IT1244 README

Abstract

In this project, we present machine learning methods aimed at classifying between healthy individuals and two distinct cancer stages: screening stage and early stage. We explored a range of machine learning models, including K-Nearest Neighbours (KNN), Decision Trees, Random Forests and Artificial Neural Networks (ANN). Through experimentation and hyperparameter tuning, we managed to use Random Forest to classify the test dataset to an recall score of 0.33 for class 2 (Screening), recall score of 0.66 for class 3 (early stage) and True Negative Rate (TNR) of 0.21 for class 0 (Healthy).

Dataset information

The training dataset is `Train_Set.csv` and the testing dataset is `Test_Set.csv`. In these 2 datasets, there are 350 columns denoting the maximum normalized frequency of DNA fragment lengths and the last column is `class_label`, indicating the cancer stage (healthy /late stage cancer/ screening stage cancer / early stage cancer / mid stage cancer). We converted the `class_label` into `numeric`, namely class 0, 1, 2, 3, 4 respectively.

The train dataset is further split into training data `train` and validation data `val`, allowing us to do validation prior to testing with test dataset `test.`

Requirements

These are the library requirements required for the project to run smoothly.

- `pandas`
- `numpy`
- `matplotlib.pyplot`
- `seaborn`
- `imblearn`
- `sklearn`
- `tensorflow`
- `xgboost`

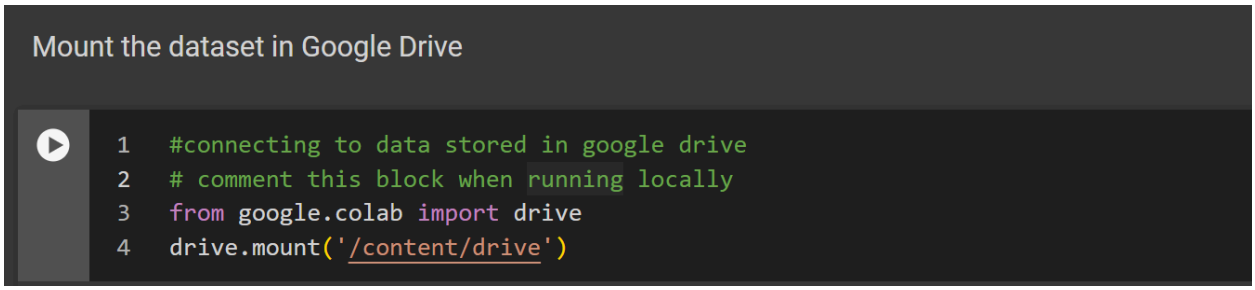
Usage

If using Google Colab,

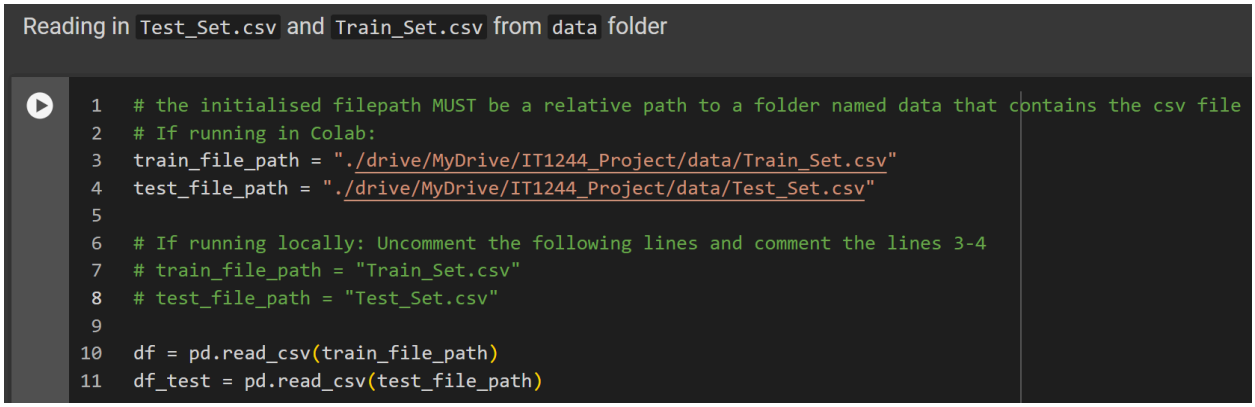
1. Upload `Train_Set.csv` and `Test_Set.csv` into a folder named data, which is in the same level as `bio_cancer_detection.ipynb`. Ensure that both items are in a folder called `IT1244_Project`.
2. Click Runtime > Run all to run the entire file.

If running locally,

1. Install all required packages.
2. Download `bio_cancer_detection.ipynb` from Colab and ensure `Train_Set.csv` and `Test_Set.csv` is in the same folder as the file.
3. Comment this block under Section 1 (Setting up and Reading in Data).



4. Uncomment lines 7 - 8 and comment lines 2 - 3 under Section 1 (Reading in Test_Set.csv and Train_Set.csv from data folder).



5. Use VSCode or Jupyter to run all the code chunks.