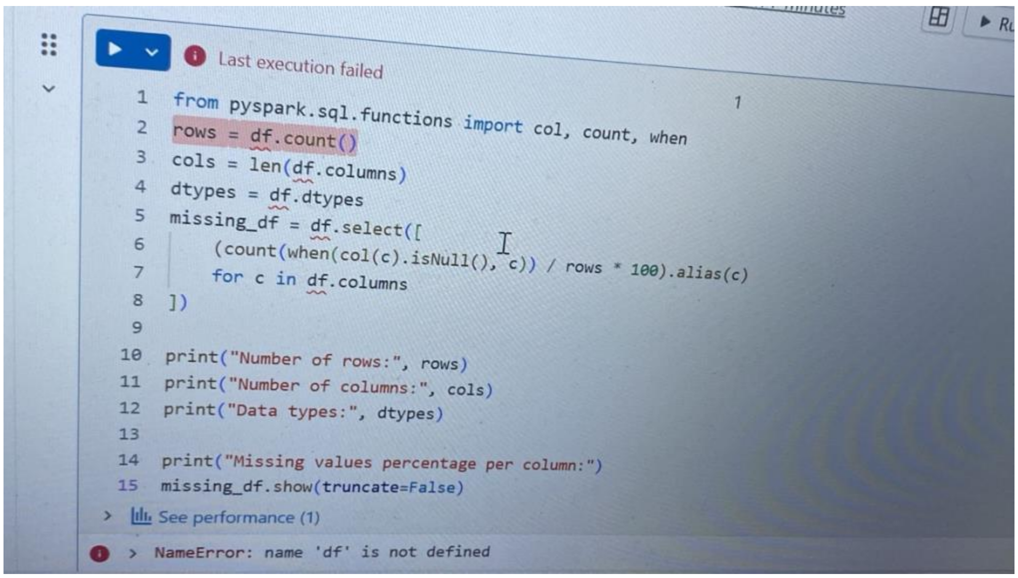
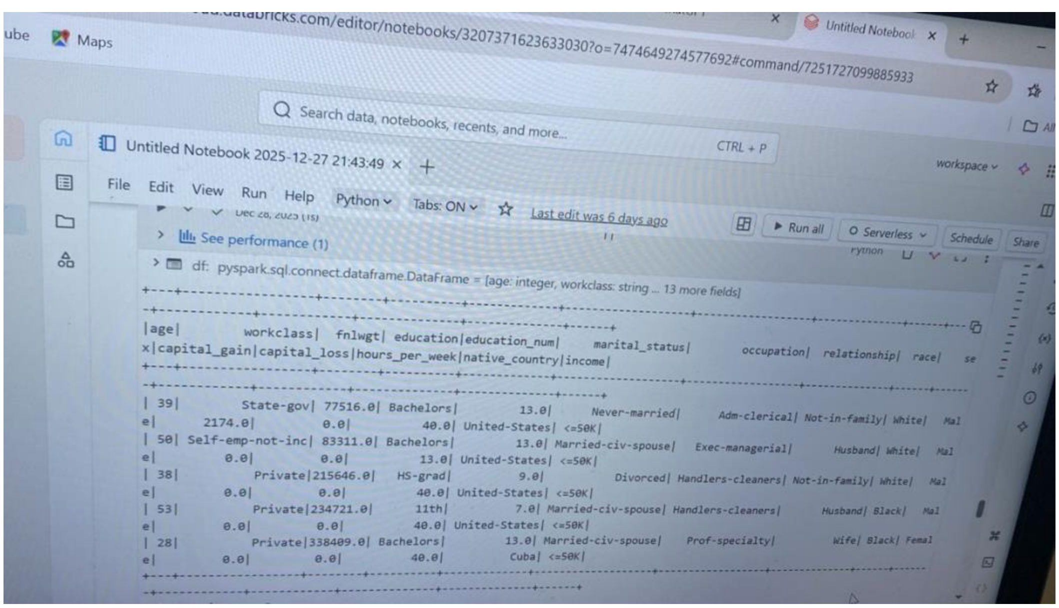
At first we use the dataset was loaded from the built-in Databricks datasets repository due to platform restrictions in the Community Edition.

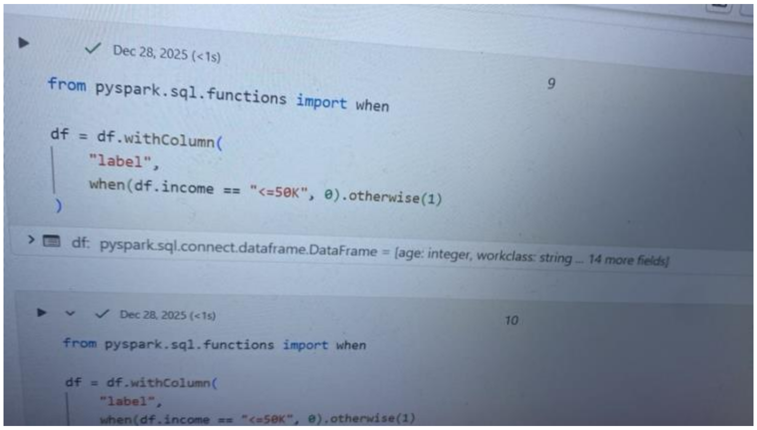
❖ In this screenshot show the code of descriptive statistics including the number of rows, number of columns, data types, and missing value percentages using Apache Spark.



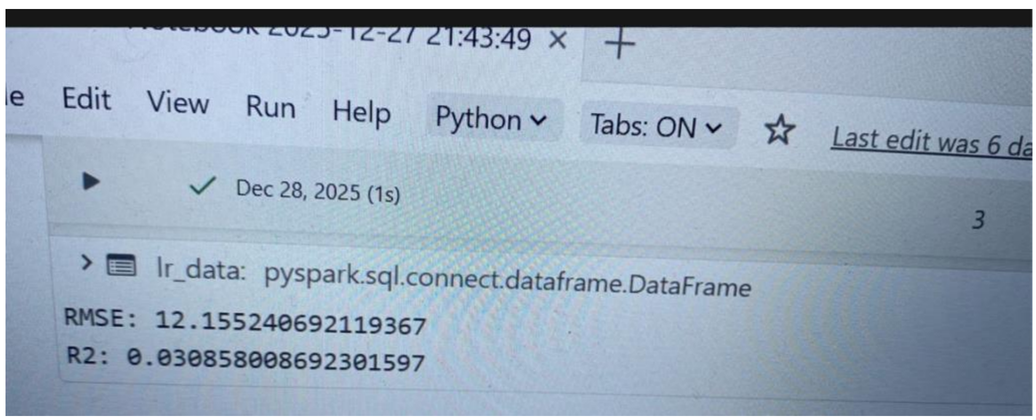
❖ This is the result of the code:



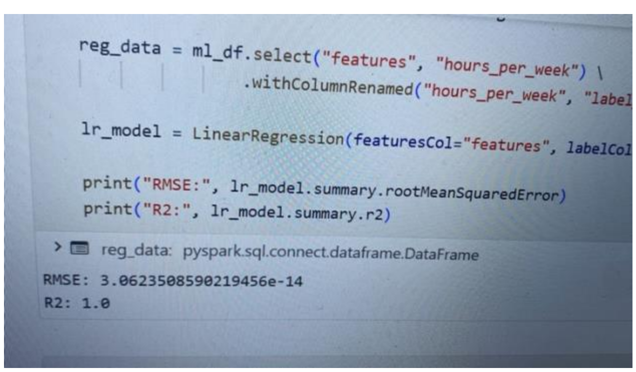
❖ Preparing the label before each ML:



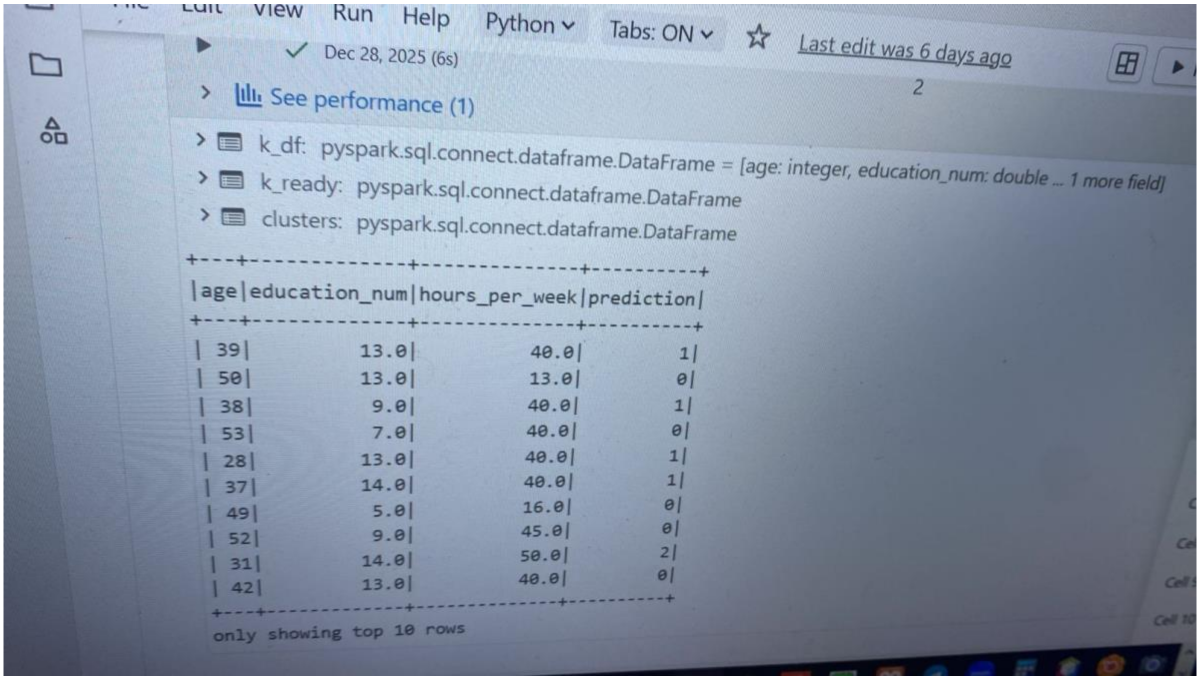
- Logistic Regression: implemented using Spark MLlib to perform binary classification. Numerical features were assembled using Vector Assembler, and the model was trained with 20 iterations. Predictions and class probabilities were generated and analyzed the screenshot below shown the result of it:



- Linear Regression: used to predict the number of working hours per week based on demographic and education-related features.



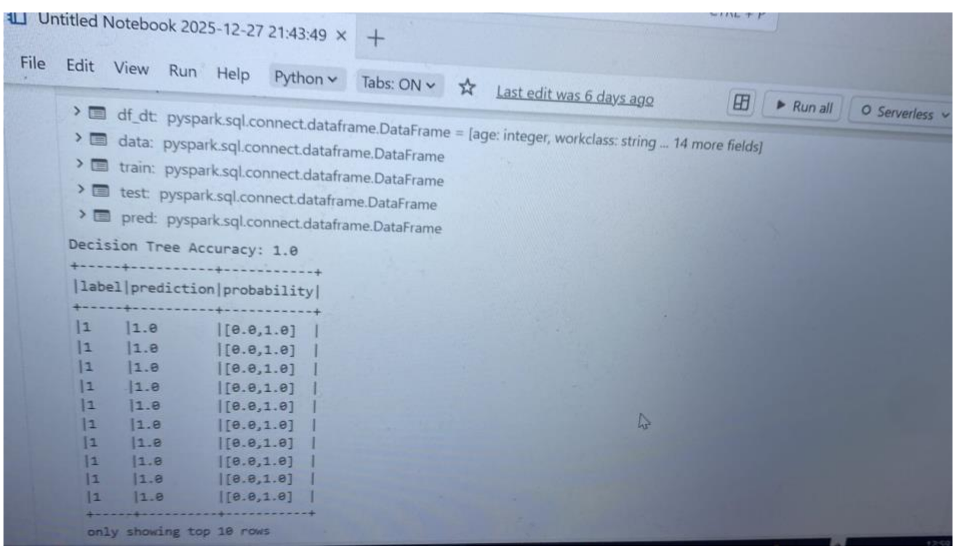
- KMeans clustering: applied to group individuals based on age, education level, and weekly working hours. The result below shown that the algorithm successfully divided the dataset into three distinct clusters.



age	education_num	hours_per_week	prediction
39	13.0	40.0	1
50	13.0	13.0	0
38	9.0	40.0	1
53	7.0	40.0	0
28	13.0	40.0	1
37	14.0	40.0	1
49	5.0	16.0	0
52	9.0	45.0	0
31	14.0	50.0	2
42	13.0	40.0	0

only showing top 10 rows

- The Decision Tree classifier achieved an accuracy of 100%. This result indicates strong classification performance; however, it may also suggest potential overfitting due to the limited number of numerical features used.



label	prediction	probability
1	1.0	[0.0,1.0]
1	1.0	[0.0,1.0]
1	1.0	[0.0,1.0]
1	1.0	[0.0,1.0]
1	1.0	[0.0,1.0]
1	1.0	[0.0,1.0]
1	1.0	[0.0,1.0]
1	1.0	[0.0,1.0]
1	1.0	[0.0,1.0]
1	1.0	[0.0,1.0]

only showing top 10 rows

❖ The result of the Experiments:

- Note: Due to Databricks Serverless limitations, the cluster size was simulated by controlling Spark parallelism using repartition(p), rather than physical machines.