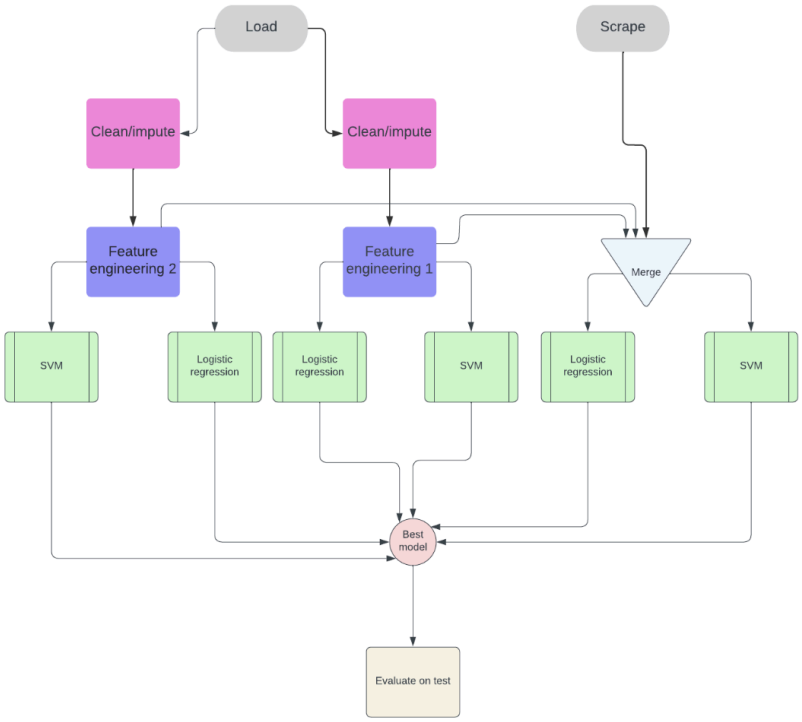
Matt Saperstein
DE300 Homework 4

Convert the codes that you created in the first three homework assignments to Airflow. There should be a single workflow with two independent branches (paths) – one for standard EDA and one for EDA with Spark. Loading/accessing the source data is the only part in common. At least the operations indicated in the figure must be included (you are free to have more of them). The top branch must be done in sklearn (reuse the code from the first assignment). The second parallel branch must be done in Spark (reuse the code from the third project assignment).

With regards to FE-1 and FE-2, pick two different feature engineering strategies that make sense.

Everything must be done in AWS. The backend database (or S3) choice is up to you. You must assume that each operation is potentially scheduled by airflow on a different server/VM.

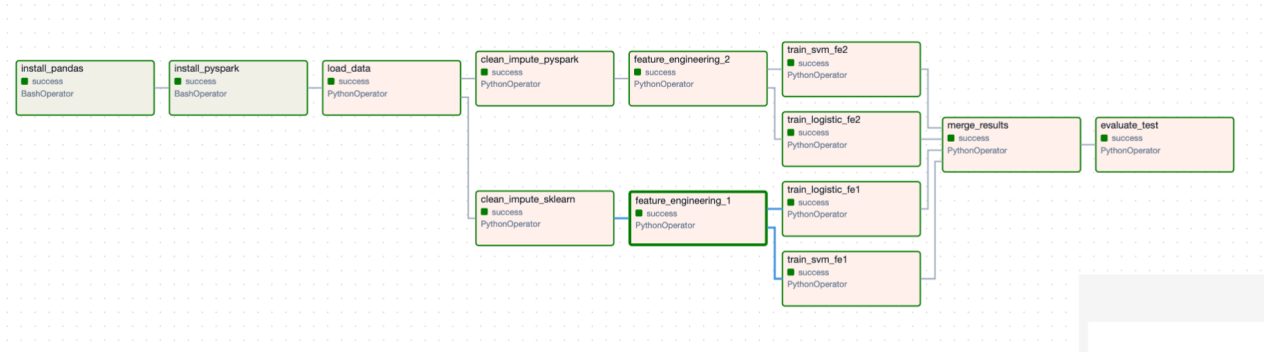


Turn the web scrapping code from the second assignment into airflow operations (the bottom portion of the diagram).

- I created two separate branches, one for EDA using sklearn and one for EDA using pyspark.
    - For the feature engineering method 1, which was for EDA using sklearn, I created a column for maximum heart rate using its formula. I named the column,

"max_HR", and used the following equation that I found on the internet: 206.9 - (0.67 * age).
    - For the feature engineering method 2, which was for EDA using pyspark, I created a column for blood pressure difference from normal blood pressure. I named this column, “bp_diff_from_norm”, and used the following equation which is based on the normal systolic blood pressure of 120 mmHg: trestbps (resting blood pressure) - 120.
- For each branch, I trained an SVM and Logistic Regression model. For each of these, I evaluated the model based on the test dataset to calculate the accuracy of the model. I then merged all the results together and identified which model was the most accurate out of the four SVM and Logistic Regression models.

Graph of Airflow Pipeline:



Results:

- SVM_FE1_accuracy: 0.7444444444444445
- Logistic_FE1_accuracy: 0.7444444444444445
- SVM_FE2_accuracy: 0.7083333333333334
- Logistic_FE2_accuracy: 0.75
    - **Best model based on accuracy: Logistic_FE2**

Sources: Used ChatGPT with help using Airflow.