**Hướng dẫn sử dụng Dash Machine**

1. **Biểu đồ dự án**

├───.vscode

├───antv\_tinhvan.com

├───app

│ ├───eda

│ │ ├───assets

│ │ ├───data

│ │ ├───utils

│ │ │ └───\_\_pycache\_\_

│ │ └───\_\_pycache\_\_

│ └───main

│ ├───assets

│ ├───component

│ │ └───\_\_pycache\_\_

│ └───\_\_pycache\_\_

└───catboost\_info

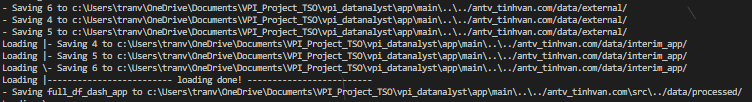
└───learn

* Cài đặt dependences ở sourch chính dự án bằng lệnh: pip install -r requements.txt
* Từ mã nguồn truy cập vào folder app (eda dành cho app eda, main dành cho app machine) , từ 1 trong 2 folder tìm đến file app.py và run chương trình.

A screenshot of a computer

Description automatically generated

* Chọn hoặc đưa file vào đầu vào sẽ là 1 hoặc nhiều file
* Phía dưới khung trông sẽ là thông báo đọc file terminal sẽ hiện file được đọc và lưu ở đâu



* **Chọn kiểu xử lý dữ liệu:** Tất cả mặc định sẽ được xử lý trên tất cả các cột với ngưỡng k= 3 ngoại trừ lọc bằng tay tất cả giá trị > 2000 (có thể ko lọc bỏ trống)
* **Scale:** các loại Scale
* **Removemissing:** Any( loại bỏ các bản ghi có chứa dù gì chỉ 1 missing), All(chỉ xóa bản ghi gần như missing hết các giá trị) , còn lại là chấp nhận mising tối đa bao nhiêu cột.
* **Orther:** Phương pháp xử lý thêm (PCA: giảm chiều dữ liệu, Poly: thêm hyper feature)
* **Chọn mô hình:** có thể chọn một hặc nhiều mô hình ( chú ý với dữ liệu để bị thiếu missing data chỉ chọn mô hình liên quan đến xgb, lgbm, catboost)
* **K-fold:** bao nhiêu fold ví dụ 1 tập dữ liệu chia 5 phần (1/5=0.2 => 20%)
* **Type model:** có áp dụng cài đặt trên hay không mặc định apply, auto sẽ train toàn bộ mô hình và đưa ra kết quả
* **Shuffle:** Đảo dữ liệu đảm bảo đa dạng đặc trưng
* **Standerize Column back to original value:** chuẩn hóa các cột bắt buộc ở trong giá trị 0-1 loại bỏ giá trị nằm ngoài khoảng
* **Huấn luyện lâu với dữ liệu lớn.**