

Programming Assignment 02

Assignment:

You are the junior member of a data mining consulting firm. Your newest client has collected data on forest fires in an area of interest and wants to do some analysis on that data, however they need some direction on what to do with their data. They have provided to you a data set on forest fires they have collected in the form of a comma separated value file and would like your firm to do some exploratory data analysis on the data. You have been given this client and have been asked to provide a report on their data and make recommendations on what can be done with the data.

Start with an overall discussion of the data set. For example, get a count of how many examples, and how many attributes the data set has.

Then analyze the attributes.

For **each** attribute do the following:

1. Determine if the attribute is Nominal, Ordinal, Interval, or Ratio data.
2. Visualize the data for each attribute.
3. Determine the possible values or range of the values for each attribute.
4. Determine if there are any data values that we might be concerned about and state why they are of concern. If/when concerns are discovered, suggest what can be done to address those concerns.

When you have completed evaluating each attribute, determine if there is a relationship between any of the attributes.

Document all findings in a report.

Partition the data set into one training data set and one test data set. Take both data sets and compare **three** of the attributes using visualizations, and where appropriate statistics, to show that the attributes from the training set have similar characteristics to the same attribute in the test set. When selecting **three** attributes; one attribute must be either Nominal or Ordinal data with no data issues, one attribute must be Interval or Ratio with no data issues, and one attribute must have data issue.

Also, in your results, give some suggestion on

- what variables the client might consider using,
- what variable the client might consider adding or creating that is not in the data set already,
- which attributes they might want to consider using as a predictor (independent) variable, and
- which attribute they might want to consider using as a predicted (dependent) variable.

Data:

Each example represents a known forest fire within a specified area of interest.

The following information on the attributes of the data set has been provided:

coord_X - x-axis spatial coordinate within a topographical map of the area of interest.

coord_Y - y-axis spatial coordinate within a topographical map of the area of interest.

month - the month in which the forest fire happened

day - the day of the week in which the forest fire happened

FFMC - Fine Fuel Moisture Code from the Fire Weather Index (FWI) System

DMC - Duff Moisture Code from the Fire Weather Index (FWI) System

DC - Draught Code from the Fire Weather Index (FWI) System

ISI - Initial Spread Index from the Fire Weather Index (FWI) System

temp - temperature in Celsius degrees

RH - relative humidity in %

wind - wind speed in km/h

rain - outside rain in mm/m²

area - the burned area of the forest in hectares