First Mid-term Assessments

Introduction to Machine Learning CLP3102 L

Problem

Flying has been the go-to mode of travel for years now; it is time-saving, affordable, and extremely convenient. According to the FAA, 2,781,971 passengers fly every day in the US, as in June 2019. Passengers reckon that flying is very safe, considering strict inspections are conducted and security measures are taken to avoid and/or mitigate any mishappenings. However, there remain a few chances of unfortunate incidents.

Imagine you have been hired by a leading airline. You are required to build Machine Learning models to anticipate and classify the severity of any airplane accident based on past incidents. With this, all airlines, even the entire aviation industry, can predict the severity of airplane accidents caused due to various factors and, correspondingly, have a plan of action to minimize the risk associated with them.

Data

The dataset comprises 3 files:

- **Train.csv:** [10000 x 12 excluding the headers] contains Training data
- **Test.csv:** [2500 x 11 excluding the headers] contains Test data
- **sample_submission.csv:** contains a sample of the format in which the Results.csv needs to be

Data Description:

Columns	Description
Accident_ID	unique id assigned to each row
Accident_Type_Code	the type of accident (factor, not numeric)
Cabin_Temperature	the last recorded temperature before the incident, measured in degrees fahrenheit

Turbulence_In_gforces	the recorded/estimated turbulence experienced during the accident
Control_Metric	an estimation of how much control the pilot had during the incident given the factors at play
Total_Safety_Complaints	number of complaints from mechanics prior to the accident
Days_Since_Inspection	how long the plane went without inspection before the incident
Safety_Score	a measure of how safe the plane was deemed to be
Violations	number of violations that the aircraft received during inspections
Severity	a description (4 level factor) on the severity of the crash **[Target]**

Requirements

- Successfully load the dataset.
- Implement one or two machine learning models such as logistic regression or decision tree and compare their performance.
- If a single model is implemented, tune the hyper-parameters and analyze the performance with different settings to choose the best configuration.
- If two models are implemented (it will give you additional points), compare their performance.
- Calculate the accuracy of the training and test sets.
- Submit the prediction results in a .csv file. If two models were implemented submit two .csv files.
- Write a 1-2 pages report explaining the analysis and work done, including
  - the number of classes found in the test data.
  - If one model was used, compare different settings and explain why the chosen setting produced better results.
  - If two models were used, compare their final performances.
  - Other relevant findings can also be reported.

Submission Format

You need to submit your assignment in a zip file that includes the following files:

1. Code files.
2. Your predictions on test data into a .csv file [Refer sample_submission.csv for format].
3. A short report (a pdf file).

Deadline

Before 28.02.2023, Please send to my email: a.toleu@ kimep.kz