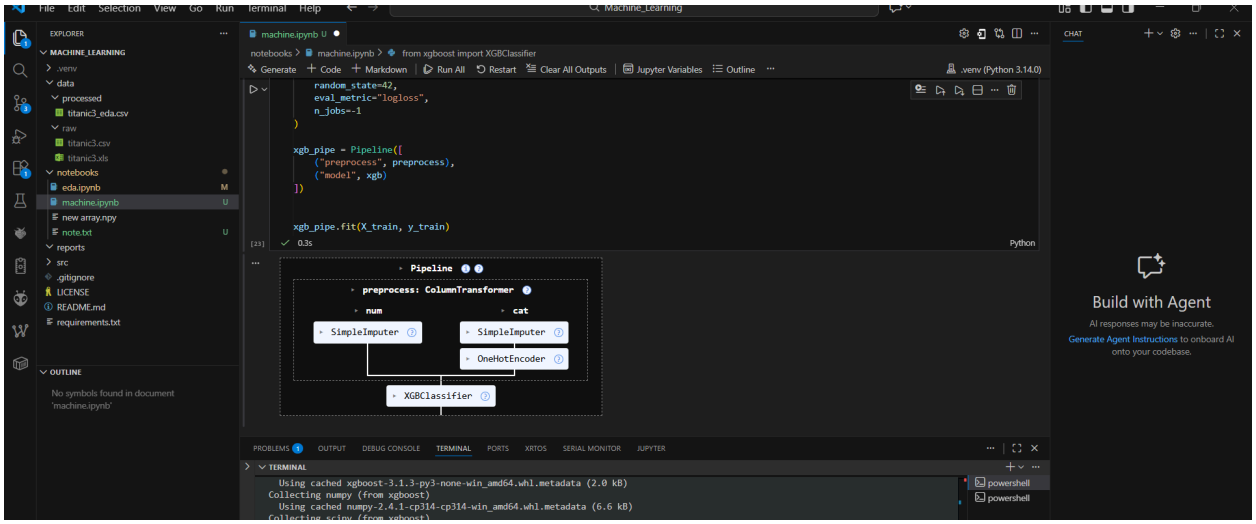
Days 3–4: (Scikit-Learn)
Did study more on machine learning, about logistic regression and linear regression

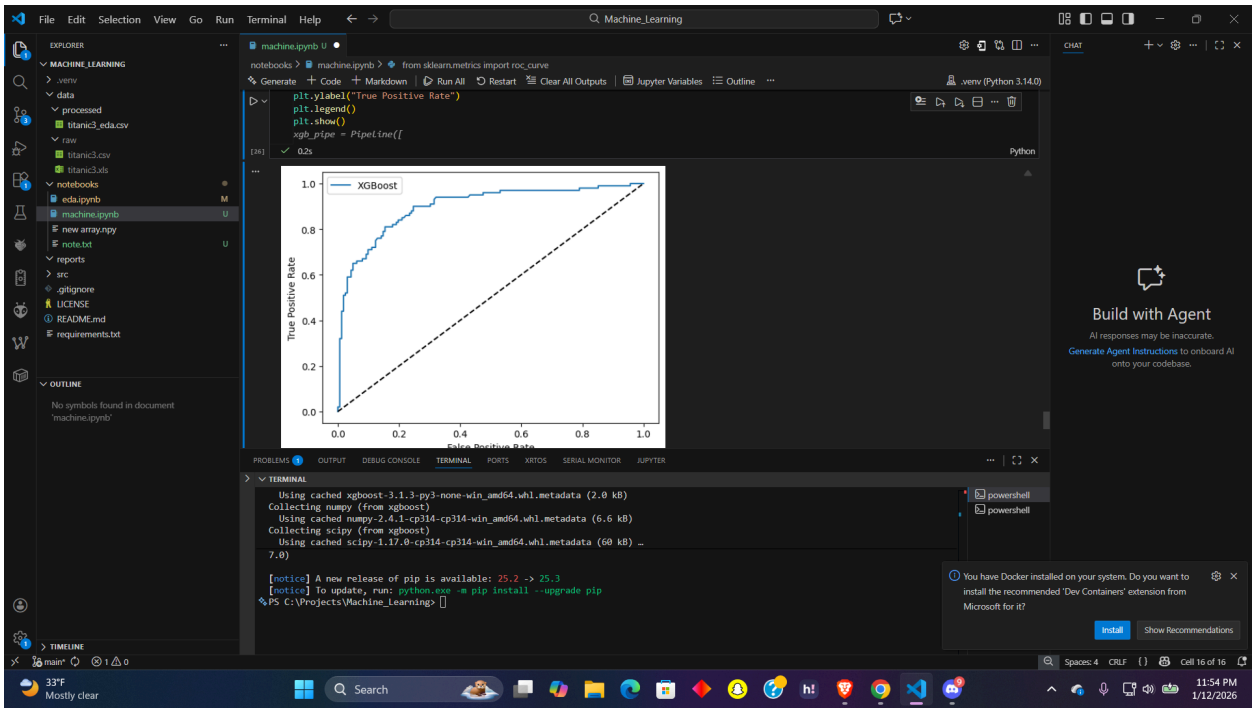


Learned about regulation, L1 and L2, variance, overfitting and underfitting,
I also learned about Bias-Variance Tradeoff, I also learned about generalization and why we must keep some data for testing separate from training data.

Days 5–6: XGBoost

Implemented an **XGBoost classifier** using a Scikit-Learn Pipeline
Trained an XGBoost model on the Titanic dataset.
Plotted the ROC curve and verified strong separation from random guessing





Day 7: Evaluation

Compared baseline models and XGBoost using probabilistic outputs
Learned why accuracy is misleading for imbalanced datasets.
Learned why accuracy is misleading for imbalanced datasets.
ROC-AUC measures ranking ability, not just correctness.