Introduction to Machine Learning Program Assignment #1

This program assignment aims to help you set up your environment for the upcoming assignments and the final project, and also help you understand the basic workflow of machine learning/data mining.

I. Problem

There are two datasets need to be analyzed. For each dataset, you have to do the following:

1. **Visualize the basic statistics** like average, standard deviation and value counts of every feature.
2. **Preprocess the data**, so that it can be used as training data & testing data.
3. Use the data to generate a **decision tree model** with any package you want.
4. Generate decision tree models to construct a **random forest model**. Please note that package-provided random forest model is **NOT** allowed in this step.
5. Use **confusion matrix**, **resubstitution validation** (i.e. identical training and testing set) and **K-fold cross validation** with K > 1 to validate the performance of your models.

II. Dataset

1. Iris dataset

   https://archive.ics.uci.edu/ml/datasets/Iris

   Including 150 number of instances with 4 attributes.

   Attribute Information:

   (1) sepal length in cm
   (2) sepal width in cm
   (3) petal length in cm
   (4) petal width in cm
   (5) class
   - Iris Setosa
   - Iris Versicolour
   - Iris Virginica

2. Google Play Store Apps

   Web scraped data of 10k Play Store apps for analyzing the Android market. You are allowed to do experiments with this dataset for different

data-preprocessing methods or different targets to predict, as long as it makes sense.

This dataset contains two parts:

(1) googleplaystore_user_reviews.csv

- App name
- Translated Review
- Sentiment
- Sentiment Polarity
- Sentiment Subjectivity

(2) googleplaystore.csv

- App
  Application name
- Category
  Category the app belongs to
- Rating
  Overall user rating of the app (as when scraped)
- Reviews
  Number of user reviews for the app (as when scraped)
- Size
  Size of the app (as when scraped)
- Installs
  Number of user downloads/installs for the app (as when scraped)
- Type
  Paid or Free
- Price
  Price of the app (as when scraped)
- Content Rating
  Age group the app is targeted at - Children / Mature 21+ / Adult
- Genres
  An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, Family genres.
- Last Updated
  Date when the app was last updated on Play Store (as when scraped)
- Current Ver

Current version of the app available on Play Store (as when scraped)

- Android Ver

  Min required Android version (as when scraped)

III. Report & Scoring

This is a team-based program assignment, so one team should only submit one report and one source code to E3. The report should contain the following for both datasets:

1. What environments the members are using (5%)
2. Basic statistic visualization of the data (20%)
3. Data preprocessing methods (20%)
4. How you generate decision tree and random forest models (20%)
5. The performance (20%)
6. Conclusion (20%)

There are some rules to follow:

1. Since this assignment is about environment setup, ALL of the members must be able to execute the code and get roughly the same results.
2. **Each member should provide one screenshot of the result in the report.**
3. C / C++ / Java / Python / Matlab are allowed to use. For visualization, Excel or other programs are allowed.
4. Report format should be **PDF**.
5. One team should only make one submission to E3 by one member.
6. No cheating and plagiarizing.