CST8390 - Business Intelligence & Data Analytic

Assignment1

By

Wentao Lu (040991469)

Thien An Dao (040902993)

April 12th, 2021

# Table of Content

[**Table of Content**](#_94hdbsn7zl3i) **1**

[**Data Analysis**](#_82505hd8acer) **2**

[1.1. Heart failure clinical records Data set](#_ks75jqvh6nxl) 2

[1.2. Algerian forest fire dataset](#_1sgyvkczahph) 2

[1.3. Data selection](#_gtvnhieszvg4) 2

[**Overall data**](#_imiiygugisy1) **3**

[**Attribute analysis**](#_5mier613i04h) **4**

[3.1. Wind speed & Rain (The cooling factor):](#_w2743dbhlfgl) 4

[2.2. Temperature & Humidity:](#_ssehq35ha54d) 5

[2.3. Humidity & Rain:](#_dxq0gb9euzd) 6

[**K - Nearest Neighbors Testing**](#_7gfdn4nxllm5) **6**

[**Applying 5NN to an instance to make the prediction.**](#_prj1y86j3hsa) **10**

# 

# 

# Data Analysis

## 1.1. Heart failure clinical records Data set

“Heart failure clinical records Data set” has shown datas that has potential lead to heart failure. These data types were recorded from 299 patients who had heart failure. The data set contains 13 attributes, six of which are categoricals and seven numericals. All six categorical attributes all are nominals: anaemia, high blood pressure, diabetes, sex, smoking, death event. On the other hand, numericals have both discrete and continuous data types. In discrete data type, there are age and time; in continuous data type, there are creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium. Nominal data types only have two values: male & female or yes & no. In discrete data types, patient age’s values cover from 40 to 95 and follow-up period (time/days) cover from 4 to 285. All five attributes in continuous data represent the amount of CPK enzyme, platelet, serum creatinine, serum sodium in the patient and the percentage of blood leaving the heart at each contraction.

## 1.2. Algerian forest fire dataset

“Algerian forest fire dataset” shows the forest fire-related data of two regions of Algeria, namely the Bejaia region located in the northeast of Algeria and the Sidi Bel-abbes region located in the northwest of Algeria. There are in total 122 instances for each of the regions, and the period is from June 2012 to September 2012. The dataset records day, month, and year and has 11 numerical attributes and one output attribute as classes. The numerical attributes record the max temperature in celsius(22 to 42), relative humidity( 21 to 90 in %), wind speed( 6 to 29 in km/h ), total rain( 0 to 16.8 in mm ), fine fuel moisture code(28.6 to 92.5), duff moisture code(1.1 to 65.9), drought code(7 to 220.4), initial spread index(0 to 18.5), buildup index(1.1 to 68) and fire weather index(0 to 31.1). The class attribute is showing two classes: fire or not fire. In total, there are 138 fire classes and 106 no-fire classes. The data type of all the numerical data is continuous.

## 1.3. Data selection

After discussion, the team agreed to choose the “Algerian forest fire dataset” as the topic to be analyzed. Global warming has been a major global issue and it already has a massive impact to not only to humans but also the wildlife. By studying the data of the “Algerian forest fire dataset”, we can identify the attributes that have the possibility to create forest fire or the attributes that can help prevent disaster. Tree is one of the most important living entities on the planet. Plants not only absorb carbon dioxide and produce oxygen, they can prevent terrain terraformation, create sound reduction, prevent flood, filter toxic in the air & many more benefits from trees. Also, forest is the home of all wildlife, losing forest can drive many to every creature from endangered to extinction. Last but not least, burning a tree not only takes away many benefits we can gain but also creates a carbon footprint. Carbon footprint has a huge negative indirect effect toward the ozone layer.

Since the dataset is built up by two sets of data representing different regions, we are taking the first set which represents the data of the Bejaia Region from June to September.

# Overall data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute** | **Minimum** | **Maximum** | **Mean** | **StdDev** |
| Temperature | 22 | 37 | 31.18 | 3.32 |
| relative humidity (RH) | 45 | 89 | 67.975 | 11.154 |
| Wind speed(Ws) | 11 | 26 | 16 | 2.849 |
| Rain | 0 | 16.8 | 0.843 | 2.409 |
| Fine fuel moisture code (FFMC) | 28.6 | 90.3 | 74.673 | 15.559 |
| Duff moisture code(DMC) | 0.7 | 54.2 | 12.315 | 11.274 |
| Drought code (DC) | 6.9 | 220.4 | 53.161 | 51.778 |
| Initial spread index (ISI) | 0 | 12.5 | 3.656 | 3.022 |
| Buildup index (BUI) | 1.1 | 67.4 | 15.426 | 14.474 |
| Fire weather index (FWI) | 0 | 30.2 | 5.578 | 6.343 |

# Attribute analysis

## 3.1. Wind speed & Rain (The cooling factor):

In this chart, the data has shown that the increase of wind speed and rain can reduce the chance of creating forest fire. Most of the forest fire cases have 0 mm of rain and the highest rain measurement for the fire case is 0.3. For wind speed, most of the fire cases are between 13 to 19 km/h. However, the highest Wind speed that has a fire case is 21 km/h, any case that has Wind speed higher than 21 km/h doesn’t involve forest fire. So, higher Wind speed and Rain have a lower chance to cost fire.

## 2.2. Temperature & Humidity:

In this chart, the data displayed Temperature and Relative humidity of the forest which have related factors with each other. Based on the chart, the higher the temperature the lower humidity in the air and the lower the temperature the higher the humidity. The fire cases concentrate between 49% to 68% in Humidity and 30 degree Celsius to 37 degree Celsius. This proves that higher Relative humidity and the lower temperature, the forest is less likely to have a fire case.

## 2.3. Humidity & Rain:

In this chart, the data displayed the relativity between Rain and Relative humidity of the forest. Between 48% to 78% of humidity and 0 mm to 0.3 mm there are no significant differences and most of the fire cases are in those ranges. There are only a few instances that the Rain measurement has a significant difference when the humidity is equal or higher than 78%. So, most of the fire cases happen when there is little or no rain and the humidity under 78% doesn’t have a major effect on the forest fire.

# K - Nearest Neighbors Testing

Correctly classified instances:

|  |  |
| --- | --- |
| K | Percentage of correctly classified instances |
| 3 | 91.8033 % |
| 5 | 94.2623% |
| 7 | 93.4426 % |
| 9 | 95.082 % |
| 11 | 93.4426 % |

|  |  |  |  |
| --- | --- | --- | --- |
| K | TPR | FPR | Number of Misclassifications |
| 3 | 0.873(not fire)  0.966(fire) | 0.034(not fire)  0.127(fire) | 10(a:8 b:2) |
| 9 | 0.937(not fire)  0.966(fire) | 0.034(not fire)  0.063(fire) | 6(a:4 b:2) |

Correctly classified instances:

|  |  |
| --- | --- |
| K | Percentage of correctly classified instances |
| 3 | 86.4865% |
| 5 | 89.1892% |
| 7 | 89.1892% |
| 9 | 89.1892% |
| 11 | 89.1892% |

|  |  |  |  |
| --- | --- | --- | --- |
| K | TPR | FPR | Number of Misclassifications |
| 3 | 0.833(not fire)  0.895(fire) | 0.105(not fire)  0.167(fire) | 5 (a:3 b:2) |
| 5/7/9 | 0.889(not fire)  0.895(fire) | 0.105(not fire)  0.111(fire) | 4 (a:2 b:2) |
| 11 | 0.833(not fire)  0.947(fire) | 0.053(not fire)  0.167(fire) | 4(a:3 b:1) |

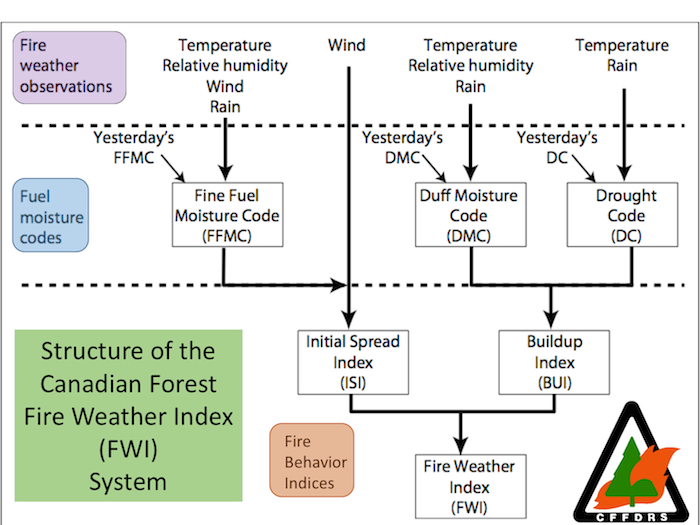
Correctly classified instances:

|  |  |
| --- | --- |
| K | Percentage of correctly classified instances |
| 3 | 94.5946% |
| 5 | 89.1892% |
| 7 | 89.1892% |
| 9 | 94.5946% |
| 11 | 97.2973% |

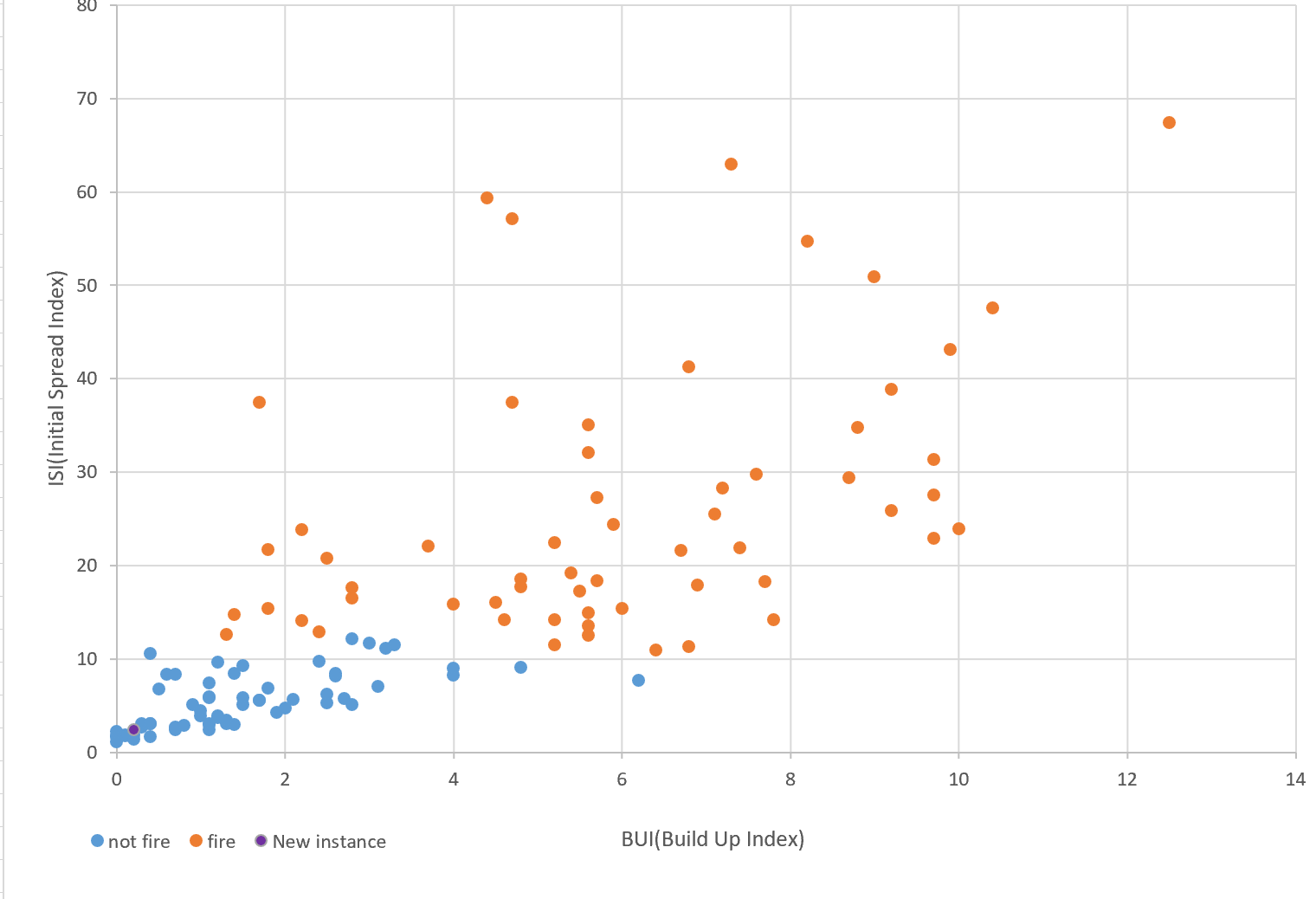
|  |  |  |  |
| --- | --- | --- | --- |
| K | TPR | FPR | Number of Misclassifications |
| 11 | 0.944(not fire)  1.000(fire) | 0.000(not fire)  0.056(fire) | 5 (a:1 b:0) |
| 5/7 | 0.778(not fire)  1.000(fire) | 0.000(not fire)  0.222(fire) | 4 (a:4 b:0) |
| 11 | 0.833(not fire)  0.947(fire) | 0.053(not fire)  0.167(fire) | 4(a:3 b:1) |

# 5. Applying 5NN to an instance to make the prediction.

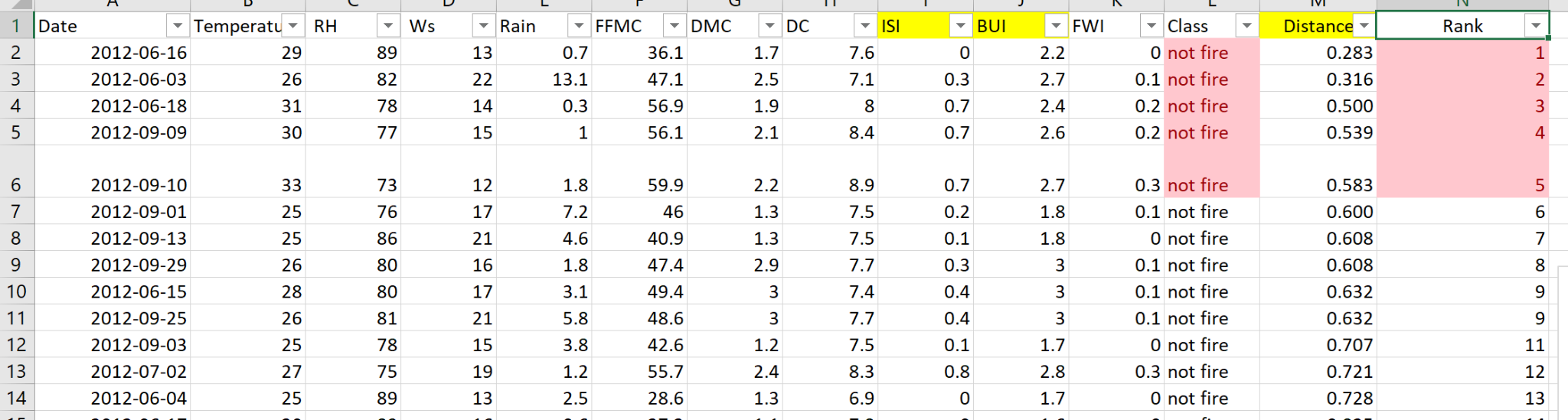
We took the very last instance as a test instance and we took Initial Spread Index and Buildup Index as our attributes according to FWI Process Flow Chart from Fire Weather Index(FWI) System[1]



The other attributes eventually lead to ISI and BUI, and then, FWI. Thus we took ISI and BUI to make a two-dimensional calculation and predict the result. The scatter plot of BUI and ISI bellow shows the distribution of fire days according to these two attributes.The blue dots mean not fire and the orange dots mean fire. The purple dot is the instance we put in and try to predict the result.



After calculating the distance of the instance dot to all the other dots, we ranked them and sorted them from smallest to largest and got the following results:



The first 5 instances, known as 5NN are all “no fire” classes. Thus, we predict that the instance should be “not fire”.

Reference list

[1]“Fire Weather Index (FWI) System,” *NWCG*, 10-Mar-2021. [Online]. Available: https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system. [Accessed: 12-Jun-2021].