Review - 2

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

WINTER SEMESTER 2023-2024

CAPSTONE PROJECT

Course code - **SWE1904**

Guide

Dr. CHANDRASEGAR. T

Student

VIVEK R - 19MIS0184

Review - 2

PANEL No: 02

TITLE OF THE PROJECT

COMPREHENSIVE APPROACH OF STATIC AND DYNAMIC DATA ANALYTICS USING AUTOML

PANAL INCHARGE

Dr. Shantharajah S P

PANAL INCHARGE

Prof. Chint alapudi V N U Bharathi Murthy

PROPOSED METHODOLOGY

- Developing automated solutions for data preprocessing, model training, algorithm selection, and hyperparameter optimization using AutoML techniques.
- Addressing class imbalance, redundant records, and missing guidelines in the CICIDS2017 dataset to improve its effectiveness as a benchmark for intrusion detection systems.
- Investigating techniques for handling IoT data and developing automated model updating procedures to maintain model performance over time.
- Comparing the performance of AutoML models with traditional machine learning
- Demonstrating the feasibility and effectiveness of AutoML in improving the efficiency and accuracy of data analytics tasks in both static and dynamic environments.

MODULE DESCRIPTION

Data Preprocessing Module

- Submodules and their functionalities:
  - Automated Encoding
  - Automated Imputation
  - Automated Normalization
  - Automated Data Balancing
  - Train-Test Split

MODULE DESCRIPTION

Model Learning Module

- Model Training
  - List of machine learning models: Naive Bayes, KNN, Random Forest, LightGBM, ANN
  - Brief description and implementation details for each model
- Model Evaluation
  - Performance metrics: Accuracy, Precision, Recall, F1-score

MODULE DESCRIPTION

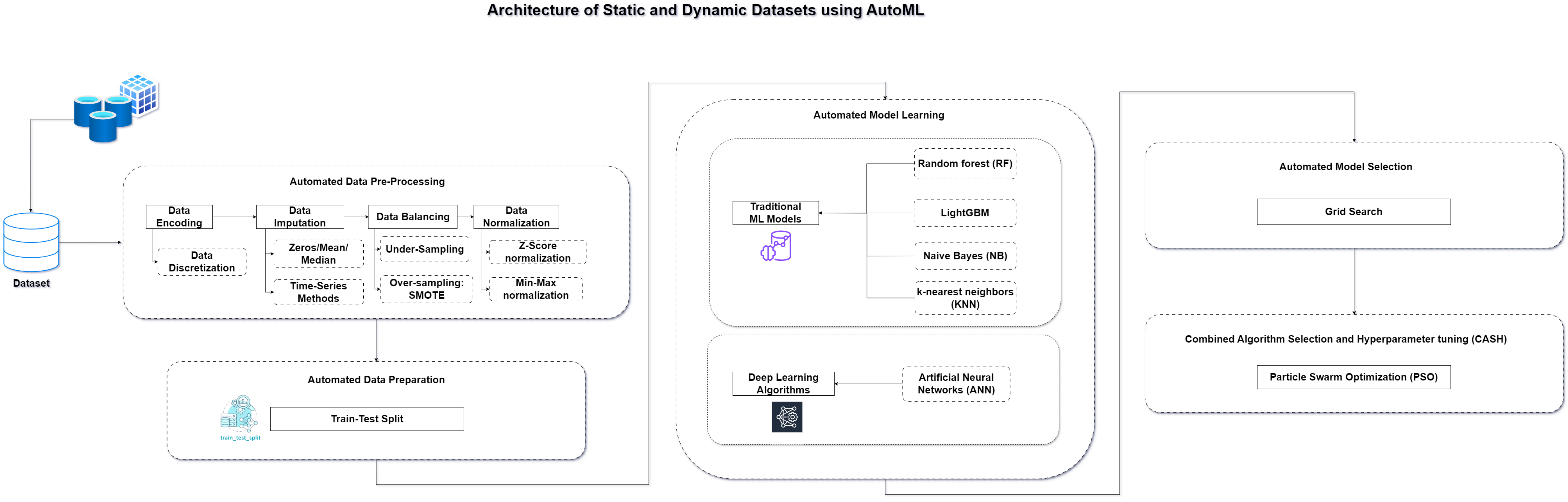
Model Selection Module

- Grid Search
  - Description and functionality

Combined Algorithm Selection and Hyperparameter Tuning (CASH) Module

- Particle Swarm Optimization (PSO)
  - Description and functionality

SYSTEM ARCHITECTURE:



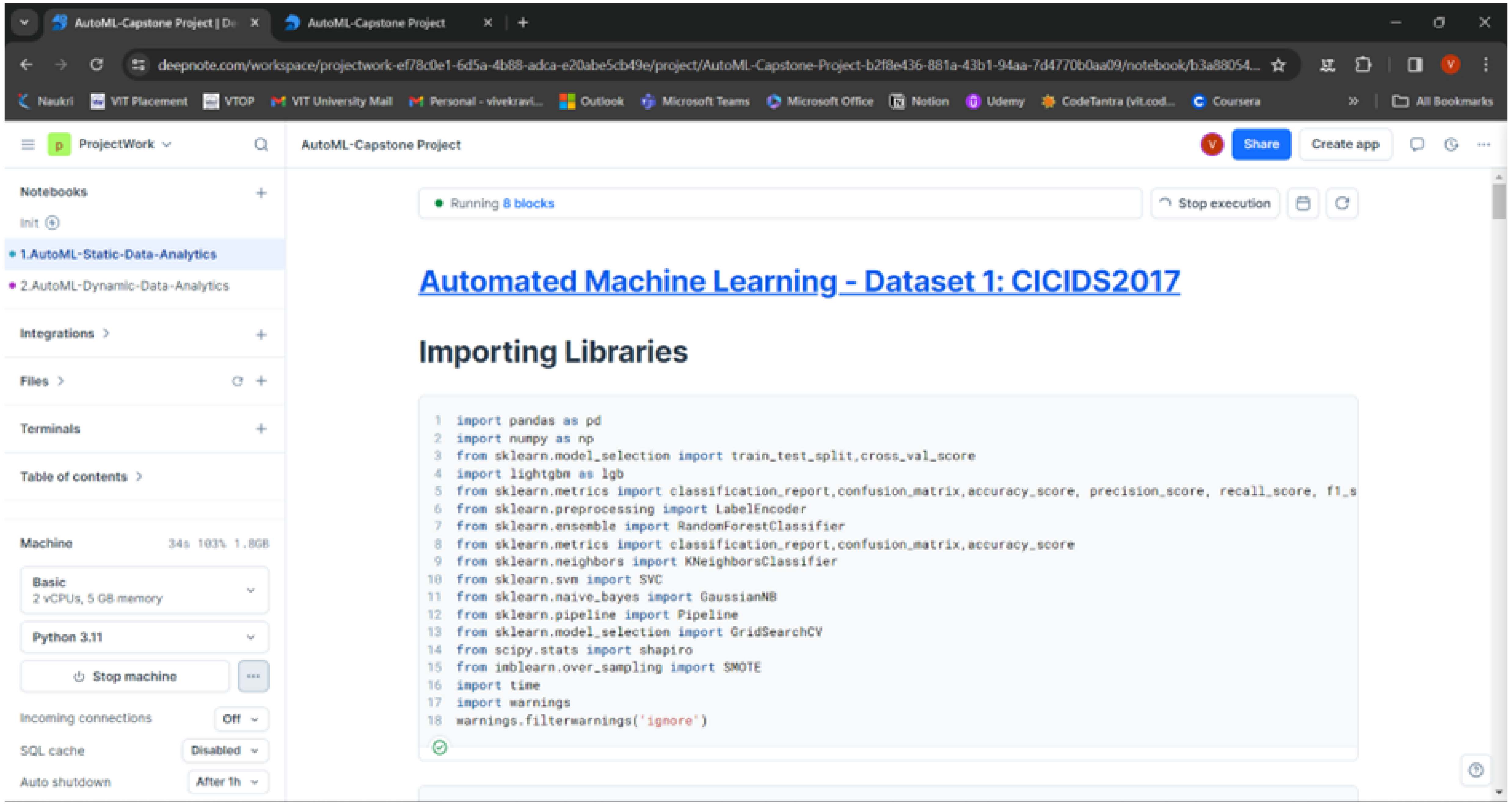
TECHNOLOGY USED

python™

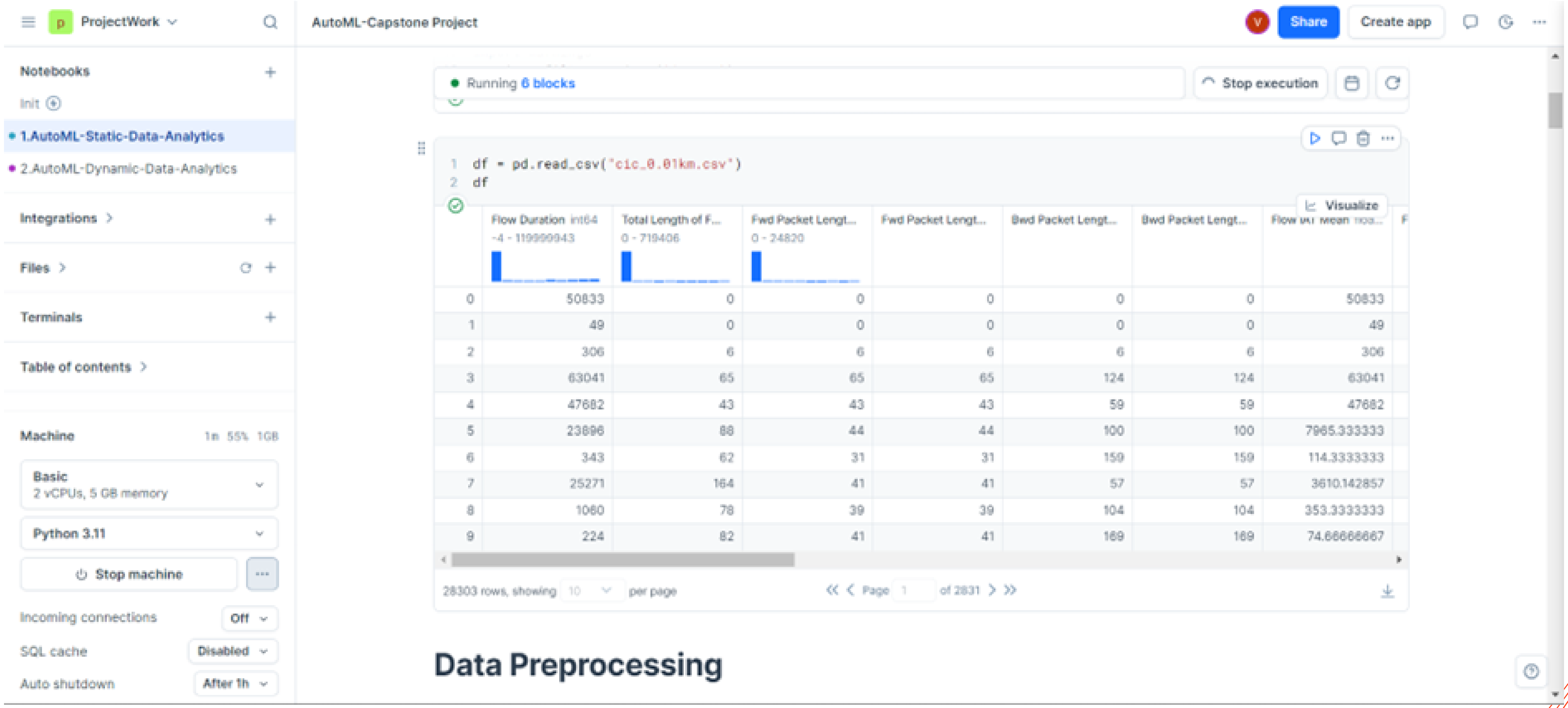
Deepnote

scikit learn

IMPLEMENTATION



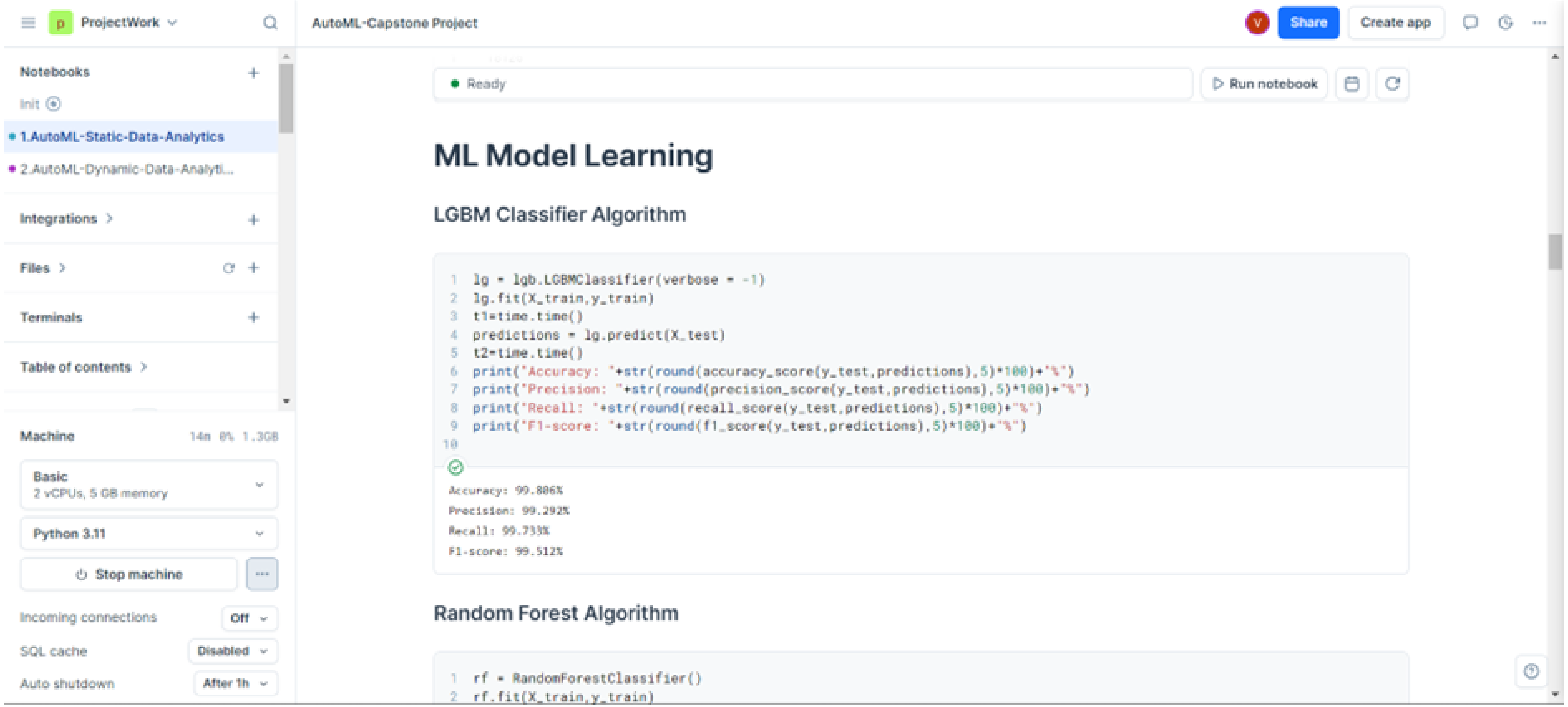
IMPLEMENTATION



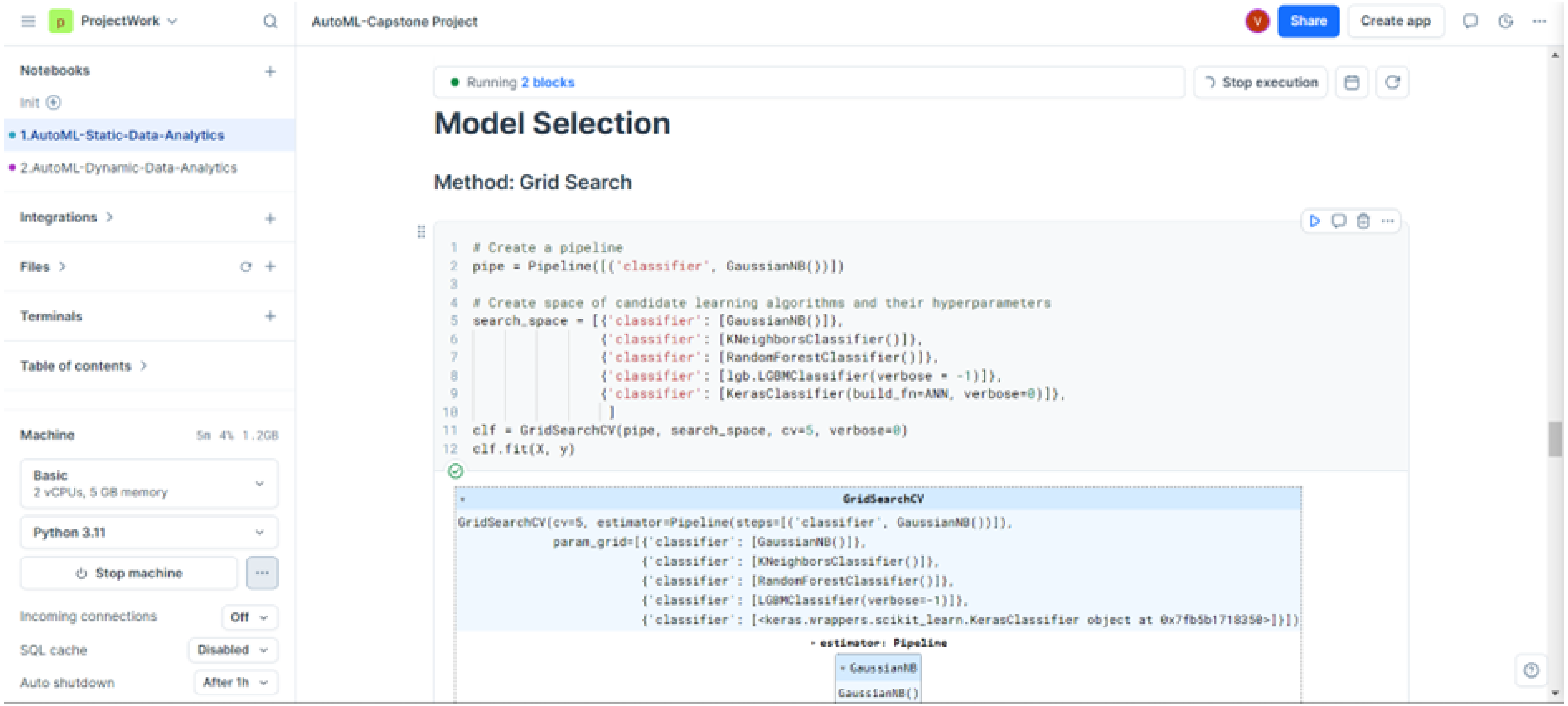

```
df = pd.read_csv('cic_0.01km.csv')
df
```

	Flow Duration int64 -4 - 119999943	Total Length of F... 0 - 719406	Fwd Packet Lengt... 0 - 24820	Fwd Packet Lengt...	Bwd Packet Lengt...	Bwd Packet Lengt...	Flow IAT Mean
0	50833	0	0	0	0	0	50833
1	49	0	0	0	0	0	49
2	306	6	6	6	6	6	306
3	63041	65	65	65	124	124	63041
4	47682	43	43	43	59	59	47682
5	23896	88	44	44	100	100	7965.333333
6	343	62	31	31	159	159	114.3333333
7	25271	164	41	41	57	57	3610.142857
8	1060	78	39	39	104	104	353.3333333
9	224	82	41	41	169	169	74.66666667

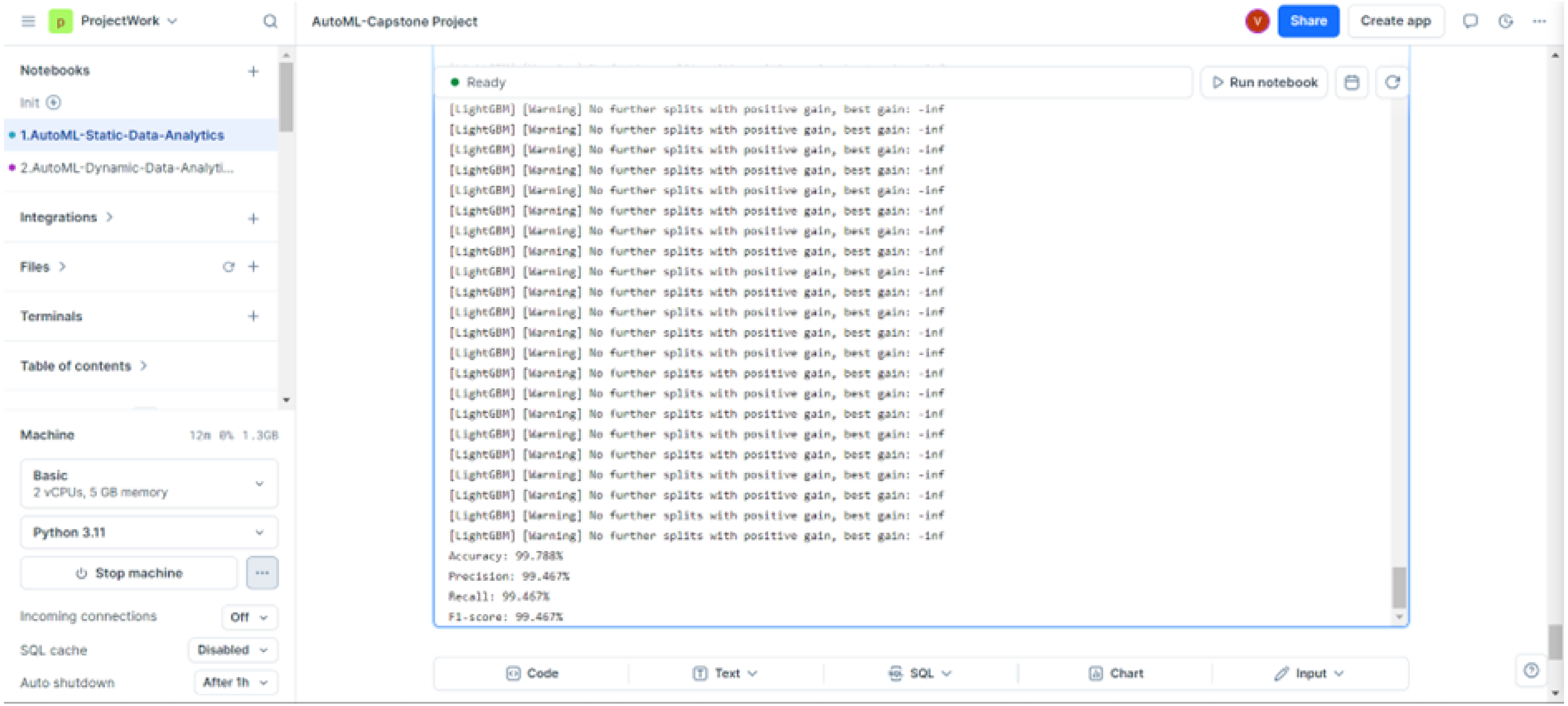
IMPLEMENTATION



IMPLEMENTATION



IMPLEMENTATION



PERFORMANCE METRICS: ACCURACY

Accuracy measures how many predictions were correct overall.

$$Acc = TP + TN / TP + TN + FP + FN$$

Static Dataset

Model Accuracy	**Percentage**
LGBM Classifier	99.753 %
Random Forest	75.729 %
Naive Bayes	98.728 %
k-nearest neighbors (KNN)	92.475 %
KerasClassifier Model	99.753 %

Dynamic Dataset

Model Accuracy	**Percentage**
LGBM Classifier	99.92%
Random Forest	99.839%
Naive Bayes	70.184%
k-nearest neighbors (KNN)	99.280%
KerasClassifier Model	92.475 %

PERFORMANCE METRICS: PRECISION

Precision measures proportion predictions that were actually correct.

Precision = TP / TP + FP

Static Dataset

Model Precision	Percentage
LGBM Classifier	99.378 %
Random Forest	99.554 %
Naive Bayes	44.891 %
k-nearest neighbors (KNN)	95.584 %
KerasClassifier Model	73.378 %

Dynamic Dataset

Model Precision	Percentage
LGBM Classifier	99.914%
Random Forest	99.83%
Naive Bayes	99.875%
k-nearest neighbors (KNN)	99.744%
KerasClassifier Model	92.475 %

PERFORMANCE METRICS: RECALL

Recall is a measure of proportion of actual positives that were correctly predicted.

Recall = TP / TP + FN

Static Dataset

Model Recall	**Percentage**
LGBM Classifier	99.788 %
Random Forest	99.753 %
Naive Bayes	97.244 %
k-nearest neighbors (KNN)	98.133 %
KerasClassifier Model	97.511 %

Dynamic Dataset

Model Recall	**Percentage**
LGBM Classifier	100.0%
Random Forest	100.0%
Naive Bayes	68.313%
k-nearest neighbors (KNN)	99.489%
KerasClassifier Model	92.475 %

PERFORMANCE METRICS: F1 SCORE

F1 score is harmonic mean of the Recall and Precision scores, therefore balancing their respective strengths

F1 = 2 x TP / 2 x TP + FP + FN

Static Dataset

Model F1 score	Percentage
LGBM Classifier	99.788 %
Random Forest	99.753 %
Naive Bayes	61.426 %
k-nearest neighbors (KNN)	96.842%
KerasClassifier Model	83.740%

Dynamic Dataset

Model F1 score	Percentage
LGBM Classifier	99.957%
Random Forest	99.914%
Naive Bayes	81.133%
k-nearest neighbors (KNN)	99.616%
KerasClassifier Model	92.475 %

REFERENCES:

[1] Yang, L., & Shami, A. (2022). IoT data analytics in dynamic environments: From an automated machine learning perspective. Engineering Applications of Artificial Intelligence, 116, 105366.

[2] Singh, A., Amutha, J., Nagar, J., Sharma, S., & Lee, C. C. (2022). AutoML-ID: Automated machine learning model for intrusion detection using wireless sensor network. Scientific Reports, 12(1), 9074.

[3] Lindstedt, H. (2022). Methods for network intrusion detection : Evaluating rule-based methods and machine learning models on the CIC-IDS2017 dataset (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-479347

[4] Garouani, M., Ahmad, A., Bouneffa, M., & Hamlich, M. (2022). AMLBID: an auto-explained automated machine learning tool for big industrial data. SoftwareX, 17, 100919.

[5] He, Y., Lin, J., Liu, Z., Wang, H., Li, L. J., & Han, S. (2018). Amc: Automl for model compression and acceleration on mobile devices. In Proceedings of the European conference on computer vision (ECCV) (pp. 784-800).

REFERENCES:

[6] He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. Knowledge-Based Systems, 212, 106622.

[7] Lee, J., Ahn, S., Kim, H., & Lee, J. R. (2022). Dynamic Hyperparameter Allocation under Time Constraints for Automated Machine Learning. Intelligent Automation & Soft Computing, 31(1).

[8] Wever, M., Tornede, A., Mohr, F., & Hüllermeier, E. (2021). AutoML for multi-label classification: Overview and empirical evaluation. IEEE transactions on pattern analysis and machine intelligence, 43(9), 3037-3054.

[9] Celik, B., Singh, P., & Vanschoren, J. (2023). Online automl: An adaptive automl framework for online learning. Machine Learning, 112(6), 1897-1921.

[10] Zhang, S., Gong, C., Wu, L., Liu, X., & Zhou, M. (2023). AutoML-GPT: Automatic Machine Learning with GPT. arXiv preprint arXiv:2305.02499

THANK YOU