# Data Science Pipeline Application - Mobot

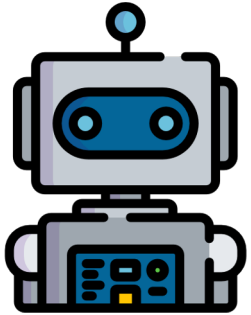
**Team Members :**
**Jane, Andy, Henry**

# Agenda

- ***What is our idea ?***
- ***Application Structure***
- ***Data Flow***

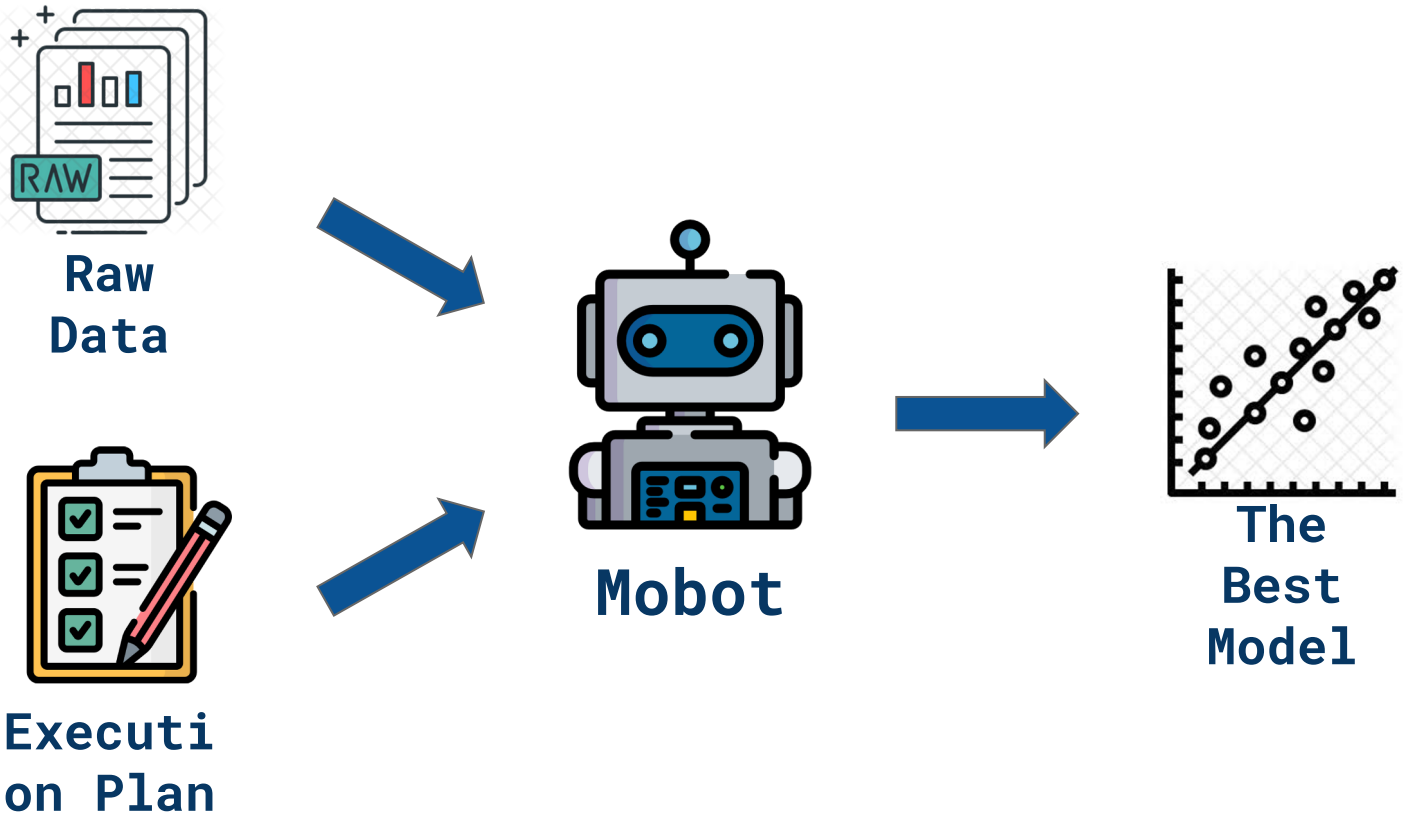
# What is our idea ?

***We want to build a Data Science Pipeline Application that could be beneficial to the future students who are interested in developing models for data analysis project.***



**Mobot**

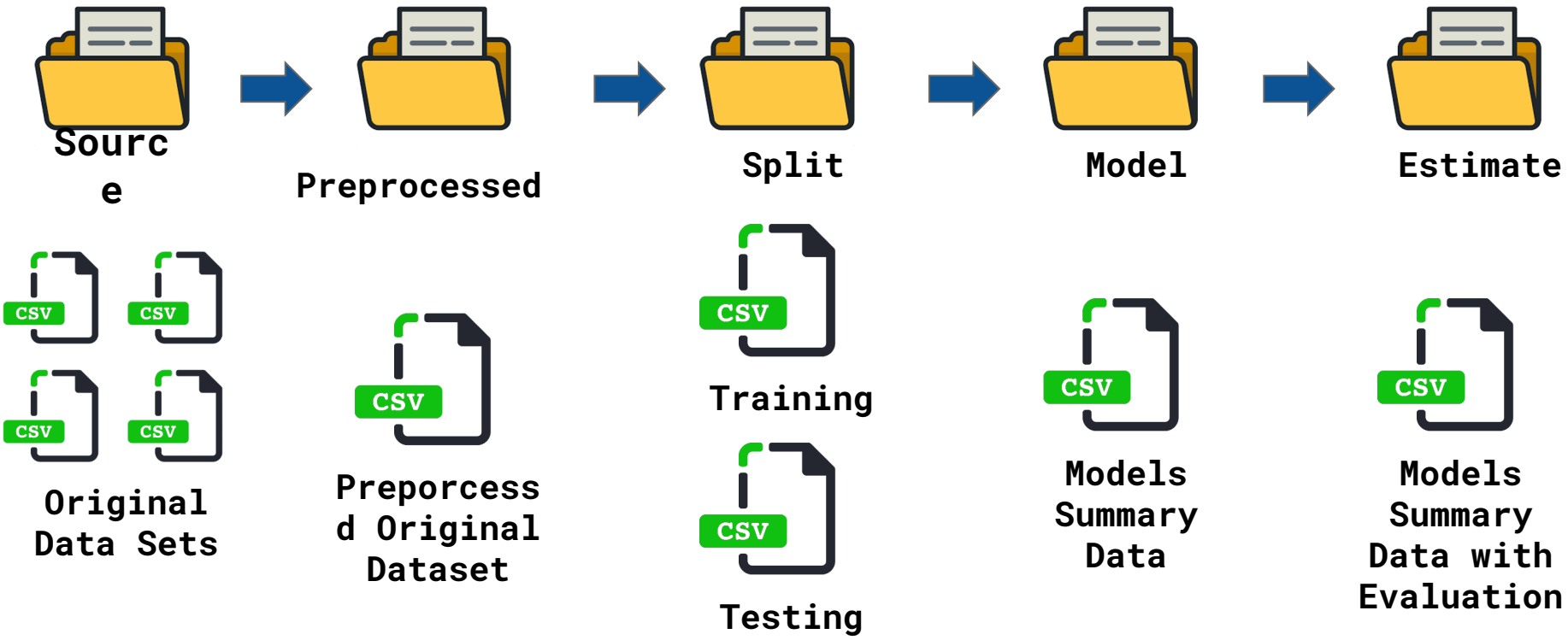
# The Structure of the Application



# The Structure of the Application

| Mission | Main Objective | Action | Detail |
|---|---|---|---|
| ***Data Collection*** | Gather all the data needed. | Put data to correct folder | The user has to collect the datasets and put it in the same folder. |
| ***Data Preprocess*** | Convert multiple original datasets into a single ready-to-run dataset. | Missing Data Imputation | The user has to decide a way to fill NA. ( Mean, Median, Zero and KDE ) |
| | | Data Transformation | The user has to decide a way to transorm data. ( Root Square, Log ) |
| | | Create a Flat Table | The user has to pick a column name as key to merge all the original datasets. |
| ***Training / Testing Data Split*** | Split processed dataset into training/testing datasets. | Train / Test Split | The user has to decide a way to train/test data split. ( Percentage, Column Value ) |
| ***Model Training*** | Run the processed training dataset with various kinds of model. | Run data with models | The user has to decide types of model to run. ( AIC, Simple Linear Regression, Stepwise ) |
| ***Model Evaluation*** | Evaulate the performance of different model running the preocessed data set. | Estimate performance of models run. | The user has to decide a way to estimate the performance of model. ( RMSE...etc ) |

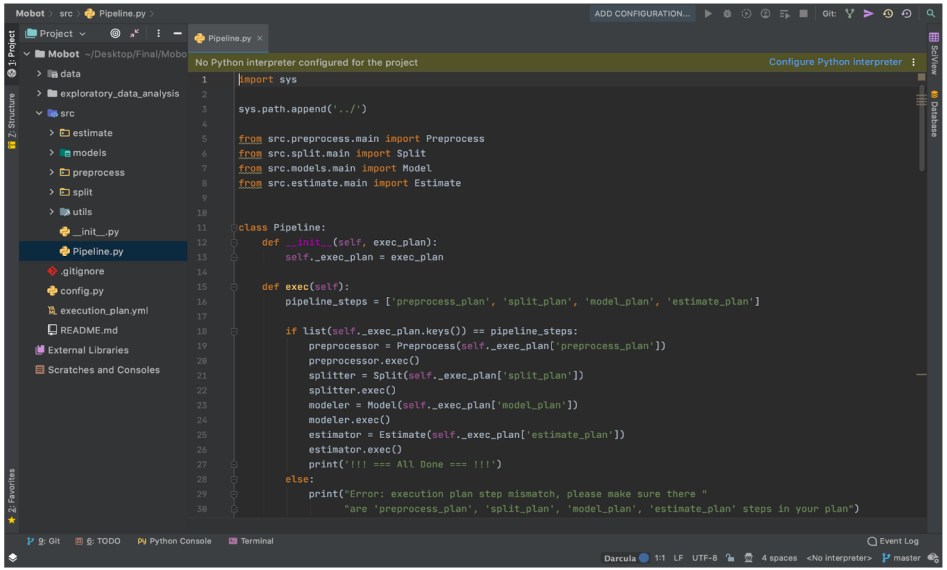
# Data Flow


Source
Preprocessed
Split
Model
Estimate
CSV
CSV
CSV
CSV
Original Data Sets
CSV
Preporcessed Original Dataset
CSV
Training
CSV
Testing
CSV
Models Summary Data
CSV
Models Summary Data with Evaluation

# Final Output Data Set Example

| | model_name | response | predictors | criteria | rmse |
|---|---|---|---|---|---|
| **0** | AIC | ['Recovery Rate'] | Physicians.density..physicians.1.000.population./Life expectancy at birth (years)/GDP - per capita (PPP) (US$) | | 0.259886543 |
| **1** | SimpleLm | ['Recovery Rate'] | Health.expenditures....of.GDP./Literacy..../Physicians.density..physicians.1.000.population./Obesity - adult prevalence rate (%)/Life expectancy at birth (years)/H_bed_density/Imigrate_Rate/Pop_Density/GDP - per capita (PPP) (US$)/Unemployment rate (%) | | 0.26027653 |
| **2** | StepWise_0.1 | ['Recovery Rate'] | Physicians.density..physicians.1.000.population./Life expectancy at birth (years)/GDP - per capita (PPP) (US$) | p_value 0.1 | 0.259886543 |
| **3** | StepWise_0.01 | ['Recovery Rate'] | Life expectancy at birth (years) | p_value 0.01 | 0.259581175 |
| **4** | StepWise_0.02 | ['Recovery Rate'] | Life expectancy at birth (years) | p_value 0.02 | 0.259581175 |

# Code Demo


Mobot > src > Pipeline.py >
ADD CONFIGURATION...
Project
Pipeline.py
Mobot ~/Desktop/Final/Mobo
No Python interpreter configured for the project
Configure Python interpreter
data
exploratory_data_analysis
src
estimate
models
preprocess
split
utils
__init__.py
Pipeline.py
.gitignore
config.py
execution_plan.yml
README.md
External Libraries
Scratches and Consoles
import sys
sys.path.append('../')
from src.preprocess.main import Preprocess
from src.split.main import Split
from src.models.main import Model
from src.estimate.main import Estimate
class Pipeline:
def __init__(self, exec_plan):
self._exec_plan = exec_plan
def exec(self):
pipeline_steps = ['preprocess_plan', 'split_plan', 'model_plan', 'estimate_plan']
if list(self._exec_plan.keys()) == pipeline_steps:
preprocessor = Preprocess(self._exec_plan['preprocess_plan'])
preprocessor.exec()
splitter = Split(self._exec_plan['split_plan'])
splitter.exec()
modeler = Model(self._exec_plan['model_plan'])
modeler.exec()
estimator = Estimate(self._exec_plan['estimate_plan'])
estimator.exec()
print('!!! === All Done === !!!')
else:
print("Error: execution plan step mismatch, please make sure there "
"are 'preprocess_plan', 'split_plan', 'model_plan', 'estimate_plan' steps in your plan")
9: Git
6: TODO
Python Console
Terminal
Event Log
Darcula
1:1
LF
UTF-8
4 spaces
<No interpreter>
master