AML - group 21

Assignment instructions
https://courseworks2.columbia.edu/courses/154941/pages/coms-w4995-aml-project

Members:

- Dieter Joubert
    - Id: dj2574
    - Cellphone: 650 450 3963
    - Email: dj2574@columbia.edu
- Keli Wang
    - Id: kw3015
    - Cellphone: 9178564778
- Ritvik Khandelwal
    - Id: rk3213
    - Cellphone: 917 251 5495
- Xinyu He
    - Id: xh2562
    - Cellphone: 347 883 3541
- Ethan Tucker
    - Id: eht2122
    - Cellphone: 216 970 5360
    - Email: eht2122@columbia.edu

Project ideas:

1. Covid-19 datasets:
    - https://paperswithcode.com/dataset/covid19-algeria-and-world-dataset
    - Too time-series focused?
2. https://paperswithcode.com/dataset/mimic-iii
3. https://paperswithcode.com/dataset/learning-to-rank-challenge
4. NBA kaggle
    - https://www.kaggle.com/datasets/nathanlauga/nba-games
5. Flight delay:
    - https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022
    - Data might have issues?
6.

Proposals:

- Dieter Joubert: Dieter project idea
- Keli Wang: proposal link (Covid-19)
- Ritvik Khandelwal: aml project proposal (Stock Prices)
- Xinyu He: https://docs.google.com/document/d/1FagCKB9wUOkDjS2BoNGQGMr_ihKPg6kHh5oR

hXQMLCs/edit?usp=sharing (Data: https://drive.google.com/drive/folders/1AIQJWjbGXxrP9KPJd5CBwx4Ko5k38uSW?usp=sharing)

- Ethan Tucker: Proposal Link (Wildfires)

Chosen proposal: https://docs.google.com/document/d/1RlbUorm_KdkI7Agq-XLYDrL7ZA7hIQEGkRhBblrj18U/edit

Next Steps

Project Deliverable #2 - Data Analysis and Visualization (due 11/07/2022)

- **Github: https://github.com/DieterJoubert/AML_group_21**
- **Slide deck: https://docs.google.com/presentation/d/1YFcsJc7321MP4g3oc8x5s9eccrGIlbsOEvp0ealHdsM/edit#slide=id.p**
- **Deliverable: 8-10 slides with notes**

  Components
- **1. Initial data exploration**
- **2. Cleaning and sampling**
- **3. Insights from data exploration and**
- **4. Machine Learning techniques proposed to be implemented**

Initial Data Exploration: (Dieter)

- Figure out how to read data or how to convert to CSV
- Helpful kaggle notebooks
  - https://www.kaggle.com/code/edhirif/predict-the-causes-of-wildfires-using-python
- Creating jupyter notebook reading and displaying initial data

Cleaning and Sampling: (Ethan)

Insights from data exploration: (Xinyu & Keli)

Machine Learning techniques proposed: (Ritvik)

Project Deliverable #3 - Report & Code (due 12/05/2022)

- **Deliverable: 3-page final report, python code on github classroom**

Slides link:
https://docs.google.com/presentation/d/1YFcsJc7321MP4g3oc8x5s9eccrGIlbsOEvp0ealHdsM/edit#slide=id.g16a2ced5261_0_30

Github:
https://github.com/DieterJoubert/AML_group_21

Tasks to do:

- Data
  - Consolidate notebooks, final cleanup (Dieter)
- Machine learning methods application
  - Multinomial Logistic Regression (Dieter)
  - SVM (Dieter)
  - Decision-tree (Ritvik)
  - Random-forest (Ritvik)
  - Boosting (Keli) (Xinyu)
  - Neural Networks (Ritvik & Keli & Xinyu)
    - Fastai library (https://docs.fast.ai/tutorial.tabular.html)
    - Basic feedforward net
    - CNN
  - Time Series Analysis (Ethan)
    - Vanilla Recurrent Neural Network (RNN)
    - Long Short Term Memory (LSTM)
- Project write-up
  - Meet Saturday (and Monday?)

Write-up:
AML group 21 - final project writeup

Relevant kaggle notebooks:
https://www.kaggle.com/code/edhirif/predict-the-causes-of-wildfires-using-python