FIT5149 S2 2019 Assessment 1
Bushfire Analysis using Meteorological Data under the Theme of IT for Social Good

Mar-2020

Marks	15% of all marks for the unit
Due Date	17:00 Friday 8 May 2020
Extension	An extension could be granted for circumstances. A special consideration application form must be submitted. Please refer to the university webpage on special consideration.
Lateness	For all assessment items handed in after the official due date, and without an agreed extension, a 10% penalty applies to the student's mark for each day after the due date (including weekends, and public holidays) for up to 5 days. **Assessment items handed in after 5 days will not be considered/marked.**
Authorship	This assignment is **an individual assignment** and the final submission must be identifiable your own work. Breaches of this requirement will result in an assignment not being accepted for assessment and many result in disciplinary action.
Submission	You are required to submit two files, one is either a Jupyter notebook or a R Markdown file, another is the PDF file generated by them. The two files must be submitted via Moodle. Students are required to accepted the terms and conditions in the Moodle submission page. **A draft submission won't be marked.**
Programming language	R in Jupyter Notebook or R Markdown

Figure 1: Australian Bushfire. This picture is from https://www.lonelyplanet.com/articles/bushfires-in-australia-travel

Introduction

In later 2019 and early 2020, Australia faced devastating bushfires started in late 2019, which swiftly got worse before rains helped contain many of the worst fires in February 2020. Bushfires are major environmental issues, creating economical and ecological damages. It is reported that Australia's catastrophic bushfire crisis has destroyed thousands of homes, burned millions hectares of forest, and taken an enormous toll on wildlife. Therefore, fast and automatic detection of bushfires at an early stage is crucial for a successful firefighting.

Traditional human surveillance is expensive and inefficient, which can also be affected by subject factors. With the advances in information technologies, a variety of data about the forest can be collected, such as remote images collected by satellites and meteorological data collected by local sensors. The collected data contains rich information about the status of the forest, the analysis of which can help us detect potential bushfires so as to make effective and efficient firefighting plan and then minimize the damage caused by the bushfires.

In this task, we are interested in exploring machine learning approaches to predict the burned area of bushfires by using meteorological data that are known to influence the wild fires. The dataset that we are going to use here were originally collected from the northeast region of Portugal between January 2000 and December 2003. It contains geographical information, fire weather indices, and the corresponding weather conditions. The aim is to build statistical models that can predict the burned area of the bushfires. Specifically, the problem you are going to solve is: Can you

- accurately **predict** the burned area of a bushfire given the collected data?
- well **explain** your prediction and the associated findings? For example, identify the key factors are strongly associated with the response variable, i.e., the burned area.

Dataset

The dataset contains 517 fire instances, each of which have 13 columns: the first 12 columns corresponding to the attributes (e.g., spatial coordinates, month, day, four fire indices, and other meteorological data) and the last column containing the burned area, i.e., the variable that we will predict. The details of the dataset can be found in the original research paper. The dataset files are stored in UCI's website below (click the hyper-link to download the data)

Forest fires data : There are two files on the website. One called "forestfires.csv" contains the data needed for the analysis, and another called "forestfires.names" contains the information about the dataset.

In order to conduct the analysis task, you should split the provided **forestfires.csv** into your own training dataset (80%) and testing dataset (20%) before building the models.

Task description

In this assessment, you will finish the following two tasks.

Prediction task

For the prediction task, the underlying problem is to predict the burned area of fires using the collected attributes. The provided dataset is well organised and cleaned. It is important that you understand each attribute.

To measure the performance of your model(s), you should firstly split the original data into training and testing datasets, fit the model to the training dataset, perform the prediction on the test dataset and finally compute some performance metrics.

In this task, you are required to develop models that can accurately predict the burned areas. To finish the task, you should

1. develop and compare 2 to 3 models;
2. describe and justify the choice of your models;
3. analyze and interpret your results

Description task

The purpose of the description task is to identify the key factors that have strong effects on the burned areas. In other words, which attribute contributes the most to your model's performance? Descriptions can be based on variable correlation analysis, regression equations, or any other statistical analysis. The descriptions and the accompanying interpretation must be comprehensible, useful and with statistic support whenever it is possible. To finish this task, you should use proper data analysis techniques to

1. identify a subset of attributes that have a significant impact on the prediction of the burned area;
2. report your identification with statistical evidence (e.g. correlations, p-values) and interpret the identified attribute subset (e.g. as to why certain attributes have certain impacts on the prediction).

Files to be submitted

There are two files required to be submitted, which are

- The **R** implementation of the tasks in one file.
  - The file **must be** either a **Jupyter notebook** or **an R Markdown file**. Besides the R code, all the discussions must also be included in the file.
  - The name of the file **must be** in one of the following formats:
    * XXXXXXXX_FIT5149_Ass1.ipynb
    * XXXXXXXX_FIT5149_Ass1.Rmd

    You should replace "XXXXXXXX" with your student ID.
- A PDF file generated by the Jupyter notebook or R Markdown. The name of the PDF file must be in the following format
  - XXXXXXXX_FIT5149_Ass1.pdf

  Before you generated the PDF file, **please clear all the outputs**. Please note that the PDF file will be used by Turnitin for the purpose of plagiarism check. It is your full responsibility to make sure that all the outputs are cleared before the PDF file is generated, as the outputs can contribute significantly to the Turnitin scores.

Please refer to the Assessment 1's Moodle page for how to submit the two files and note that **If you do not follow the above way to name your submission, your submission will not be marked and will receive 0 mark directly**.

Additional learning resources

This assessment is based on the paper "*A Data Mining Approach to Predict Forest Fires using Meteorological data*" at http://www3.dsi.uminho.pt/pcortez/fires.pdf.

There are some existing analyses using the same set of data, such as

- http://www.columbia.edu/~yh2693/ForestFire.html
- https://www.kaggle.com/elikplim/predict-the-burned-area-of-forest-fires

for your reference only. However, please be aware of University's policy on academic integrity. **Monash University takes academic misconduct very seriously. You can learn from the above materials and understand the principle of how the analysis was done. However, you must finish this assessment with your own work.**