Author: Cong Liu

Date: 11/24/2019

**0. Get The Data**

There are 517 rows and 13 columns in the dataset.

**2. Business Understanding**

What is the most serious and overlooked problem of our time? We think it's the environment.

A major environmental problem is the occurrence of forest fires (also known as wildfires), which damage and destroy the forest environment, causing economic and ecological damage to humans, as well as physical and mental damage. Although people have been trying to control the fire, the effect is not ideal. We believe that the reason why the consequences of fire accidents are so serious is that people cannot make a good judgment and predict the time and place of fire. In order to help with the problem of fire, we decided to make a data mining and analysis report related to the problem of fire-prone areas.

We chose "Forest Fires Dataset" from the UCI Machine Learning Repository2;

We use this dataset because we found that Portugal is highly affected by forest fires, thus, the dataset is suitable for our analysis. The dataset we will use covers meteorological and spatiotemporal data for forest fires in Portugal’s Montesinho Natural Park between 2000 and 2003, with 13 attributes. Our projects going to build a model to predict the burned area of forest fires and to estimate the fire case level for different attributes values by cleaning, analyzing and predicting the dataset. In addition, another reason we do this analysis is because we want to use other factors to predict the areas where fires are likely to occur and the severity of the fires

We plan to divide the fire prone area into three levels (small, median and large) and then use it as the label. Thus, the data mining approach will be classification and the models will include Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Support Vector Machine, Multi-layer Perceptron, Decision tree, Random Forest and K Nearest Neighbors. The software will be Scikit-Learn.

**3. Data Understanding and Preprocessing**

Data Source:

Source Citation: [Cortez and Morais, 2007] P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at: [Web Link]

Source URL: UCI Machine Learning Repository2 - Forest Fires Data Set

Retrieved from:

<https://archive.ics.uci.edu/ml/datasets/Forest+Fires>

Attribute Information:

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9

2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9

3. month - month of the year: 'jan' to 'dec'

4. day - day of the week: 'mon' to 'sun'

5. FFMC - FFMC index from the FWI system: 18.7 to 96.20

6. DMC - DMC index from the FWI system: 1.1 to 291.3

7. DC - DC index from the FWI system: 7.9 to 860.6

8. ISI - ISI index from the FWI system: 0.0 to 56.10

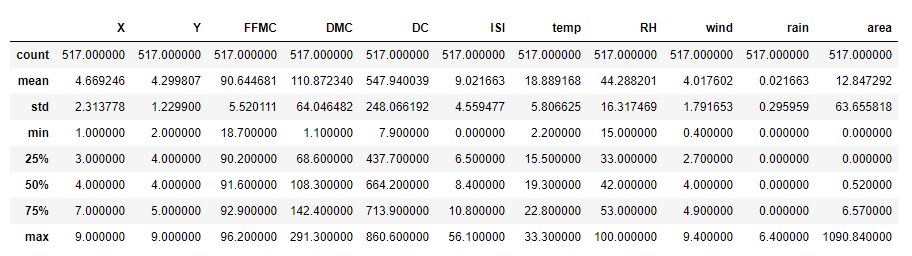
9. temp - temperature in Celsius degrees: 2.2 to 33.30

10. RH - relative humidity in %: 15.0 to 100

11. wind - wind speed in km/h: 0.40 to 9.40

12. rain - outside rain in mm/m2 : 0.0 to 6.4

13. area - the burned area of the forest (in ha): 0.00 to 1090.84

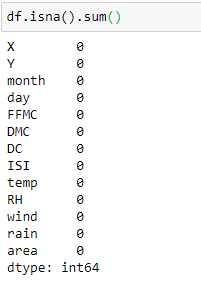


Total count of all variables is 517;

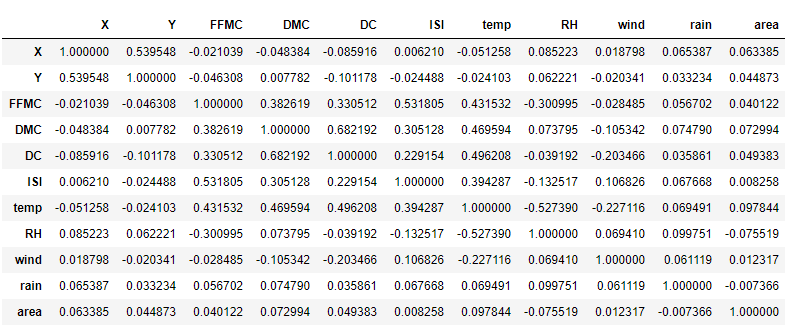
The range of attribute X is between 1 and 9, median is 4, mean is 4.669246 with a standard deviation 2.313778.

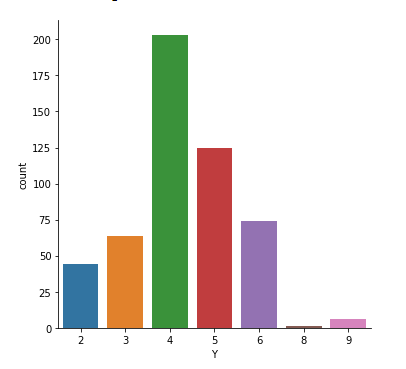
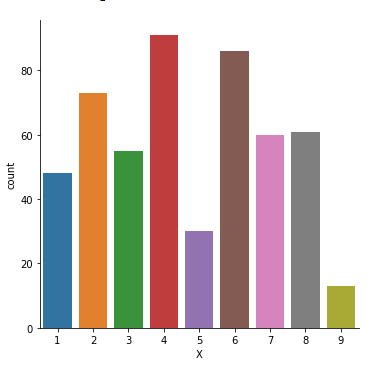
The range of Y is between 1 and 9, median is 4, mean is 4.299807 with a standard deviation 1.2299.

Month and Day are not numerical attributes;

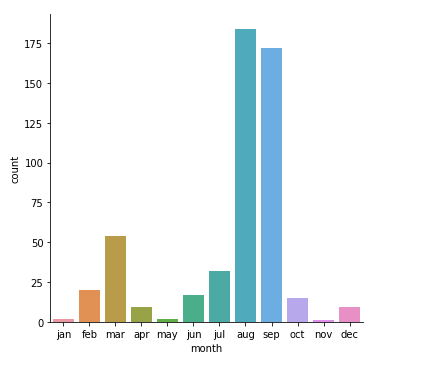


There are not any missing values in the dataset.

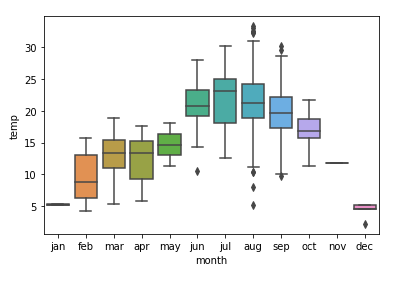




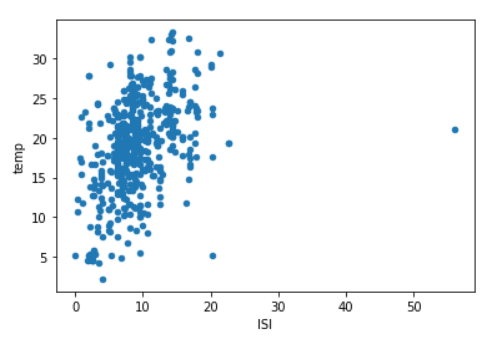
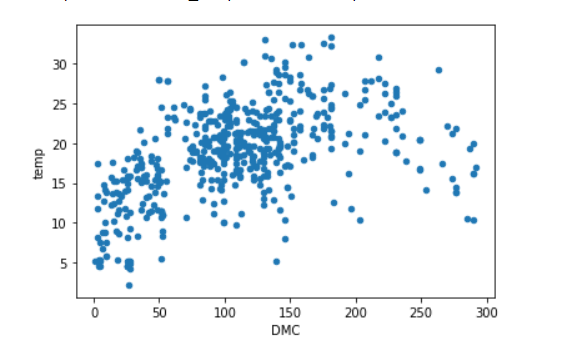
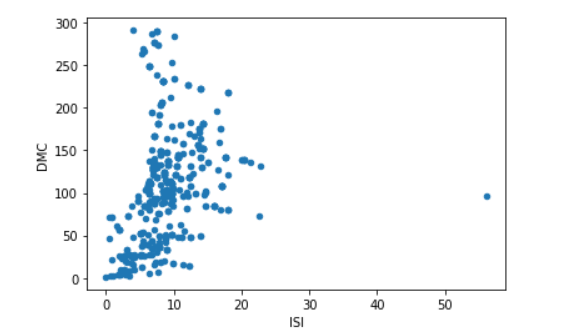
Base on the plot graph above, we can find that the fire case always happened in Y=4,5, 6 and X = 1,2,3,4,6,7,8.



Based on the plot graph above, we can see that the fire cases are obviously appeared in August and September. Then we want to know whether other factors in these two months are different from other months;



Now we find that there is a strong relationship between month and temperature; Thus, we want to know if there are any other attributes can be implied by temp;

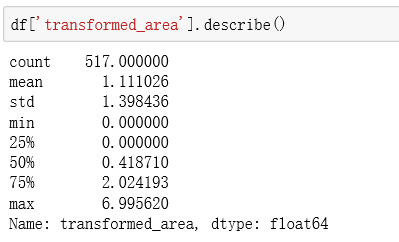
 

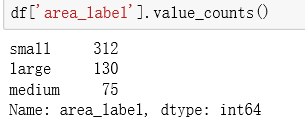
Based on these graphs above, we can find that there exists strong relationship between DMC vs. temp, ISI vs. temp, and DMC vs ISI; Thus, we decide to use temp imply these attributes and delete them from the dataset.

Also, X and Y are unique and not related to our analysis and thus, we delete them from dataset.

Data preprocessing:

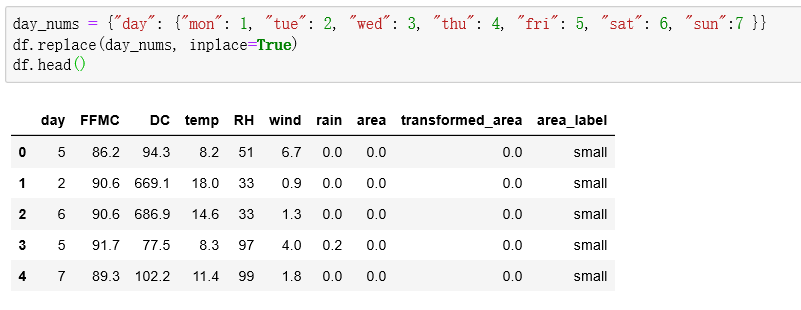
Firstly, we did a logarithm transform for the area attribute because its values are very skewed towards 0.0. Since our goal is to predict the degree of the burned area of the forest fires, we need to classify the continuous values of area into different class labels (small, medium, and large) to determine the levels of the forest fires. Based on the summary statistics of the transformed area, we decide to classify the values less than 1.111 as small, the values greater than 1.111 but less than 2.024 as medium, and the values greater than 2.024 as large. Then, we can see there are 312 observations being classified as small, 130 observations being classified as medium and 75 observations being classified as large.





Next, since the day attribute is categorical, we decide to convert and replace them to numeric values.

The data frame after transforming is shown below.



**4. Modeling and Assessment**

Our goal is to build a model to predict the burned area of the forest fires (the attribute "area" in the original dataset) and we have classified the forest fire into three levels (small, medium, large), thus the data mining approach is classification. Our models include Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Support Vector Machine, Multi-layer Perceptron, Decision tree, Random Forest and K Nearest Neighbors. We will choose the model that gives the best performance on the dataset.

The evaluation approach we will use is K-fold cross-validation and the K value we choose is 10. The reason why we choose cross-validation instead of train/test split is because it can give us more accurate estimate of accuracy and we can make more "efficient" use of data since every observation is used for both training and testing.

For each classifier, we will use 10-fold cross-validation to split the train and test sets and then calculate the training and testing accuracy scores to evaluate the performance of the classifier. By comparing their accuracy scores, we will choose a best classifier among them. The software we will use is the Python package scikit-learn because it provides us a convenient use of all the seven classifiers mentioned above.

The accuracy results are shown below.

**Gaussian Naïve Bayes**

The 10-fold cross-validation training results for Gaussian Naïve Bayes are:

[0.59611231 0.59698276 0.59569892 0.60215054 0.59871245 0.5944206

0.59656652 0.60300429 0.59656652 0.25321888]

The average training accuracy is: 0.563

The 10-fold cross-validation testing results for Gaussian Naïve Bayes are:

[0.57407407 0.60377358 0.59615385 0.59615385 0.60784314 0.58823529

0.58823529 0.60784314 0.58823529 0.29411765]

The average testing accuracy is: 0.564

**Bernoulli Naïve Bayes**

The 10-fold cross-validation training results for Bernoulli Naïve Bayes are:

[0.60475162 0.60344828 0.60430108 0.60430108 0.60300429 0.60300429

0.60300429 0.60300429 0.60300429 0.60300429]

The average training accuracy is: 0.603

The 10-fold cross-validation testing results for Bernoulli Naïve Bayes are:

[0.59259259 0.60377358 0.59615385 0.59615385 0.60784314 0.60784314

0.60784314 0.60784314 0.60784314 0.60784314]

The average testing accuracy is: 0.604

**Support Vector Machine**

The 10-fold cross-validation training results for Support Vector Machine are:

[0.22030238 0.61422414 0.59569892 0.31612903 0.57296137 0.42703863

0.60300429 0.61373391 0.56223176 0.59871245]

The avearge training accuracy is: 0.512

The 10-fold cross-validation testing results for Support Vector Machine are:

[0.2037037 0.60377358 0.55769231 0.36538462 0.64705882 0.52941176

0.60784314 0.60784314 0.54901961 0.60784314]

The avearge testing accuracy is: 0.528

**Multi-layer Perceptron**

The 10-fold cross-validation training results for Multi-layer Perceptron are:

[0.52699784 0.34267241 0.61505376 0.62580645 0.49356223 0.60944206

0.6223176 0.62446352 0.54506438 0.58583691]

The average training accuracy is: 0.559

The 10-fold cross-validation testing results for Multi-layer Perceptron are:

[0.37037037 0.24528302 0.57692308 0.59615385 0.49019608 0.60784314

0.56862745 0.58823529 0.60784314 0.60784314]

The average testing accuracy is: 0.526

**Decision Tree**

The 10-fold cross-validation training results for Decision Tree are:

[0.96760259 0.96982759 0.97204301 0.96344086 0.96781116 0.96137339

0.96781116 0.972103 0.96566524 0.96137339]

The average training accuracy is: 0.967

The 10-fold cross-validation testing results for Decision Tree are:

[0.38888889 0.32075472 0.46153846 0.48076923 0.43137255 0.37254902

0.29411765 0.21568627 0.39215686 0.29411765]

The average testing accuracy is: 0.365

**Random Forest**

The 10-fold cross-validation training results for Random Forest are:

[0.96760259 0.96982759 0.97204301 0.96344086 0.96781116 0.96137339

0.96781116 0.972103 0.96566524 0.96137339]

The average training accuracy is: 0.967

The 10-fold cross-validation testing results for Random Forest are:

[0.53703704 0.58490566 0.51923077 0.5 0.58823529 0.49019608

0.45098039 0.2745098 0.37254902 0.49019608]

The average testing accuracy is: 0.481

**K Nearest Neighbors**

The 10-fold cross-validation training results for K Nearest Neighbors are:

[0.6349892 0.64224138 0.66666667 0.64086022 0.64592275 0.66309013

0.65450644 0.65879828 0.6695279 0.6695279 ]

The average training accuracy is: 0.655

The 10-fold cross-validation testing results for K Nearest Neighbors are:

[0.40740741 0.47169811 0.46153846 0.46153846 0.43137255 0.37254902

0.35294118 0.31372549 0.41176471 0.39215686]

The average testing accuracy is: 0.408

Based on the average training accuracy scores of the seven classifiers, we can see both the Random Forest and Decision Tree fit the training data very well and have the highest average training accuracy, which is 0.967. However, if we combine the training and testing accuracy scores and compare, we can find many classifiers suffer the problem of overfitting with a higher training accuracy but a lower testing accuracy. And the overfitting is most severe on the Random Forest and Decision Tree, with testing accuracy scores 0.481 and 0.365 respectively. Overall, we find the performance of the Bernoulli Naïve Bayes is best, with a 0.603 training accuracy and a 0.604 testing accuracy. Thus, we decide to choose the Bernoulli Naïve Bayes Classifier as our final model.

**5. Conclusion**

In conclusion, we believe the work we have done so far has achieved the goal to build a model to predict the burned area of forest fires somehow. It is worth noting that the authors in the original study stated that the SVM model can predict better. Compared to their results, in our lab, we find that the SVM model does have a decent performance, but the Gaussian Naïve Bayes and Bernoulli Naïve Bayes model can outperform it a little. However, to be honest, we wouldn’t say this result is what we expected or satisfactory since a 0.604 accuracy score on the testing data is still not good enough. It’s interesting how severe the overfitting problem is for the Decision Tree and the Random Forest model. We may want to do more feature selection or hyperparameter tuning on them in the future. Also, we think there may be many other classifiers that we haven’t covered working better and we will try to model with them.

During this project, we had some struggles on transforming the continuous attribute “area” into categorial values and determining a threshold for the three different level. Later we decided to use the mean as threshold for small, the median as threshold for medium and the third quantile as threshold for large.