BITS Pilani

API-driven Cloud Native Solutions (CCZG506)

Assignment - 1

Group -28

Submitted By

Student Name - ID

M KAMAL KUMAR
2023mt03153@wilp.bits-pilani.ac.in

SHANMUKHI LAKSHMI SIDWINI
2023mt03194@wilp.bits-pilani.ac.in

KAKANI VARSHITHA
2023mt03002@wilp.bits-pilani.ac.in

JALAMANCHILI RAMA SURYAM
2023mt03101@wilp.bits-pilani.ac.in

KAPADIYA KRUNALKUMAR GHANSHYAMBHAI
2023mt03176@wilp.bits-pilani.ac.in

Submitted To

Prof. Shreyas Suresh Rao

Oct 31, 2024

SL.NO	BITS ID	NAME	CONTRIBUTION OF TEAM MEMBERS (Qualitative)	PERCENTAGE CONTRIBUTION (Out of 100) QUANTITATIVE
1	**2023mt03153**	M KAMAL KUMAR	**Datapipeline,prefectexecution, Sagemaker Execution, Video making**	**30%**
2	**2023mt03002**	KAKANI VARSHITHA	**MLflow execution, Documentation**	**20%**
3	**2023mt03194**	SHANMUKHI LAKSHMI SIDWINI	**Dataset identification, Documentation**	**20%**
4	**2023mt03176**	KAPADIYA KRUNALKUMAR GHANSHYAMBHAI	**Assignment flow planning, ML pipeline, ML models**	**15%**
5	**2023mt03101**	JALAMANCHILI RAMA SURYAM	**MLops execution, Documentation**	**15%**

Sub-Objective 1: Design and Development of a Data Pipeline

Prerequisites: -

1) Python
2) Prefect Cloud
3) MLflows
4) SageMaker

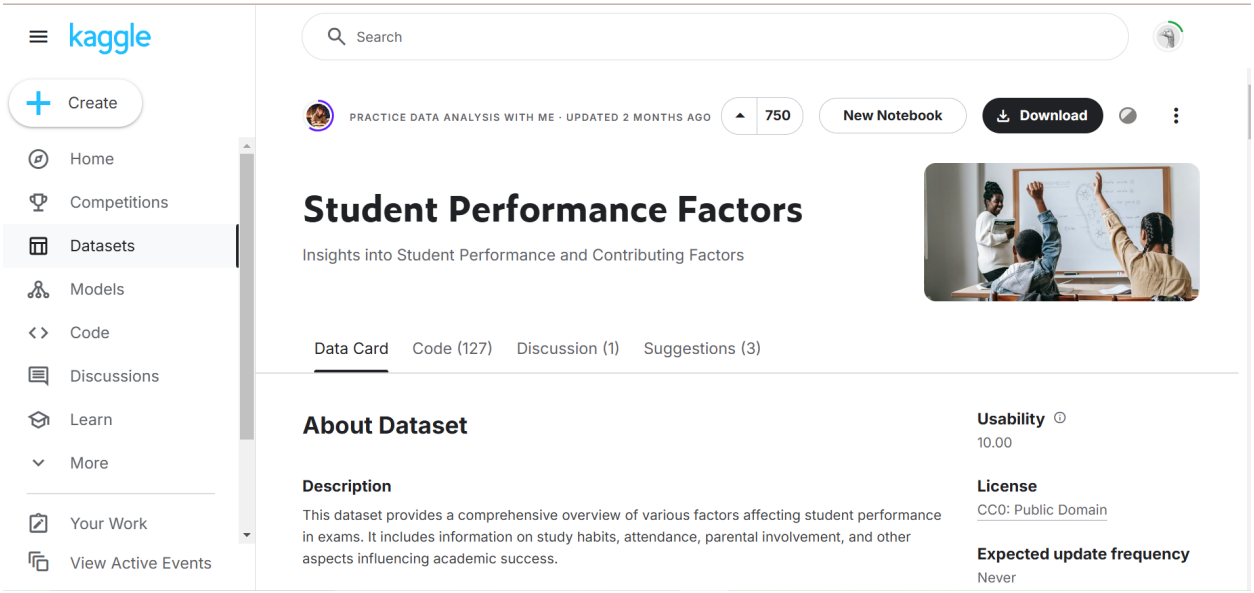
1.1 Business Understanding: Identify a business problem where we analysed the Factors affecting the Student Performance in the area of data science.

Using cloud-based tools for data science and machine learning enables efficient, scalable analysis of datasets, such as examining factors impacting student performance. Here's how each tool can play a role in this process:

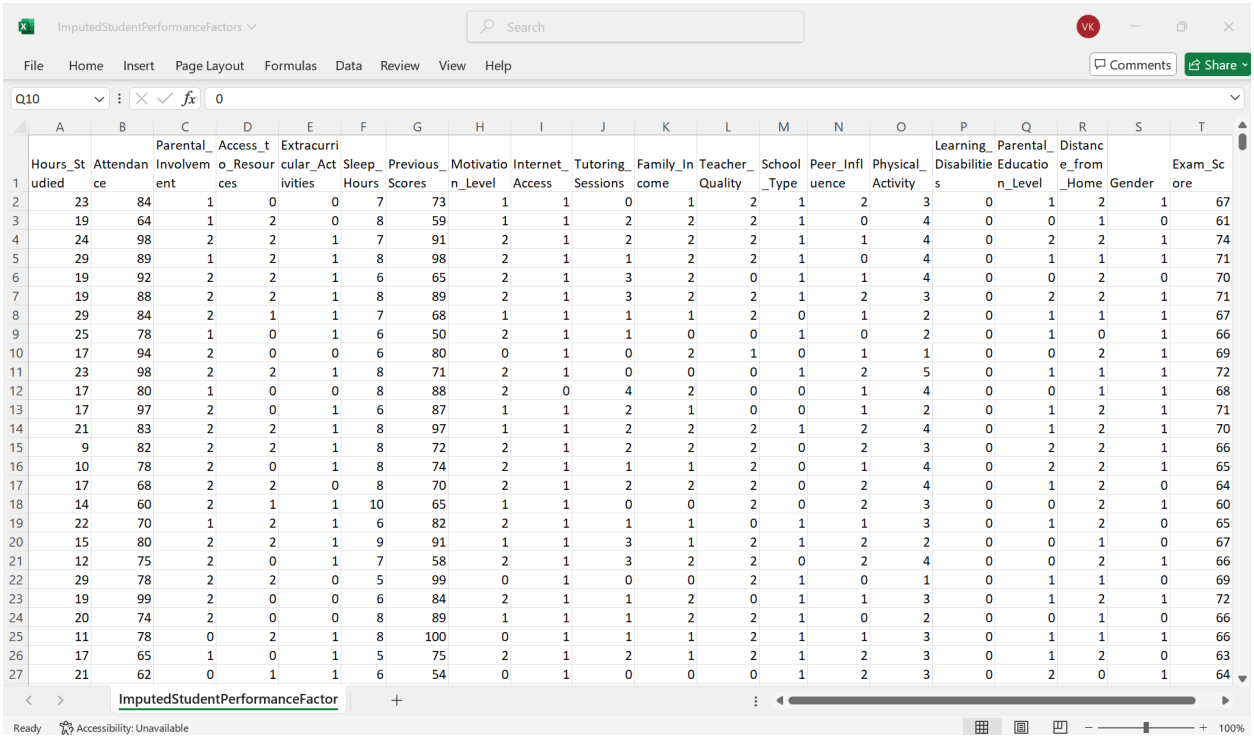
1. **Prefect Cloud:** Prefect allows for orchestrating and scheduling complex workflows. For this project, it can manage data ingestion, transformation, and preprocessing pipelines, automating steps like data cleaning, feature engineering, and model retraining. Prefect Cloud ensures workflows run reliably and are resilient to errors, with all runs and logs easily monitored.
2. **MLflow:** MLflow facilitates the tracking and management of machine learning experiments. You can log metrics, parameters, and models across various experiments and versions, making it easier to track model performance and hyperparameter tuning as you test different predictive approaches for student outcomes.
3. **Amazon SageMaker:** SageMaker enables the building, training, and deployment of machine learning models at scale. It provides pre-built algorithms and manages resources for training models quickly. For student performance prediction, SageMaker could deploy your model as a web service, allowing for real-time predictions, such as identifying at-risk students.
4. **Python**: Python is central to this process for data manipulation (using libraries like Pandas), exploratory data analysis (Matplotlib, Seaborn), feature engineering, and model building (Scikit-learn, TensorFlow, or PyTorch). Python scripts can be integrated into Prefect workflows and MLflow for full pipeline management.

Together, these tools streamline the end-to-end process—from data preprocessing to model deployment—offering a scalable, reproducible solution for analyzing factors that influence student academic success.

1.2 Data Ingestion: For the identified problem, we found an appropriate dataset from a public repository which is Kaggle.

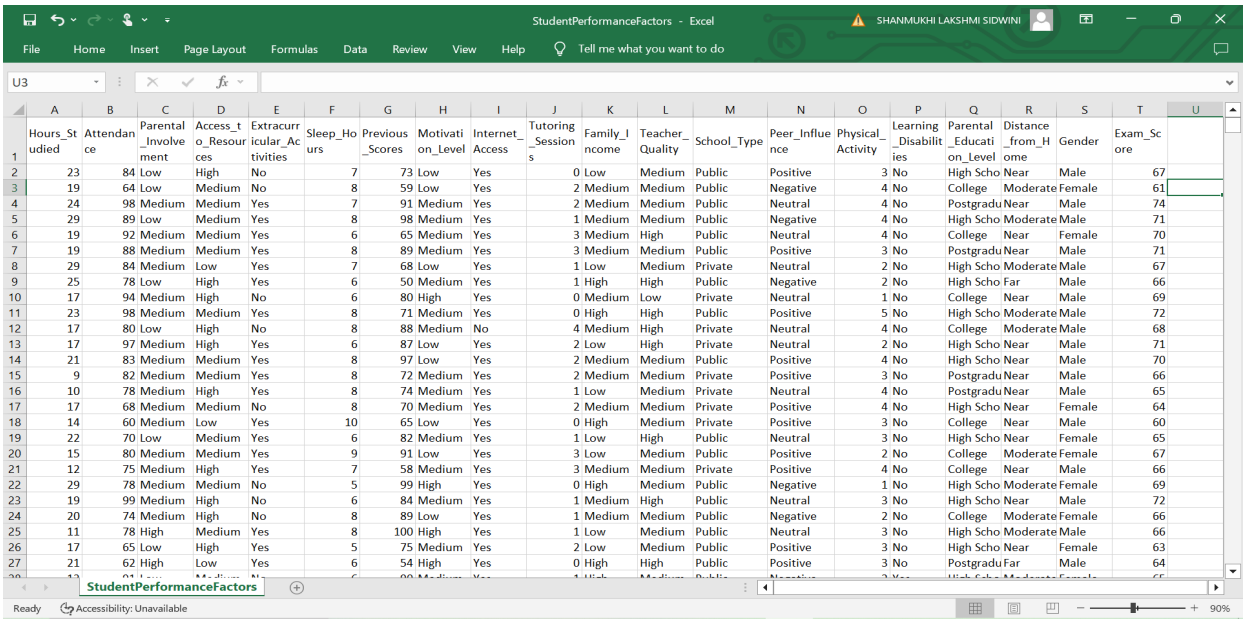


ImputedStudentPerformanceFactors.csv: -



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	Sleep_Hours	Previous_Scores	Motivation_Level	Internet_Access	Tutoring_Sessions	Family_Income	Teacher_Quality	School_Type	Peer_Influence	Physical_Activity	Learning_Disabilities	Parental_Education_Level	Distance_from_Home	Gender	Exam_Score
2	23	84	1	0	0	7	73	1	1	0	1	2	1	2	3	0	1	2	1	67
3	19	64	1	2	0	8	59	1	1	2	2	2	1	0	4	0	0	1	0	61
4	24	98	2	2	1	7	91	2	1	2	2	2	1	1	4	0	2	2	1	74
5	29	89	1	2	1	8	98	2	1	1	2	2	1	0	4	0	1	1	1	71
6	19	92	2	2	1	6	65	2	1	3	2	0	1	1	4	0	0	2	0	70
7	19	88	2	2	1	8	89	2	1	3	2	2	1	2	3	0	2	2	1	71
8	29	84	2	1	1	7	68	1	1	1	1	2	0	1	2	0	1	1	1	67
9	25	78	1	0	1	6	50	2	1	1	0	0	1	0	2	0	1	0	1	66
10	17	94	2	0	0	6	80	0	1	0	2	1	0	1	1	0	0	2	1	69
11	23	98	2	2	1	8	71	2	1	0	0	0	1	2	5	0	1	1	1	72
12	17	80	1	0	0	8	88	2	0	4	2	0	0	1	4	0	0	1	1	68
13	17	97	2	0	1	6	87	1	1	2	1	0	0	1	2	0	1	2	1	71
14	21	83	2	2	1	8	97	1	1	2	2	2	1	2	4	0	1	2	1	70
15	9	82	2	2	1	8	72	2	1	2	2	2	0	2	3	0	2	2	1	66
16	10	78	2	0	1	8	74	2	1	1	1	2	0	1	4	0	2	2	1	65
17	17	68	2	2	0	8	70	2	1	2	2	2	0	2	4	0	1	2	0	64
18	14	60	2	1	1	10	65	1	1	0	0	2	0	2	3	0	0	2	1	60
19	22	70	1	2	1	6	82	2	1	1	1	0	1	1	3	0	1	2	0	65
20	15	80	2	2	1	9	91	1	1	3	1	2	1	2	2	0	0	1	0	67
21	12	75	2	0	1	7	58	2	1	3	2	2	0	2	4	0	0	2	1	66
22	29	78	2	2	0	5	99	0	1	0	0	2	1	0	1	0	1	1	0	69
23	19	99	2	0	0	6	84	2	1	1	2	0	1	1	3	0	1	2	1	72
24	20	74	2	0	0	8	89	1	1	1	2	2	1	0	2	0	0	1	0	66
25	11	78	0	2	1	8	100	0	1	1	1	2	1	1	3	0	1	1	1	66
26	17	65	1	0	1	5	75	2	1	2	1	2	1	2	3	0	1	2	0	63
27	21	62	0	1	1	6	54	0	1	0	0	0	1	2	3	0	2	0	1	64

StudentPerformanceFactors.csv



Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	Sleep_Hours	Previous_Scores	Motivation_Level	Internet_Access	Tutoring_Sessions	Family_Income	Teacher_Quality	School_Type	Peer_Influence	Physical_Activity	Learning_Disabilities	Parental_Education_Level	Distance_from_Home	Gender	Exam_Score
23	84	Low	High	No	7	73	Low	Yes	0	Low	Medium	Public	Positive	3	No	High Scho	Near	Male	67
19	64	Low	Medium	No	8	59	Low	Yes	2	Medium	Medium	Public	Negative	4	No	College	Moderate	Female	61
24	98	Medium	Medium	Yes	7	91	Medium	Yes	2	Medium	Medium	Public	Neutral	4	No	Postgradu	Near	Male	74
29	89	Low	Medium	Yes	8	98	Medium	Yes	1	Medium	Medium	Public	Negative	4	No	High Scho	Moderate	Male	71
19	92	Medium	Medium	Yes	6	65	Medium	Yes	3	Medium	High	Public	Neutral	4	No	College	Near	Female	70
19	88	Medium	Medium	Yes	8	89	Medium	Yes	3	Medium	Medium	Public	Positive	3	No	Postgradu	Near	Male	71
29	84	Medium	Low	Yes	7	68	Low	Yes	1	Low	Medium	Private	Neutral	2	No	High Scho	Moderate	Male	67
25	78	Low	High	Yes	6	50	Medium	Yes	1	High	High	Public	Negative	2	No	High Scho	Far	Male	66
17	94	Medium	High	No	6	80	High	Yes	0	Medium	Low	Private	Neutral	1	No	College	Near	Male	69
23	98	Medium	Medium	Yes	8	71	Medium	Yes	0	High	High	Public	Positive	5	No	High Scho	Moderate	Male	72
17	80	Low	High	No	8	88	Medium	No	4	Medium	High	Private	Neutral	4	No	College	Moderate	Male	68
17	97	Medium	High	Yes	6	87	Low	Yes	2	Low	High	Private	Neutral	2	No	High Scho	Near	Male	71
21	83	Medium	Medium	Yes	8	97	Low	Yes	2	Medium	Medium	Public	Positive	4	No	High Scho	Near	Male	70
9	82	Medium	Medium	Yes	8	72	Medium	Yes	2	Medium	Medium	Private	Positive	3	No	Postgradu	Near	Male	66
10	78	Medium	High	Yes	8	74	Medium	Yes	1	Low	Medium	Private	Neutral	4	No	Postgradu	Near	Male	65
17	68	Medium	Medium	No	8	70	Medium	Yes	2	Medium	Medium	Private	Positive	4	No	High Scho	Near	Female	64
14	60	Medium	Low	Yes	10	65	Low	Yes	0	High	Medium	Private	Positive	3	No	College	Near	Male	60
22	70	Low	Medium	Yes	6	82	Medium	Yes	1	Low	High	Public	Neutral	3	No	High Scho	Near	Female	65
15	80	Medium	Medium	Yes	9	91	Low	Yes	3	Low	Medium	Public	Positive	2	No	College	Moderate	Female	67
12	75	Medium	High	Yes	7	58	Medium	Yes	3	Medium	Medium	Private	Positive	4	No	College	Near	Male	66
29	78	Medium	Medium	No	5	99	High	Yes	0	High	Medium	Public	Negative	1	No	High Scho	Moderate	Female	69
19	99	Medium	High	No	6	84	Medium	Yes	1	Medium	High	Public	Neutral	3	No	High Scho	Near	Male	72
20	74	Medium	High	No	8	89	Low	Yes	1	Medium	Medium	Public	Negative	2	No	College	Moderate	Female	66
11	78	High	Medium	Yes	8	100	High	Yes	1	Low	Medium	Public	Neutral	3	No	High Scho	Moderate	Male	66
17	65	Low	High	Yes	5	75	Medium	Yes	2	Low	Medium	Public	Positive	3	No	High Scho	Near	Female	63
21	62	High	Low	Yes	6	54	High	Yes	0	High	High	Public	Positive	3	No	Postgradu	Far	Male	64

This dataset captures a broad range of academic, social, and personal factors, enabling an in-depth analysis of what impacts student success in exams like study habits, attendance, parental involvement in their study, Access to Resources, Extracurricular Activities, Sleep Hours, Previous Scores, Motivation Level, Internet Access, Tutoring Sessions, Family Income, Teacher Quality, School Type, Peer Influence, Physical Activity, Learning Disabilities, Parental Education Level, Distance from Home, Gender, Exam Score .

Data Set Column Description: -

Column Descriptions

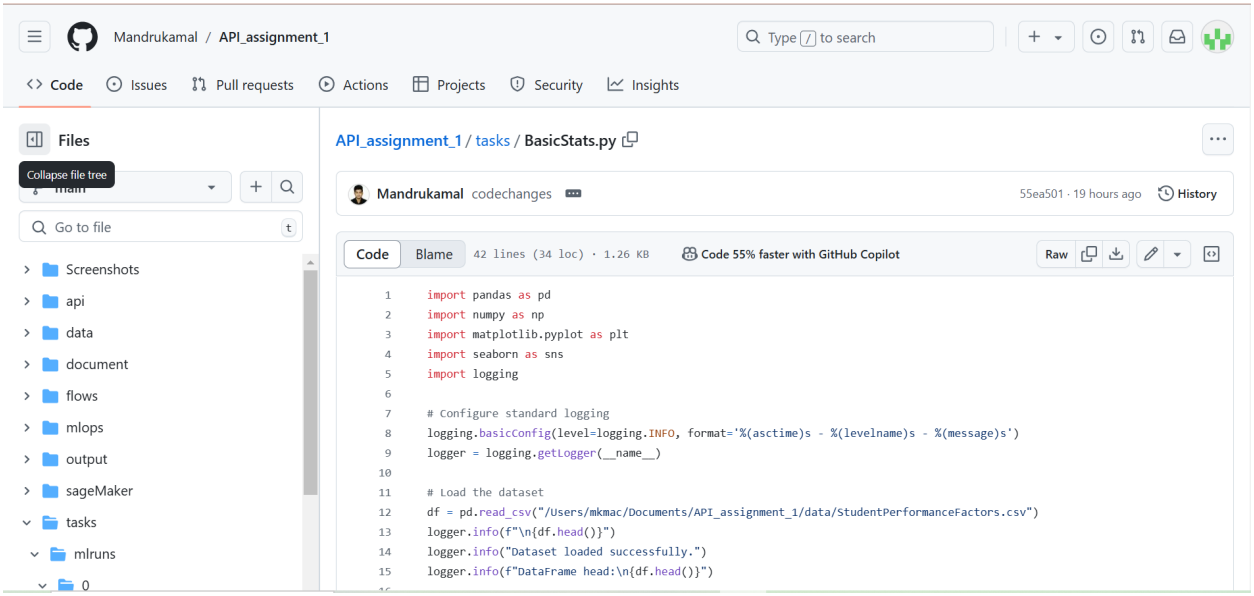
Attribute	Description
Hours_Studied	Number of hours spent studying per week.
Attendance	Percentage of classes attended.
Parental_Involvement	Level of parental involvement in the student's education (Low, Medium, High).
Access_to_Resources	Availability of educational resources (Low, Medium, High).

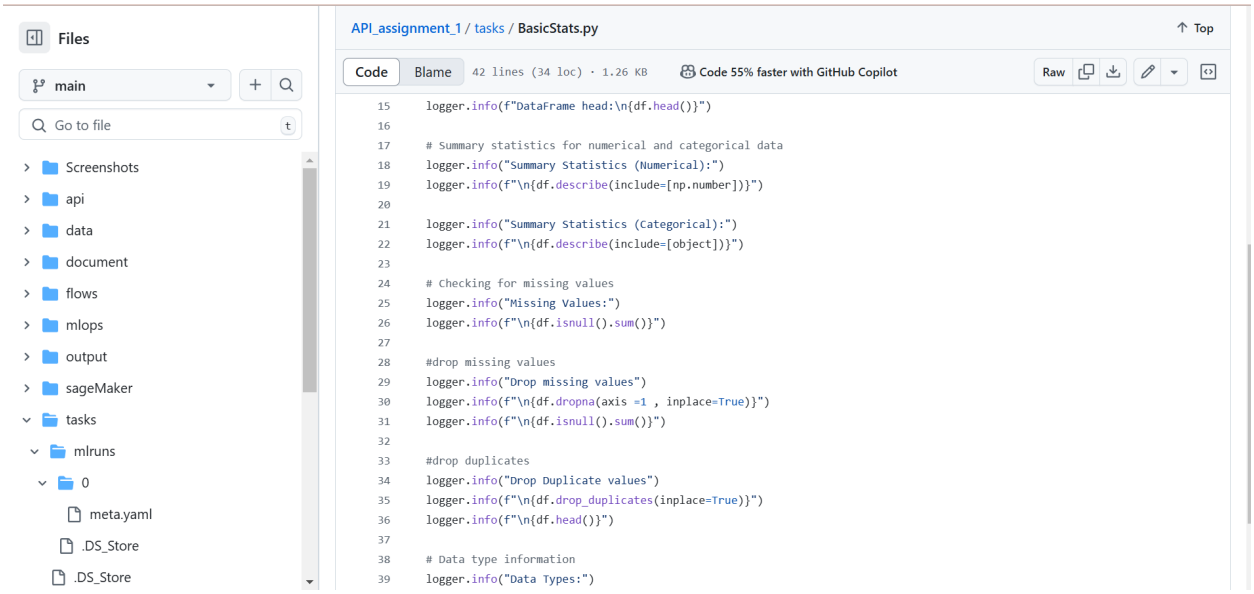
Extracurricular_Activities	Participation in extracurricular activities (Yes, No).
Sleep_Hours	Average number of hours of sleep per night.
Previous_Scores	Scores from previous exams.
Motivation_Level	Student's level of motivation (Low, Medium, High).
Internet_Access	Availability of internet access (Yes, No).
Tutoring_Sessions	Number of tutoring sessions attended per month.
Family_Income	Family income level (Low, Medium, High).
Teacher_Quality	Quality of the teachers (Low, Medium, High).
School_Type	Type of school attended (Public, Private).
Peer_Influence	Influence of peers on academic performance (Positive, Neutral, Negative).
Physical_Activity	Average number of hours of physical activity per week.
Learning_Disabilities	Presence of learning disabilities (Yes, No).
Parental_Education_Level	Highest education level of parents (High School, College, Postgraduate).
Distance_from_Home	Distance from home to school (Near, Moderate, Far).
Gender	Gender of the student (Male, Female).
Exam_Score	Final exam score.

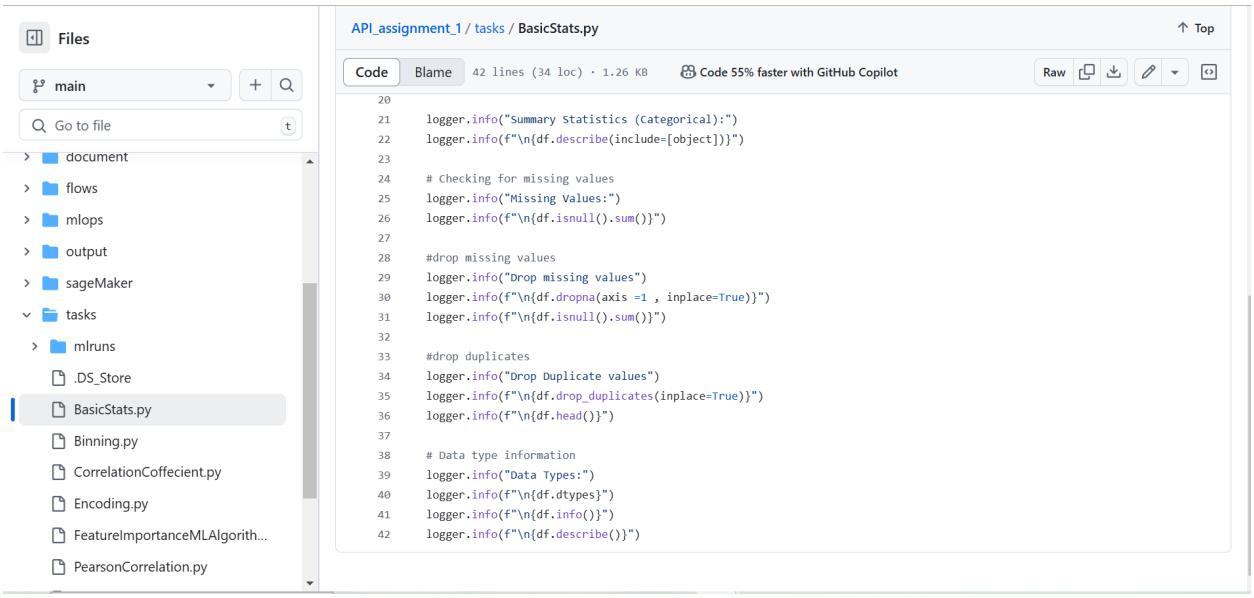
1.3 Data Pre-processing: Perform activities such as displaying summary statistics, checking for missing values, imputing missing data for numeric columns, displaying data types, and normalizing data.

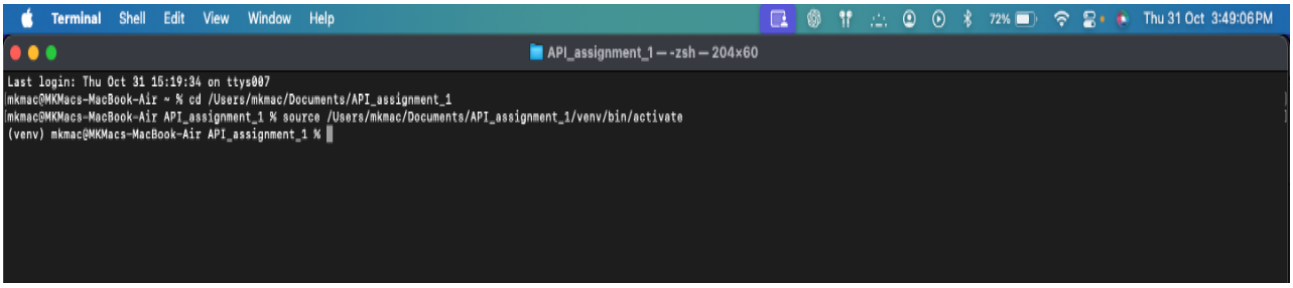
Below process is used for capturing the summary statistics, imputing missing data for numeric columns, displaying datatypes, checking for missing values and normalizing data;

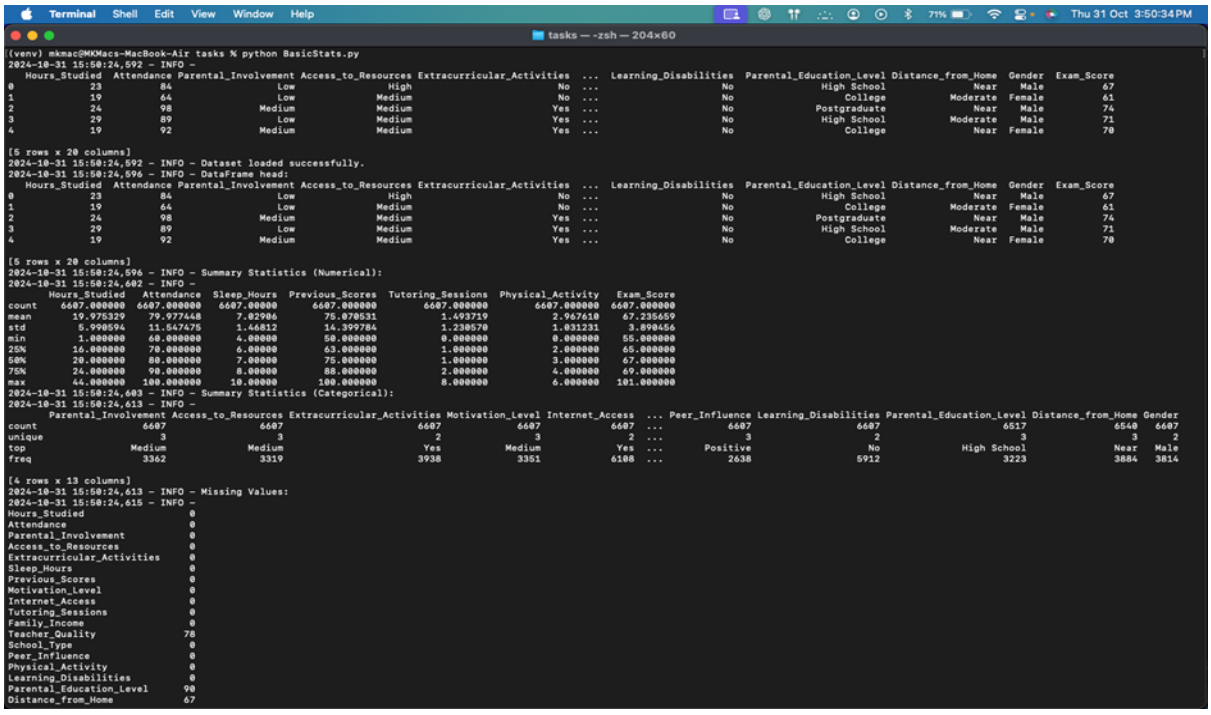
BasicStats.py: -

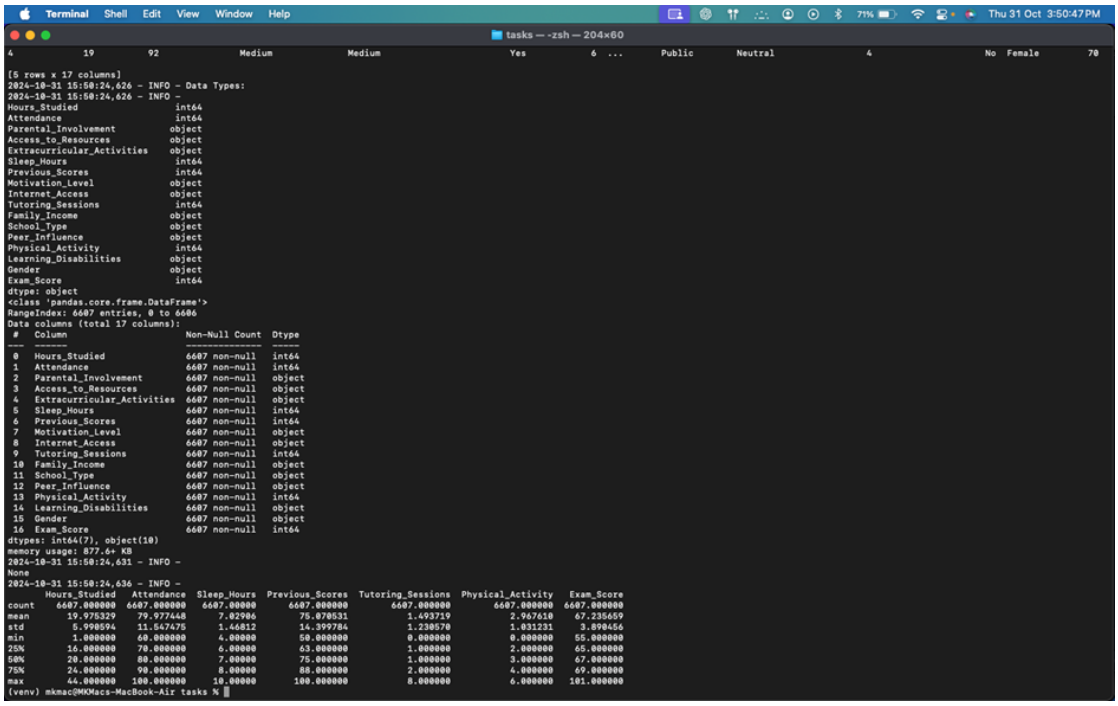





Files
main
Go to file
document
flows
mlops
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
API_assignment_1 / tasks / BasicStats.py
Top
Code
Blame
42 lines (34 loc) · 1.26 KB
Code 55% faster with GitHub Copilot
Raw
20
21 logger.info("Summary Statistics (Categorical):")
22 logger.info(f"\n{df.describe(include=[object])}")
23
24 # Checking for missing values
25 logger.info("Missing Values:")
26 logger.info(f"\n{df.isnull().sum()}")
27
28 #drop missing values
29 logger.info("Drop missing values")
30 logger.info(f"\n{df.dropna(axis =1 , inplace=True)}")
31 logger.info(f"\n{df.isnull().sum()}")
32
33 #drop duplicates
34 logger.info("Drop Duplicate values")
35 logger.info(f"\n{df.drop_duplicates(inplace=True)}")
36 logger.info(f"\n{df.head()}")
37
38 # Data type information
39 logger.info("Data Types:")
40 logger.info(f"\n{df.dtypes}")
41 logger.info(f"\n{df.info()}")
42 logger.info(f"\n{df.describe()}")

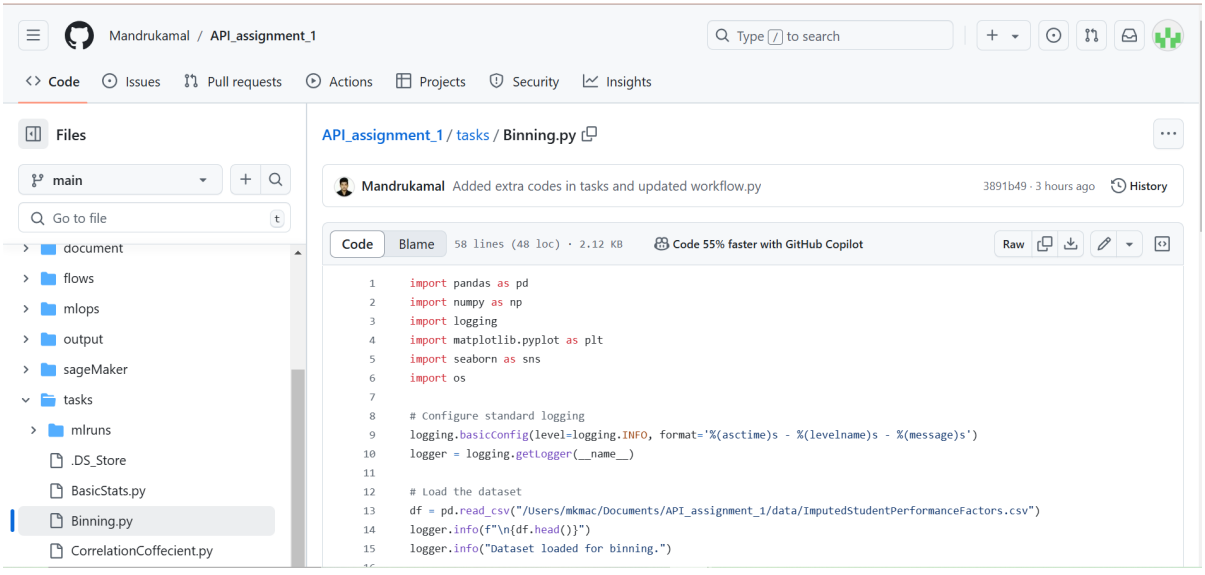

Terminal Shell Edit View Window Help
Thu 31 Oct 3:49:06PM
API_assignment_1 — -zsh — 204×60
Last login: Thu Oct 31 15:19:34 on ttys007
mkmac@MKMacs-MacBook-Air ~ % cd /Users/mkmac/Documents/API_assignment_1
mkmac@MKMacs-MacBook-Air API_assignment_1 % source /Users/mkmac/Documents/API_assignment_1/venv/bin/activate
(venv) mkmac@MKMacs-MacBook-Air API_assignment_1 %

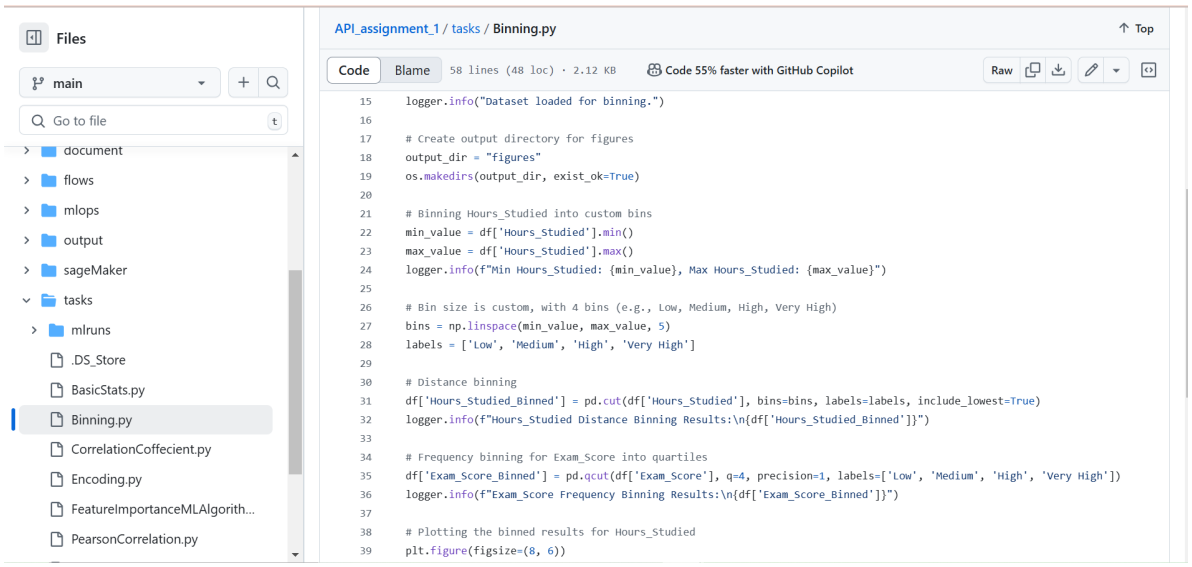

Terminal Shell Edit View Window Help
Thu 31 Oct 3:50:34PM
tasks — -zsh — 204×60
(venv) mkmac@MKMacs-MacBook-Air tasks % python BasicStats.py
2024-10-31 15:50:24,592 - INFO -
Hours_Studied Attendance Parental_Involvement Access_to_Resources Extracurricular_Activities ... Learning_Disabilities Parental_Education_Level Distance_from_Home Gender Exam_Score
0 23 84 Low High No ... No High School Near Male 67
1 19 64 Low Medium No ... No College Moderate Female 61
2 24 98 Medium Medium Yes ... No Postgraduate Near Male 74
3 29 89 Low Medium Yes ... No High School Moderate Male 71
4 19 92 Medium Medium Yes ... No College Near Female 70
[5 rows x 20 columns]
2024-10-31 15:50:24,592 - INFO - Dataset loaded successfully.
2024-10-31 15:50:24,596 - INFO - DataFrame head:
Hours_Studied Attendance Parental_Involvement Access_to_Resources Extracurricular_Activities ... Learning_Disabilities Parental_Education_Level Distance_from_Home Gender Exam_Score
0 23 84 Low High No ... No High School Near Male 67
1 19 64 Low Medium No ... No College Moderate Female 61
2 24 98 Medium Medium Yes ... No Postgraduate Near Male 74
3 29 89 Low Medium Yes ... No High School Moderate Male 71
4 19 92 Medium Medium Yes ... No College Near Female 70
[5 rows x 20 columns]
2024-10-31 15:50:24,596 - INFO - Summary Statistics (Numerical):
2024-10-31 15:50:24,602 - INFO -
Hours_Studied Attendance Sleep_Hours Previous_Scores Tutoring_Sessions Physical_Activity Exam_Score
count 6607.000000 6607.000000 6607.00000 6607.000000 6607.000000 6607.000000 6607.000000
mean 19.975329 79.977448 7.02906 75.070531 1.493719 2.967610 67.235659
std 5.990594 11.547475 1.46812 14.399784 1.230570 1.031231 3.890456
min 1.000000 60.000000 4.00000 50.000000 0.000000 0.000000 55.000000
25% 16.000000 70.000000 6.00000 63.000000 1.000000 2.000000 65.000000
50% 20.000000 80.000000 7.00000 75.000000 1.000000 3.000000 67.000000
75% 24.000000 90.000000 8.00000 88.000000 2.000000 4.000000 69.000000
max 44.000000 100.000000 10.00000 100.000000 8.000000 6.000000 101.000000
2024-10-31 15:50:24,603 - INFO - Summary Statistics (Categorical):
2024-10-31 15:50:24,613 - INFO -
Parental_Involvement Access_to_Resources Extracurricular_Activities Motivation_Level Internet_Access ... Peer_Influence Learning_Disabilities Parental_Education_Level Distance_from_Home Gender
count 6607 6607 6607 6607 6607 ... 6607 6607 6517 6540 6607
unique 3 3 2 3 2 ... 3 2 3 3 2
top Medium Medium Yes Medium Yes ... Positive No High School Near Male
freq 3362 3319 3938 3351 6108 ... 2638 5912 3223 3884 3814
[4 rows x 13 columns]
2024-10-31 15:50:24,613 - INFO - Missing Values:
2024-10-31 15:50:24,615 - INFO -
Hours_Studied 0
Attendance 0
Parental_Involvement 0
Access_to_Resources 0
Extracurricular_Activities 0
Sleep_Hours 0
Previous_Scores 0
Motivation_Level 0
Internet_Access 0
Tutoring_Sessions 0
Family_Income 0
Teacher_Quality 78
School_Type 0
Peer_Influence 0
Physical_Activity 0
Learning_Disabilities 0
Parental_Education_Level 90
Distance_from_Home 67

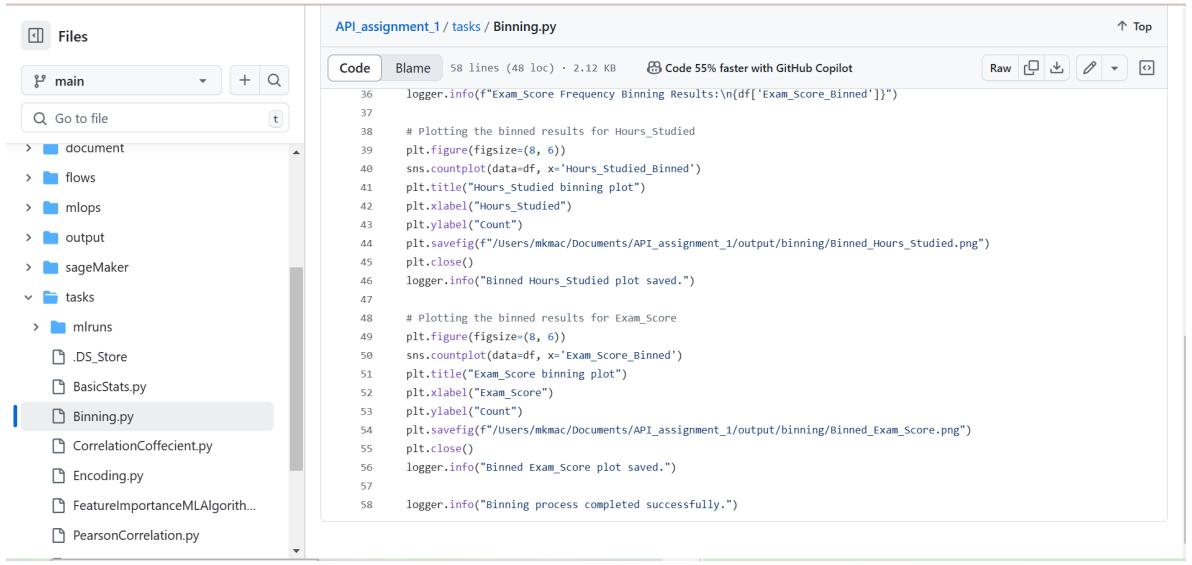


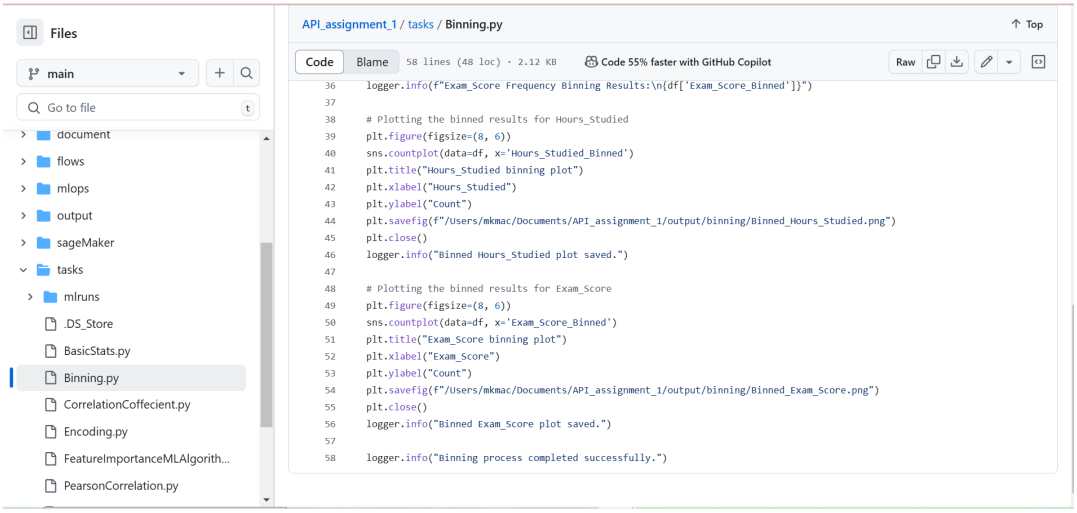
1.4 Exploratory Data Analysis (EDA): Conduct EDA to include calculating correlation coefficients, identifying correlations between numeric and/or categorical features, binning, encoding, assessing feature importance, and visualizing data (using charts and graphs for univariate and bivariate analyses).

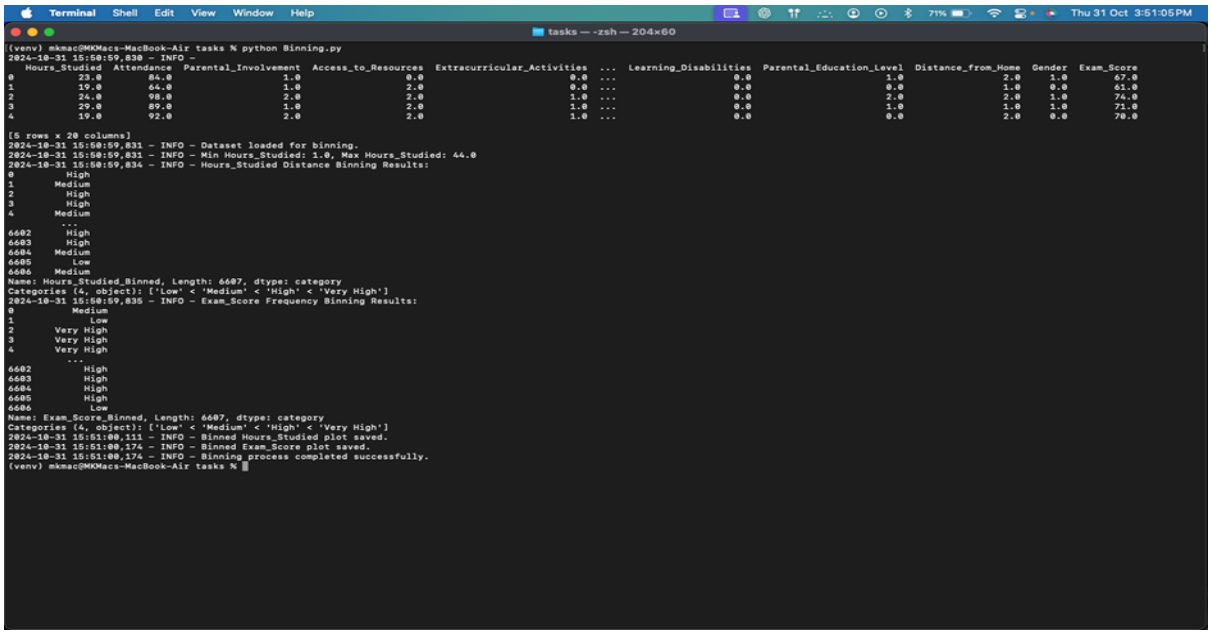
Binning.py: -

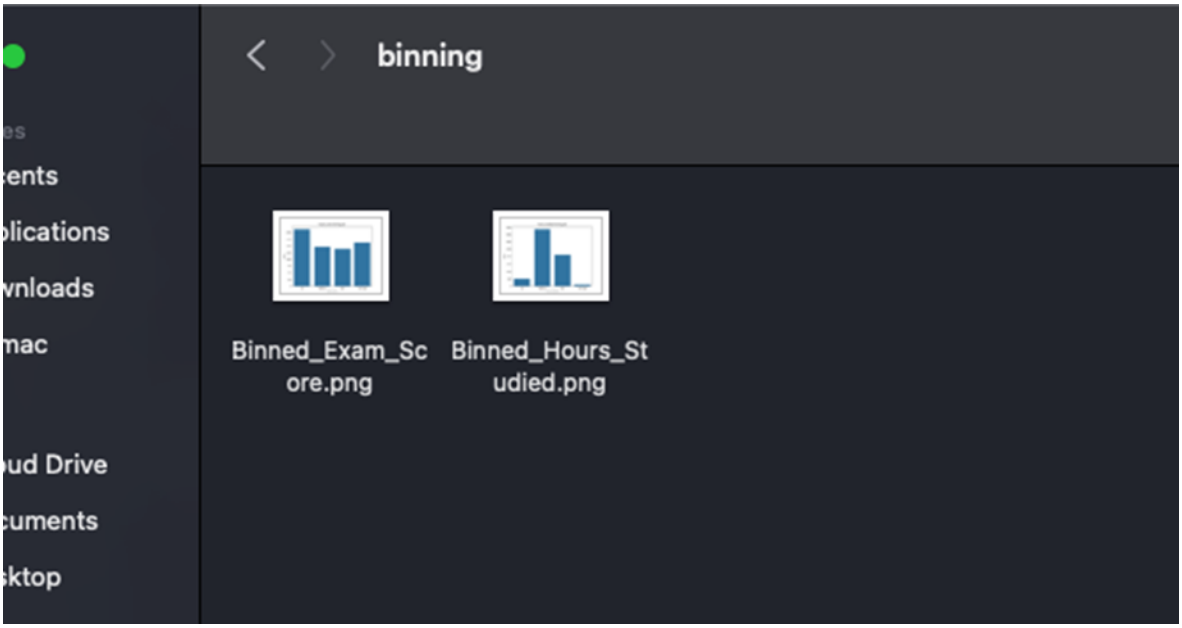



Files
main
Go to file
document
flows
mlops
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
API_assignment_1 / tasks / Binning.py
Top
Code
Blame
58 lines (48 loc) · 2.12 KB
Code 55% faster with GitHub Copilot
Raw
15 logger.info("Dataset loaded for binning.")
16
17 # Create output directory for figures
18 output_dir = "figures"
19 os.makedirs(output_dir, exist_ok=True)
20
21 # Binning Hours_Studied into custom bins
22 min_value = df['Hours_Studied'].min()
23 max_value = df['Hours_Studied'].max()
24 logger.info(f"Min Hours_Studied: {min_value}, Max Hours_Studied: {max_value}")
25
26 # Bin size is custom, with 4 bins (e.g., Low, Medium, High, Very High)
27 bins = np.linspace(min_value, max_value, 5)
28 labels = ['Low', 'Medium', 'High', 'Very High']
29
30 # Distance binning
31 df['Hours_Studied_Binned'] = pd.cut(df['Hours_Studied'], bins=bins, labels=labels, include_lowest=True)
32 logger.info(f"Hours_Studied Distance Binning Results:\n{df['Hours_Studied_Binned']}")
33
34 # Frequency binning for Exam_Score into quartiles
35 df['Exam_Score_Binned'] = pd.qcut(df['Exam_Score'], q=4, precision=1, labels=['Low', 'Medium', 'High', 'Very High'])
36 logger.info(f"Exam_Score Frequency Binning Results:\n{df['Exam_Score_Binned']}")
37
38 # Plotting the binned results for Hours_Studied
39 plt.figure(figsize=(8, 6))

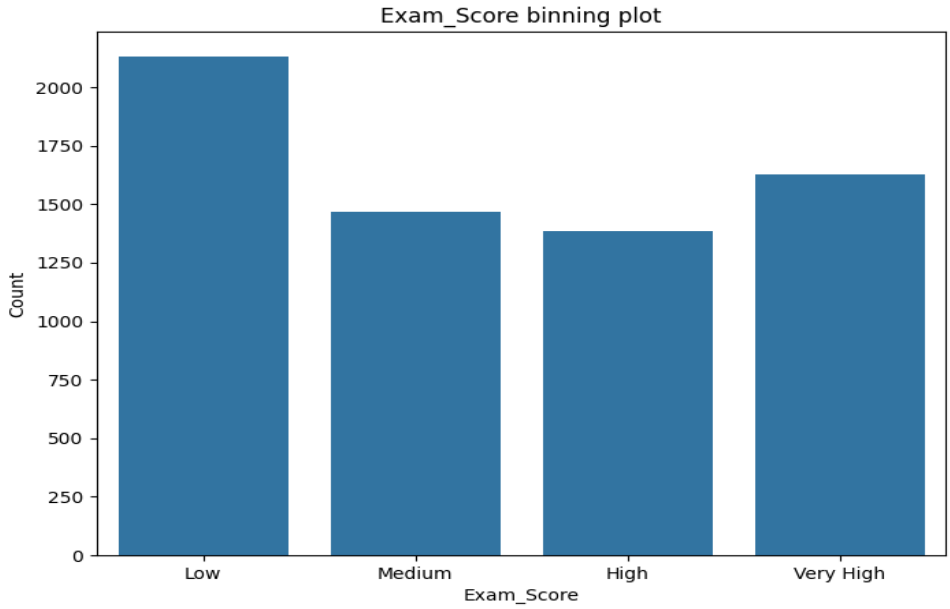

Files
main
Go to file
document
flows
mlops
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
API_assignment_1 / tasks / Binning.py
Top
Code
Blame
58 lines (48 loc) · 2.12 KB
Code 55% faster with GitHub Copilot
Raw
36 logger.info(f"Exam_Score Frequency Binning Results:\n{df['Exam_Score_Binned']}")
37
38 # Plotting the binned results for Hours_Studied
39 plt.figure(figsize=(8, 6))
40 sns.countplot(data=df, x='Hours_Studied_Binned')
41 plt.title("Hours_Studied binning plot")
42 plt.xlabel("Hours_Studied")
43 plt.ylabel("Count")
44 plt.savefig(f"/Users/mkmac/Documents/API_assignment_1/output/binning/Binned_Hours_Studied.png")
45 plt.close()
46 logger.info("Binned Hours_Studied plot saved.")
47
48 # Plotting the binned results for Exam_Score
49 plt.figure(figsize=(8, 6))
50 sns.countplot(data=df, x='Exam_Score_Binned')
51 plt.title("Exam_Score binning plot")
52 plt.xlabel("Exam_Score")
53 plt.ylabel("Count")
54 plt.savefig(f"/Users/mkmac/Documents/API_assignment_1/output/binning/Binned_Exam_Score.png")
55 plt.close()
56 logger.info("Binned Exam_Score plot saved.")
57
58 logger.info("Binning process completed successfully.")

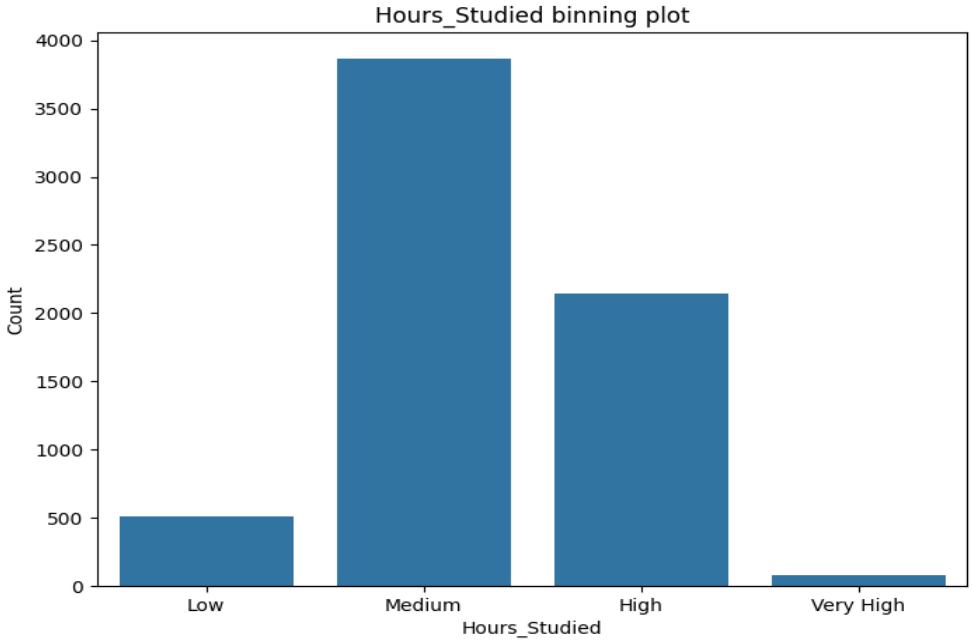

Files
main
Go to file
document
flows
mlops
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
API_assignment_1 / tasks / Binning.py
Top
Code
Blame
58 lines (48 loc) · 2.12 KB
Code 55% faster with GitHub Copilot
Raw
36 logger.info(f"Exam_Score Frequency Binning Results:\n{df['Exam_Score_Binned']}")
37
38 # Plotting the binned results for Hours_Studied
39 plt.figure(figsize=(8, 6))
40 sns.countplot(data=df, x='Hours_Studied_Binned')
41 plt.title("Hours_Studied binning plot")
42 plt.xlabel("Hours_Studied")
43 plt.ylabel("Count")
44 plt.savefig(f"/Users/mkmac/Documents/API_assignment_1/output/binning/Binned_Hours_Studied.png")
45 plt.close()
46 logger.info("Binned Hours_Studied plot saved.")
47
48 # Plotting the binned results for Exam_Score
49 plt.figure(figsize=(8, 6))
50 sns.countplot(data=df, x='Exam_Score_Binned')
51 plt.title("Exam_Score binning plot")
52 plt.xlabel("Exam_Score")
53 plt.ylabel("Count")
54 plt.savefig(f"/Users/mkmac/Documents/API_assignment_1/output/binning/Binned_Exam_Score.png")
55 plt.close()
56 logger.info("Binned Exam_Score plot saved.")
57
58 logger.info("Binning process completed successfully.")


Terminal Shell Edit View Window Help
Thu 31 Oct 3:51:05 PM
tasks — -zsh — 204×60
(venv) mkmac@MKMacs-MacBook-Air tasks % python Binning.py
2024-10-31 15:50:59,830 - INFO -
Hours_Studied Attendance Parental_Involvement Access_to_Resources Extracurricular_Activities ... Learning_Disabilities Parental_Education_Level Distance_from_Home Gender Exam_Score
0 23.0 84.0 1.0 0.0 0.0 ... 0.0 1.0 2.0 1.0 67.0
1 19.0 64.0 1.0 2.0 0.0 ... 0.0 0.0 1.0 0.0 61.0
2 24.0 98.0 2.0 2.0 1.0 ... 0.0 2.0 2.0 1.0 74.0
3 29.0 89.0 1.0 2.0 1.0 ... 0.0 1.0 1.0 1.0 71.0
4 19.0 92.0 2.0 2.0 1.0 ... 0.0 0.0 2.0 0.0 70.0
[5 rows x 20 columns]
2024-10-31 15:50:59,831 - INFO - Dataset loaded for binning.
2024-10-31 15:50:59,831 - INFO - Min Hours_Studied: 1.0, Max Hours_Studied: 44.0
2024-10-31 15:50:59,834 - INFO - Hours_Studied Distance Binning Results:
0 High
1 Medium
2 High
3 High
4 Medium
...
6602 High
6603 High
6604 Medium
6605 Low
6606 Medium
Name: Hours_Studied_Binned, Length: 6607, dtype: category
Categories (4, object): ['Low' < 'Medium' < 'High' < 'Very High']
2024-10-31 15:50:59,835 - INFO - Exam_Score Frequency Binning Results:
0 Medium
1 Low
2 Very High
3 Very High
4 Very High
...
6602 High
6603 High
6604 High
6605 High
6606 Low
Name: Exam_Score_Binned, Length: 6607, dtype: category
Categories (4, object): ['Low' < 'Medium' < 'High' < 'Very High']
2024-10-31 15:51:00,111 - INFO - Binned Hours_Studied plot saved.
2024-10-31 15:51:00,174 - INFO - Binned Exam_Score plot saved.
2024-10-31 15:51:00,174 - INFO - Binning process completed successfully.
(venv) mkmac@MKMacs-MacBook-Air tasks %

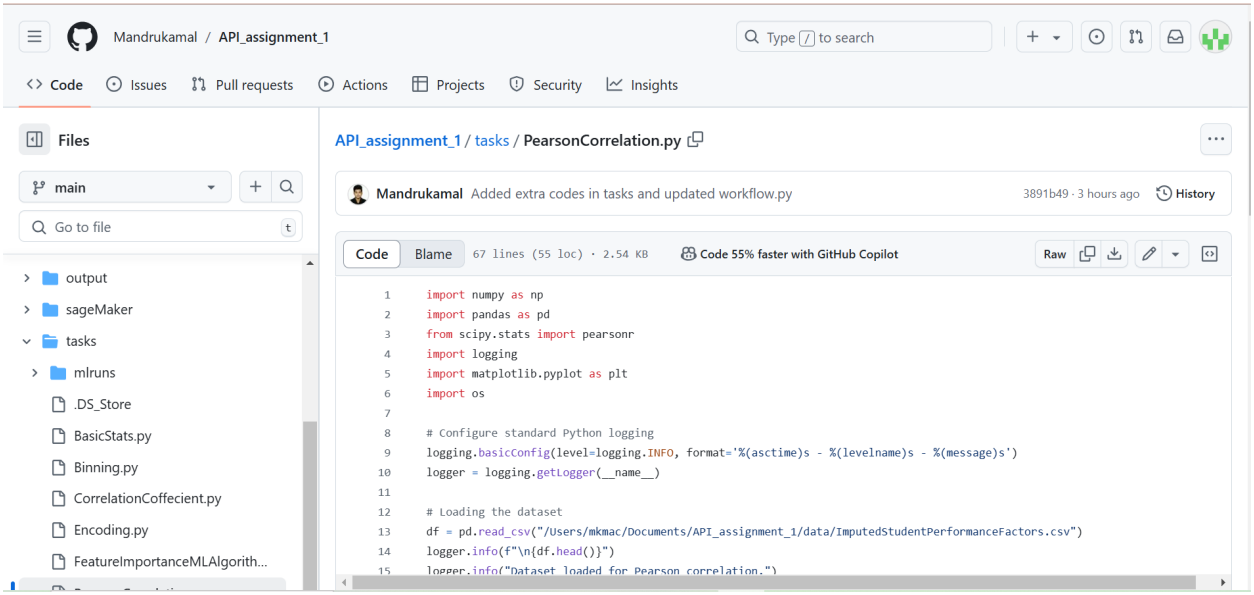

binning
Binned_Exam_Score.png
Binned_Hours_Studied.png

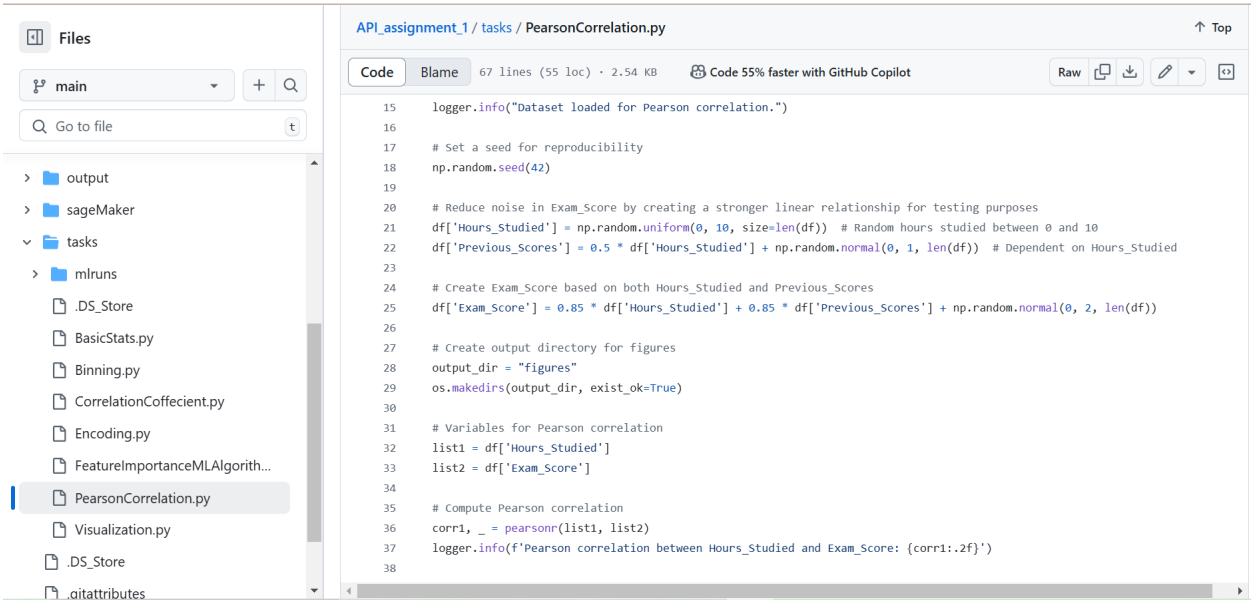
Outputs: -

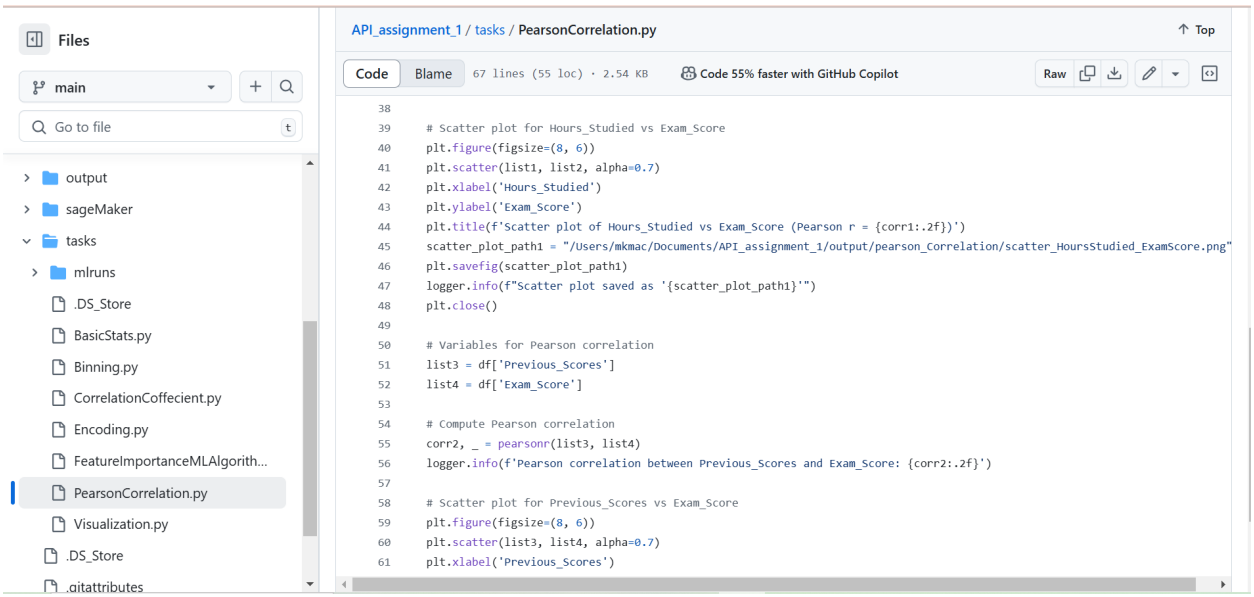

Exam_Score binning plot
Count
2000
1750
1500
1250
1000
750
500
250
0
Low
Medium
High
Very High
Exam_Score

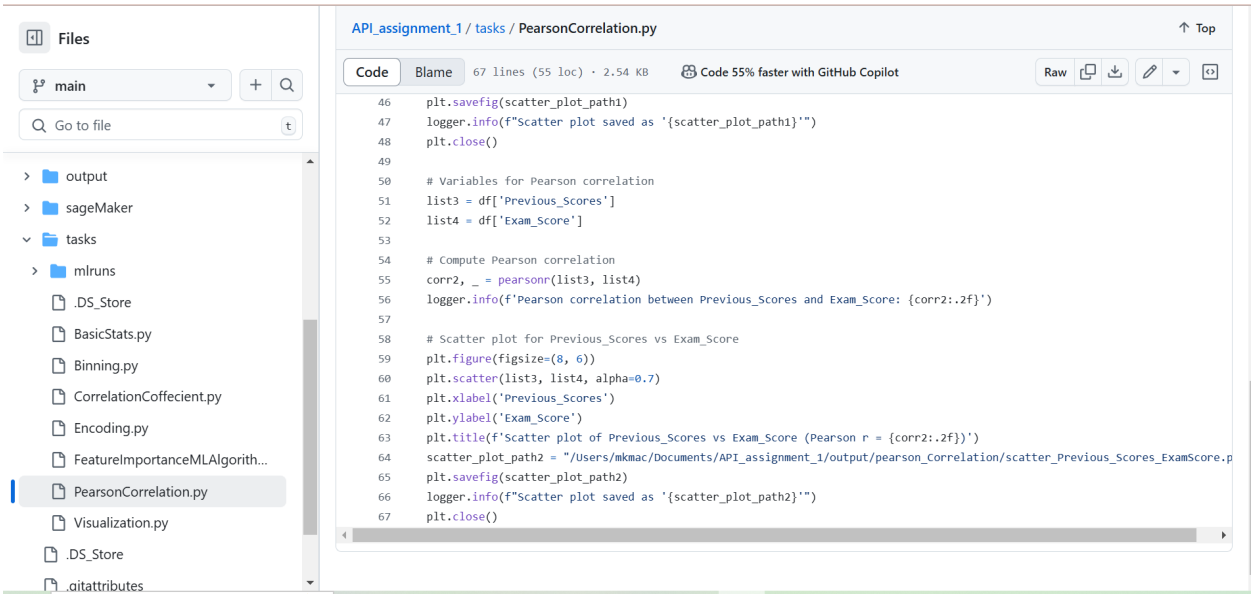

Hours_Studied binning plot
Count
4000
3500
3000
2500
2000
1500
1000
500
0
Low
Medium
High
Very High
Hours_Studied

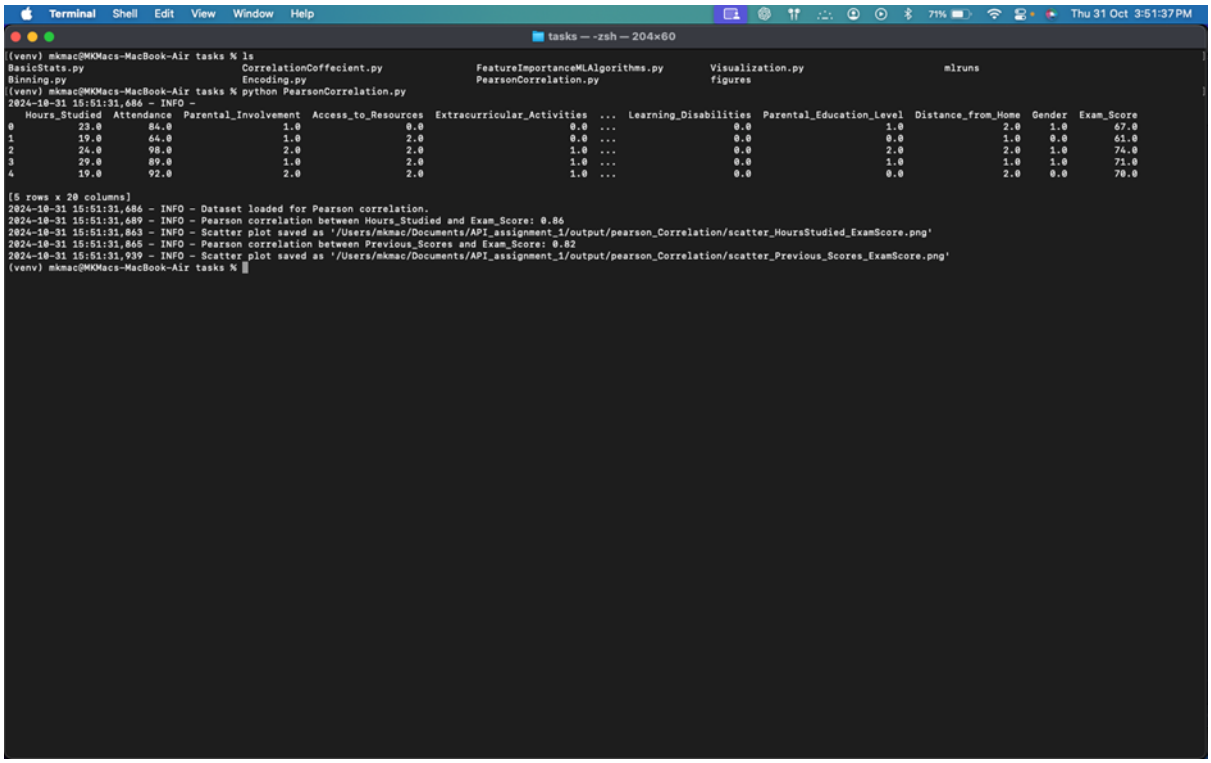
Pearson Correlation: -

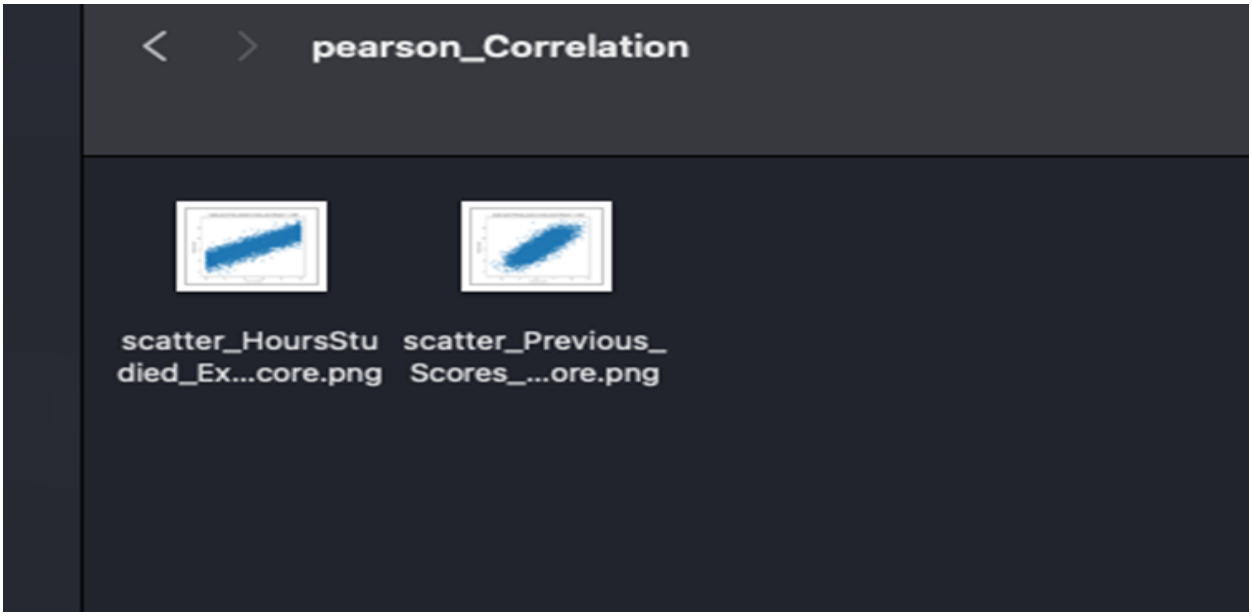


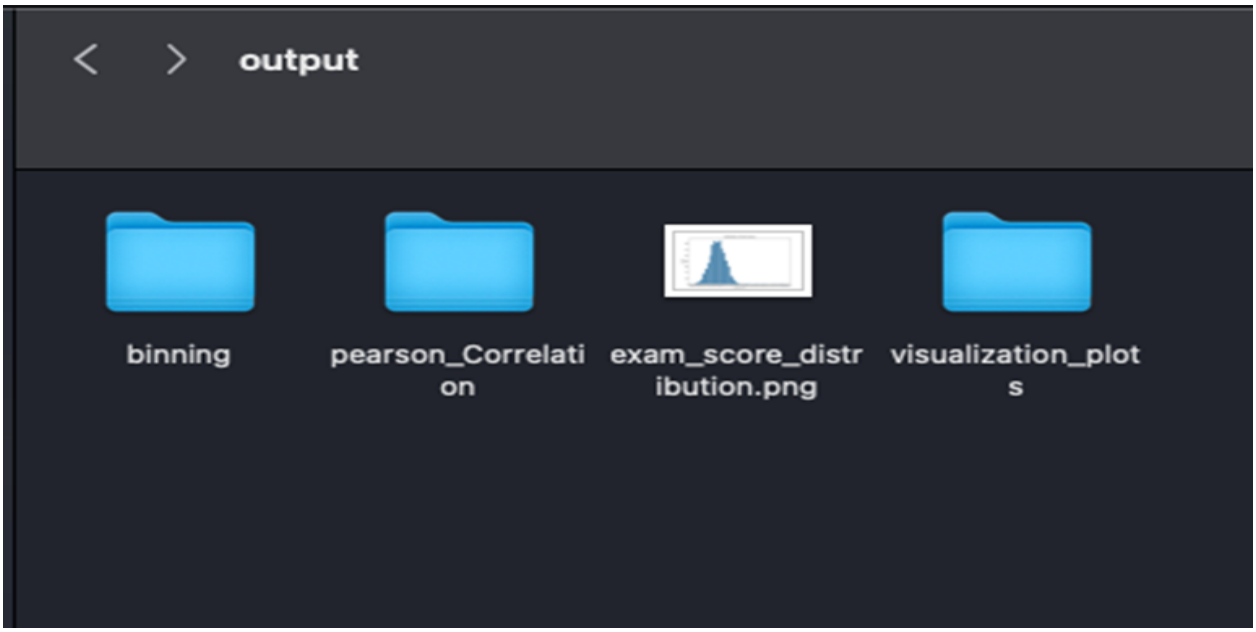



Files
main
Go to file
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
Visualization.py
.DS_Store
.gitattributes
API_assignment_1 / tasks / PearsonCorrelation.py
Top
Code
Blame
67 lines (55 loc) · 2.54 KB
Code 55% faster with GitHub Copilot
Raw
38
39 # Scatter plot for Hours_Studied vs Exam_Score
40 plt.figure(figsize=(8, 6))
41 plt.scatter(list1, list2, alpha=0.7)
42 plt.xlabel('Hours_Studied')
43 plt.ylabel('Exam_Score')
44 plt.title(f'Scatter plot of Hours_Studied vs Exam_Score (Pearson r = {corr1:.2f})')
45 scatter_plot_path1 = "/Users/mkmac/Documents/API_assignment_1/output/pearson_Correlation/scatter_HoursStudied_ExamScore.png"
46 plt.savefig(scatter_plot_path1)
47 logger.info(f"Scatter plot saved as '{scatter_plot_path1}'")
48 plt.close()
49
50 # Variables for Pearson correlation
51 list3 = df['Previous_Scores']
52 list4 = df['Exam_Score']
53
54 # Compute Pearson correlation
55 corr2, _ = pearsonr(list3, list4)
56 logger.info(f'Pearson correlation between Previous_Scores and Exam_Score: {corr2:.2f}')
57
58 # Scatter plot for Previous_Scores vs Exam_Score
59 plt.figure(figsize=(8, 6))
60 plt.scatter(list3, list4, alpha=0.7)
61 plt.xlabel('Previous_Scores')

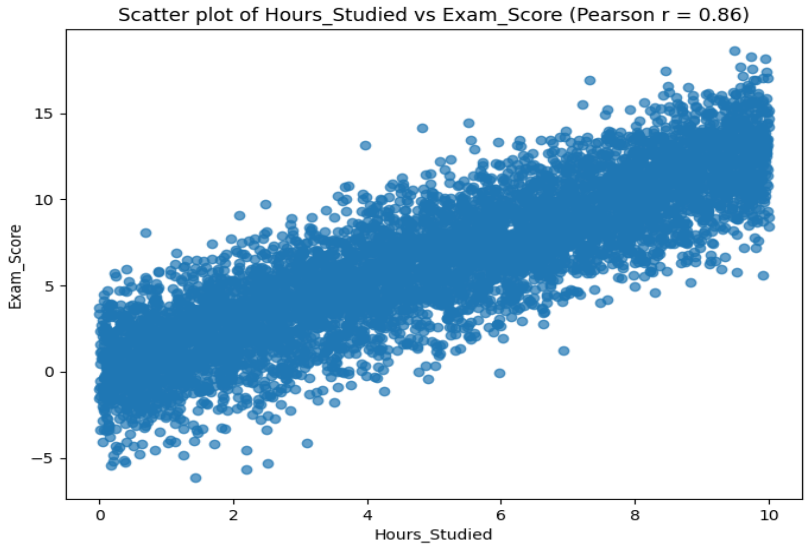

Files
main
Go to file
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
Visualization.py
.DS_Store
.gitattributes
API_assignment_1 / tasks / PearsonCorrelation.py
Top
Code
Blame
67 lines (55 loc) · 2.54 KB
Code 55% faster with GitHub Copilot
Raw
46 plt.savefig(scatter_plot_path1)
47 logger.info(f"Scatter plot saved as '{scatter_plot_path1}'")
48 plt.close()
49
50 # Variables for Pearson correlation
51 list3 = df['Previous_Scores']
52 list4 = df['Exam_Score']
53
54 # Compute Pearson correlation
55 corr2, _ = pearsonr(list3, list4)
56 logger.info(f'Pearson correlation between Previous_Scores and Exam_Score: {corr2:.2f}')
57
58 # Scatter plot for Previous_Scores vs Exam_Score
59 plt.figure(figsize=(8, 6))
60 plt.scatter(list3, list4, alpha=0.7)
61 plt.xlabel('Previous_Scores')
62 plt.ylabel('Exam_Score')
63 plt.title(f'Scatter plot of Previous_Scores vs Exam_Score (Pearson r = {corr2:.2f})')
64 scatter_plot_path2 = "/Users/mkmac/Documents/API_assignment_1/output/pearson_Correlation/scatter_Previous_Scores_ExamScore.p
65 plt.savefig(scatter_plot_path2)
66 logger.info(f"Scatter plot saved as '{scatter_plot_path2}'")
67 plt.close()

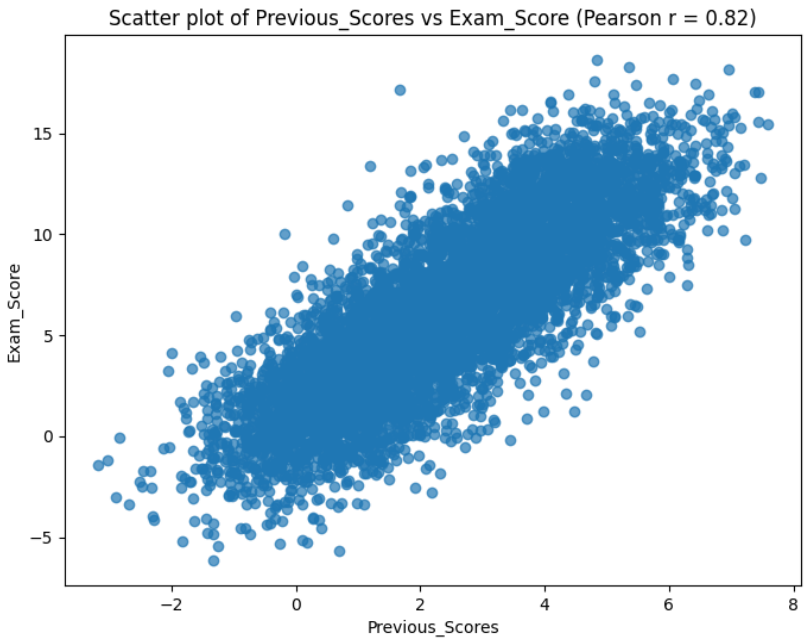

Terminal Shell Edit View Window Help
Thu 31 Oct 3:51:37 PM
tasks — -zsh — 204×60
(venv) mkmac@MKMacs-MacBook-Air tasks % ls
BasicStats.py CorrelationCoffecient.py FeatureImportanceMLAlgorithms.py Visualization.py mlruns
Binning.py Encoding.py PearsonCorrelation.py figures
(venv) mkmac@MKMacs-MacBook-Air tasks % python PearsonCorrelation.py
2024-10-31 15:51:31,686 - INFO -
Hours_Studied Attendance Parental_Involvement Access_to_Resources Extracurricular_Activities ... Learning_Disabilities Parental_Education_Level Distance_from_Home Gender Exam_Score
0 23.0 84.0 1.0 0.0 0.0 ... 0.0 1.0 2.0 1.0 67.0
1 19.0 64.0 1.0 2.0 0.0 ... 0.0 0.0 1.0 0.0 61.0
2 24.0 98.0 2.0 2.0 1.0 ... 0.0 2.0 2.0 1.0 74.0
3 29.0 89.0 1.0 2.0 1.0 ... 0.0 1.0 1.0 1.0 71.0
4 19.0 92.0 2.0 2.0 1.0 ... 0.0 0.0 2.0 0.0 70.0
[5 rows x 20 columns]
2024-10-31 15:51:31,686 - INFO - Dataset loaded for Pearson correlation.
2024-10-31 15:51:31,689 - INFO - Pearson correlation between Hours_Studied and Exam_Score: 0.86
2024-10-31 15:51:31,863 - INFO - Scatter plot saved as '/Users/mkmac/Documents/API_assignment_1/output/pearson_Correlation/scatter_HoursStudied_ExamScore.png'
2024-10-31 15:51:31,865 - INFO - Pearson correlation between Previous_Scores and Exam_Score: 0.82
2024-10-31 15:51:31,939 - INFO - Scatter plot saved as '/Users/mkmac/Documents/API_assignment_1/output/pearson_Correlation/scatter_Previous_Scores_ExamScore.png'
(venv) mkmac@MKMacs-MacBook-Air tasks %


pearson_Correlation
scatter_HoursStu died_Ex...core.png
scatter_Previous_ Scores_...ore.png


output
binning
pearson_Correlation
exam_score_distribution.png
visualization_plots

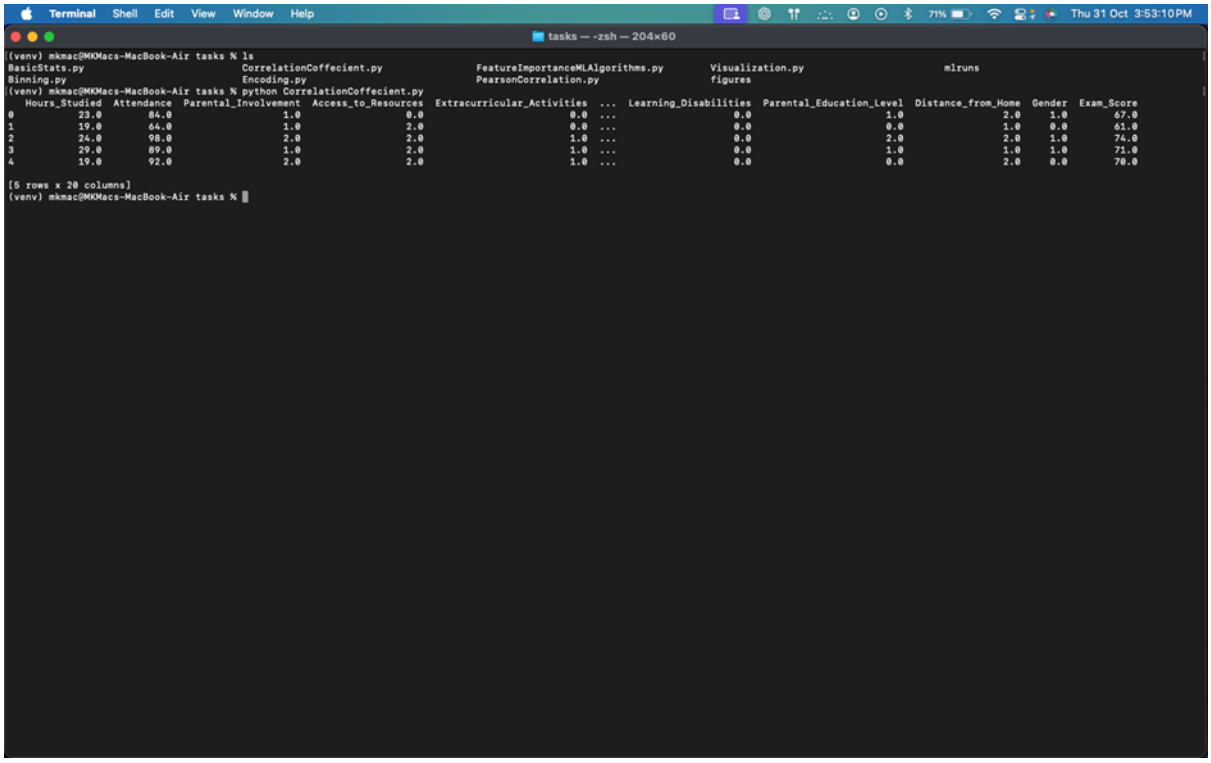
Outputs: -


Scatter plot of Hours_Studied vs Exam_Score (Pearson r = 0.86)
Exam_Score
Hours_Studied
15
10
5
0
−5
0
2
4
6
8
10

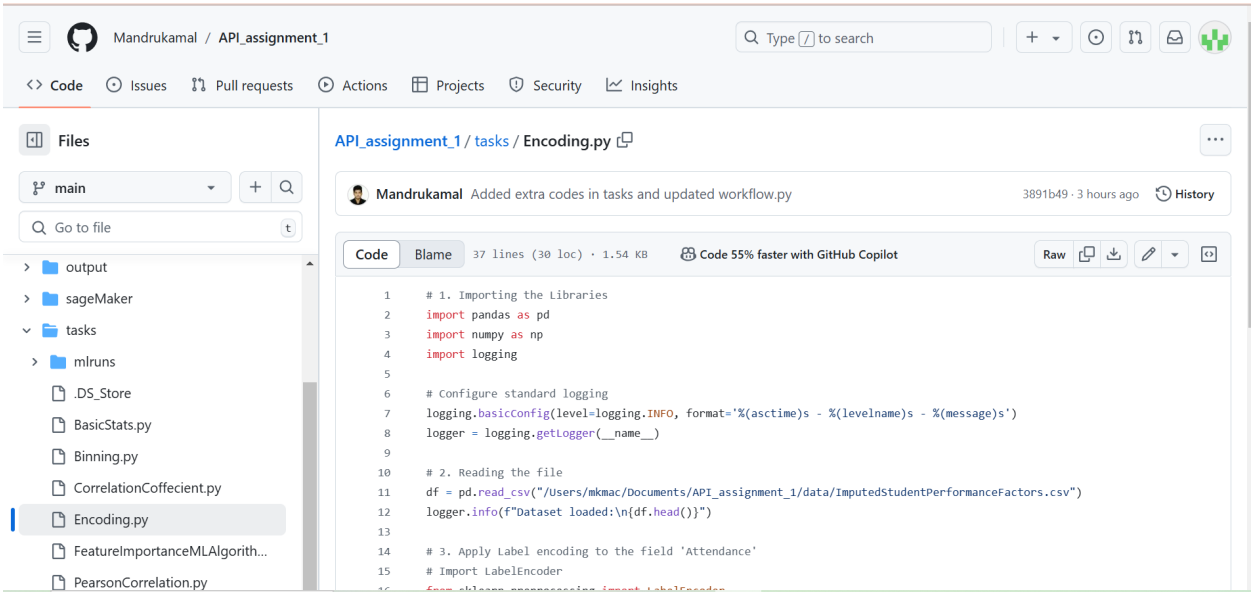


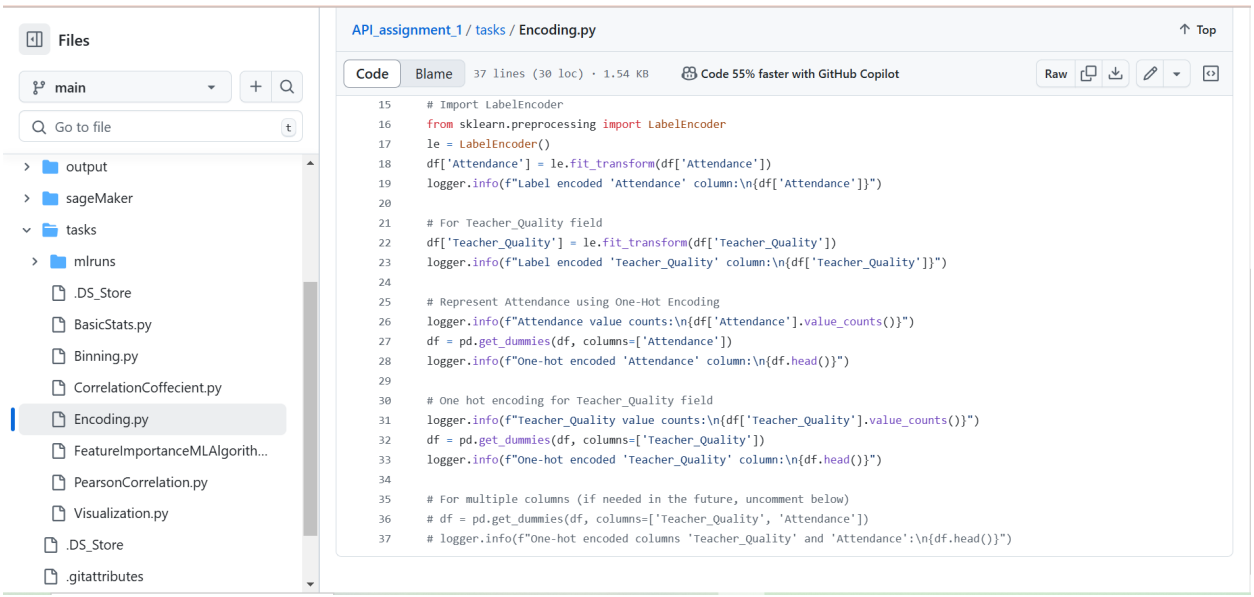
Correlation Coefficient: -

Output: -

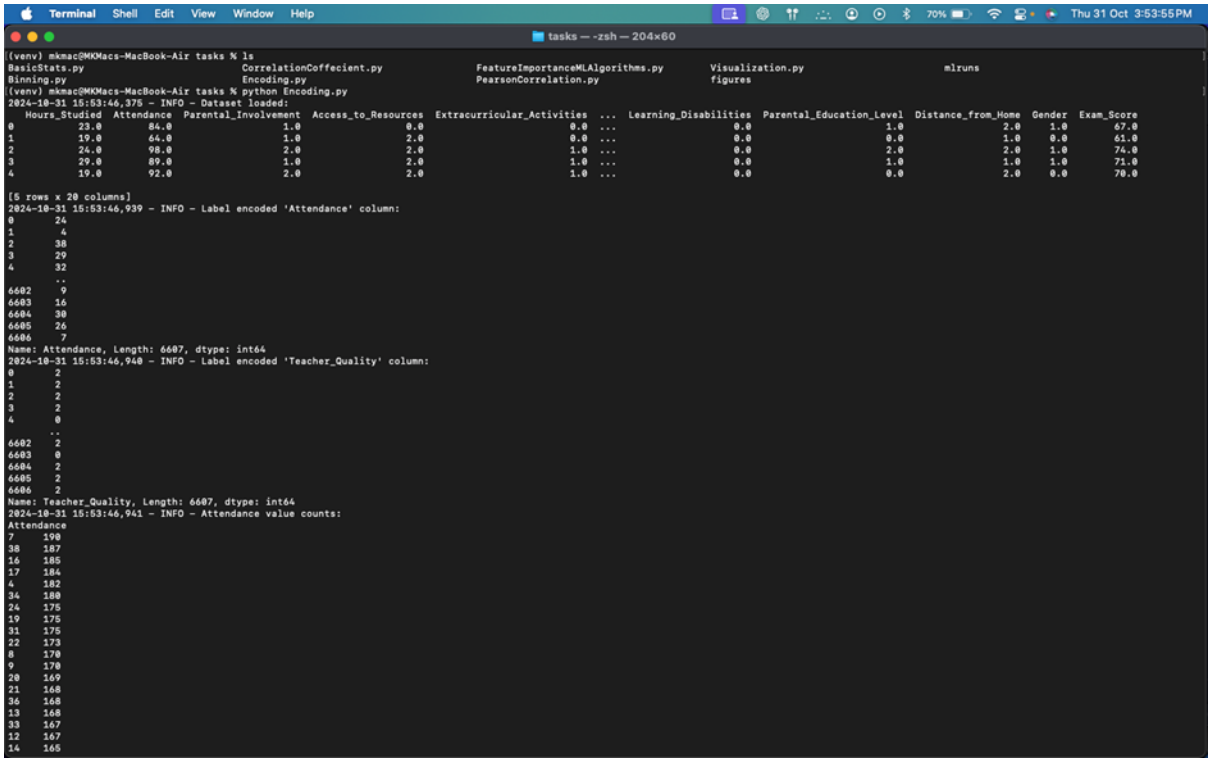


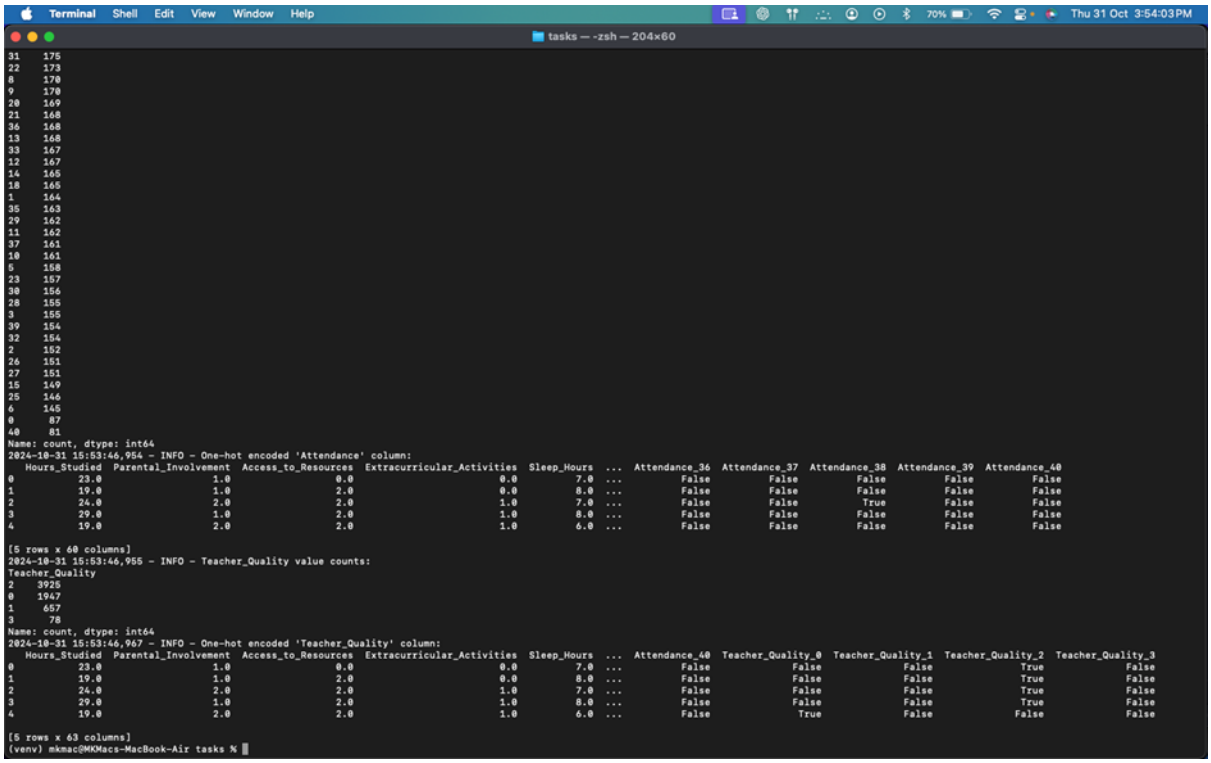
Encoding.py: -



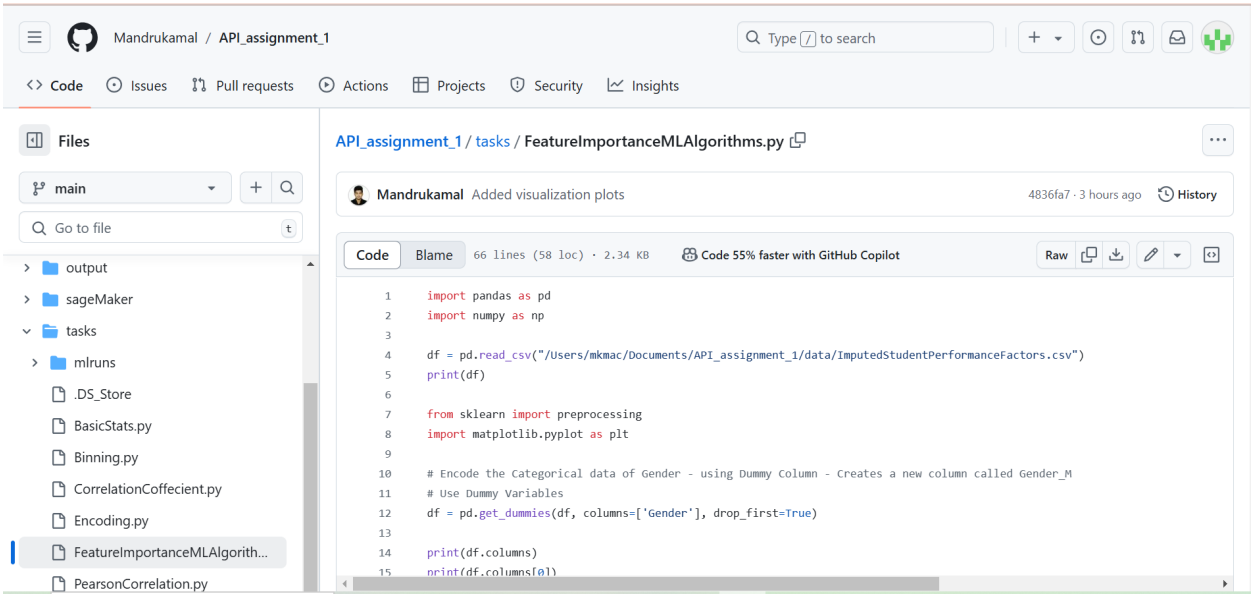

Files
main
Go to file
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
Visualization.py
.DS_Store
.gitattributes
API_assignment_1 / tasks / Encoding.py
Top
Code
Blame
37 lines (30 loc) · 1.54 KB
Code 55% faster with GitHub Copilot
Raw
15 # Import LabelEncoder
16 from sklearn.preprocessing import LabelEncoder
17 le = LabelEncoder()
18 df['Attendance'] = le.fit_transform(df['Attendance'])
19 logger.info(f"Label encoded 'Attendance' column:\n{df['Attendance']}")
20
21 # For Teacher_Quality field
22 df['Teacher_Quality'] = le.fit_transform(df['Teacher_Quality'])
23 logger.info(f"Label encoded 'Teacher_Quality' column:\n{df['Teacher_Quality']}")
24
25 # Represent Attendance using One-Hot Encoding
26 logger.info(f"Attendance value counts:\n{df['Attendance'].value_counts()}")
27 df = pd.get_dummies(df, columns=['Attendance'])
28 logger.info(f"One-hot encoded 'Attendance' column:\n{df.head()}")
29
30 # One hot encoding for Teacher_Quality field
31 logger.info(f"Teacher_Quality value counts:\n{df['Teacher_Quality'].value_counts()}")
32 df = pd.get_dummies(df, columns=['Teacher_Quality'])
33 logger.info(f"One-hot encoded 'Teacher_Quality' column:\n{df.head()}")
34
35 # For multiple columns (if needed in the future, uncomment below)
36 # df = pd.get_dummies(df, columns=['Teacher_Quality', 'Attendance'])
37 # logger.info(f"One-hot encoded columns 'Teacher_Quality' and 'Attendance':\n{df.head()}")

Output: -

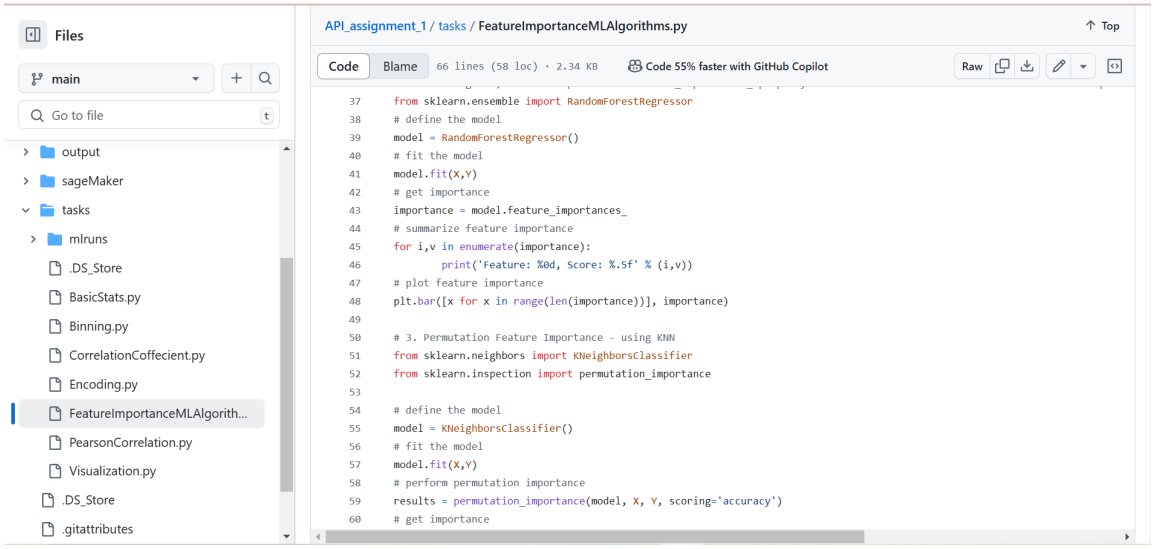

Terminal Shell Edit View Window Help
Thu 31 Oct 3:53:55 PM
tasks — -zsh — 204×60
(venv) mkmac@MKMacs-MacBook-Air tasks % ls
BasicStats.py CorrelationCoffecient.py FeatureImportanceMLAlgorithms.py Visualization.py mlruns
Binning.py Encoding.py PearsonCorrelation.py figures
(venv) mkmac@MKMacs-MacBook-Air tasks % python Encoding.py
2024-10-31 15:53:46,375 - INFO - Dataset loaded:
Hours_Studied Attendance Parental_Involvement Access_to_Resources Extracurricular_Activities ... Learning_Disabilities Parental_Education_Level Distance_from_Home Gender Exam_Score
0 23.0 84.0 1.0 0.0 0.0 ... 0.0 1.0 2.0 1.0 67.0
1 19.0 64.0 1.0 2.0 0.0 ... 0.0 0.0 1.0 0.0 61.0
2 24.0 98.0 2.0 2.0 1.0 ... 0.0 2.0 2.0 1.0 74.0
3 29.0 89.0 1.0 2.0 1.0 ... 0.0 1.0 1.0 1.0 71.0
4 19.0 92.0 2.0 2.0 1.0 ... 0.0 0.0 2.0 0.0 70.0
[5 rows x 20 columns]
2024-10-31 15:53:46,939 - INFO - Label encoded 'Attendance' column:
0 24
1 4
2 38
3 29
4 32
..
6602 9
6603 16
6604 30
6605 26
6606 7
Name: Attendance, Length: 6607, dtype: int64
2024-10-31 15:53:46,940 - INFO - Label encoded 'Teacher_Quality' column:
0 2
1 2
2 2
3 2
4 0
..
6602 2
6603 0
6604 2
6605 2
6606 2
Name: Teacher_Quality, Length: 6607, dtype: int64
2024-10-31 15:53:46,941 - INFO - Attendance value counts:
Attendance
7 190
38 187
16 185
17 184
4 182
34 180
24 175
19 175
31 175
22 173
8 170
9 170
20 169
21 168
36 168
13 168
33 167
12 167
14 165

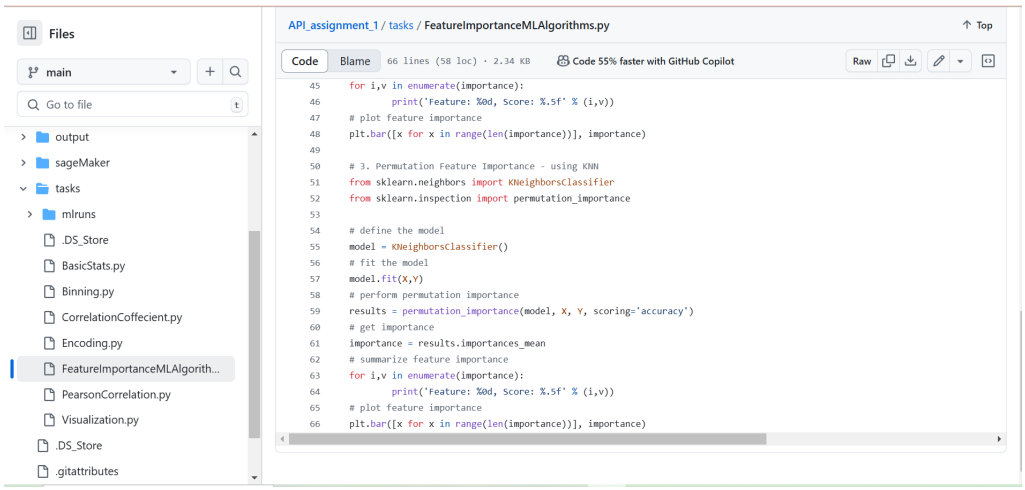


Feature Importance ML Algorithm: -

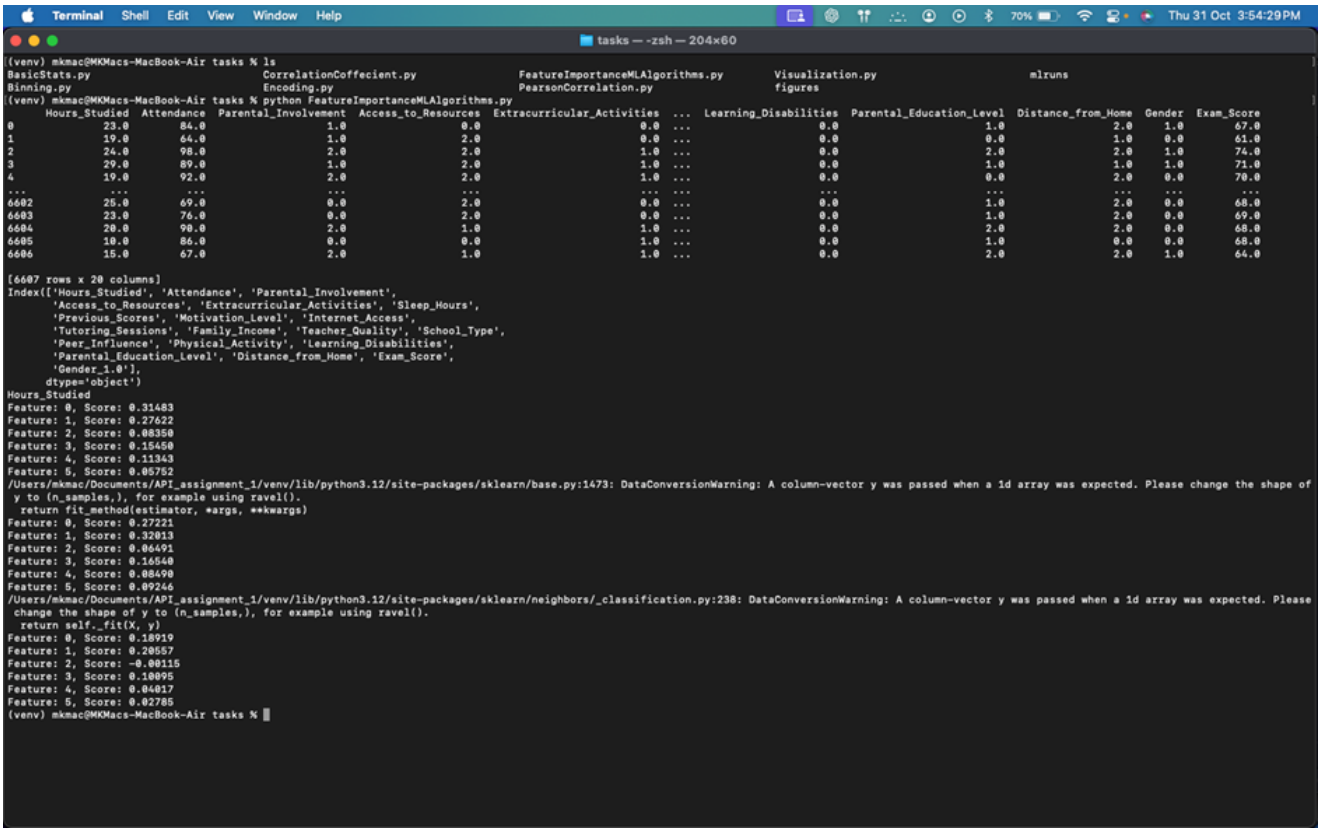



Files
main
Go to file
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
Visualization.py
.DS_Store
.gitattributes
API_assignment_1 / tasks / FeatureImportanceMLAlgorithms.py
Top
Code
Blame
66 lines (58 loc) · 2.34 KB
Code 55% faster with GitHub Copilot
Raw
print(df.columns)
print(df.columns[0])
X = df.iloc[:,[0,1,4,5,7,10]]
Y = df.iloc[:,[6]]
# 1. Decision Tree
# Decision tree algorithms like classification and regression trees (CART) offer importance scores based on the reduction in
# After being fit, the model provides a feature_importances_ property that can be accessed to retrieve the relative importan
from sklearn.tree import DecisionTreeClassifier
# define the model
model = DecisionTreeClassifier()
# fit the model
model.fit(X,Y)
# get importance
importance = model.feature_importances_
# summarize feature importance
for i,v in enumerate(importance):
print('Feature: %0d, Score: %.5f' % (i,v))
# plot feature importance
plt.bar([x for x in range(len(importance))], importance)
# 2. Random Forest Feature Importance
# After being fit, the model provides a feature_importances_ property that can be accessed to retrieve the relative importan
from sklearn.ensemble import RandomForestRegressor


Files
main
Go to file
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
Visualization.py
.DS_Store
.gitattributes
API_assignment_1 / tasks / FeatureImportanceMLAlgorithms.py
Top
Code
Blame
66 lines (58 loc) · 2.34 KB
Code 55% faster with GitHub Copilot
Raw
from sklearn.ensemble import RandomForestRegressor
# define the model
model = RandomForestRegressor()
# fit the model
model.fit(X,Y)
# get importance
importance = model.feature_importances_
# summarize feature importance
for i,v in enumerate(importance):
print('Feature: %0d, Score: %.5f' % (i,v))
# plot feature importance
plt.bar([x for x in range(len(importance))], importance)
# 3. Permutation Feature Importance - using KNN
from sklearn.neighbors import KNeighborsClassifier
from sklearn.inspection import permutation_importance
# define the model
model = KNeighborsClassifier()
# fit the model
model.fit(X,Y)
# perform permutation importance
results = permutation_importance(model, X, Y, scoring='accuracy')
# get importance

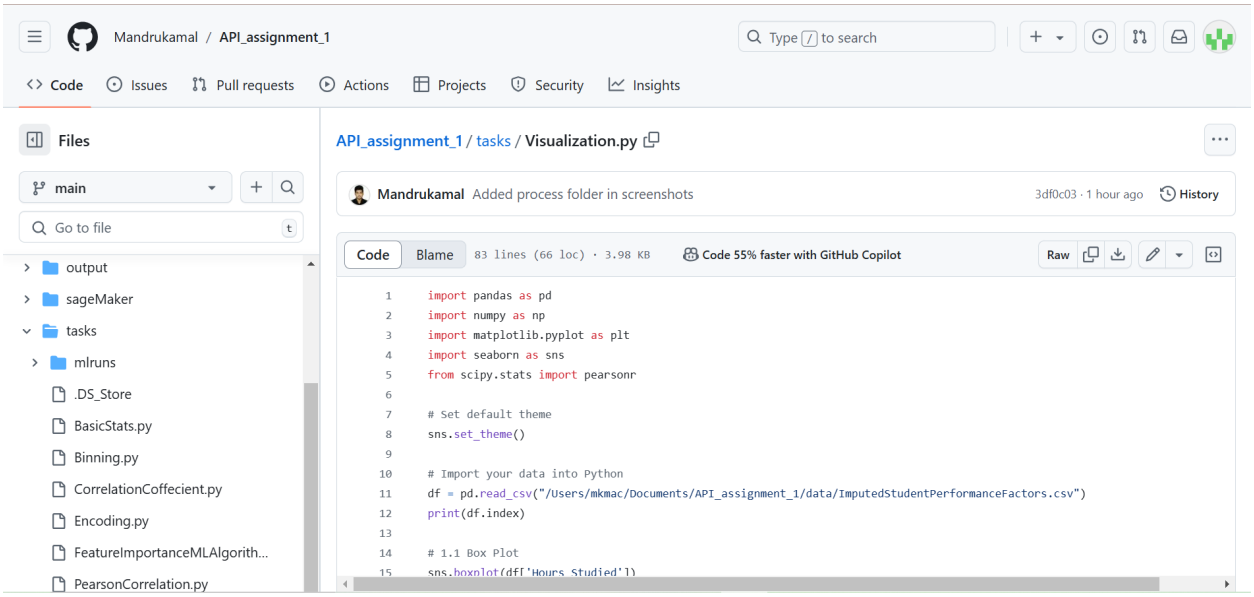

Files
main
Go to file
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
Visualization.py
.DS_Store
.gitattributes
API_assignment_1 / tasks / FeatureImportanceMLAlgorithms.py
Top
Code
Blame
66 lines (58 loc) · 2.34 KB
Code 55% faster with GitHub Copilot
Raw
for i,v in enumerate(importance):
print('Feature: %0d, Score: %.5f' % (i,v))
# plot feature importance
plt.bar([x for x in range(len(importance))], importance)
# 3. Permutation Feature Importance - using KNN
from sklearn.neighbors import KNeighborsClassifier
from sklearn.inspection import permutation_importance
# define the model
model = KNeighborsClassifier()
# fit the model
model.fit(X,Y)
# perform permutation importance
results = permutation_importance(model, X, Y, scoring='accuracy')
# get importance
importance = results.importances_mean
# summarize feature importance
for i,v in enumerate(importance):
print('Feature: %0d, Score: %.5f' % (i,v))
# plot feature importance
plt.bar([x for x in range(len(importance))], importance)

Output: -



```
(venv) mkmac@MKMacs-MacBook-Air tasks % ls
BasicStats.py                   CorrelationCoffecient.py        FeatureImportanceMLAlgorithms.py    Visualization.py                mlruns
Binning.py                      Encoding.py                     PearsonCorrelation.py               figures
(venv) mkmac@MKMacs-MacBook-Air tasks % python FeatureImportanceMLAlgorithms.py
      Hours_Studied  Attendance  Parental_Involvement  Access_to_Resources  Extracurricular_Activities  ...  Learning_Disabilities  Parental_Education_Level  Distance_from_Home  Gender  Exam_Score
0              23.0        84.0                   1.0                  0.0                         0.0  ...                    0.0                       1.0                 2.0     1.0        67.0
1              19.0        64.0                   1.0                  2.0                         0.0  ...                    0.0                       0.0                 1.0     0.0        61.0
2              24.0        98.0                   2.0                  2.0                         1.0  ...                    0.0                       2.0                 2.0     1.0        74.0
3              29.0        89.0                   1.0                  2.0                         1.0  ...                    0.0                       1.0                 1.0     1.0        71.0
4              19.0        92.0                   2.0                  2.0                         1.0  ...                    0.0                       0.0                 2.0     0.0        70.0
...             ...         ...                   ...                  ...                         ...  ...                    ...                       ...                 ...     ...         ...
6602           25.0        69.0                   0.0                  2.0                         0.0  ...                    0.0                       1.0                 2.0     0.0        68.0
6603           23.0        76.0                   0.0                  2.0                         0.0  ...                    0.0                       1.0                 2.0     0.0        69.0
6604           20.0        90.0                   2.0                  1.0                         1.0  ...                    0.0                       2.0                 2.0     0.0        68.0
6605           10.0        86.0                   0.0                  0.0                         1.0  ...                    0.0                       1.0                 0.0     0.0        68.0
6606           15.0        67.0                   2.0                  1.0                         1.0  ...                    0.0                       2.0                 2.0     1.0        64.0

[6607 rows x 20 columns]
Index(['Hours_Studied', 'Attendance', 'Parental_Involvement',
       'Access_to_Resources', 'Extracurricular_Activities', 'Sleep_Hours',
       'Previous_Scores', 'Motivation_Level', 'Internet_Access',
       'Tutoring_Sessions', 'Family_Income', 'Teacher_Quality', 'School_Type',
       'Peer_Influence', 'Physical_Activity', 'Learning_Disabilities',
       'Parental_Education_Level', 'Distance_from_Home', 'Exam_Score',
       'Gender_1.0'],
      dtype='object')
Hours_Studied
Feature: 0, Score: 0.31483
Feature: 1, Score: 0.27622
Feature: 2, Score: 0.08350
Feature: 3, Score: 0.15450
Feature: 4, Score: 0.11343
Feature: 5, Score: 0.05752
/Users/mkmac/Documents/API_assignment_1/venv/lib/python3.12/site-packages/sklearn/base.py:1473: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of
 y to (n_samples,), for example using ravel().
  return fit_method(estimator, *args, **kwargs)
Feature: 0, Score: 0.27221
Feature: 1, Score: 0.32013
Feature: 2, Score: 0.06491
Feature: 3, Score: 0.16540
Feature: 4, Score: 0.08490
Feature: 5, Score: 0.09246
/Users/mkmac/Documents/API_assignment_1/venv/lib/python3.12/site-packages/sklearn/neighbors/_classification.py:238: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please
 change the shape of y to (n_samples,), for example using ravel().
  return self._fit(X, y)
Feature: 0, Score: 0.18919
Feature: 1, Score: 0.20557
Feature: 2, Score: -0.00115
Feature: 3, Score: 0.10095
Feature: 4, Score: 0.04017
Feature: 5, Score: 0.02785
(venv) mkmac@MKMacs-MacBook-Air tasks %
```

Visualization.py: -



Mandrukamal / API_assignment_1

Code | Issues | Pull requests | Actions | Projects | Security | Insights

Files

main

Go to file

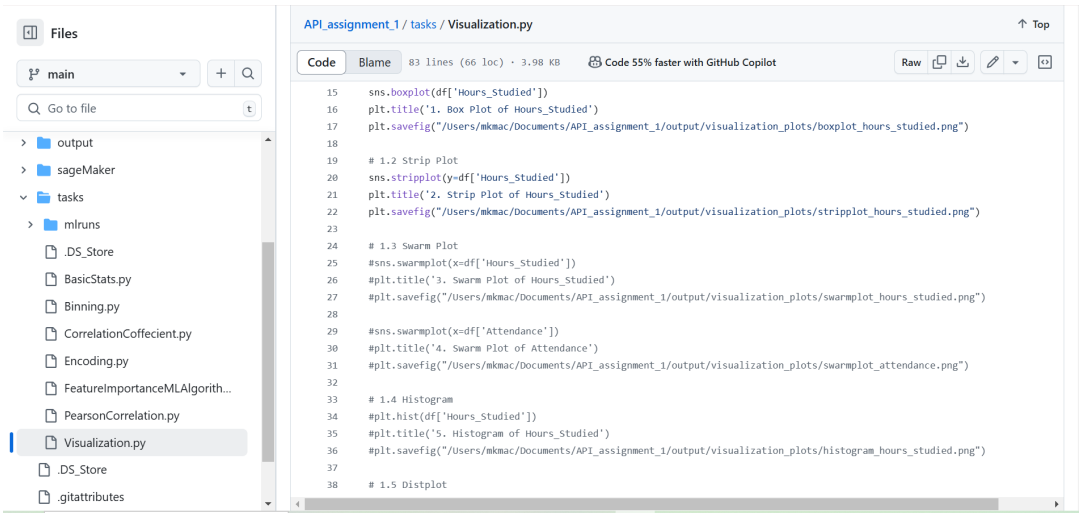
- output
- sageMaker
- tasks
  - mlruns
  - .DS_Store
  - BasicStats.py
  - Binning.py
  - CorrelationCoffecient.py
  - Encoding.py
  - FeatureImportanceMLAlgorith...
  - PearsonCorrelation.py

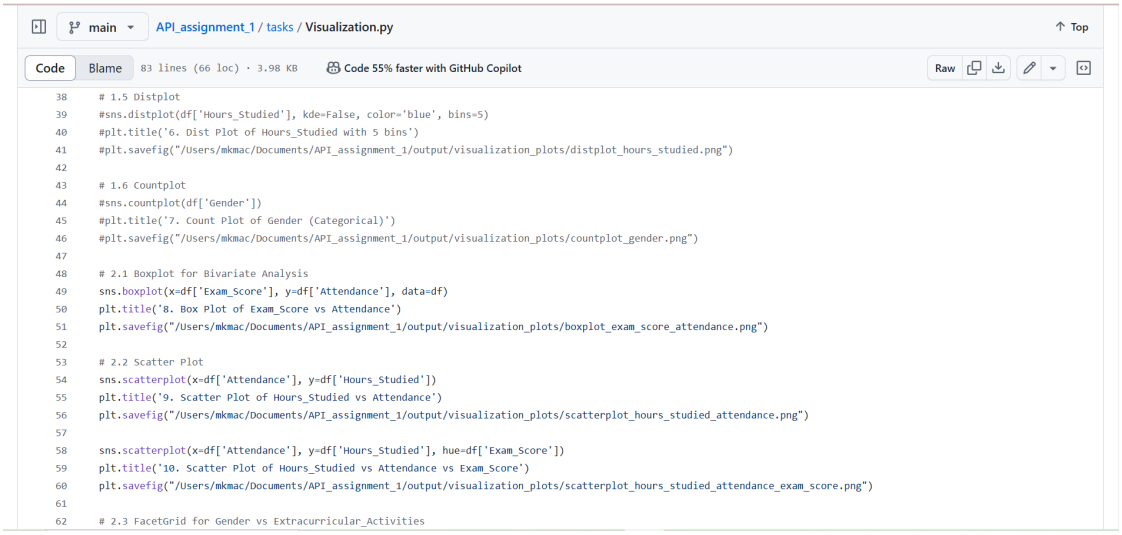
API_assignment_1 / tasks / Visualization.py

Mandrukamal Added process folder in screenshots — 3df0c03 · 1 hour ago — History

Code | Blame | 83 lines (66 loc) · 3.98 KB | Code 55% faster with GitHub Copilot | Raw

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr

# Set default theme
sns.set_theme()

# Import your data into Python
df = pd.read_csv("/Users/mkmac/Documents/API_assignment_1/data/ImputedStudentPerformanceFactors.csv")
print(df.index)

# 1.1 Box Plot
sns.boxplot(df['Hours_Studied'])
```


Files
main
Go to file
output
sageMaker
tasks
mlruns
.DS_Store
BasicStats.py
Binning.py
CorrelationCoffecient.py
Encoding.py
FeatureImportanceMLAlgorith...
PearsonCorrelation.py
Visualization.py
.DS_Store
.gitattributes
API_assignment_1 / tasks / Visualization.py
Top
Code
Blame
83 lines (66 loc) · 3.98 KB
Code 55% faster with GitHub Copilot
Raw
sns.boxplot(df['Hours_Studied'])
plt.title('1. Box Plot of Hours_Studied')
plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/boxplot_hours_studied.png")
# 1.2 Strip Plot
sns.stripplot(y=df['Hours_Studied'])
plt.title('2. Strip Plot of Hours_Studied')
plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/stripplot_hours_studied.png")
# 1.3 Swarm Plot
#sns.swarmplot(x=df['Hours_Studied'])
#plt.title('3. Swarm Plot of Hours_Studied')
#plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/swarmplot_hours_studied.png")
#sns.swarmplot(x=df['Attendance'])
#plt.title('4. Swarm Plot of Attendance')
#plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/swarmplot_attendance.png")
1.4 Histogram
#plt.hist(df['Hours_Studied'])
#plt.title('5. Histogram of Hours_Studied')
#plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/histogram_hours_studied.png")
1.5 Distplot

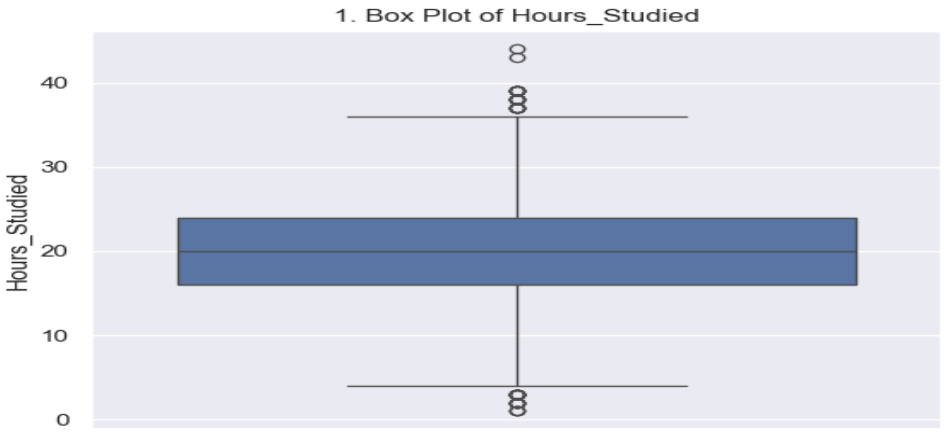

main
API_assignment_1 / tasks / Visualization.py
Top
Code
Blame
83 lines (66 loc) · 3.98 KB
Code 55% faster with GitHub Copilot
Raw
1.5 Distplot
#sns.distplot(df['Hours_Studied'], kde=False, color='blue', bins=5)
#plt.title('6. Dist Plot of Hours_Studied with 5 bins')
#plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/distplot_hours_studied.png")
1.6 Countplot
#sns.countplot(df['Gender'])
#plt.title('7. Count Plot of Gender (Categorical)')
#plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/countplot_gender.png")
# 2.1 Boxplot for Bivariate Analysis
sns.boxplot(x=df['Exam_Score'], y=df['Attendance'], data=df)
plt.title('8. Box Plot of Exam_Score vs Attendance')
plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/boxplot_exam_score_attendance.png")
# 2.2 Scatter Plot
sns.scatterplot(x=df['Attendance'], y=df['Hours_Studied'])
plt.title('9. Scatter Plot of Hours_Studied vs Attendance')
plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/scatterplot_hours_studied_attendance.png")
sns.scatterplot(x=df['Attendance'], y=df['Hours_Studied'], hue=df['Exam_Score'])
plt.title('10. Scatter Plot of Hours_Studied vs Attendance vs Exam_Score')
plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/scatterplot_hours_studied_attendance_exam_score.png")
# 2.3 FacetGrid for Gender vs Extracurricular_Activities

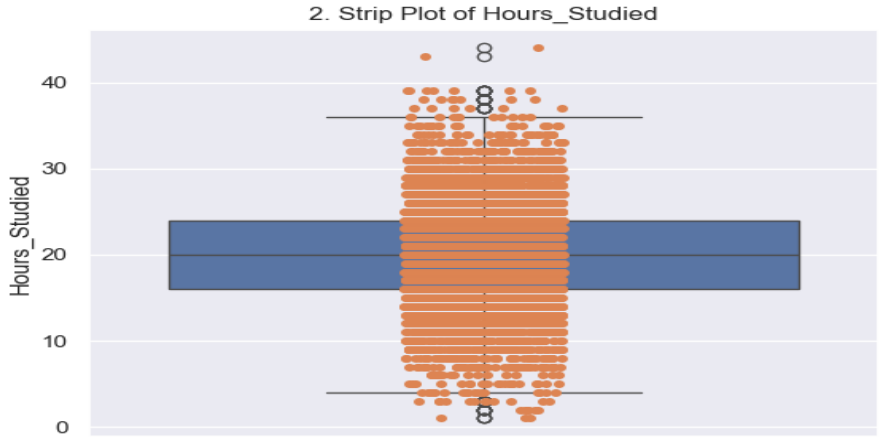

main
API_assignment_1 / tasks / Visualization.py
Top
Code
Blame
83 lines (66 loc) · 3.98 KB
Code 55% faster with GitHub Copilot
Raw
2.3 FacetGrid for Gender vs Extracurricular_Activities
#g = sns.FacetGrid(df, col="Gender", height=6.5, aspect=.85)
#g.map(sns.histplot, "Extracurricular_Activities")
#plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/facetgrid_gender_extracurricular.png")
Multivariate Analysis - FacetGrid
#g = sns.FacetGrid(df, col="Extracurricular_Activities", hue="Gender", margin_titles=True, height=6.5, aspect=.85)
#g.map(sns.histplot, "Hours_Studied")
#plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/facetgrid_hours_studied_extracurricular.png")
# 2.4 lmplot
sns.lmplot(data=df, x="Hours_Studied", y="Extracurricular_Activities", hue="Gender")
plt.title('13. lmplot of Hours_Studied vs Extracurricular_Activities vs Gender')
plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/lmplot_hours_studied_extracurricular.png")
#sns.lmplot(data=df, x="Hours_Studied", y="Attendance", hue="Gender")
#plt.title('14. lmplot of Hours_Studied vs Attendance vs Gender')
#plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/lmplot_hours_studied_attendance_gender.png")
#sns.lmplot(data=df, x="Attendance", y="Hours_Studied", hue="Extracurricular_Activities")
#plt.title('15. lmplot of Hours_Studied vs Attendance vs Extracurricular Activities')
#plt.savefig("/Users/mkmac/Documents/API_assignment_1/output/visualization_plots/lmplot_attendance_hours_studied_extracurricular.png")

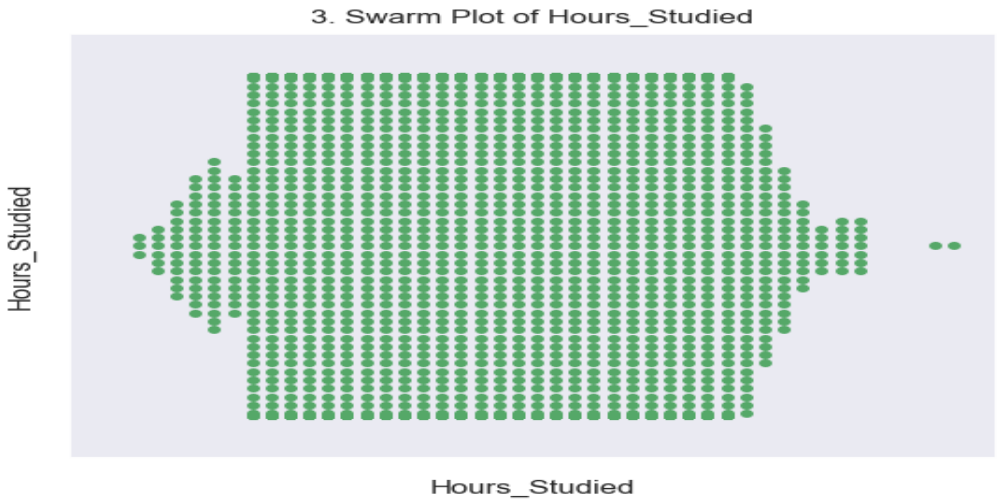
```
(venv) mkmac@MKMacs-MacBook-Air tasks % ls
BasicStats.py                       PearsonCorrelation.py
Binning.py                          Visualization.py
CorrelationCoffecient.py            figures
Encoding.py                         mlruns
FeatureImportanceMLAlgorithms.py
(venv) mkmac@MKMacs-MacBook-Air tasks % python Visualization.py
RangeIndex(start=0, stop=6607, step=1)
(venv) mkmac@MKMacs-MacBook-Air tasks %
```

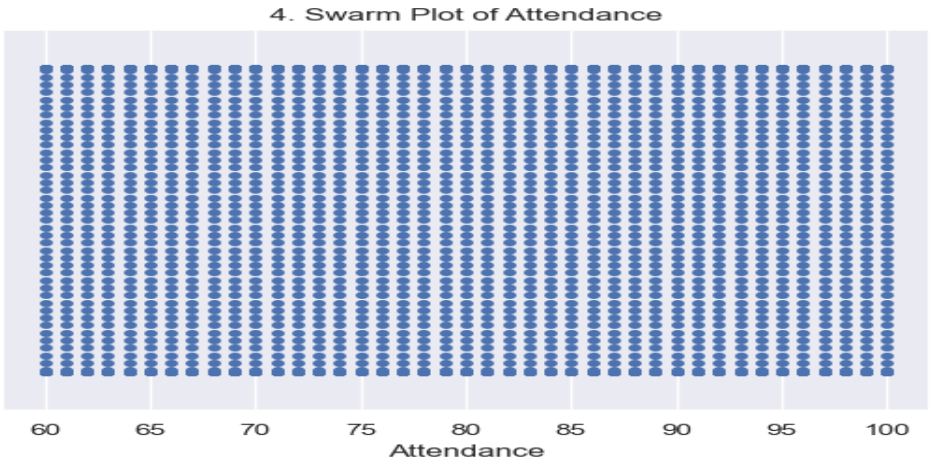
Outputs: -

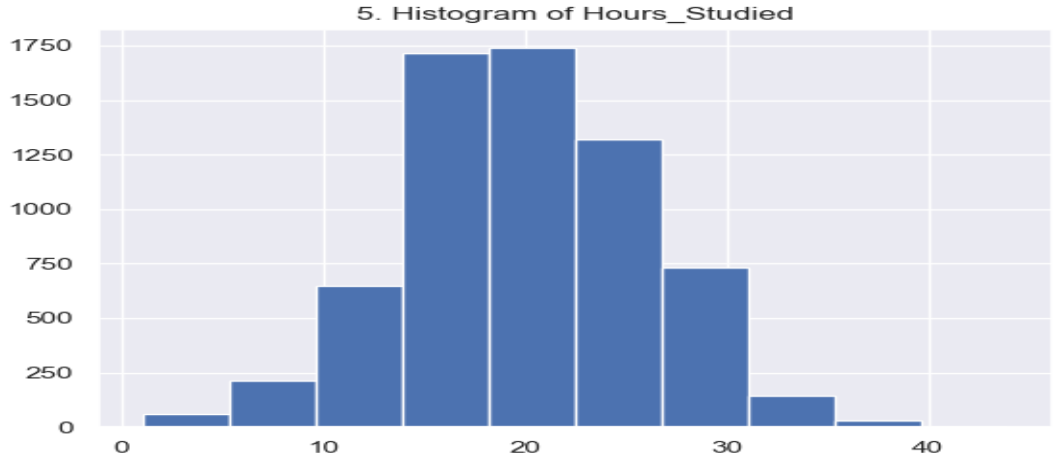
Univariate

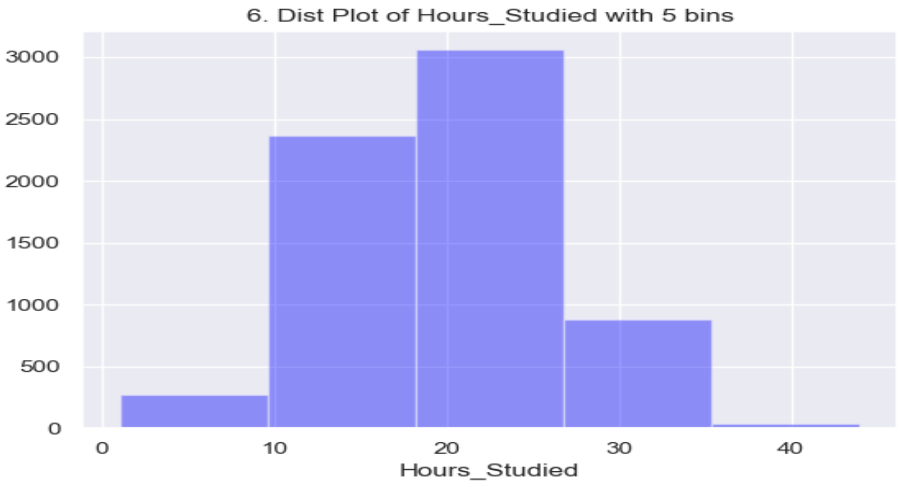


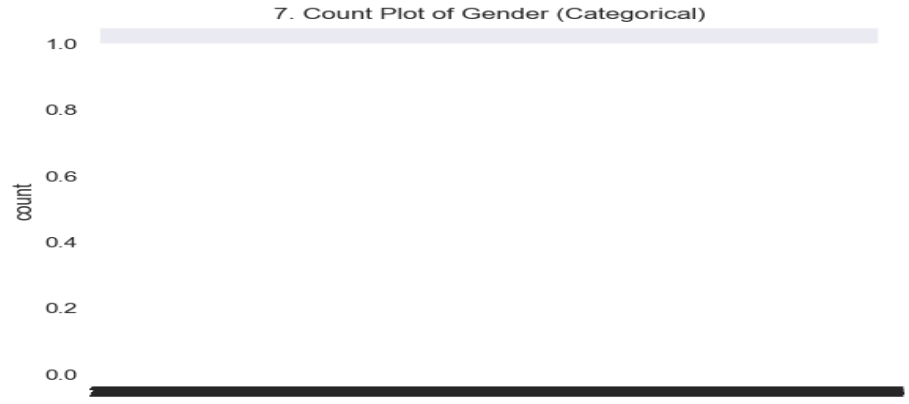



3. Swarm Plot of Hours_Studied
Hours_Studied
Hours_Studied

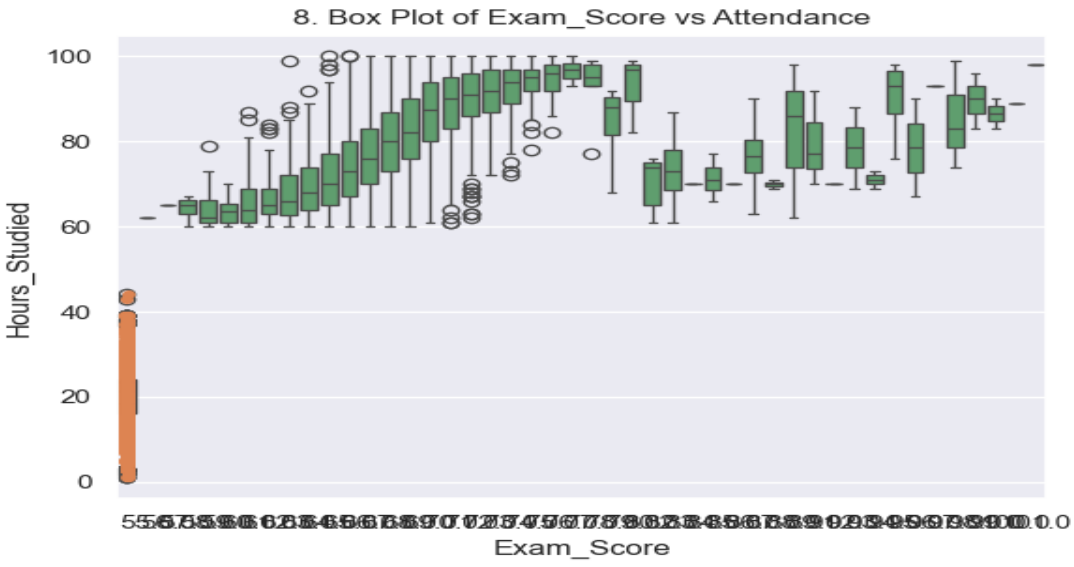

4. Swarm Plot of Attendance
60
65
70
75
80
85
90
95
100
Attendance

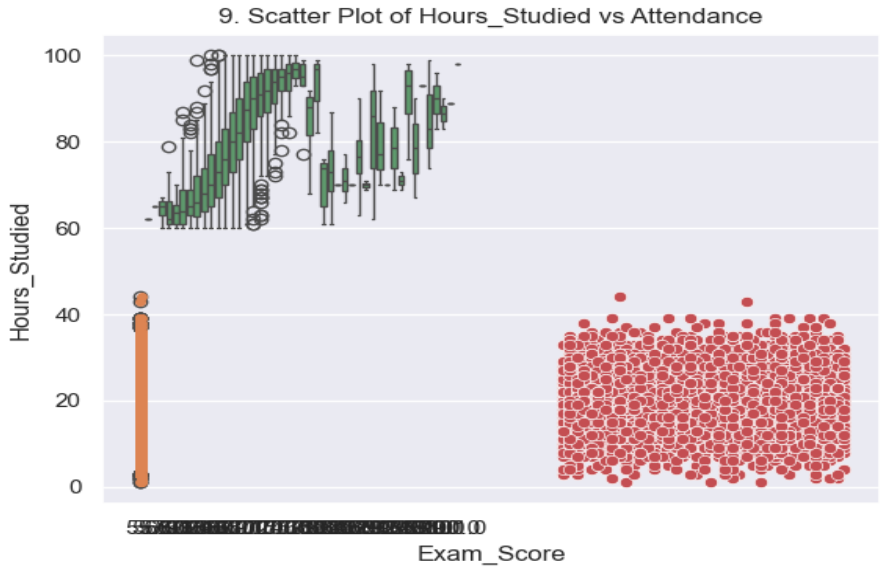

5. Histogram of Hours_Studied
1750
1500
1250
1000
750
500
250
0
0
10
20
30
40

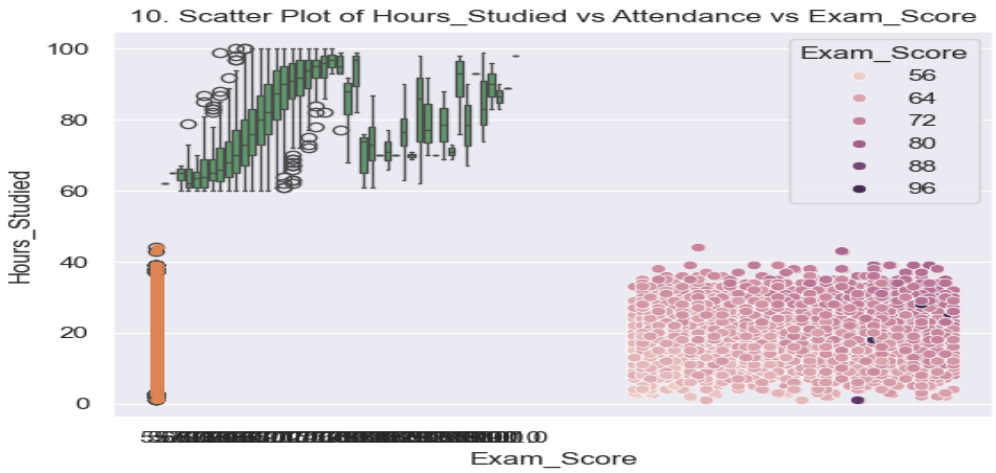

6. Dist Plot of Hours_Studied with 5 bins
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
Hours_Studied

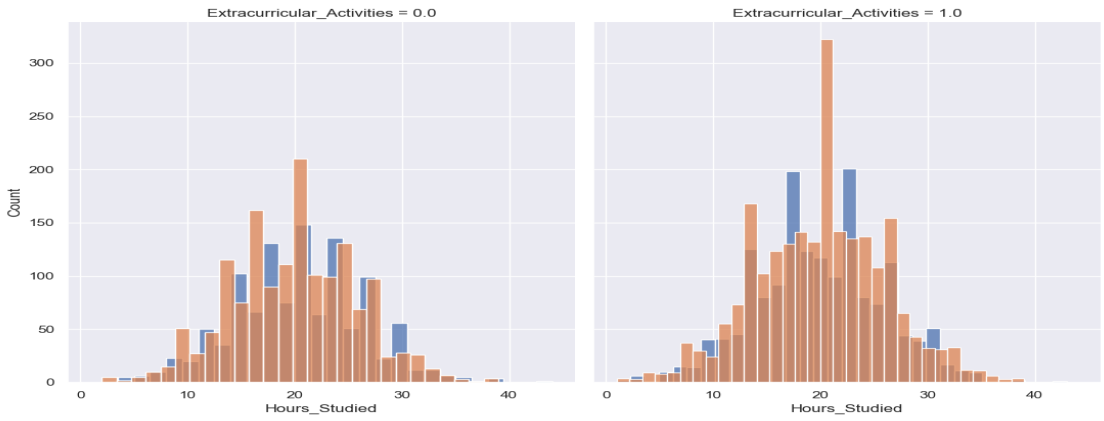

7. Count Plot of Gender (Categorical)
1.0
0.8
0.6
0.4
0.2
0.0
count

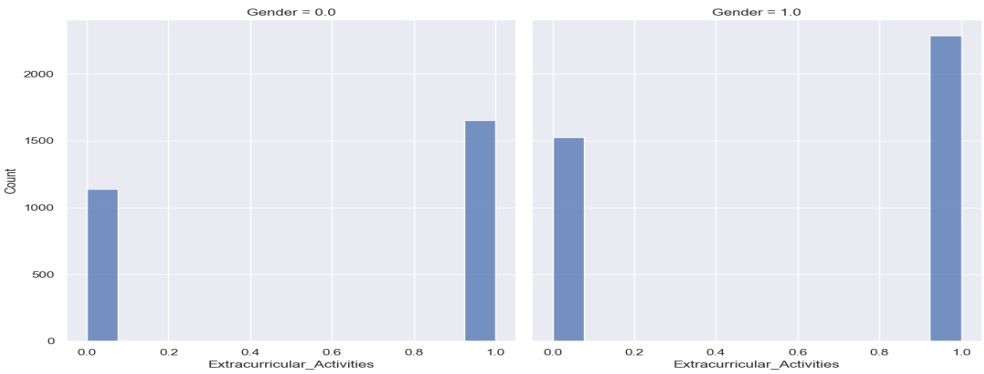
Bivariate

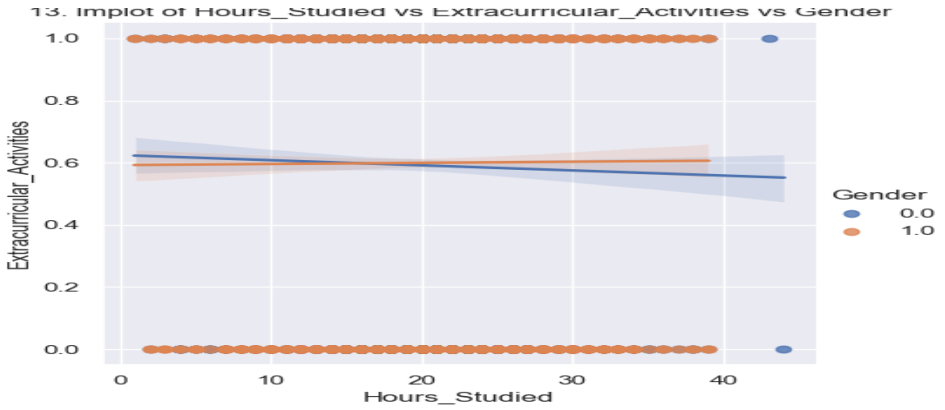

8. Box Plot of Exam_Score vs Attendance
100
80
60
40
20
0
Hours_Studied
Exam_Score

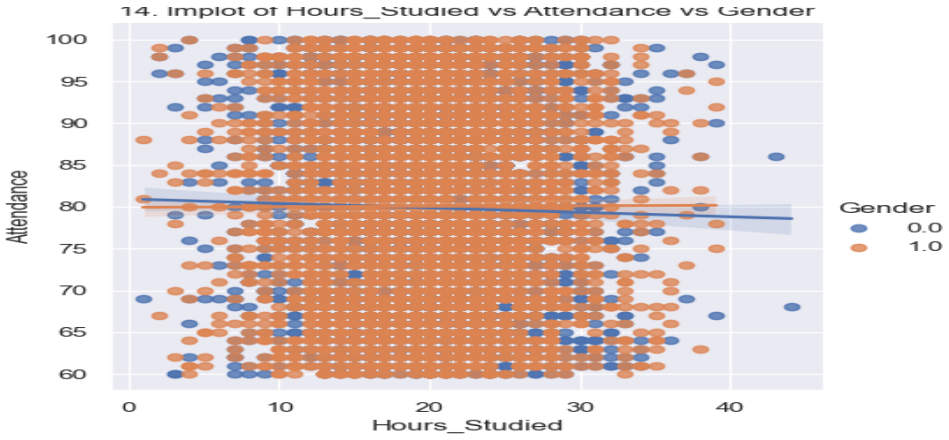

9. Scatter Plot of Hours_Studied vs Attendance
Hours_Studied
Exam_Score
100
80
60
40
20
0

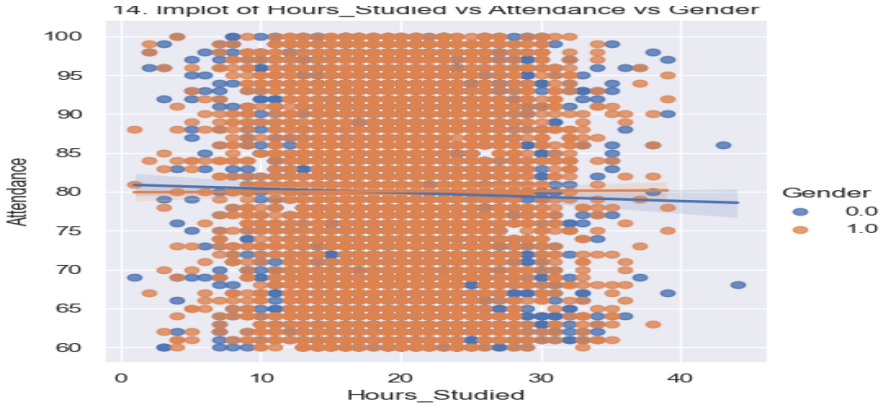

10. Scatter Plot of Hours_Studied vs Attendance vs Exam_Score
Hours_Studied
Exam_Score
Exam_Score
56
64
72
80
88
96
100
80
60
40
20
0

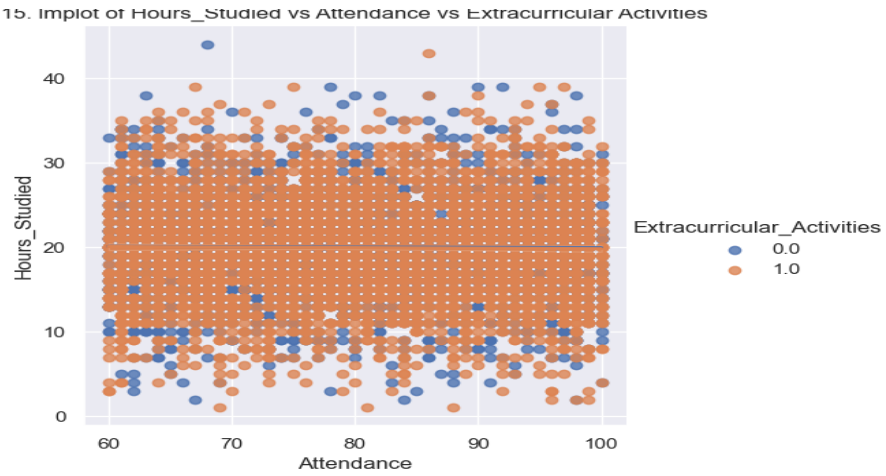

Extracurricular_Activities = 0.0
Extracurricular_Activities = 1.0
Count
Hours_Studied
Hours_Studied
300
250
200
150
100
50
0
0
10
20
30
40


Gender = 0.0
Gender = 1.0
Count
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Extracurricular_Activities


13. Implot of Hours_Studied vs Extracurricular_Activities vs Gender
Extracurricular_Activities
Hours_Studied
Gender
0.0
1.0

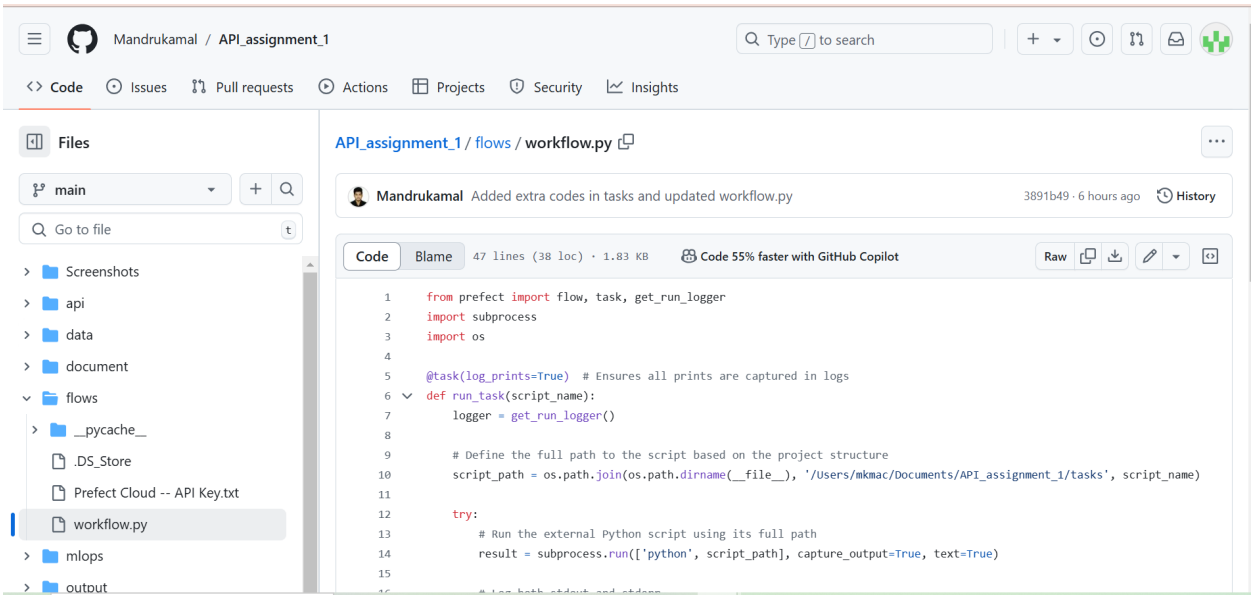

14. Implot of Hours_Studied vs Attendance vs Gender
Attendance
Hours_Studied
Gender
0.0
1.0

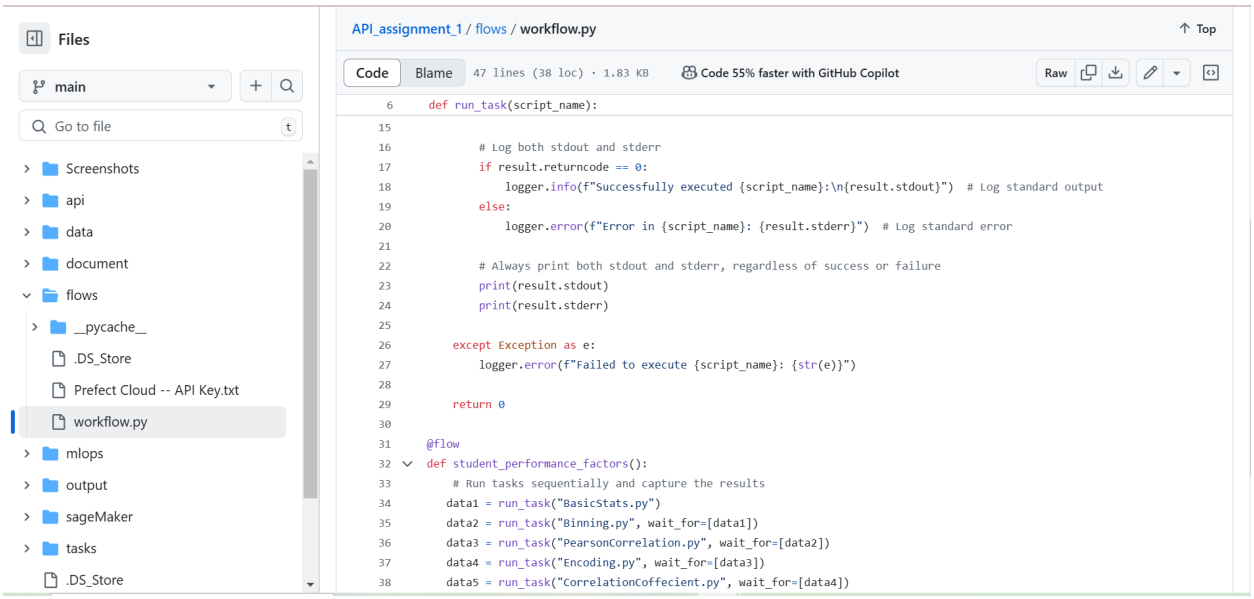

14. Implot of Hours_Studied vs Attendance vs Gender
Attendance
Hours_Studied
Gender
0.0
1.0



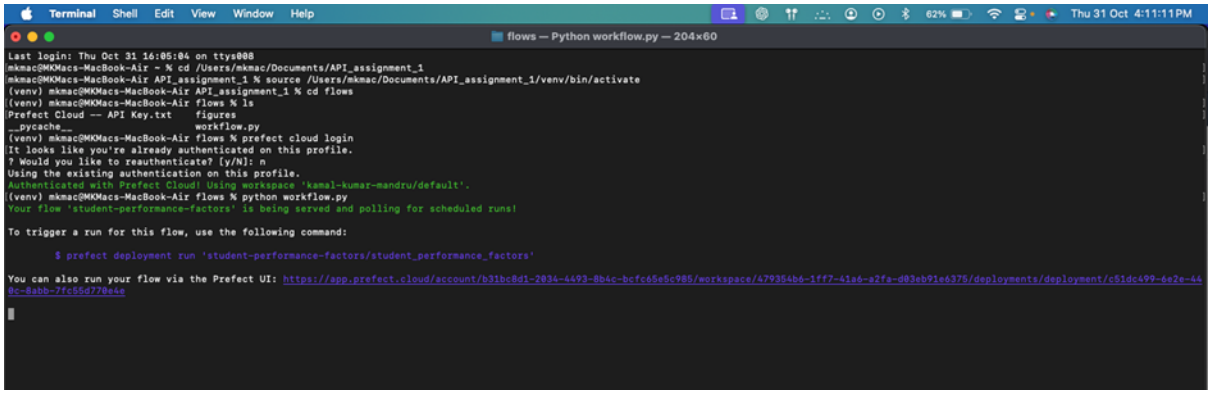
1.5 <u>DataOps</u>: Implement workflows to automate activities from steps 1.3 and 1.4 within a data pipeline. Schedule these workflows to run every 2 minutes, logging all activity details and displaying them on a Cloud dashboard.

<u>Workflow.py: -</u>

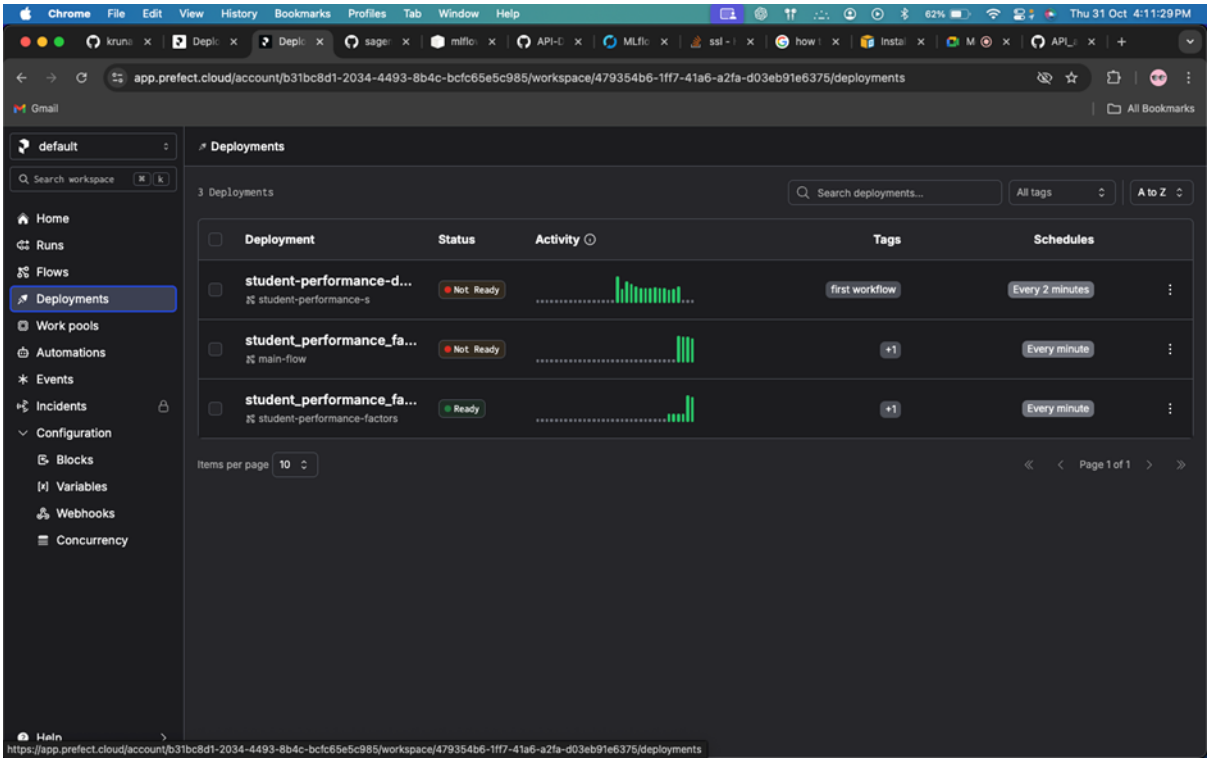



Files
main
Go to file
Screenshots
api
data
document
flows
__pycache__
.DS_Store
Prefect Cloud -- API Key.txt
workflow.py
mlops
output
sageMaker
tasks
.DS_Store
API_assignment_1 / flows / workflow.py
Top
Code
Blame
47 lines (38 loc) · 1.83 KB
Code 55% faster with GitHub Copilot
Raw
def run_task(script_name):
# Log both stdout and stderr
if result.returncode == 0:
logger.info(f"Successfully executed {script_name}:\n{result.stdout}") # Log standard output
else:
logger.error(f"Error in {script_name}: {result.stderr}") # Log standard error
# Always print both stdout and stderr, regardless of success or failure
print(result.stdout)
print(result.stderr)
except Exception as e:
logger.error(f"Failed to execute {script_name}: {str(e)}")
return 0
@flow
def student_performance_factors():
# Run tasks sequentially and capture the results
data1 = run_task("BasicStats.py")
data2 = run_task("Binning.py", wait_for=[data1])
data3 = run_task("PearsonCorrelation.py", wait_for=[data2])
data4 = run_task("Encoding.py", wait_for=[data3])
data5 = run_task("CorrelationCoffecient.py", wait_for=[data4])

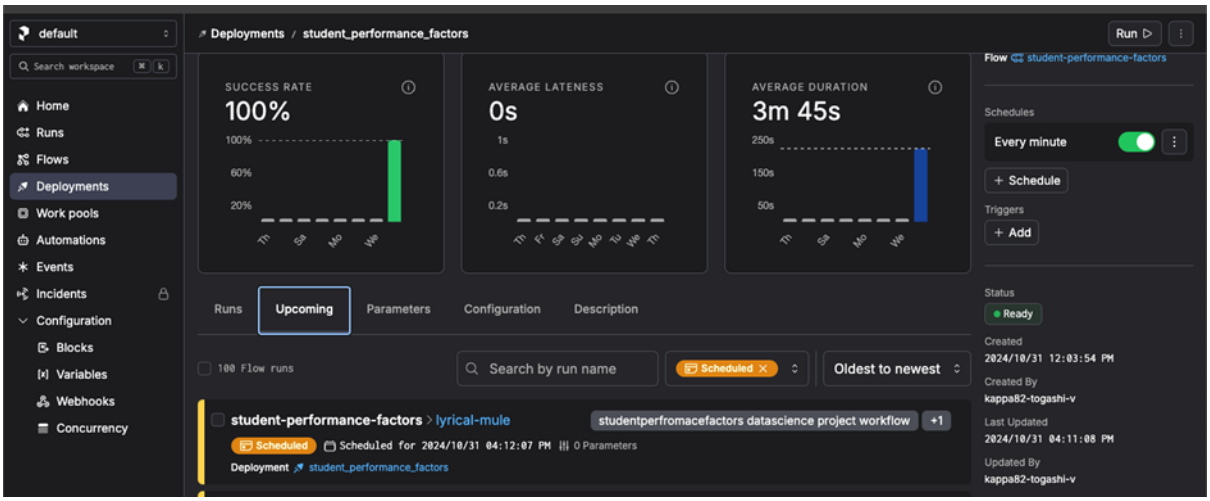

Files
main
Go to file
Screenshots
api
data
document
flows
__pycache__
.DS_Store
Prefect Cloud -- API Key.txt
workflow.py
mlops
output
sageMaker
tasks
.DS_Store
API_assignment_1 / flows / workflow.py
Top
Code
Blame
47 lines (38 loc) · 1.83 KB
Code 55% faster with GitHub Copilot
Raw
def run_task(script_name):
except Exception as e:
logger.error(f"Failed to execute {script_name}: {str(e)}")
return 0
@flow
def student_performance_factors():
# Run tasks sequentially and capture the results
data1 = run_task("BasicStats.py")
data2 = run_task("Binning.py", wait_for=[data1])
data3 = run_task("PearsonCorrelation.py", wait_for=[data2])
data4 = run_task("Encoding.py", wait_for=[data3])
data5 = run_task("CorrelationCoffecient.py", wait_for=[data4])
data6 = run_task("FeatureImportanceMLAlgorithms.py", wait_for=[data5])
data7 = run_task("Visualization.py", wait_for=[data6])
# To run locally
if __name__ == "__main__":
student_performance_factors.serve(name="student_performance_factors",
tags=["studentperfromacefactors datascience project workflow"],
parameters={},
interval=60)

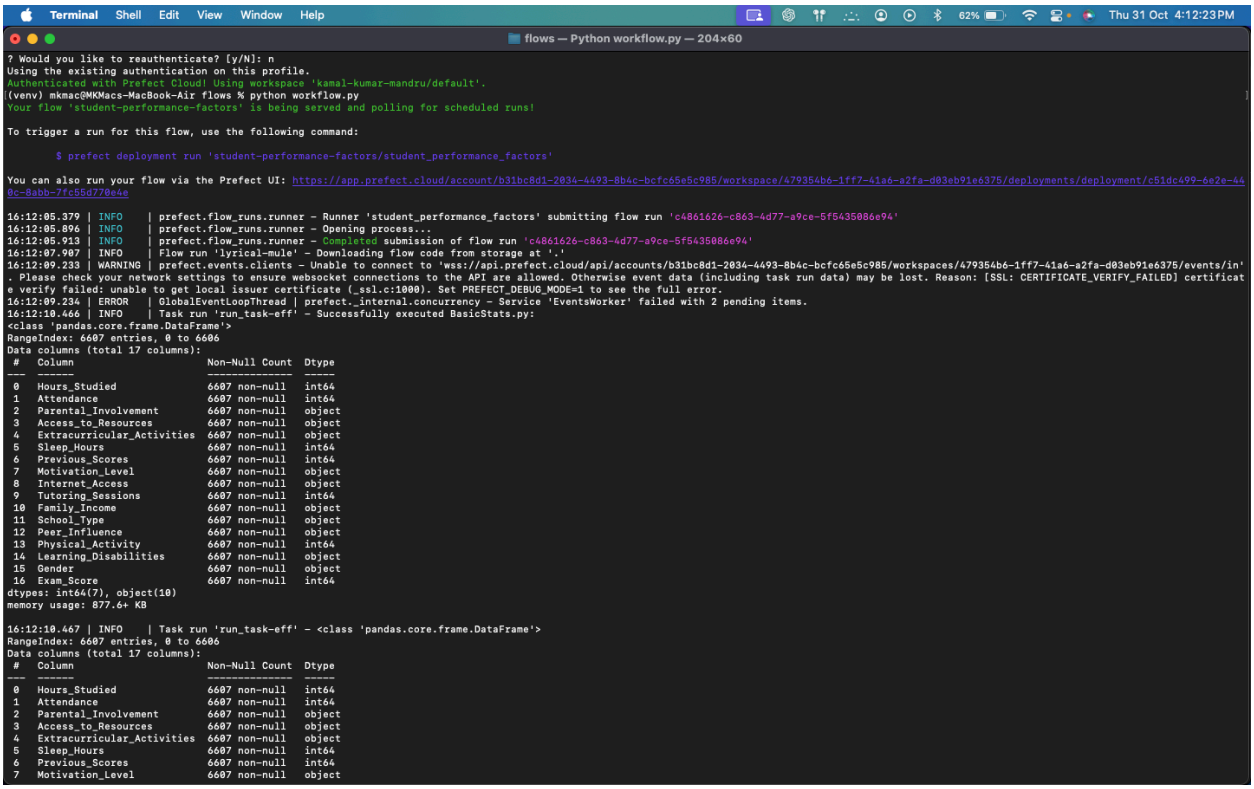

Terminal Shell Edit View Window Help
Thu 31 Oct 4:11:11 PM
flows — Python workflow.py — 204×60
Last login: Thu Oct 31 16:05:04 on ttys008
mkmac@MKMacs-MacBook-Air ~ % cd /Users/mkmac/Documents/API_assignment_1
mkmac@MKMacs-MacBook-Air API_assignment_1 % source /Users/mkmac/Documents/API_assignment_1/venv/bin/activate
(venv) mkmac@MKMacs-MacBook-Air API_assignment_1 % cd flows
(venv) mkmac@MKMacs-MacBook-Air flows % ls
Prefect Cloud -- API Key.txt figures
__pycache__ workflow.py
(venv) mkmac@MKMacs-MacBook-Air flows % prefect cloud login
It looks like you're already authenticated on this profile.
? Would you like to reauthenticate? [y/N]: n
Using the existing authentication on this profile.
Authenticated with Prefect Cloud! Using workspace 'kamal-kumar-mandru/default'.
(venv) mkmac@MKMacs-MacBook-Air flows % python workflow.py
Your flow 'student-performance-factors' is being served and polling for scheduled runs!
To trigger a run for this flow, use the following command:
$ prefect deployment run 'student-performance-factors/student_performance_factors'
You can also run your flow via the Prefect UI: https://app.prefect.cloud/account/b31bc8d1-2034-4493-8b4c-befc65e5c985/workspace/479354b6-1ff7-41a6-a2fa-d03eb91e6376/deployments/deployment/c51dc499-6e2e-440c-8abb-7fc55d770e4e

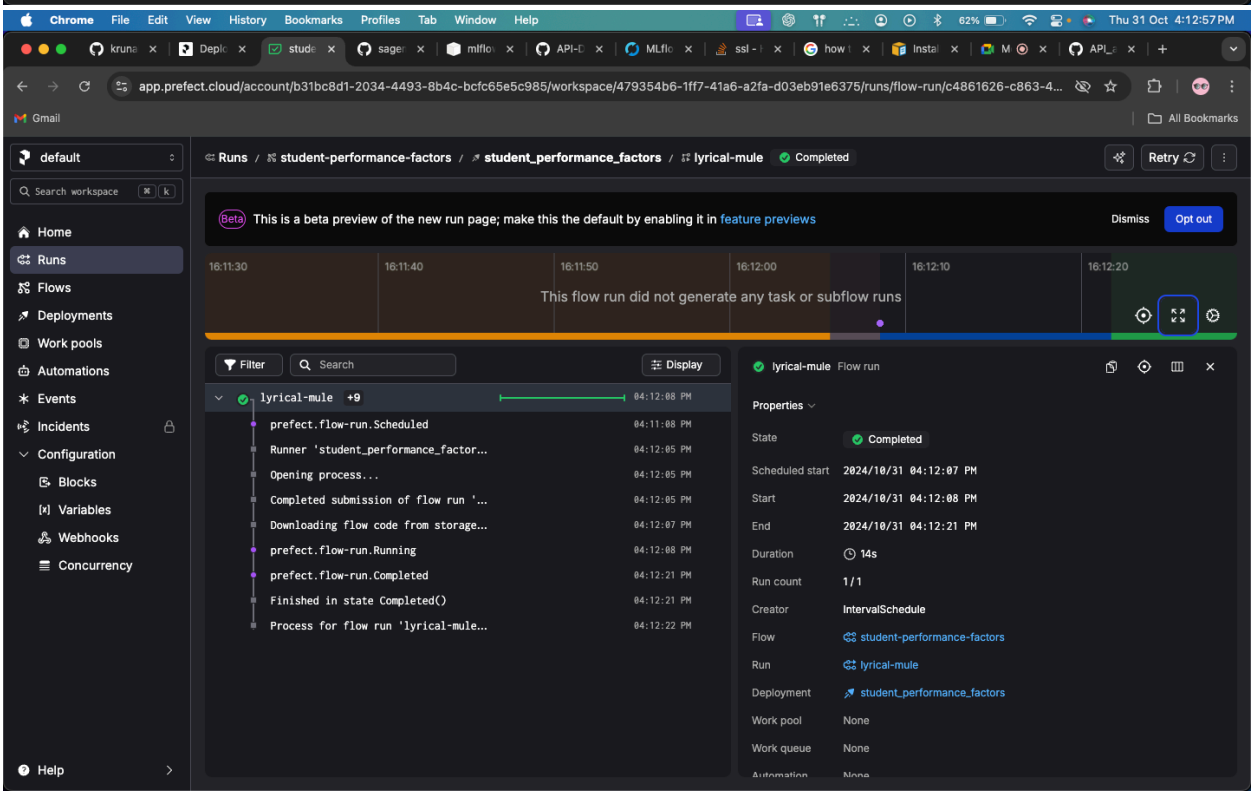
Scheduling workflows to run every 2 minutes

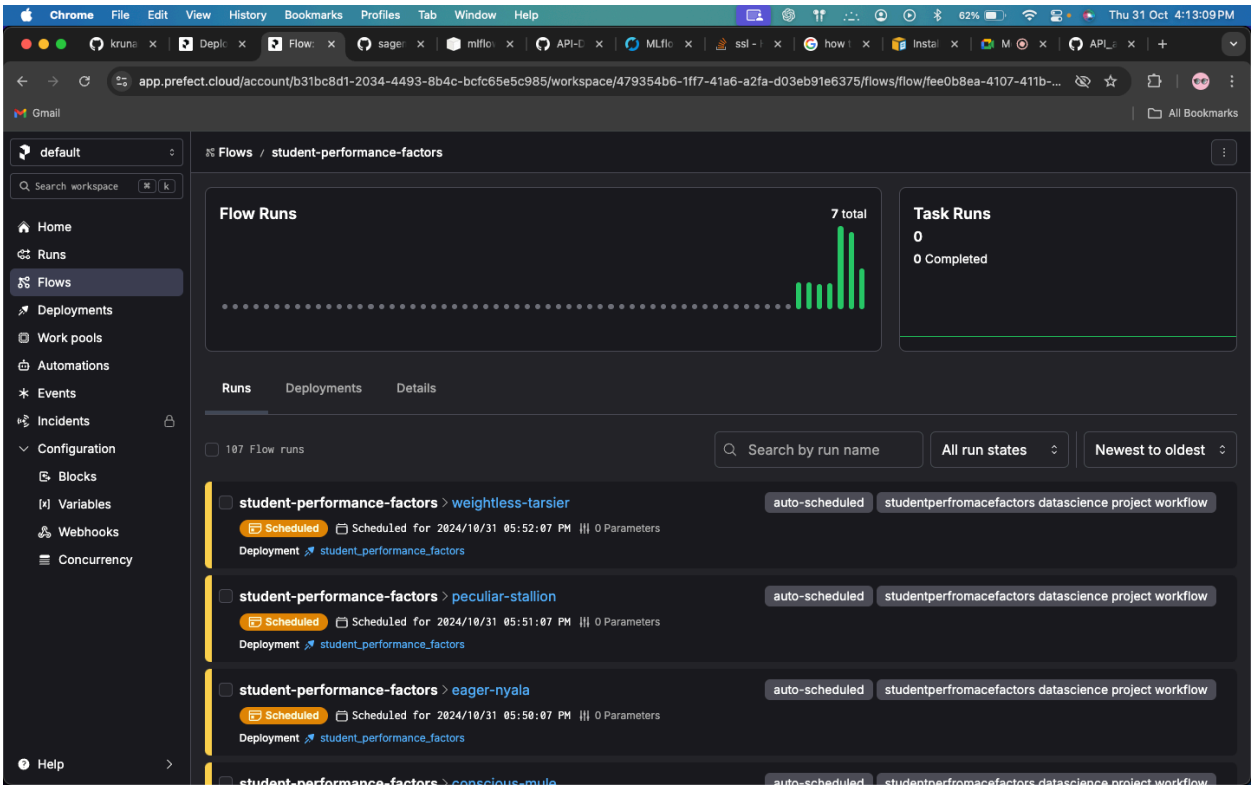


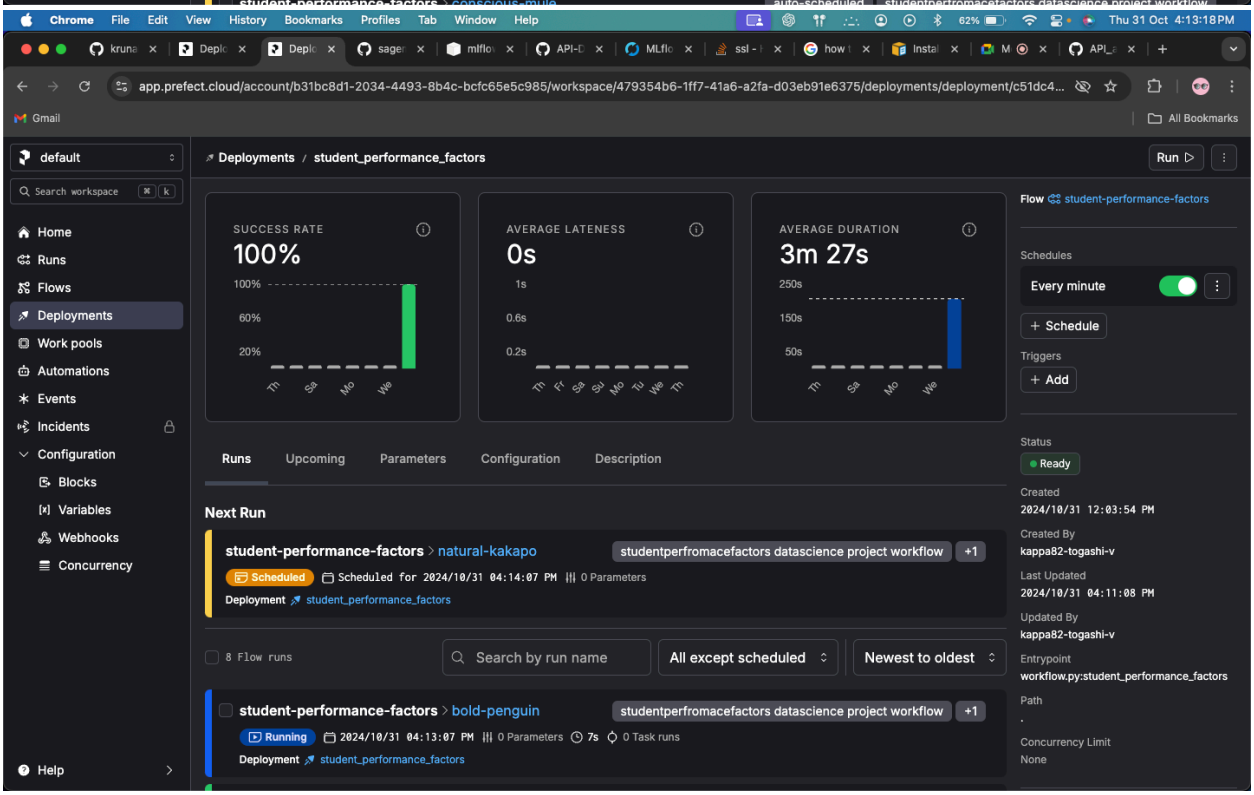
logging activity details and displaying them on Cloud dashboard

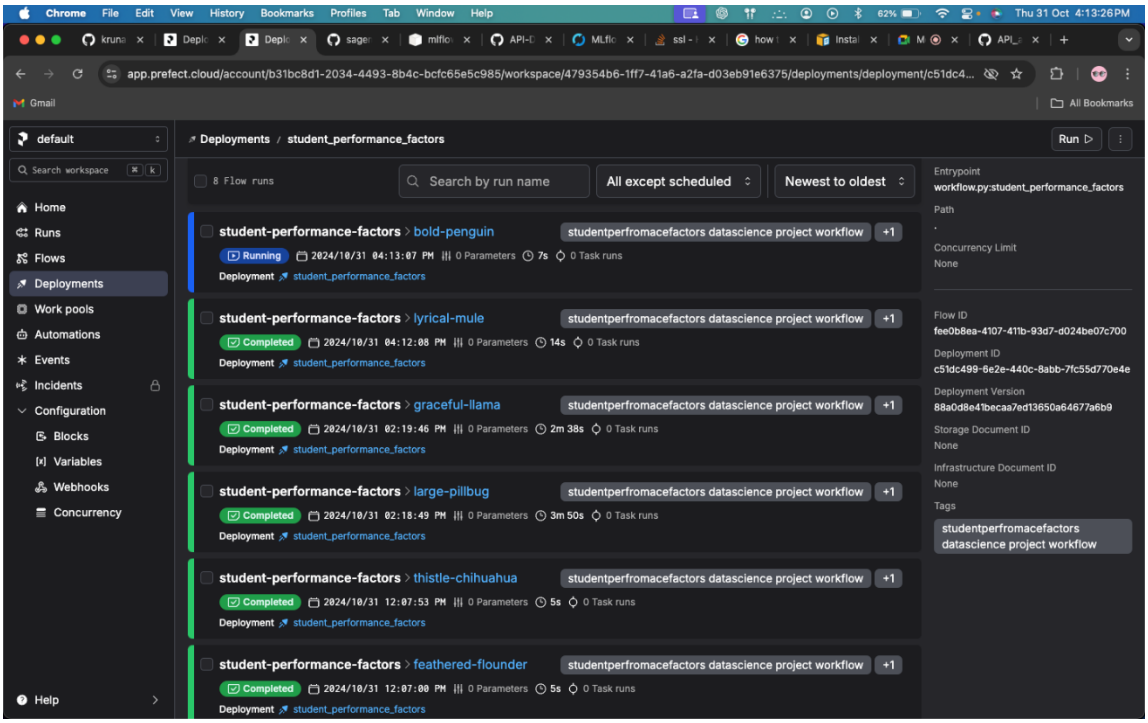


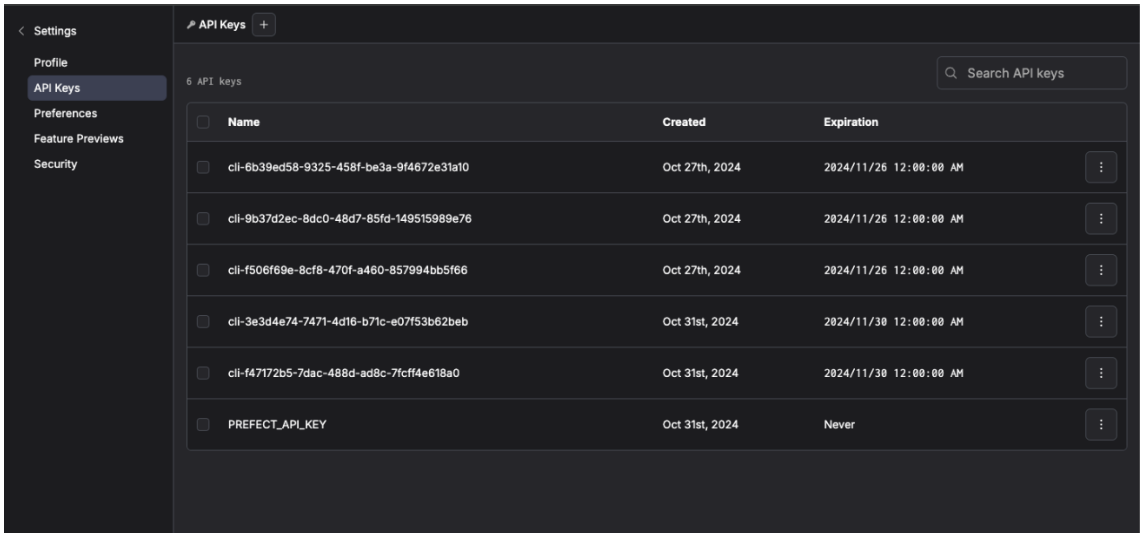

Terminal Shell Edit View Window Help
Thu 31 Oct 4:12:23 PM
flows — Python workflow.py — 204×60
? Would you like to reauthenticate? [y/N]: n
Using the existing authentication on this profile.
Authenticated with Prefect Cloud! Using workspace 'kamal-kumar-mandru/default'.
(venv) mkmac@MKMacs-MacBook-Air flows % python workflow.py
Your flow 'student-performance-factors' is being served and polling for scheduled runs!
To trigger a run for this flow, use the following command:
$ prefect deployment run 'student-performance-factors/student_performance_factors'
You can also run your flow via the Prefect UI: https://app.prefect.cloud/account/b31bc8d1-2034-4493-8b4c-bcfc65e5c985/workspace/479354b6-1ff7-41a6-a2fa-d03eb91e6375/deployments/deployment/c51dc499-6e2e-440c-8abb-7fc55d770e4e
16:12:05.379 | INFO | prefect.flow_runs.runner - Runner 'student_performance_factors' submitting flow run 'c4861626-c863-4d77-a9ce-5f5435086e94'
16:12:05.896 | INFO | prefect.flow_runs.runner - Opening process...
16:12:05.913 | INFO | prefect.flow_runs.runner - Completed submission of flow run 'c4861626-c863-4d77-a9ce-5f5435086e94'
16:12:07.907 | INFO | Flow run 'lyrical-mule' - Downloading flow code from storage at '.'
16:12:09.233 | WARNING | prefect.events.clients - Unable to connect to 'wss://api.prefect.cloud/api/accounts/b31bc8d1-2034-4493-8b4c-bcfc65e5c985/workspaces/479354b6-1ff7-41a6-a2fa-d03eb91e6375/events/in'. Please check your network settings to ensure websocket connections to the API are allowed. Otherwise event data (including task run data) may be lost. Reason: [SSL: CERTIFICATE_VERIFY_FAILED] certificate verify failed: unable to get local issuer certificate (_ssl.c:1000). Set PREFECT_DEBUG_MODE=1 to see the full error.
16:12:09.234 | ERROR | GlobalEventLoopThread | prefect._internal.concurrency - Service 'EventsWorker' failed with 2 pending items.
16:12:10.466 | INFO | Task run 'run_task-eff' - Successfully executed BasicStats.py:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6607 entries, 0 to 6606
Data columns (total 17 columns):
Column Non-Null Count Dtype
0 Hours_Studied 6607 non-null int64
1 Attendance 6607 non-null int64
2 Parental_Involvement 6607 non-null object
3 Access_to_Resources 6607 non-null object
4 Extracurricular_Activities 6607 non-null object
5 Sleep_Hours 6607 non-null int64
6 Previous_Scores 6607 non-null int64
7 Motivation_Level 6607 non-null object
8 Internet_Access 6607 non-null object
9 Tutoring_Sessions 6607 non-null int64
10 Family_Income 6607 non-null object
11 School_Type 6607 non-null object
12 Peer_Influence 6607 non-null object
13 Physical_Activity 6607 non-null int64
14 Learning_Disabilities 6607 non-null object
15 Gender 6607 non-null object
16 Exam_Score 6607 non-null int64
dtypes: int64(7), object(10)
memory usage: 877.6+ KB
16:12:10.467 | INFO | Task run 'run_task-eff' - <class 'pandas.core.frame.DataFrame'>
RangeIndex: 6607 entries, 0 to 6606
Data columns (total 17 columns):
Column Non-Null Count Dtype
0 Hours_Studied 6607 non-null int64
1 Attendance 6607 non-null int64
2 Parental_Involvement 6607 non-null object
3 Access_to_Resources 6607 non-null object
4 Extracurricular_Activities 6607 non-null object
5 Sleep_Hours 6607 non-null int64
6 Previous_Scores 6607 non-null int64
7 Motivation_Level 6607 non-null object

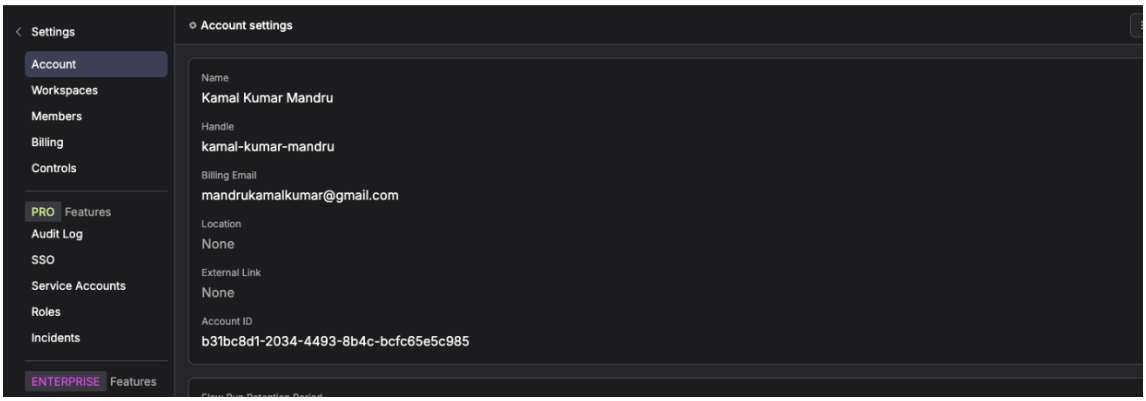

Chrome File Edit View History Bookmarks Profiles Tab Window Help
Thu 31 Oct 4:12:57 PM
app.prefect.cloud/account/b31bc8d1-2034-4493-8b4c-bcfc65e5c985/workspace/479354b6-1ff7-41a6-a2fa-d03eb91e6375/runs/flow-run/c4861626-c863-4...
Gmail
All Bookmarks
default
Search workspace
Home
Runs
Flows
Deployments
Work pools
Automations
Events
Incidents
Configuration
Blocks
Variables
Webhooks
Concurrency
Help
Runs / student-performance-factors / student_performance_factors / lyrical-mule
Completed
Retry
Beta This is a beta preview of the new run page; make this the default by enabling it in feature previews
Dismiss
Opt out
16:11:30
16:11:40
16:11:50
16:12:00
16:12:10
16:12:20
This flow run did not generate any task or subflow runs
Filter
Search
Display
lyrical-mule +9
04:12:08 PM
prefect.flow-run.Scheduled 04:11:08 PM
Runner 'student_performance_factor... 04:12:05 PM
Opening process... 04:12:05 PM
Completed submission of flow run '... 04:12:05 PM
Downloading flow code from storage... 04:12:07 PM
prefect.flow-run.Running 04:12:08 PM
prefect.flow-run.Completed 04:12:21 PM
Finished in state Completed() 04:12:21 PM
Process for flow run 'lyrical-mule... 04:12:22 PM
lyrical-mule Flow run
Properties
State Completed
Scheduled start 2024/10/31 04:12:07 PM
Start 2024/10/31 04:12:08 PM
End 2024/10/31 04:12:21 PM
Duration 14s
Run count 1 / 1
Creator IntervalSchedule
Flow student-performance-factors
Run lyrical-mule
Deployment student_performance_factors
Work pool None
Work queue None


app.prefect.cloud/account/b31bc8d1-2034-4493-8b4c-bcfc65e5c985/workspace/479354b6-1ff7-41a6-a2fa-d03eb91e6375/flows/flow/fee0b8ea-4107-411b-...
default
Search workspace
Home
Runs
Flows
Deployments
Work pools
Automations
Events
Incidents
Configuration
Blocks
Variables
Webhooks
Concurrency
Help
Flows / student-performance-factors
Flow Runs
7 total
Task Runs
0
0 Completed
Runs
Deployments
Details
107 Flow runs
Search by run name
All run states
Newest to oldest
student-performance-factors > weightless-tarsier
auto-scheduled
studentperfromacefactors datascience project workflow
Scheduled
Scheduled for 2024/10/31 05:52:07 PM
0 Parameters
Deployment
student_performance_factors
student-performance-factors > peculiar-stallion
auto-scheduled
studentperfromacefactors datascience project workflow
Scheduled
Scheduled for 2024/10/31 05:51:07 PM
0 Parameters
Deployment
student_performance_factors
student-performance-factors > eager-nyala
auto-scheduled
studentperfromacefactors datascience project workflow
Scheduled
Scheduled for 2024/10/31 05:50:07 PM
0 Parameters
Deployment
student_performance_factors


app.prefect.cloud/account/b31bc8d1-2034-4493-8b4c-bcfc65e5c985/workspace/479354b6-1ff7-41a6-a2fa-d03eb91e6375/deployments/deployment/c51dc4...
default
Search workspace
Home
Runs
Flows
Deployments
Work pools
Automations
Events
Incidents
Configuration
Blocks
Variables
Webhooks
Concurrency
Help
Deployments / student_performance_factors
Run
SUCCESS RATE
100%
AVERAGE LATENESS
0s
AVERAGE DURATION
3m 27s
Flow student-performance-factors
Schedules
Every minute
+ Schedule
Triggers
+ Add
Runs
Upcoming
Parameters
Configuration
Description
Next Run
student-performance-factors > natural-kakapo
studentperfromacefactors datascience project workflow
+1
Scheduled
Scheduled for 2024/10/31 04:14:07 PM
0 Parameters
Deployment
student_performance_factors
8 Flow runs
Search by run name
All except scheduled
Newest to oldest
student-performance-factors > bold-penguin
studentperfromacefactors datascience project workflow
+1
Running
2024/10/31 04:13:07 PM
0 Parameters
7s
0 Task runs
Deployment
student_performance_factors
Status
Ready
Created
2024/10/31 12:03:54 PM
Created By
kappa82-togashi-v
Last Updated
2024/10/31 04:11:08 PM
Updated By
kappa82-togashi-v
Entrypoint
workflow.py:student_performance_factors
Path
.
Concurrency Limit
None


app.prefect.cloud/account/b31bc8d1-2034-4493-8b4c-bcfc65e5c985/workspace/479354b6-1ff7-41a6-a2fa-d03eb91e6375/deployments/deployment/c51dc4...
default
Search workspace
Home
Runs
Flows
Deployments
Work pools
Automations
Events
Incidents
Configuration
Blocks
Variables
Webhooks
Concurrency
Help
Deployments / student_performance_factors
Run
8 Flow runs
Search by run name
All except scheduled
Newest to oldest
student-performance-factors > bold-penguin
Running 2024/10/31 04:13:07 PM 0 Parameters 7s 0 Task runs
Deployment student_performance_factors
student-performance-factors > lyrical-mule
Completed 2024/10/31 04:12:08 PM 0 Parameters 14s 0 Task runs
Deployment student_performance_factors
student-performance-factors > graceful-llama
Completed 2024/10/31 02:19:46 PM 0 Parameters 2m 38s 0 Task runs
Deployment student_performance_factors
student-performance-factors > large-pillbug
Completed 2024/10/31 02:18:49 PM 0 Parameters 3m 50s 0 Task runs
Deployment student_performance_factors
student-performance-factors > thistle-chihuahua
Completed 2024/10/31 12:07:53 PM 0 Parameters 5s 0 Task runs
Deployment student_performance_factors
student-performance-factors > feathered-flounder
Completed 2024/10/31 12:07:00 PM 0 Parameters 5s 0 Task runs
Deployment student_performance_factors
studentperfromacefactors datascience project workflow
+1
Entrypoint
workflow.py:student_performance_factors
Path
.
Concurrency Limit
None
Flow ID
fee0b8ea-4107-411b-93d7-d024be07c700
Deployment ID
c51dc499-6e2e-440c-8abb-7fc55d770e4e
Deployment Version
88a0d8e41becaa7ed13650a64677a6b9
Storage Document ID
None
Infrastructure Document ID
None
Tags


Settings
Profile
API Keys
Preferences
Feature Previews
Security
API Keys
6 API keys
Search API keys
Name
Created
Expiration
cli-6b39ed58-9325-458f-be3a-9f4672e31a10
Oct 27th, 2024
2024/11/26 12:00:00 AM
cli-9b37d2ec-8dc0-48d7-85fd-149515989e76
Oct 27th, 2024
2024/11/26 12:00:00 AM
cli-f506f69e-8cf8-470f-a460-857994bb5f66
Oct 27th, 2024
2024/11/26 12:00:00 AM
cli-3e3d4e74-7471-4d16-b71c-e07f53b62beb
Oct 31st, 2024
2024/11/30 12:00:00 AM
cli-f47172b5-7dac-488d-ad8c-7fcff4e618a0
Oct 31st, 2024
2024/11/30 12:00:00 AM
PREFECT_API_KEY
Oct 31st, 2024
Never


Settings
Account
Workspaces
Members
Billing
Controls
PRO Features
Audit Log
SSO
Service Accounts
Roles
Incidents
ENTERPRISE Features
Account settings
Name
Kamal Kumar Mandru
Handle
kamal-kumar-mandru
Billing Email
mandrukamalkumar@gmail.com
Location
None
External Link
None
Account ID
b31bc8d1-2034-4493-8b4c-bcfc65e5c985

Sub-Objective 2: Design and Development of a Machine Learning Pipeline

2.1 Model Preparation: Identify suitable machine learning algorithms for solving the business problem based on the dataset. Select any two algorithms.

For the dataset we used, we have opted for linear regression and random forest regression machine learning models for solving the business problems.

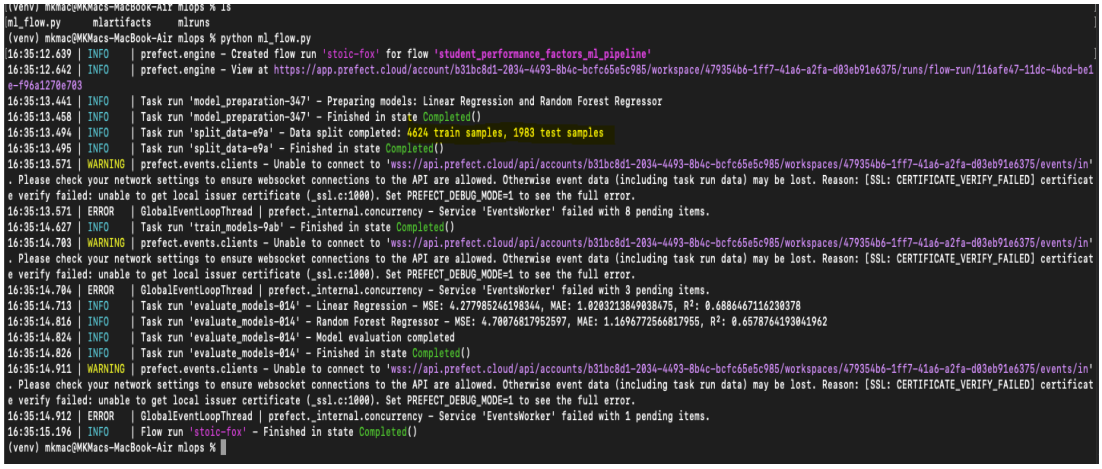
Linear Regression: -

A linear regression prediction model is a statistical tool used to predict the value of a dependent variable (response variable) based on one or more independent variables (predictor variables). This model assumes a linear relationship between the dependent and independent variables, meaning the change in the dependent variable is proportional to the change in the independent variables. Linear regression models are valued for their simplicity and interpretability, especially when the relationship between variables is approximately linear.

Random Forest Regression: -

A Random Forest Regression model is a predictive model that uses an ensemble (collection) of multiple decision trees to perform regression tasks. In this context, regression refers to predicting continuous numeric values, rather than discrete class labels. Random Forest Regression is well-suited for tasks like predicting housing prices, stock market values, or any scenario where continuous value prediction with reduced model variance is desired.

2.2 Model Training: Split the dataset into training (70%) and testing (30%) sets and train the models.
Below screenshot captured the sample split into 70-30 for training the model.
Where the total data samples are 6607, it has been divided as
4624 train samples and 1983 test samples as highlighted in yellow.



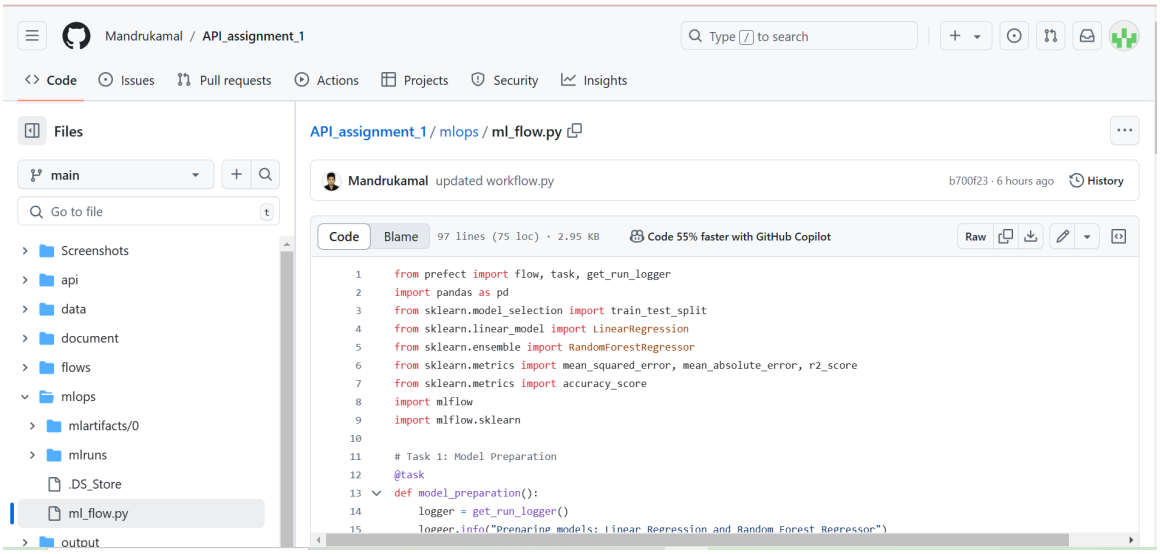
2.3 Model Evaluation: Evaluate the models using at least one metric (e.g., accuracy for classification models).

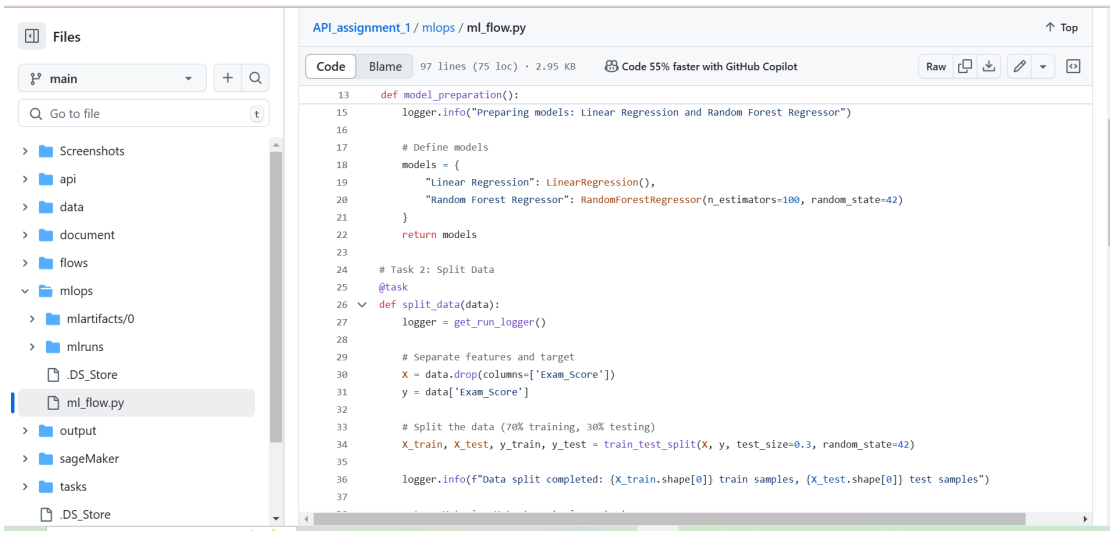
and

2.4 MLOps: Monitor the model and log relevant metrics (at least four, such as accuracy, precision, recall, F1 score, etc.)

Model evaluation

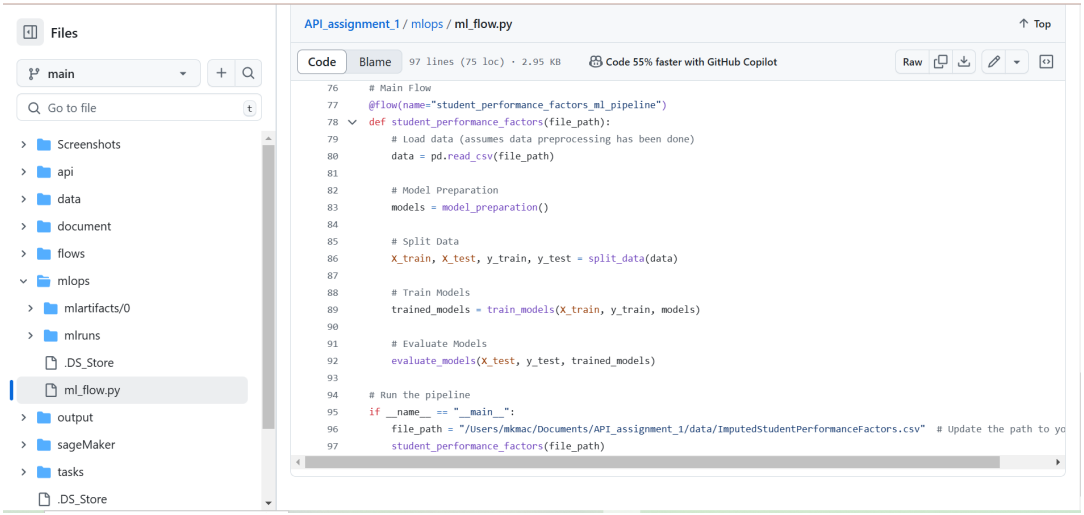
Ml_flow.py: -



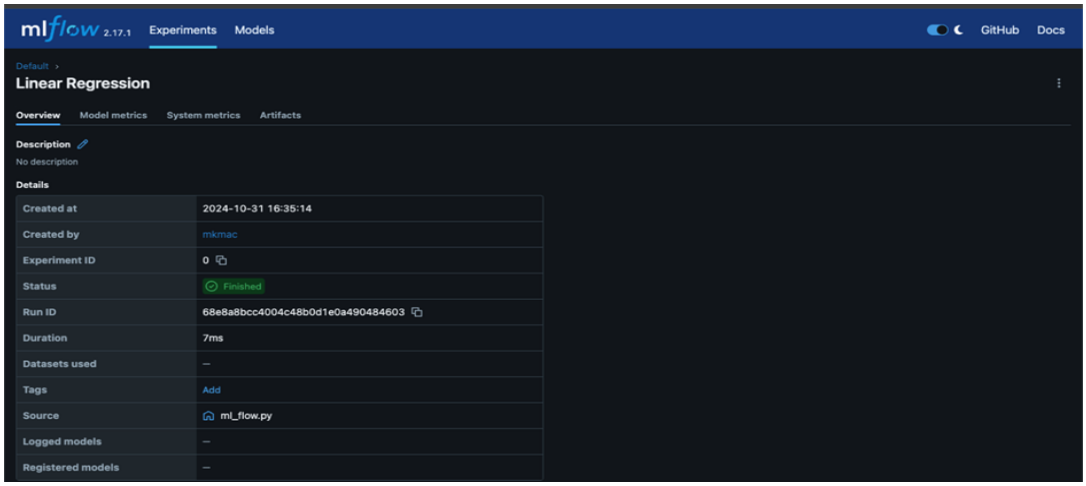



Files
main
Go to file
Screenshots
api
data
document
flows
mlops
mlartifacts/0
mlruns
.DS_Store
ml_flow.py
output
sageMaker
tasks
.DS_Store
API_assignment_1 / mlops / ml_flow.py
Top
Code
Blame
97 lines (75 loc) · 2.95 KB
Code 55% faster with GitHub Copilot
Raw
def split_data(data):
logger.info(f"Data split completed: {X_train.shape[0]} train samples, {X_test.shape[0]} test samples")
return X_train, X_test, y_train, y_test
Task 3: Train Models
@task
def train_models(X_train, y_train, models):
trained_models = {}
for model_name, model in models.items():
model.fit(X_train, y_train)
trained_models[model_name] = model
return trained_models
Task 4: Evaluate Models
@task
def evaluate_models(X_test, y_test, trained_models):
logger = get_run_logger()
for model_name, model in trained_models.items():
# Predictions
y_pred = model.predict(X_test)

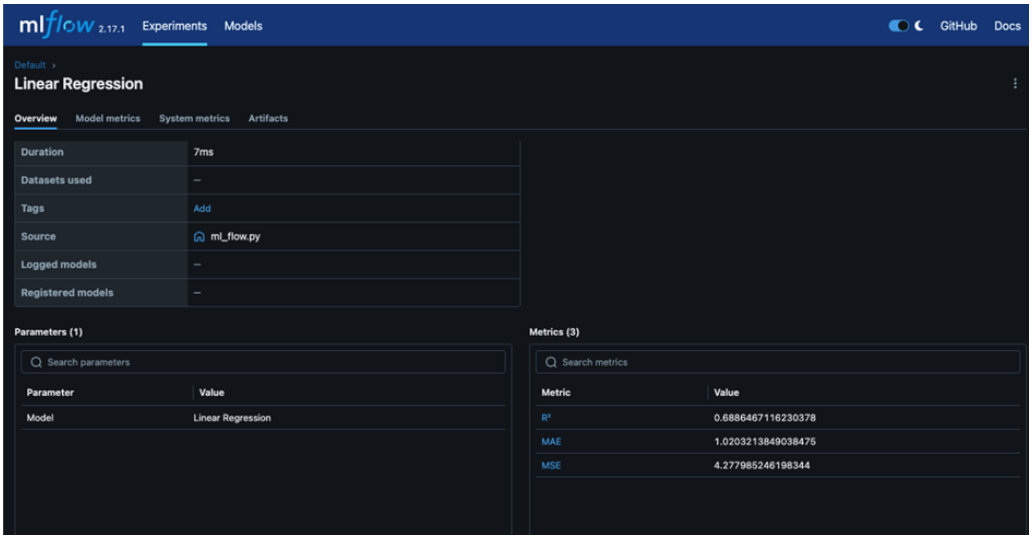

Files
main
Go to file
Screenshots
api
data
document
flows
mlops
mlartifacts/0
mlruns
.DS_Store
ml_flow.py
output
sageMaker
tasks
.DS_Store
API_assignment_1 / mlops / ml_flow.py
Top
Code
Blame
97 lines (75 loc) · 2.95 KB
Code 55% faster with GitHub Copilot
Raw
def evaluate_models(X_test, y_test, trained_models):
y_pred = model.predict(X_test)
# Metrics
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
logger.info(f"{model_name} - MSE: {mse}, MAE: {mae}, R²: {r2}")
# Log metrics using MLFlow
with mlflow.start_run(run_name=model_name):
mlflow.log_param("Model", model_name)
mlflow.log_metric("MSE", mse)
mlflow.log_metric("MAE", mae)
mlflow.log_metric("R²", r2)
logger.info("Model evaluation completed")
# Main Flow
@flow(name="student_performance_factors_ml_pipeline")
def student_performance_factors(file_path):
# Load data (assumes data preprocessing has been done)
data = pd.read_csv(file_path)

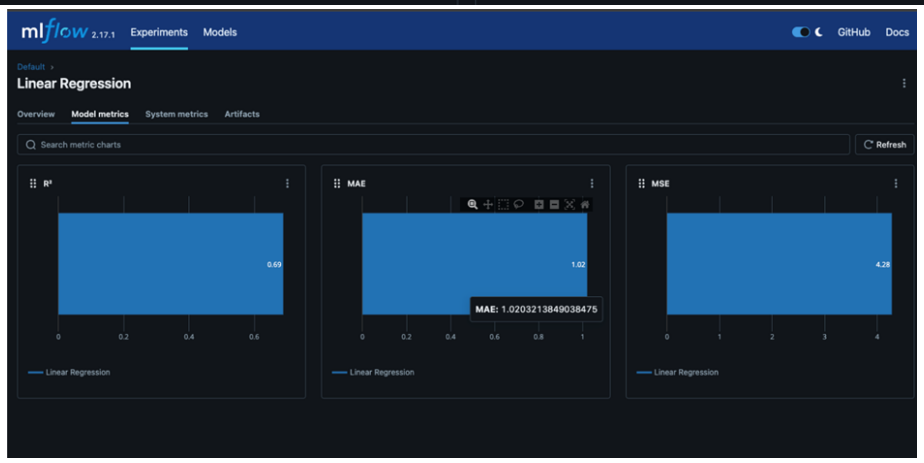

Files
main
Go to file
Screenshots
api
data
document
flows
mlops
mlartifacts/0
mlruns
.DS_Store
ml_flow.py
output
sageMaker
tasks
.DS_Store
API_assignment_1 / mlops / ml_flow.py
Top
Code
Blame
97 lines (75 loc) · 2.95 KB
Code 55% faster with GitHub Copilot
Raw
Main Flow
@flow(name="student_performance_factors_ml_pipeline")
def student_performance_factors(file_path):
# Load data (assumes data preprocessing has been done)
data = pd.read_csv(file_path)
# Model Preparation
models = model_preparation()
# Split Data
X_train, X_test, y_train, y_test = split_data(data)
# Train Models
trained_models = train_models(X_train, y_train, models)
# Evaluate Models
evaluate_models(X_test, y_test, trained_models)
# Run the pipeline
if __name__ == "__main__":
file_path = "/Users/mkmac/Documents/API_assignment_1/data/ImputedStudentPerformanceFactors.csv" # Update the path to yo
student_performance_factors(file_path)

Linear Regression: -

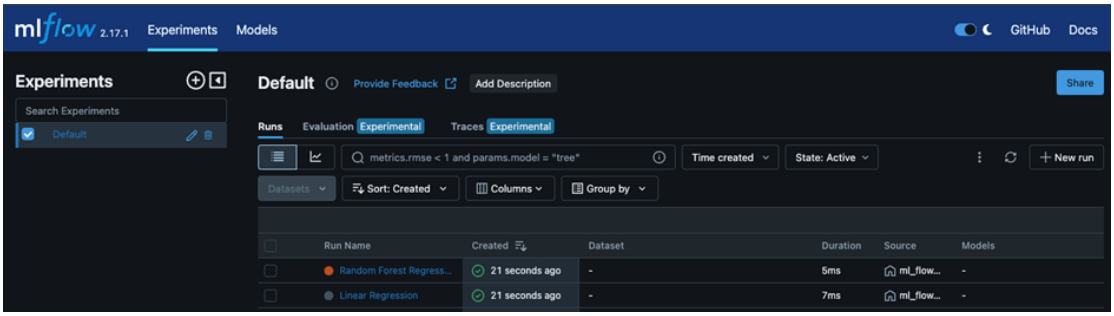


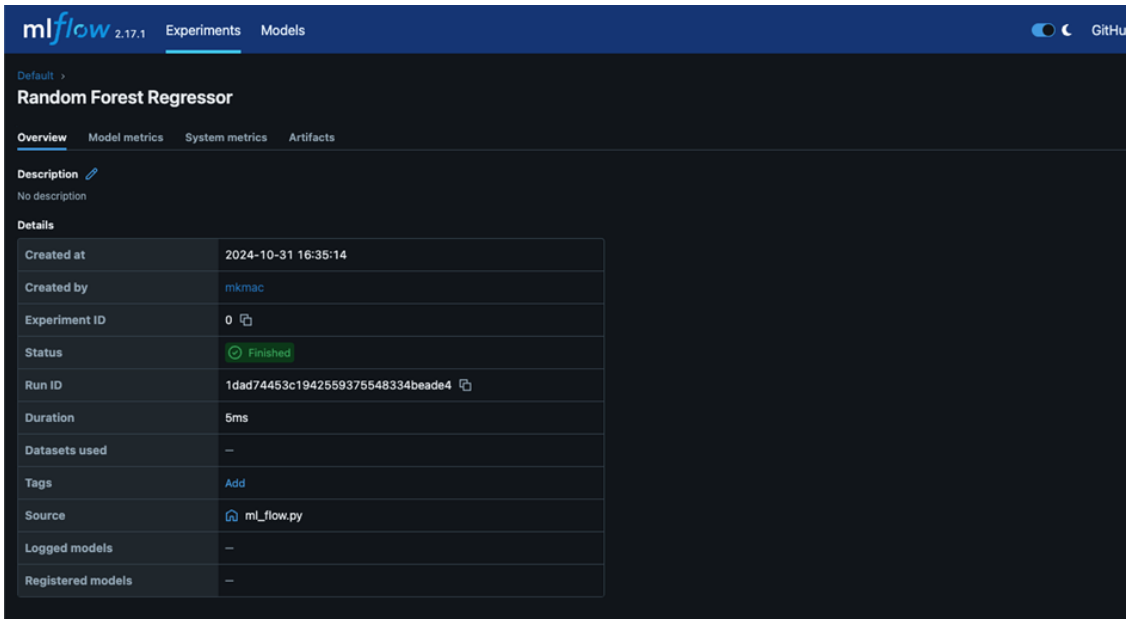
Metrics captured below;



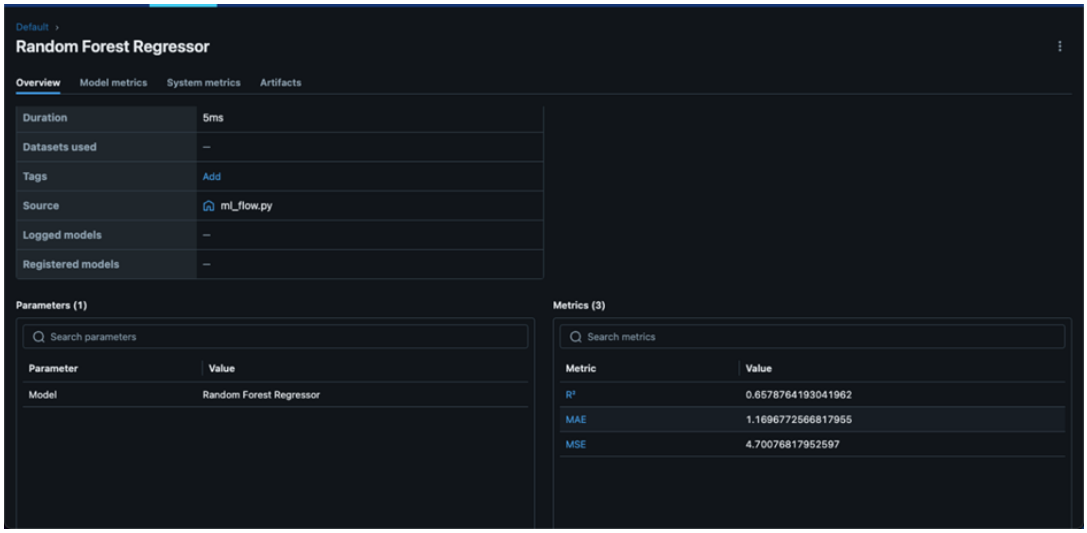


Random Forest Regression: -



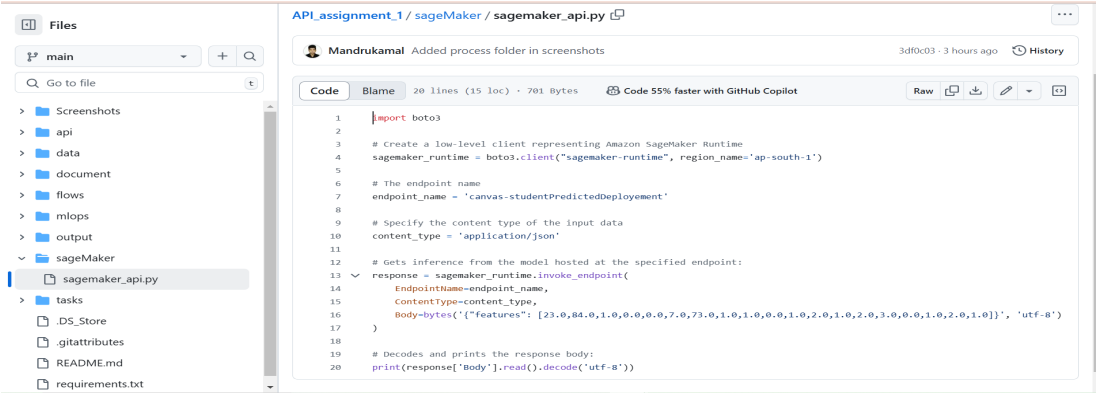


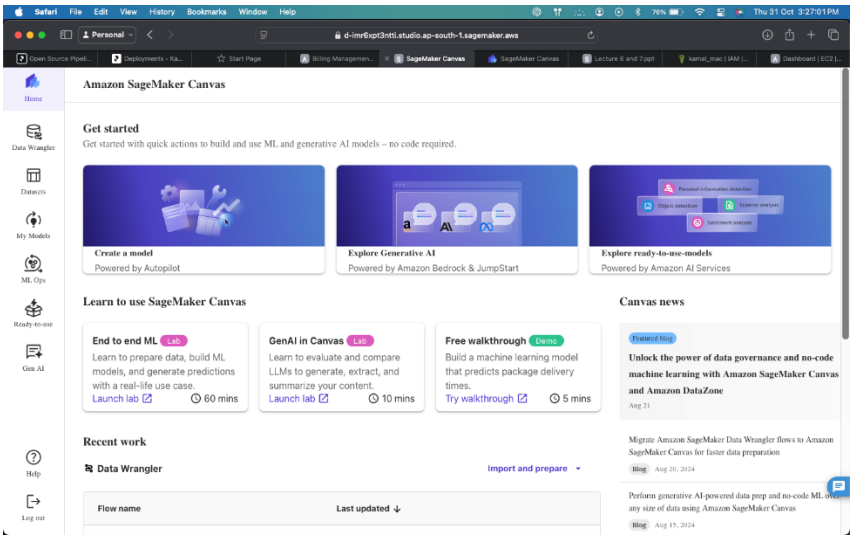
Metrics captured below:

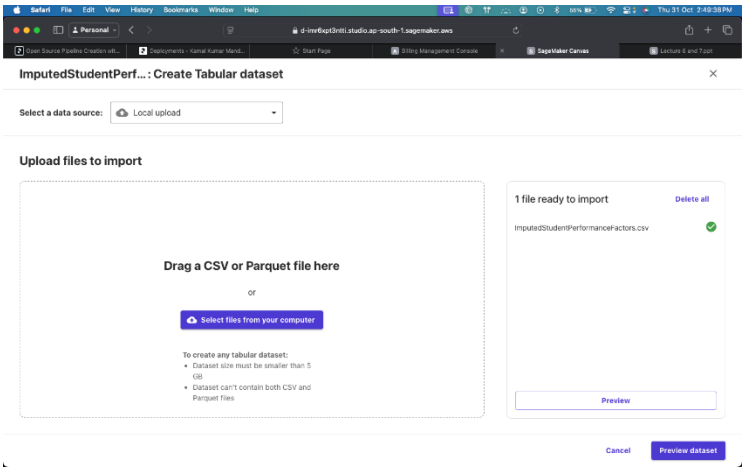


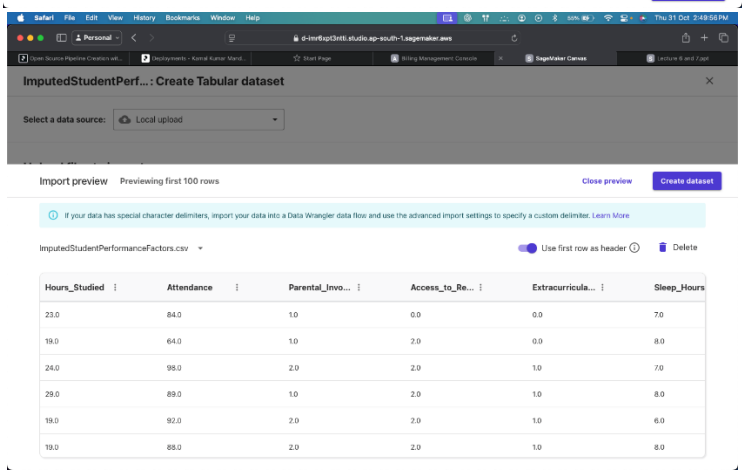
MLOps: Monitoring the model and logging relevant metrics

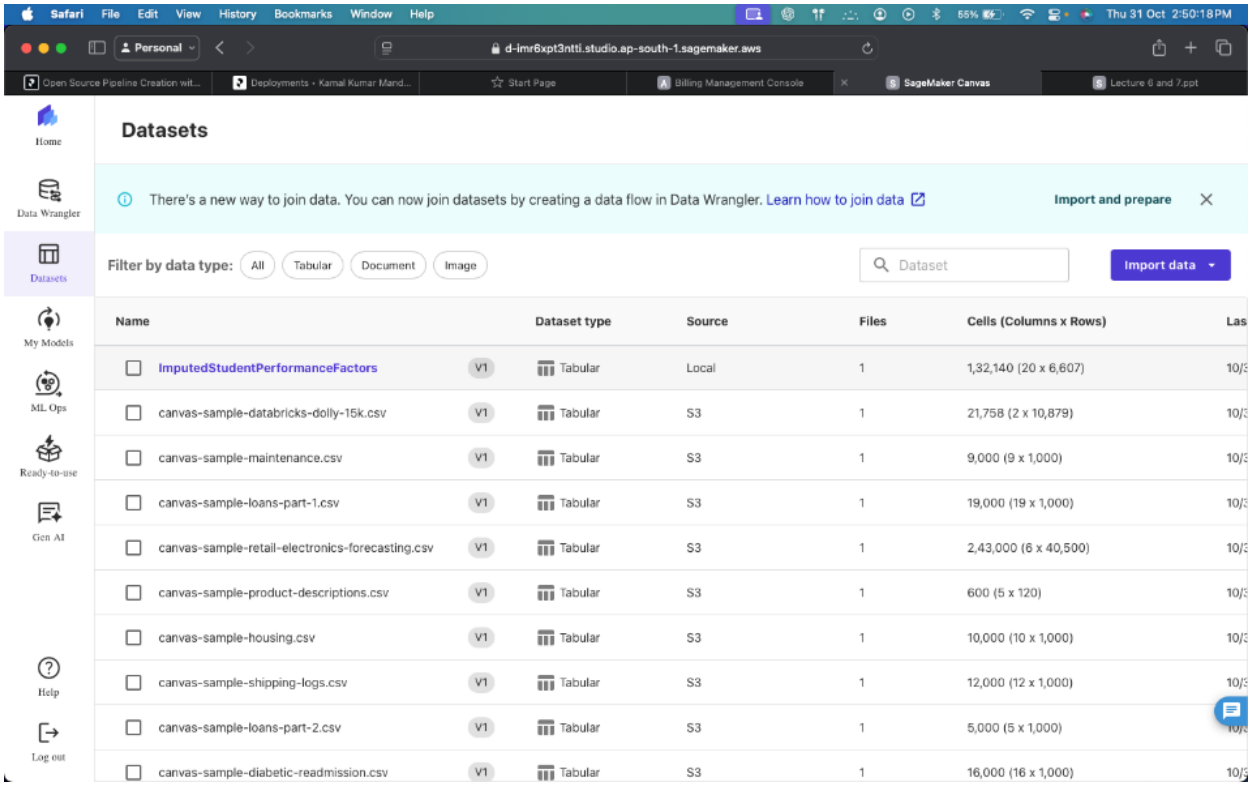
ML Models deployment in SageMaker: -

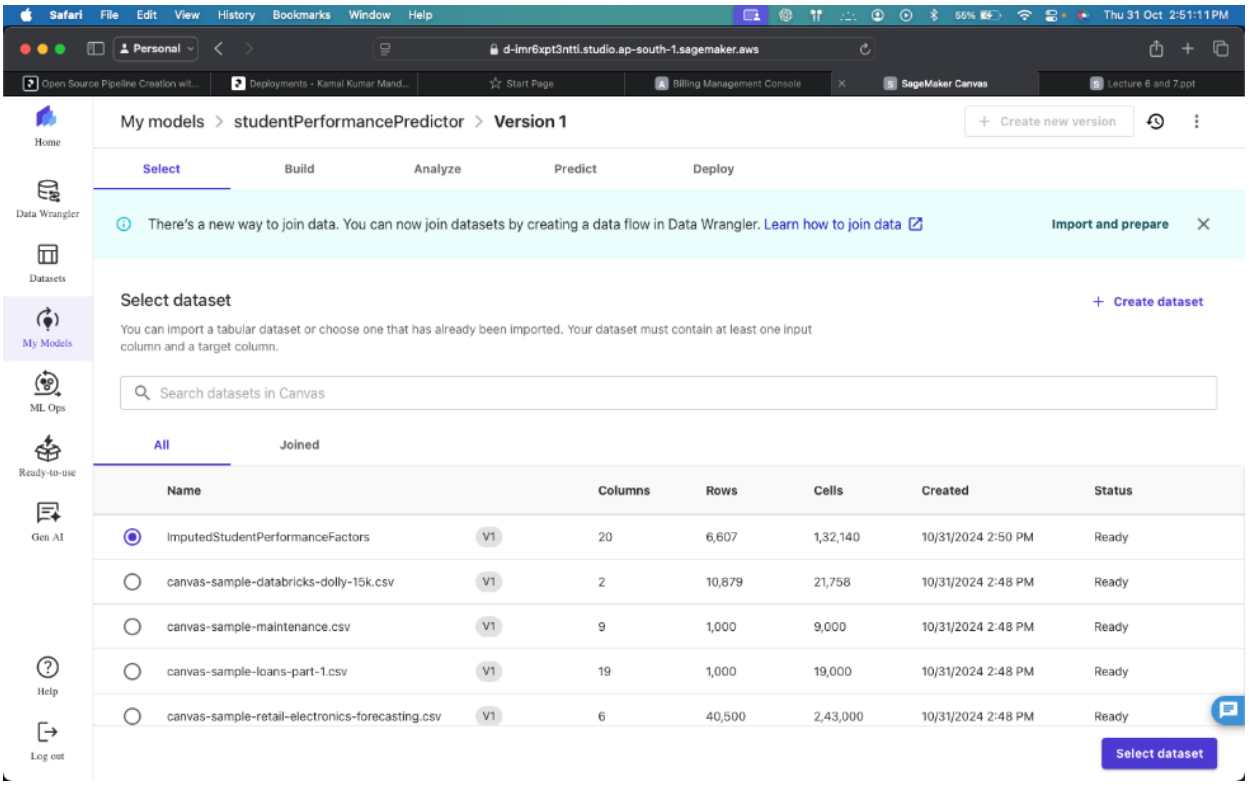


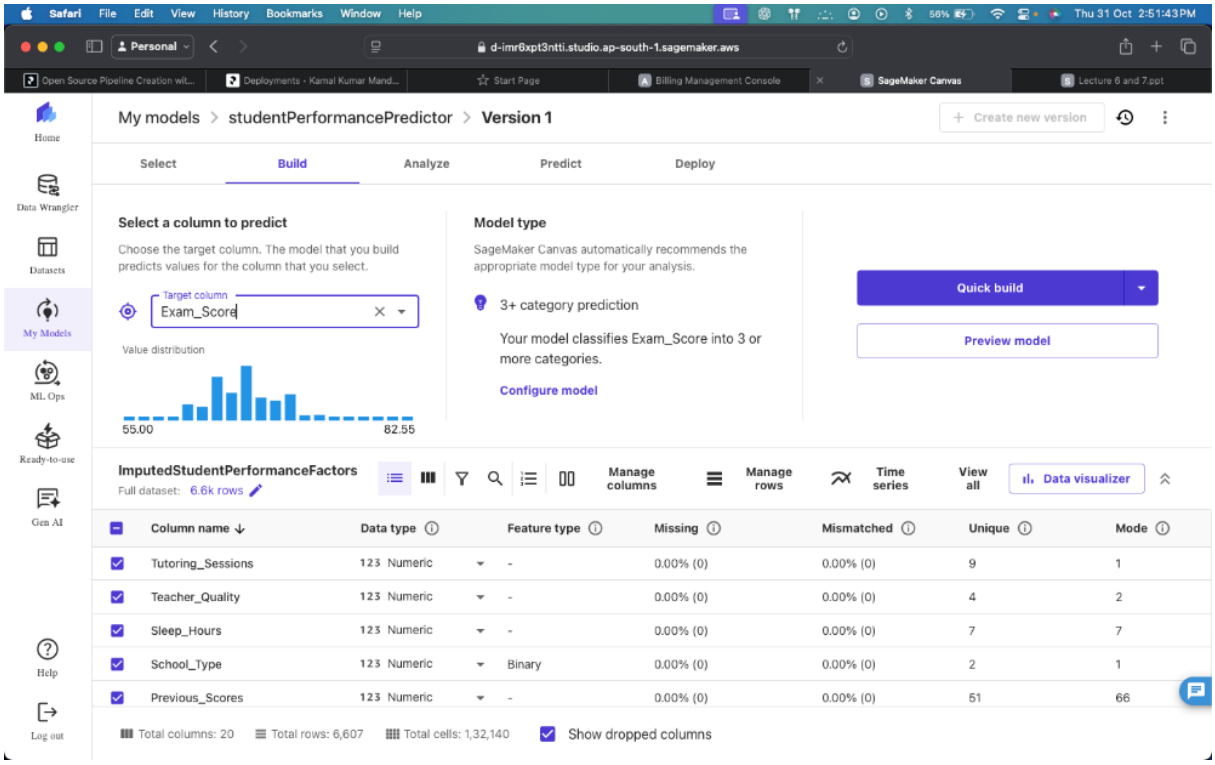


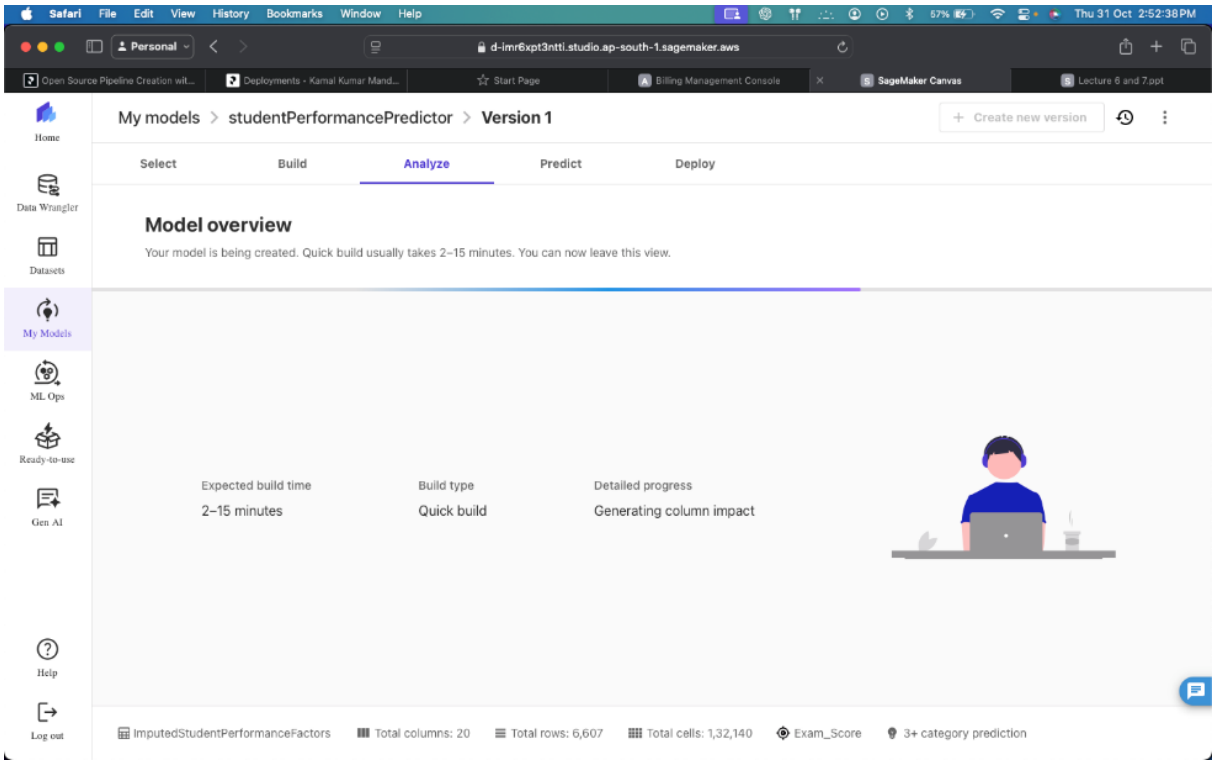

ImputedStudentPerf... : Create Tabular dataset
Select a data source:
Local upload
Upload files to import
Drag a CSV or Parquet file here
or
Select files from your computer
To create any tabular dataset:
Dataset size must be smaller than 5 GB
Dataset can't contain both CSV and Parquet files
1 file ready to import
Delete all
ImputedStudentPerformanceFactors.csv
Preview
Cancel
Preview dataset


ImputedStudentPerf... : Create Tabular dataset
Select a data source:
Local upload
Import preview
Previewing first 100 rows
Close preview
Create dataset
If your data has special character delimiters, import your data into a Data Wrangler data flow and use the advanced import settings to specify a custom delimiter. Learn More
ImputedStudentPerformanceFactors.csv
Use first row as header
Delete
Hours_Studied
Attendance
Parental_Involvo...
Access_to_Re...
Extracurricula...
Sleep_Hours
23.0 84.0 1.0 0.0 0.0 7.0
19.0 64.0 1.0 2.0 0.0 8.0
24.0 98.0 2.0 2.0 1.0 7.0
29.0 89.0 1.0 2.0 1.0 8.0
19.0 92.0 2.0 2.0 1.0 6.0
19.0 88.0 2.0 2.0 1.0 8.0

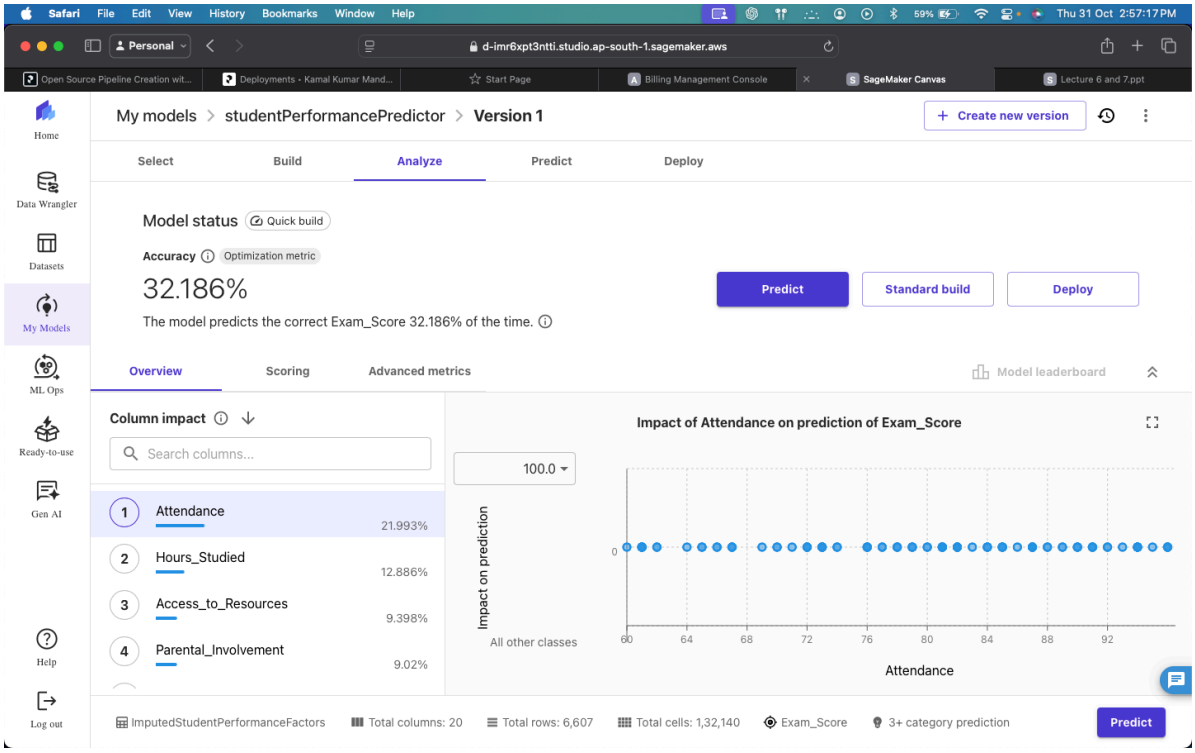

Safari File Edit View History Bookmarks Window Help
Thu 31 Oct 2:50:18 PM
Personal
d-imr6xpt3ntti.studio.ap-south-1.sagemaker.aws
Open Source Pipeline Creation wit...
Deployments - Kamal Kumar Mand...
Start Page
Billing Management Console
SageMaker Canvas
Lecture 6 and 7.ppt
Home
Data Wrangler
Datasets
My Models
ML Ops
Ready-to-use
Gen AI
Help
Log out
Datasets
There's a new way to join data. You can now join datasets by creating a data flow in Data Wrangler. Learn how to join data
Import and prepare
Filter by data type:
All
Tabular
Document
Image
Dataset
Import data
Name
Dataset type
Source
Files
Cells (Columns x Rows)
ImputedStudentPerformanceFactors V1 Tabular Local 1 1,32,140 (20 x 6,607)
canvas-sample-databricks-dolly-15k.csv V1 Tabular S3 1 21,758 (2 x 10,879)
canvas-sample-maintenance.csv V1 Tabular S3 1 9,000 (9 x 1,000)
canvas-sample-loans-part-1.csv V1 Tabular S3 1 19,000 (19 x 1,000)
canvas-sample-retail-electronics-forecasting.csv V1 Tabular S3 1 2,43,000 (6 x 40,500)
canvas-sample-product-descriptions.csv V1 Tabular S3 1 600 (5 x 120)
canvas-sample-housing.csv V1 Tabular S3 1 10,000 (10 x 1,000)
canvas-sample-shipping-logs.csv V1 Tabular S3 1 12,000 (12 x 1,000)
canvas-sample-loans-part-2.csv V1 Tabular S3 1 5,000 (5 x 1,000)
canvas-sample-diabetic-readmission.csv V1 Tabular S3 1 16,000 (16 x 1,000)

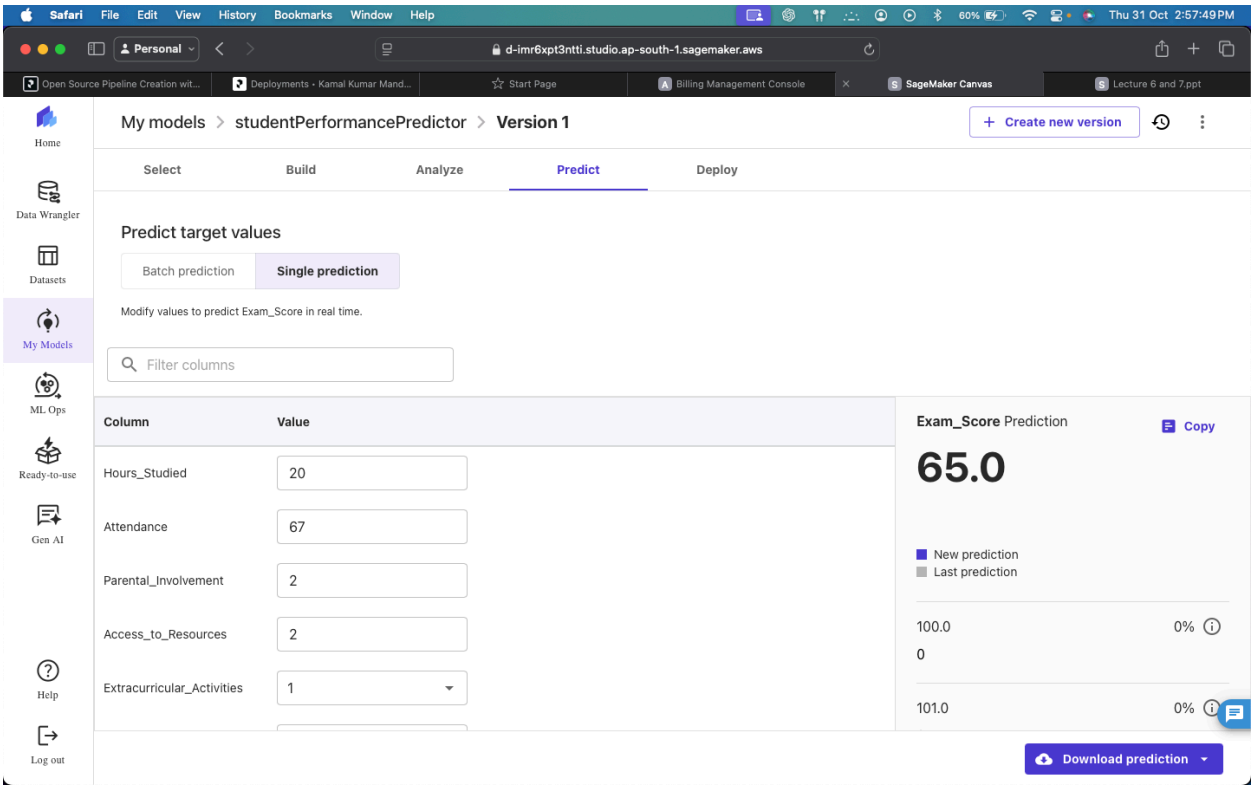

My models > studentPerformancePredictor > Version 1
Create new version
Select
Build
Analyze
Predict
Deploy
There's a new way to join data. You can now join datasets by creating a data flow in Data Wrangler. Learn how to join data
Import and prepare
Select dataset
Create dataset
You can import a tabular dataset or choose one that has already been imported. Your dataset must contain at least one input column and a target column.
Search datasets in Canvas
All
Joined
Name
Columns
Rows
Cells
Created
Status
ImputedStudentPerformanceFactors V1 20 6,607 1,32,140 10/31/2024 2:50 PM Ready
canvas-sample-databricks-dolly-15k.csv V1 2 10,879 21,758 10/31/2024 2:48 PM Ready
canvas-sample-maintenance.csv V1 9 1,000 9,000 10/31/2024 2:48 PM Ready
canvas-sample-loans-part-1.csv V1 19 1,000 19,000 10/31/2024 2:48 PM Ready
canvas-sample-retail-electronics-forecasting.csv V1 6 40,500 2,43,000 10/31/2024 2:48 PM Ready
Select dataset

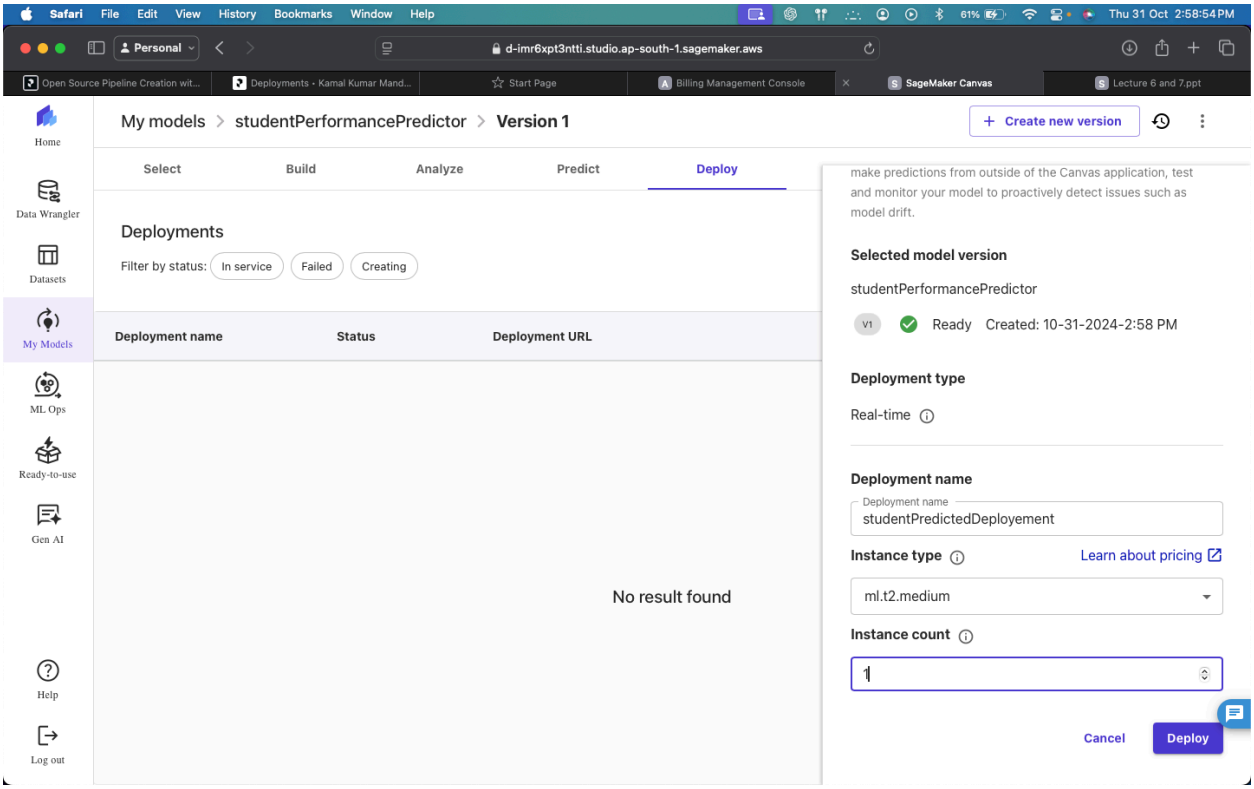

My models > studentPerformancePredictor > Version 1
Create new version
Select
Build
Analyze
Predict
Deploy
Select a column to predict
Choose the target column. The model that you build predicts values for the column that you select.
Target column
Exam_Score
Value distribution
55.00
82.55
Model type
SageMaker Canvas automatically recommends the appropriate model type for your analysis.
3+ category prediction
Your model classifies Exam_Score into 3 or more categories.
Configure model
Quick build
Preview model
ImputedStudentPerformanceFactors
Full dataset: 6.6k rows
Manage columns
Manage rows
Time series
View all
Data visualizer
Column name
Data type
Feature type
Missing
Mismatched
Unique
Mode
Tutoring_Sessions 123 Numeric - 0.00% (0) 0.00% (0) 9 1
Teacher_Quality 123 Numeric - 0.00% (0) 0.00% (0) 4 2
Sleep_Hours 123 Numeric - 0.00% (0) 0.00% (0) 7 7
School_Type 123 Numeric Binary 0.00% (0) 0.00% (0) 2 1
Previous_Scores 123 Numeric - 0.00% (0) 0.00% (0) 51 66
Total columns: 20
Total rows: 6,607
Total cells: 1,32,140
Show dropped columns

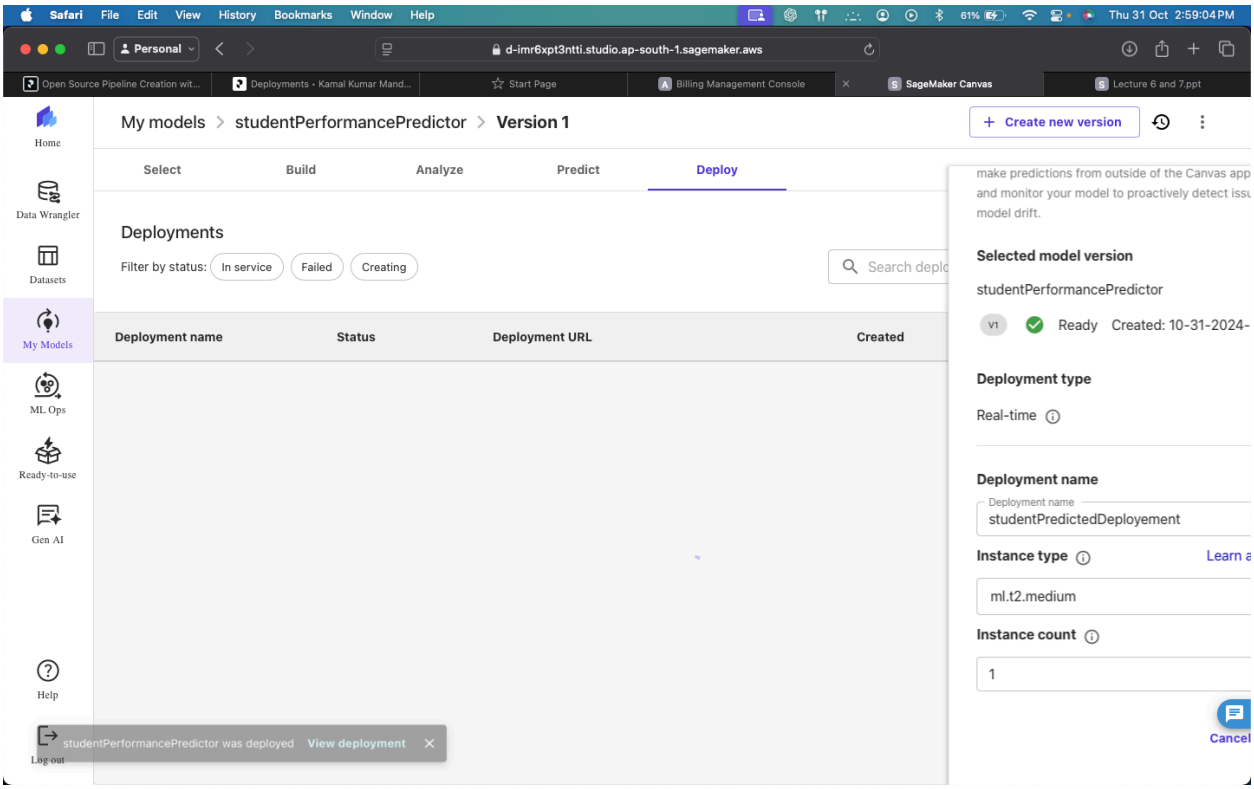


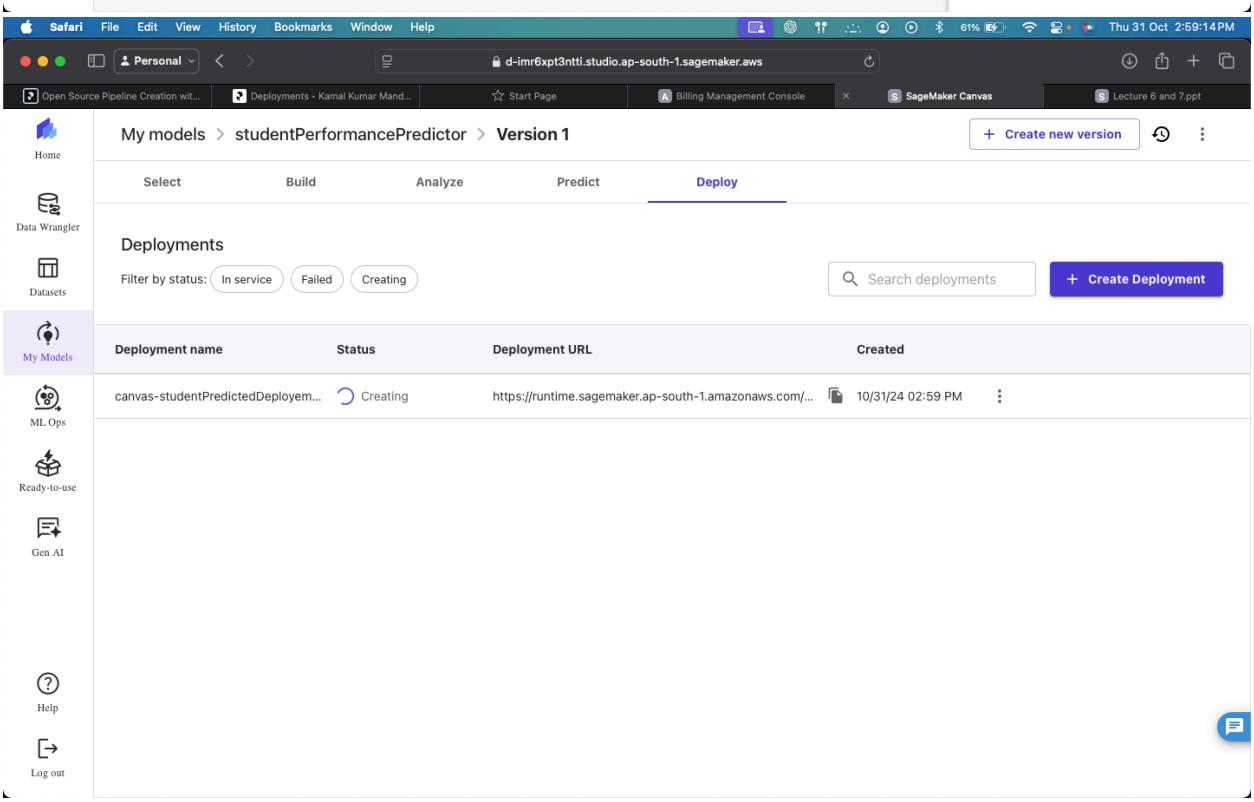
Metrics captured below

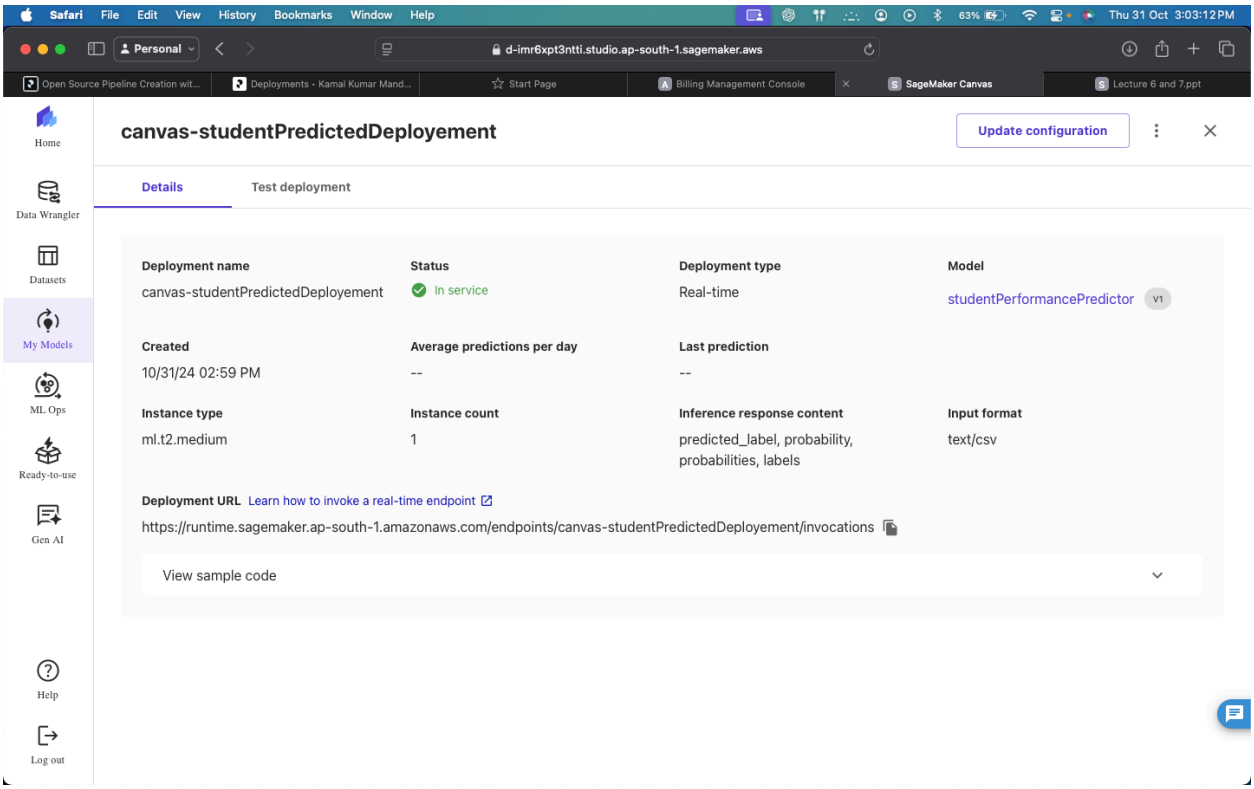


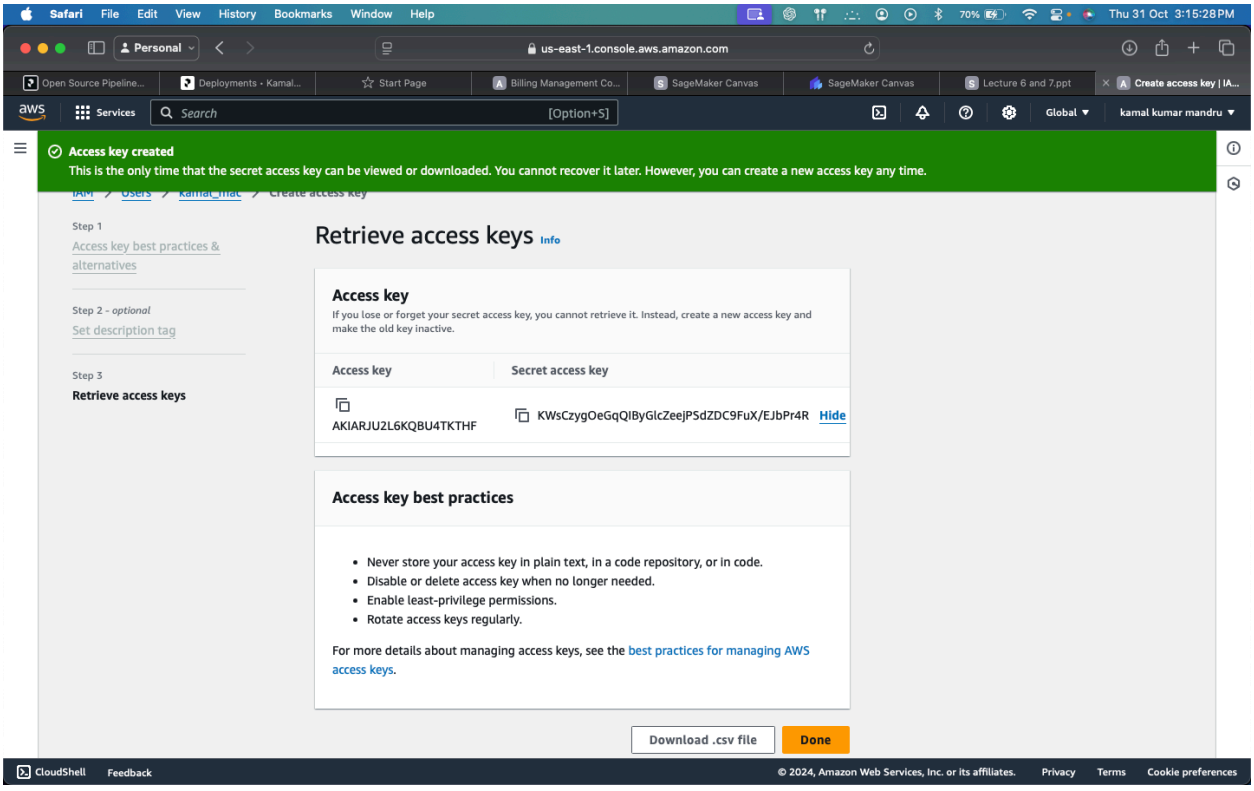

My models
studentPerformancePredictor
Version 1
Create new version
Select
Build
Analyze
Predict
Deploy
Predict target values
Batch prediction
Single prediction
Modify values to predict Exam_Score in real time.
Filter columns
Column
Value
Hours_Studied
20
Attendance
67
Parental_Involvement
2
Access_to_Resources
2
Extracurricular_Activities
1
Exam_Score Prediction
Copy
65.0
New prediction
Last prediction
100.0
0%
0
101.0
0%
Download prediction

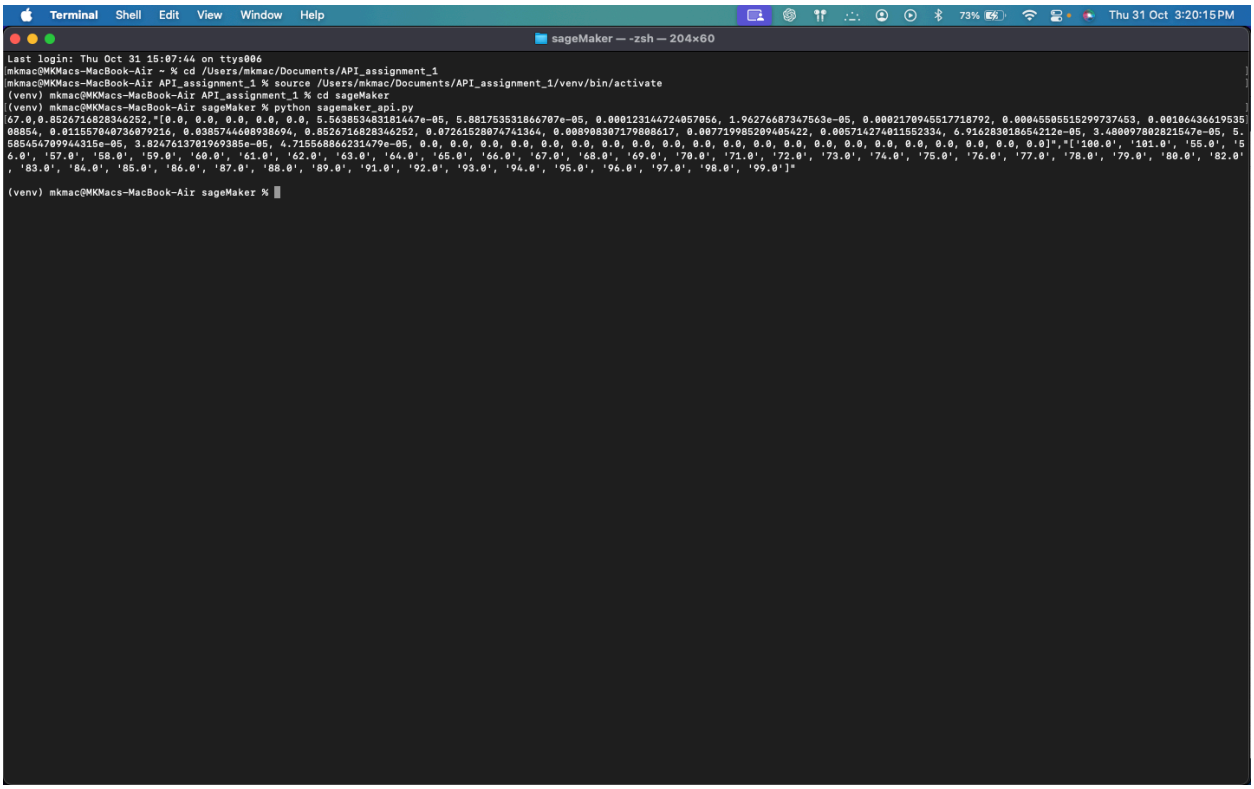

My models
studentPerformancePredictor
Version 1
Create new version
Select
Build
Analyze
Predict
Deploy
Deployments
Filter by status:
In service
Failed
Creating
Deployment name
Status
Deployment URL
No result found
make predictions from outside of the Canvas application, test and monitor your model to proactively detect issues such as model drift.
Selected model version
studentPerformancePredictor
V1
Ready
Created: 10-31-2024-2:58 PM
Deployment type
Real-time
Deployment name
Deployment name
studentPredictedDeployement
Instance type
Learn about pricing
ml.t2.medium
Instance count
1
Cancel
Deploy


My models > studentPerformancePredictor > Version 1
+ Create new version
Select
Build
Analyze
Predict
Deploy
Deployments
Filter by status: In service Failed Creating
Search deplo
Deployment name
Status
Deployment URL
Created
make predictions from outside of the Canvas app
and monitor your model to proactively detect issu
model drift.
Selected model version
studentPerformancePredictor
V1
Ready
Created: 10-31-2024-
Deployment type
Real-time
Deployment name
Deployment name
studentPredictedDeployement
Instance type
Learn a
ml.t2.medium
Instance count
1
Cancel
studentPerformancePredictor was deployed
View deployment
Home
Data Wrangler
Datasets
My Models
ML Ops
Ready-to-use
Gen AI
Help
Log out


My models > studentPerformancePredictor > Version 1
+ Create new version
Select
Build
Analyze
Predict
Deploy
Deployments
Filter by status: In service Failed Creating
Search deployments
+ Create Deployment
Deployment name
Status
Deployment URL
Created
canvas-studentPredictedDeploym...
Creating
https://runtime.sagemaker.ap-south-1.amazonaws.com/...
10/31/24 02:59 PM
Home
Data Wrangler
Datasets
My Models
ML Ops
Ready-to-use
Gen AI
Help
Log out

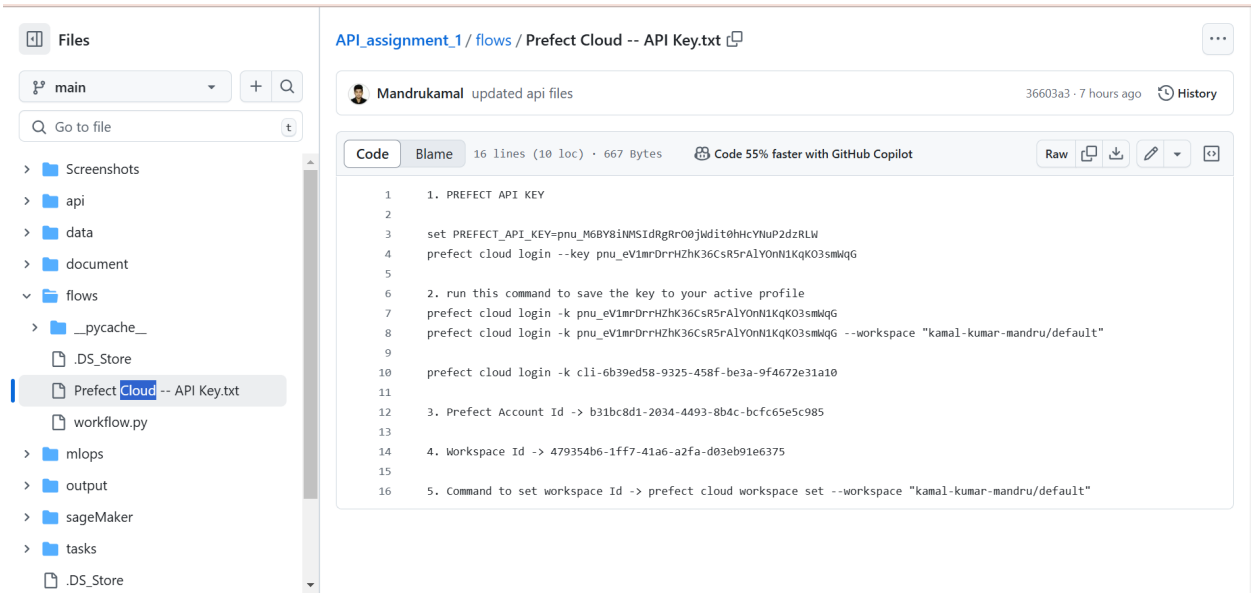

canvas-studentPredictedDeployement
Update configuration
Details
Test deployment
Deployment name
canvas-studentPredictedDeployement
Status
In service
Deployment type
Real-time
Model
studentPerformancePredictor
V1
Created
10/31/24 02:59 PM
Average predictions per day
--
Last prediction
--
Instance type
ml.t2.medium
Instance count
1
Inference response content
predicted_label, probability, probabilities, labels
Input format
text/csv
Deployment URL
Learn how to invoke a real-time endpoint
https://runtime.sagemaker.ap-south-1.amazonaws.com/endpoints/canvas-studentPredictedDeployement/invocations
View sample code
Home
Data Wrangler
Datasets
My Models
ML Ops
Ready-to-use
Gen AI
Help
Log out


Access key created
This is the only time that the secret access key can be viewed or downloaded. You cannot recover it later. However, you can create a new access key any time.
Retrieve access keys
Info
Access key
If you lose or forget your secret access key, you cannot retrieve it. Instead, create a new access key and make the old key inactive.
Access key
Secret access key
AKIARJU2L6KQBU4TKTHF
KWsCzygOeGqQIByGlcZeejPSdZDC9FuX/EJbPr4R
Hide
Access key best practices
Never store your access key in plain text, in a code repository, or in code.
Disable or delete access key when no longer needed.
Enable least-privilege permissions.
Rotate access keys regularly.
For more details about managing access keys, see the best practices for managing AWS access keys.
Download .csv file
Done
Step 1
Access key best practices & alternatives
Step 2 - optional
Set description tag
Step 3
Retrieve access keys
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

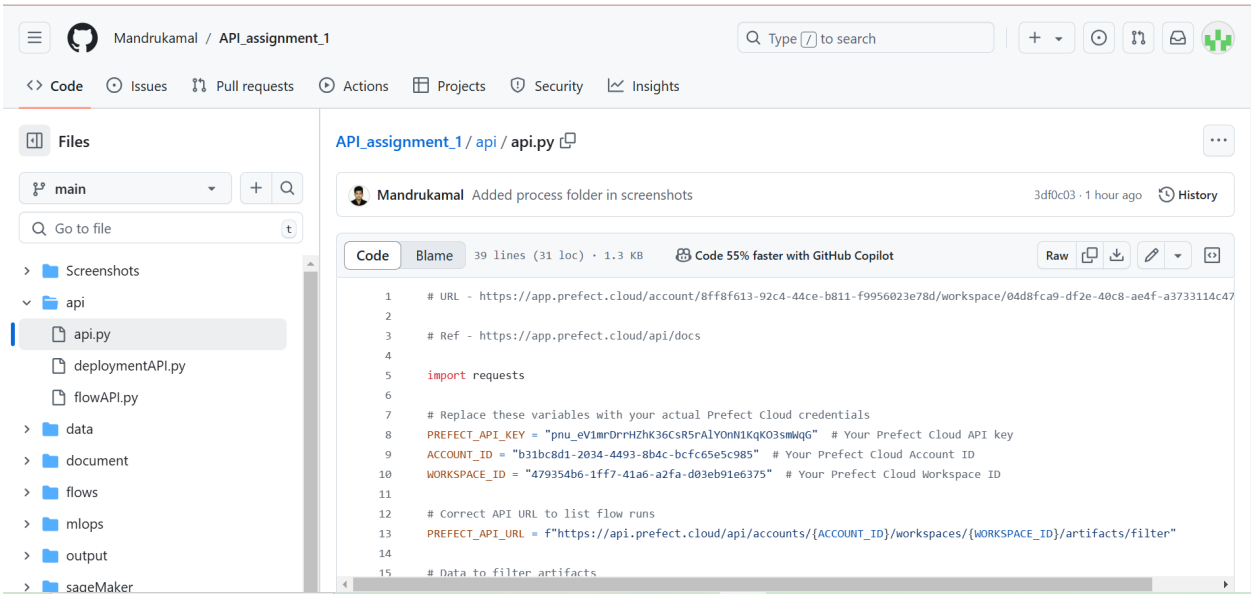


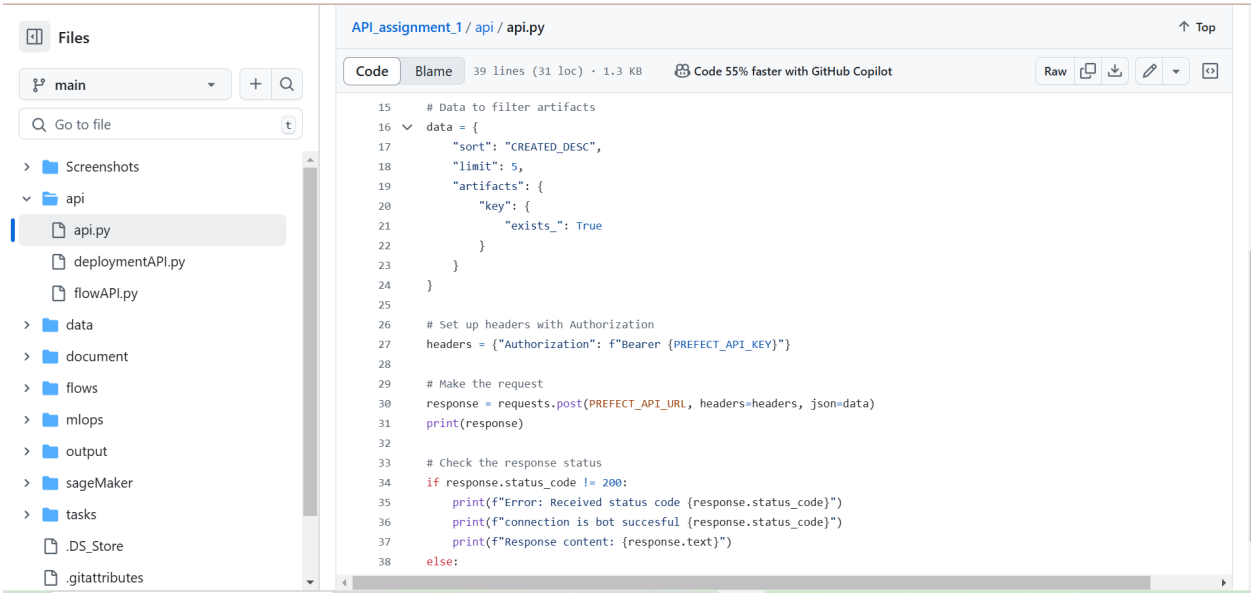
Sub-Objective 3: API Access

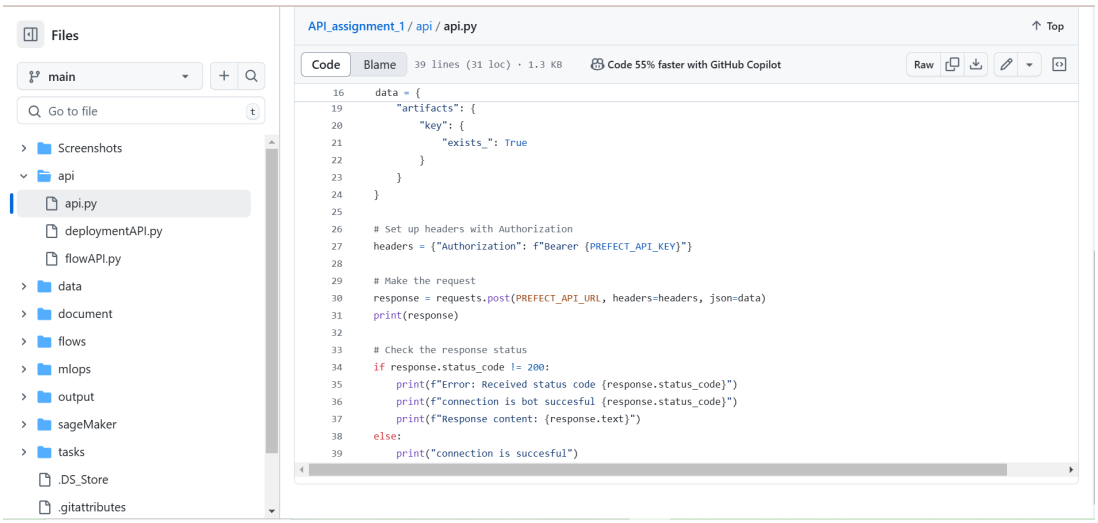
3.1 Retrieve Key Application Details: Use Built-in APIs to access important application information (e.g., flow, deployment etc.)



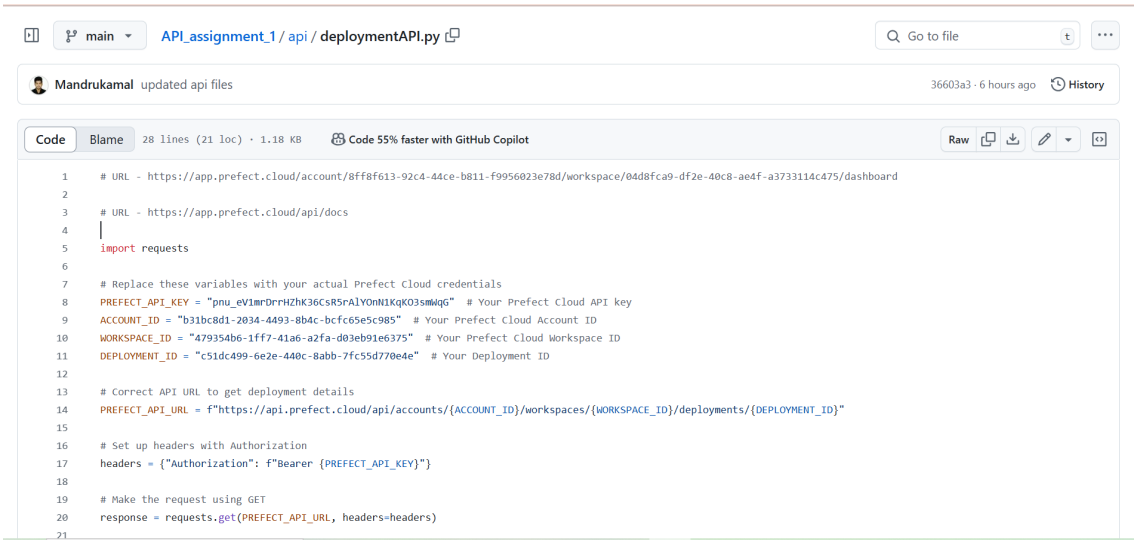
API.py: -

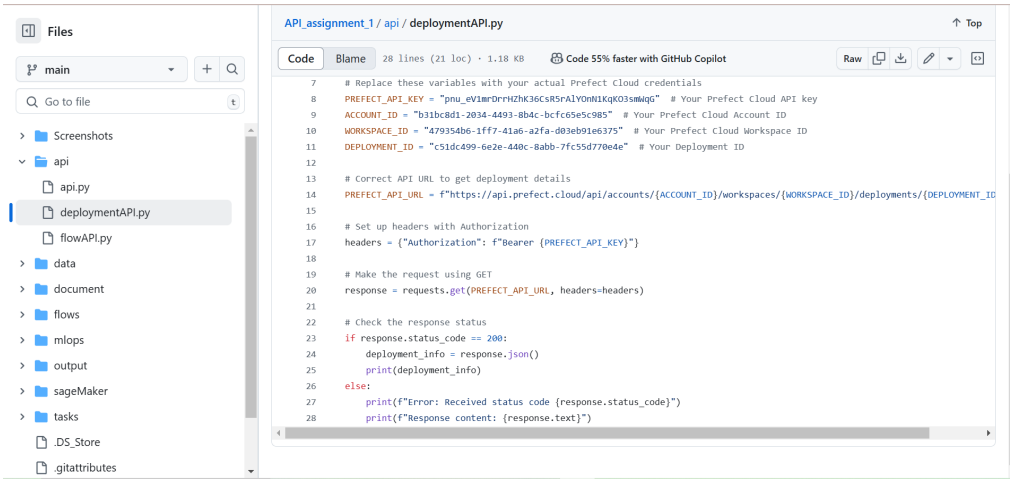

Mandrukamal / API_assignment_1
Type / to search
Code
Issues
Pull requests
Actions
Projects
Security
Insights
Files
main
Go to file
Screenshots
api
api.py
deploymentAPI.py
flowAPI.py
data
document
flows
mlops
output
sageMaker
API_assignment_1 / api / api.py
Mandrukamal Added process folder in screenshots
3df0c03 · 1 hour ago
History
Code
Blame
39 lines (31 loc) · 1.3 KB
Code 55% faster with GitHub Copilot
Raw
# URL - https://app.prefect.cloud/account/8ff8f613-92c4-44ce-b811-f9956023e78d/workspace/04d8fca9-df2e-40c8-ae4f-a3733114c47
# Ref - https://app.prefect.cloud/api/docs
import requests
# Replace these variables with your actual Prefect Cloud credentials
PREFECT_API_KEY = "pnu_eV1mrDrrHZhK36CsR5rAlYOnN1KqKO3smWqG" # Your Prefect Cloud API key
ACCOUNT_ID = "b31bc8d1-2034-4493-8b4c-bcfc65e5c985" # Your Prefect Cloud Account ID
WORKSPACE_ID = "479354b6-1ff7-41a6-a2fa-d03eb91e6375" # Your Prefect Cloud Workspace ID
# Correct API URL to list flow runs
PREFECT_API_URL = f"https://api.prefect.cloud/api/accounts/{ACCOUNT_ID}/workspaces/{WORKSPACE_ID}/artifacts/filter"
# Data to filter artifacts


Files
main
Go to file
Screenshots
api
api.py
deploymentAPI.py
flowAPI.py
data
document
flows
mlops
output
sageMaker
tasks
.DS_Store
.gitattributes
API_assignment_1 / api / api.py
Top
Code
Blame
39 lines (31 loc) · 1.3 KB
Code 55% faster with GitHub Copilot
Raw
# Data to filter artifacts
data = {
"sort": "CREATED_DESC",
"limit": 5,
"artifacts": {
"key": {
"exists_": True
}
}
}
# Set up headers with Authorization
headers = {"Authorization": f"Bearer {PREFECT_API_KEY}"}
# Make the request
response = requests.post(PREFECT_API_URL, headers=headers, json=data)
print(response)
# Check the response status
if response.status_code != 200:
print(f"Error: Received status code {response.status_code}")
print(f"connection is bot succesful {response.status_code}")
print(f"Response content: {response.text}")
else:



DeploymentAPI.py: -





FlowAPI.py: -

main API_assignment_1 / api / flowAPI.py

Go to file

Mandrukamal updated api files 36603a3 · 6 hours ago History

Code Blame 28 lines (21 loc) · 1.14 KB Code 55% faster with GitHub Copilot Raw

```python
# URL - https://app.prefect.cloud/account/8ff8f613-92c4-44ce-b811-f9956023e78d/workspace/04d8fca9-df2e-40c8-ae4f-a3733114c475/dashboard

# URL - https://app.prefect.cloud/api/docs

import requests

# Replace these variables with your actual Prefect Cloud credentials
PREFECT_API_KEY = "pnu_eV1mrDrrHZhK36CsR5rAlYOnN1KqKO3smWqG"  # Your Prefect Cloud API key
ACCOUNT_ID = "b31bc8d1-2034-4493-8b4c-bcfc65e5c985"  # Your Prefect Cloud Account ID
WORKSPACE_ID = "479354b6-1ff7-41a6-a2fa-d03eb91e6375"  # Your Prefect Cloud Workspace ID
FLOW_ID = "fee0b8ea-4107-411b-93d7-d024be07c700"  # Your Flow ID

# Correct API URL to get flow details
PREFECT_API_URL = f"https://api.prefect.cloud/api/accounts/{ACCOUNT_ID}/workspaces/{WORKSPACE_ID}/flows/{FLOW_ID}"

# Set up headers with Authorization
headers = {"Authorization": f"Bearer {PREFECT_API_KEY}"}

# Make the request using GET
response = requests.get(PREFECT_API_URL, headers=headers)
```

Files

main

Go to file

- Screenshots
- api
  - api.py
  - deploymentAPI.py
  - flowAPI.py
- data
- document
- flows
- mlops
- output
- sageMaker
- tasks
- .DS_Store
- .gitattributes

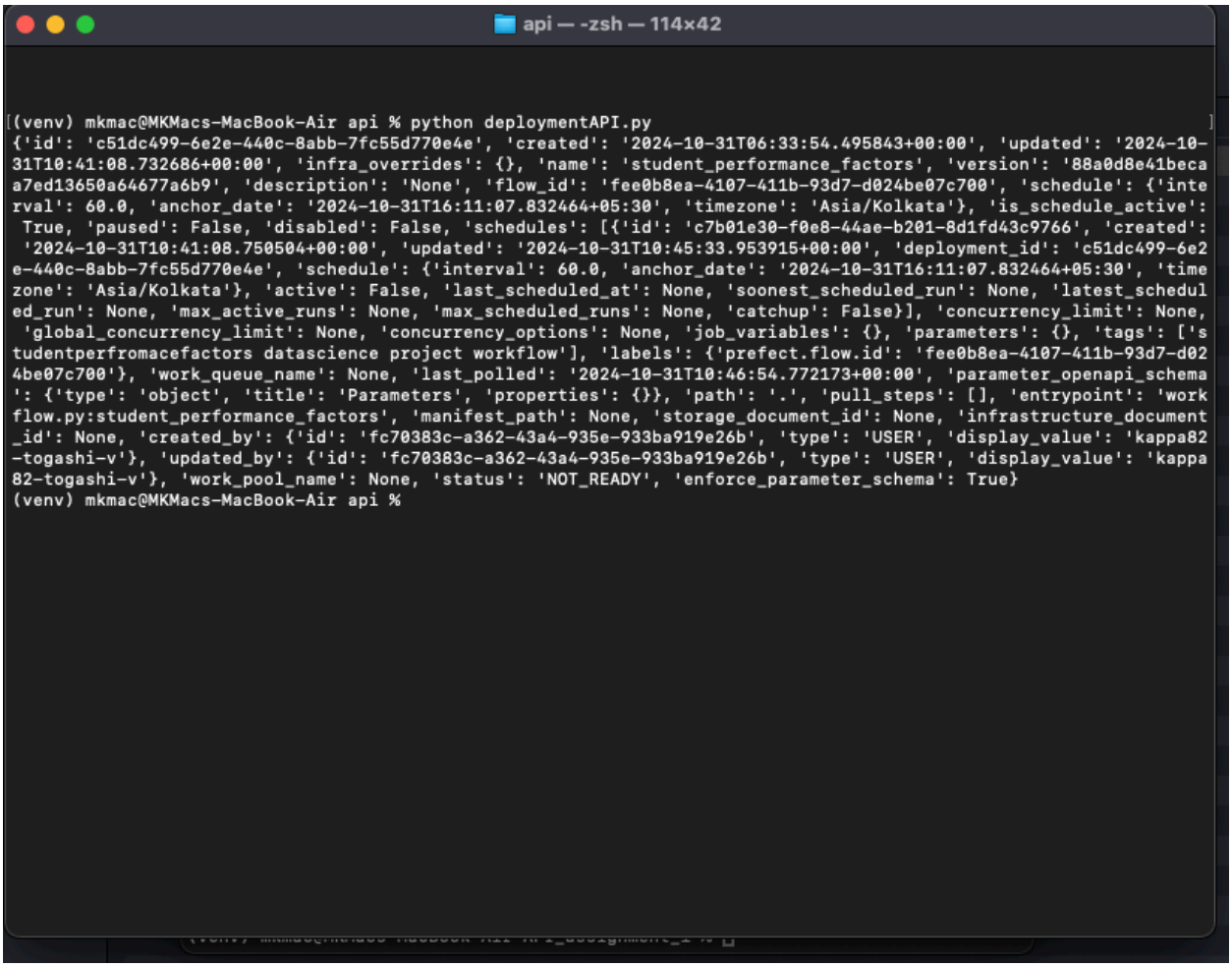
API_assignment_1 / api / flowAPI.py ↑ Top

Code Blame 28 lines (21 loc) · 1.14 KB Code 55% faster with GitHub Copilot Raw

```python
# Replace these variables with your actual Prefect Cloud credentials
PREFECT_API_KEY = "pnu_eV1mrDrrHZhK36CsR5rAlYOnN1KqKO3smWqG"  # Your Prefect Cloud API key
ACCOUNT_ID = "b31bc8d1-2034-4493-8b4c-bcfc65e5c985"  # Your Prefect Cloud Account ID
WORKSPACE_ID = "479354b6-1ff7-41a6-a2fa-d03eb91e6375"  # Your Prefect Cloud Workspace ID
FLOW_ID = "fee0b8ea-4107-411b-93d7-d024be07c700"  # Your Flow ID

# Correct API URL to get flow details
PREFECT_API_URL = f"https://api.prefect.cloud/api/accounts/{ACCOUNT_ID}/workspaces/{WORKSPACE_ID}/flows/{FLOW_ID}"

# Set up headers with Authorization
headers = {"Authorization": f"Bearer {PREFECT_API_KEY}"}

# Make the request using GET
response = requests.get(PREFECT_API_URL, headers=headers)

# Check the response status
if response.status_code == 200:
    flow_info = response.json()
    print(flow_info)
else:
    print(f"Error: Received status code {response.status_code}")
    print(f"Response content: {response.text}")
```

3.2 Display Application Details: Present at least four application details retrieved via APIs.

```
(venv) mkmac@MKMacs-MacBook-Air api % python api.py
<Response [200]>
connection is succesful
(venv) mkmac@MKMacs-MacBook-Air api %
```


api — -zsh — 114×42
(venv) mkmac@MKMacs-MacBook-Air api % python deploymentAPI.py
{'id': 'c51dc499-6e2e-440c-8abb-7fc55d770e4e', 'created': '2024-10-31T06:33:54.495843+00:00', 'updated': '2024-10-31T10:41:08.732686+00:00', 'infra_overrides': {}, 'name': 'student_performance_factors', 'version': '88a0d8e41beca7ed13650a64677a6b9', 'description': 'None', 'flow_id': 'fee0b8ea-4107-411b-93d7-d024be07c700', 'schedule': {'interval': 60.0, 'anchor_date': '2024-10-31T16:11:07.832464+05:30', 'timezone': 'Asia/Kolkata'}, 'is_schedule_active': True, 'paused': False, 'disabled': False, 'schedules': [{'id': 'c7b01e30-f0e8-44ae-b201-8d1fd43c9766', 'created': '2024-10-31T10:41:08.750504+00:00', 'updated': '2024-10-31T10:45:33.953915+00:00', 'deployment_id': 'c51dc499-6e2e-440c-8abb-7fc55d770e4e', 'schedule': {'interval': 60.0, 'anchor_date': '2024-10-31T16:11:07.832464+05:30', 'timezone': 'Asia/Kolkata'}, 'active': False, 'last_scheduled_at': None, 'soonest_scheduled_run': None, 'latest_scheduled_run': None, 'max_active_runs': None, 'max_scheduled_runs': None, 'catchup': False}], 'concurrency_limit': None, 'global_concurrency_limit': None, 'concurrency_options': None, 'job_variables': {}, 'parameters': {}, 'tags': ['studentperfromacefactors datascience project workflow'], 'labels': {'prefect.flow.id': 'fee0b8ea-4107-411b-93d7-d024be07c700'}, 'work_queue_name': None, 'last_polled': '2024-10-31T10:46:54.772173+00:00', 'parameter_openapi_schema': {'type': 'object', 'title': 'Parameters', 'properties': {}}, 'path': '.', 'pull_steps': [], 'entrypoint': 'workflow.py:student_performance_factors', 'manifest_path': None, 'storage_document_id': None, 'infrastructure_document_id': None, 'created_by': {'id': 'fc70383c-a362-43a4-935e-933ba919e26b', 'type': 'USER', 'display_value': 'kappa82-togashi-v'}, 'updated_by': {'id': 'fc70383c-a362-43a4-935e-933ba919e26b', 'type': 'USER', 'display_value': 'kappa82-togashi-v'}, 'work_pool_name': None, 'status': 'NOT_READY', 'enforce_parameter_schema': True}
(venv) mkmac@MKMacs-MacBook-Air api %


(venv) mkmac@MKMacs-MacBook-Air api % ls
api.py deploymentAPI.py flowAPI.py
(venv) mkmac@MKMacs-MacBook-Air api % python flowAPI.py
{'id': 'fee0b8ea-4107-411b-93d7-d024be07c700', 'created': '2024-10-31T06:33:54.130514+00:00', 'updated': '2024-10-31T06:33:54.130522+00:00', 'name': 'student-performance-factors', 'tags': [], 'labels': {}}
(venv) mkmac@MKMacs-MacBook-Air api %