TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP HỒ CHÍ MINH

MÔN HỌC: KHOA HỌC DỮ LIỆU

BÁO CÁO ĐỒ ÁN CUỐI KỲ

Nhóm 21

1612272 - Trần Nhật Huy

1512222 - Nguyễn Duy Hưng

I. Họp nhóm và phân công công việc

- Lần 1

  Ngày: 24/11/2019

  Địa điểm: Facebook

Công việc	Ngày thực hiện	Người làm
Crawl dữ liệu	26/11/2019	Nguyễn Duy Hưng
Viết báo cáo	30/11/2019	Trần Nhật Huy

- Lần 2

  Ngày: 02/01/2020

  Địa điểm: Facebook

Công việc	Ngày thực hiện	Người làm
Thống nhất kế hoạch làm	02/01/2020	Nguyễn Duy Hưng Trần Nhật Huy
Viết kế hoạch và chỉnh sửa	03/01/2020	Nguyễn Duy Hưng Trần Nhật Huy
Thí nghiệm 1	05/01/2020	Trần Nhật Huy
Thí nghiệm 2	05/01/2020	Nguyễn Duy Hưng
Thí nghiệm 3	05/01/2020	Trần Nhật Huy
Thí nghiệm 4, 5, 6	06/01/2020	Nguyễn Duy Hưng
Thí nghiệm 07	06/01/2020	Trần Nhật Huy
Viết báo cáo	06/01/2020	Nguyễn Duy Hưng
Slide	06/01/2020	Trần Nhật Huy

II. Bài toán đặt ra và hướng giải quyết

1. Bài toán

- Nhóm muốn dự đoán độ ẩm của ngày hôm sau dựa vào dữ liệu của ngày hôm trước
- Input: Dữ liệu thời tiết của ngày hôm trước (temperature, humidity, …)
- Output: Humidity của ngày hôm sau
- Lợi ích: Chúng ta có thể dự đoán trước được humidity của một ngày nào đó. Phục vụ cho nông nghiệp hoặc dự báo thời tiết
- Nguồn gốc: Nhóm tự đặt ra câu hỏi này

2. Thu thập dữ liệu

- Dữ liệu của nhóm được lấy từ API Dark Sky từ 01-01-2010 đến 31-12-2011 ở thành phố Hồ Chí Minh, Việt Nam

3. Tiền xử lý dữ liệu

- Điền giá trị thiếu, nhóm sử dụng SimpleImputer với các strategy lần lượt là mean, median, most_frequent
- Lựa chọn đặc trưng nhóm sử dụng PCA, SelectKBest, Pearson, Feature Importance + với tự tạo thêm đặc trưng ở một số thí nghiệm
- Chuẩn hóa dữ liệu nhóm sử dụng StandardScalar

4. Các thí nghiệm đã thực hiện

- Thí nghiệm 1

  Nhóm tách thuộc tính time thành day và month. Bỏ thuộc tính visibility, apparentTemperature. Dùng PCA với n_components từ [1:số lượng cột].

  Độ lỗi MAE trên tập validation là 0.054. Không tốt vì std của cột Humidity là 0.08.
- Thí nghiệm 2

  Nhóm giữ nguyên các thuộc tính ban đầu, không thêm hay xóa bất kỳ cột nào. Độ lỗi MAE trên tập validation là 0.055
- Thí nghiệm 3

Nhóm lựa chọn đặc trưng bằng Pearson’s Correlation

Độ lỗi MAE trên tập validation là 0.056

- Thí nghiệm 4

  Nhóm lựa chọn đặc trưng bằng SelectKBest

  Độ lỗi MAE trên tập validation là 0.056

- Thí nghiệm 5

  Nhóm lựa chọn đặc trưng bằng Feature Importance

  Độ lỗi MAE trên tập validation là 0.057

- Thí nghiệm 6

  Nhóm sử dụng mô hình Neural Network

  Độ lỗi MAE trên tập validation là 0.07 với alpha=0.1 và hidden layer là 1

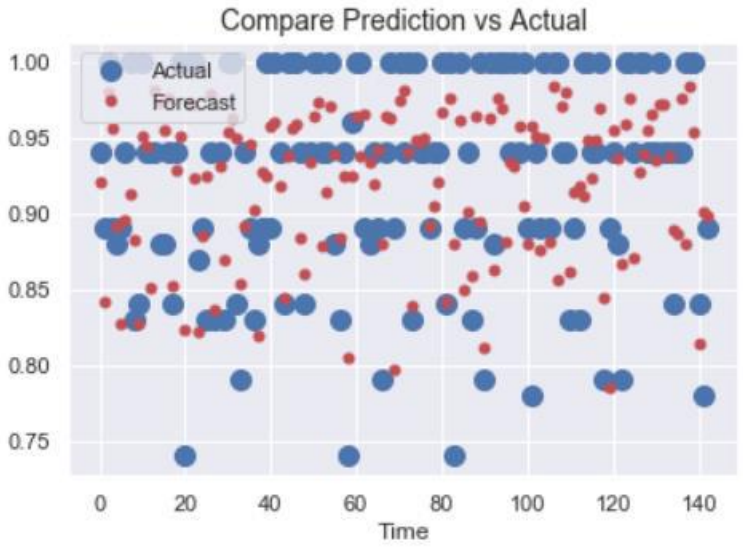
  Độ lỗi của mô hình này còn lớn hơn trước. Có thể bị overfitting

- Thí nghiệm 7

  Nhóm dự đoán có thể sử dụng dữ liệu của một ngày chưa đủ nên thử thêm độ ẩm của k ngày trước đó. Nhóm thử nghiệm với k = 7

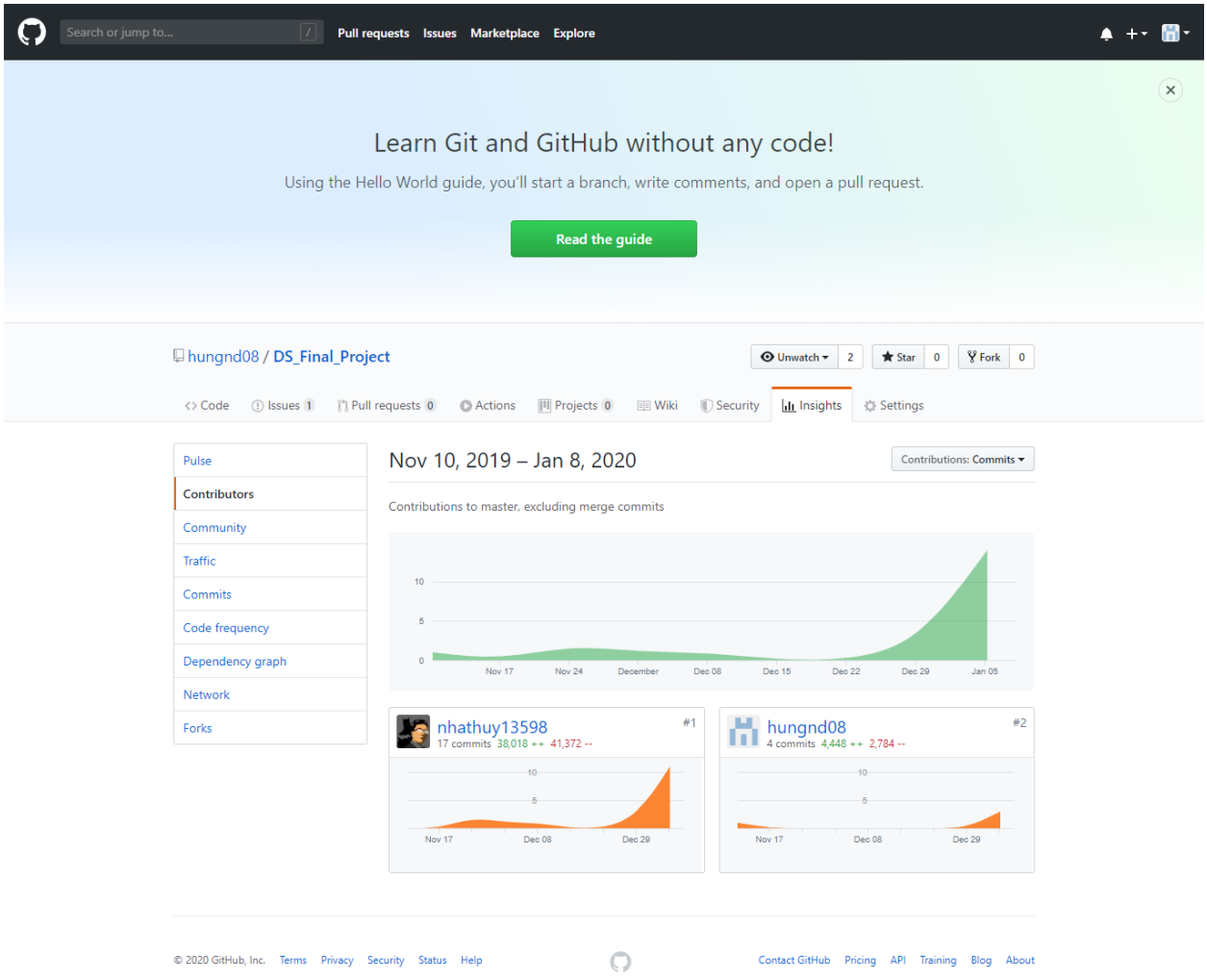
  Độ lỗi MAE trên tập validation là 0.045

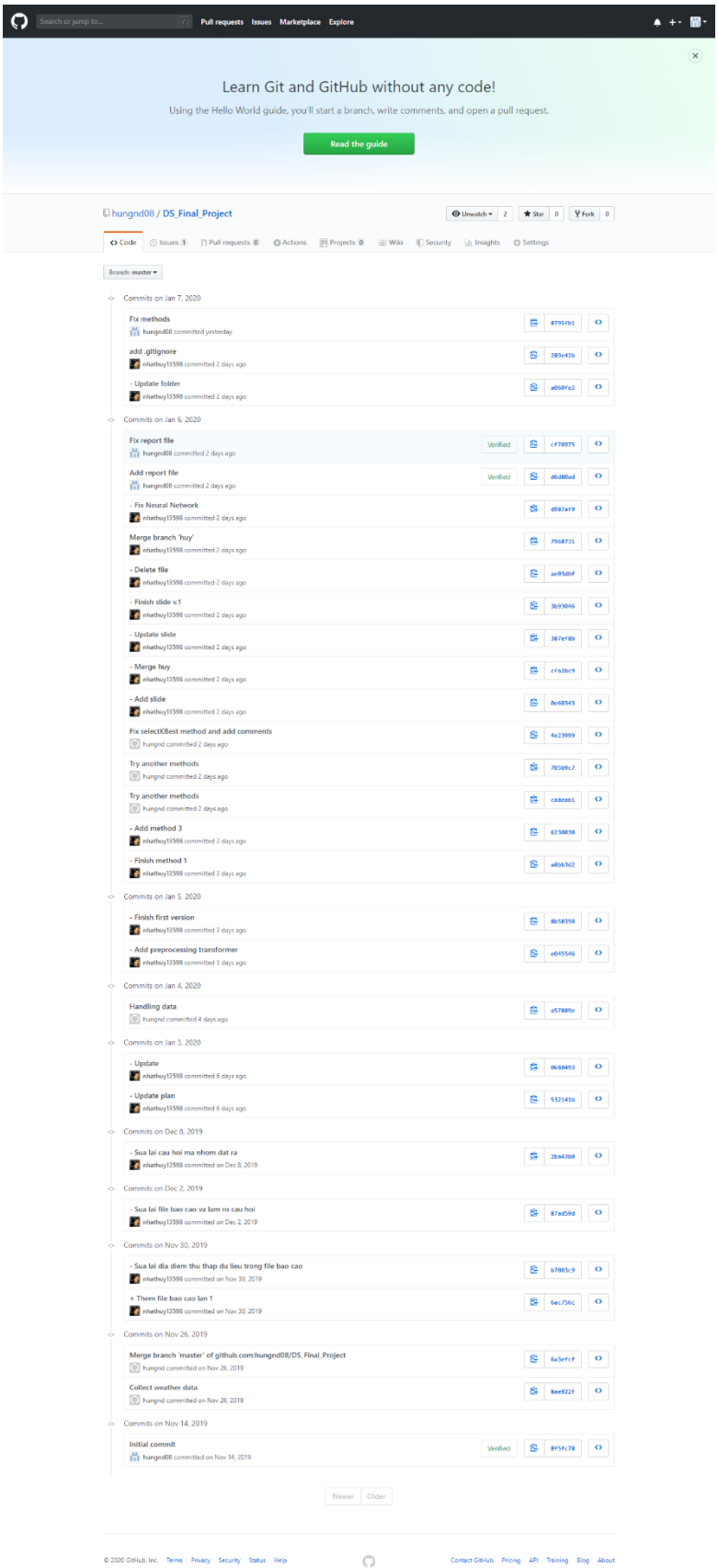
III. Tổng kết và đánh giá đóng góp của các thành viên trong nhóm



- Độ lỗi là 0.045
- Các thí nghiệm mà nhóm thực hiện có kết quả không được tốt
- Nhóm nghĩ có thể nguyên nhân là:
  - Dữ liệu có nhiễu nhưng nhóm xử lý chưa tốt
  - Mô hình không phù hợp, bị underfitting
- Link github: https://github.com/hungnd08/DS_Final_Project
- Đánh giá đóng góp của các thành viên trong nhóm

Ở mục contributors account “hungnd08” bị thiếu mất 6 commit vì em kết nối với project bằng SSH key và trong file .gitconfig trên máy cấu hình name em để là “hungnd” và email của em khác với email đăng ký account trên github nên khi commit github nó không tính vô mục contributors, phía dưới em có chụp thêm phần lịch sử commit để thầy xem ạ.


Search or jump to...
Pull requests
Issues
Marketplace
Explore
Learn Git and GitHub without any code!
Using the Hello World guide, you'll start a branch, write comments, and open a pull request.
Read the guide
hungnd08 / DS_Final_Project
Unwatch 2
Star 0
Fork 0
Code
Issues 1
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
Settings
Pulse
Contributors
Community
Traffic
Commits
Code frequency
Dependency graph
Network
Forks
Nov 10, 2019 – Jan 8, 2020
Contributions: Commits
Contributions to master, excluding merge commits
10
5
0
Nov 17
Nov 24
December
Dec 08
Dec 15
Dec 22
Dec 29
Jan 05
nhathuy13598
#1
17 commits 38,018 ++ 41,372 --
hungnd08
#2
4 commits 4,448 ++ 2,784 --
© 2020 GitHub, Inc.
Terms
Privacy
Security
Status
Help
Contact GitHub
Pricing
API
Training
Blog
About


Search or jump to...
Pull requests
Issues
Marketplace
Explore
Learn Git and GitHub without any code!
Using the Hello World guide, you'll start a branch, write comments, and open a pull request.
Read the guide
hungnd08 / DS_Final_Project
Unwatch 2
Star 0
Fork 0
Code
Issues 1
Pull requests 0
Actions
Projects 0
Wiki
Security
Insights
Settings
Branch: master
Commits on Jan 7, 2020
Fix methods
hungnd08 committed yesterday
8795f81
add .gitignore
nhathuy13598 committed 2 days ago
289e42b
- Update folder
nhathuy13598 committed 2 days ago
a060fe2
Commits on Jan 6, 2020
Fix report file
hungnd08 committed 2 days ago
Verified
cf76975
Add report file
hungnd08 committed 2 days ago
Verified
d6d80ad
- Fix Neural Network
nhathuy13598 committed 2 days ago
d987af9
Merge branch 'huy'
nhathuy13598 committed 2 days ago
7968731
- Delete file
nhathuy13598 committed 2 days ago
ae85dbf
- Finish slide v.1
nhathuy13598 committed 2 days ago
3b93046
- Update slide
nhathuy13598 committed 2 days ago
387ef8b
- Merge huy
nhathuy13598 committed 2 days ago
cfa2bc9
- Add slide
nhathuy13598 committed 2 days ago
0e68343
Fix selectKBest method and add comments
hungnd committed 2 days ago
4e23999
Try another methods
hungnd committed 2 days ago
705b9c7
Try another methods
hungnd committed 2 days ago
ca8eab1
- Add method 3
nhathuy13598 committed 2 days ago
6210050
- Finish method 1
nhathuy13598 committed 3 days ago
a8bb3d2
Commits on Jan 5, 2020
- Finish first version
nhathuy13598 committed 3 days ago
8b58350
- Add preprocessing transformer
nhathuy13598 committed 3 days ago
e045546
Commits on Jan 4, 2020
Handling data
hungnd committed 4 days ago
a57880e
Commits on Jan 3, 2020
- Update
nhathuy13598 committed 6 days ago
0684493
- Update plan
nhathuy13598 committed 6 days ago
532141b
Commits on Dec 8, 2019
- Sua lai cau hoi ma nhom dat ra
nhathuy13598 committed on Dec 8, 2019
2ba43b0
Commits on Dec 2, 2019
- Sua lai file bao cao va lam ro cau hoi
nhathuy13598 committed on Dec 2, 2019
87ad59d
Commits on Nov 30, 2019
- Sua lai dia diem thu thap du lieu trong file bao cao
nhathuy13598 committed on Nov 30, 2019
b7083c9
+ Them file bao cao lan 1
nhathuy13598 committed on Nov 30, 2019
6ec756c
Commits on Nov 26, 2019
Merge branch 'master' of github.com:hungnd08/DS_Final_Project
hungnd committed on Nov 26, 2019
6e3efcf
Collect weather data
hungnd committed on Nov 26, 2019
8ee922f
Commits on Nov 14, 2019
Initial commit
hungnd08 committed on Nov 14, 2019
Verified
8f5fc78
Newer
Older
© 2020 GitHub, Inc.
Terms
Privacy
Security
Status
Help
Contact GitHub
Pricing
API
Training
Blog
About

-