Forest Fire Report

Advanced Data Analysis Group-2

Bhavesh Kumar Chigullapalli
Computer Science Engineering
Indian Institute Of Information Technology
Sricity, Chittoor
bhaveshkumar.c19@iiits.in

Navdeep Konkipudi
Computer Science Engineering
Indian Institute Of Information Technology
Sricity, Chittoor
navdeep.k19@iiits.in

***Abstract*—This document is a report on forest fire data analysis. The dataset includes 244 instances that regroup a data of two regions of Algeria. Analysis and prediction is done on the forest fire in Bejaia and Sidi-Bel Abbes regions of Algeria. Achieved F1-score 0.96% on both Bejaia and Abbes region Datasets.**

***Index Terms*—Data Preprocessing, Exploratory Data Analysis (EDA), Linear Discriminant Analysis, Hypothesis testing, Hotelling-T^2, Simultaneous Confidence Interval, Inference Test, Logistic Regression**

I. Introduction

The dataset used in this document is algerian forest fire dataset which included the data from two regions in algeria - Sidi-Bel Abbes and Bejaia. The data was from the year 2012 June, July, August. The attributes include - Temperature, Relative Humidity, Rainfall (in mm), Windspeed and FWI components and the records were classified into two classes namely - fire and not fire. This document involves EDA on the Algerian forest fire Dataset, followed by visualizing of the data and hypothesis testing using Hotelling-T^2, Simultaneous Confidence Intervals(both fire & no fire case), LDA, Classification using Logistic Regression and Inference testing of the two regions.

II. Methodology

A. Data Preprocessing

The dataset included 244 instances, of which 122 belonged to Bejaia Region and other 122 were of Sidi-BEl Abbes Region. Initially the data was read and split into two dataframes namely bejaia dataframe and Abbes Dataframe. Then Data Analysis was performed on each of them individually. Data Cleaning was performed on the raw data such as finding/filling missing values, removing whitespaces in attributes, label encoding categorical attribute. It was found that bejaia dataset was a balanced dataset. Where as Abbes dataset was an imbalanced one, so F1 score was used as evalutaion metric for abbes.

B. Exploratory Data Analysis

The data was grouped by month and forest fires were counted. It was found that forest fires were irrelevant of month. hence month attribute was removed from the data. Then Correlation analysis was performed on the data and few attributes which had high correlation with others were removed. We plotted each attribute against each other attribute using pair plot. Also it was found in the pair plot that the distribution was normal in both classes. The Categorical attribute Classes was Label Encoded. Box plots were also used to get information on each attribute.

C. Simultaneous Confidence Intervals & Hypothesis Testing

- The data was separated based on the classes and simultaneous confidence interval of each attribute was found out on each case with a confidence of 90%.
- In hypothesis Testing, the Null Hypothesis was - In a region with mean [Temp, RH, Rain] = [33, 64, 0.015] a fire will start 90% of the time. This was proven to be right on Bejaia region's dataset and was rejected on Abbes region's dataset.

D. Linear Discriminant Analysis

Applied LDA on the data and reduced the feature vectors from 6 dimensions to 1 dimension. Visualized the result of LDA of both classes and observed that the distributions were highly separable with minimal overlap. This indicates that classification algorithm might perform very well on this data.

E. Classification

Data was split into 80% Train and 20% test data and logistic regression was performed. The classification algorithm was performed both before and after performing LDA. A classification Report and ROC curve were plotted.

F. Inference regarding two populations

An Assumption was made that the two populations from which abbes data and bejaia data were extracted have same covariance matrix.An hypothesis was made that both populations have same mean. Spool value, Crtical value was calculated with a confidence of 90% and Hotelling-T^2 was compared with the critical value. We reject the hypothesis but found out that With a 0.01% confidence the hypothesis was tested to be true which gave rise to a very slim possibility that both the region's data is from the same population.

III. Results

Algeria Regions	Evaluation Metrics		
	Precision	***Recall***	***f1-score***
Bejaia	0.97	0.95	0.96
Bejaia-LDA	0.92	0.92	0.92
Abbes	0.95	0.97	0.96
Abbes-LDA	0.97	0.95	0.96

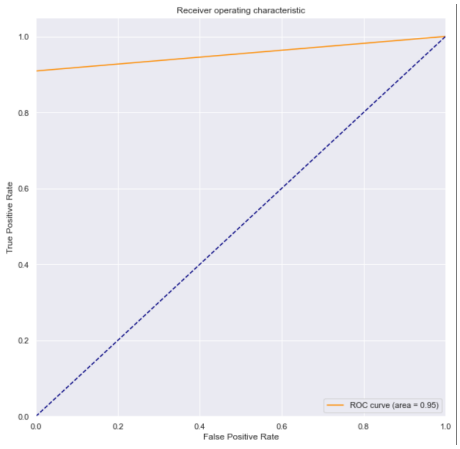

Fig. 1. ROC-AUC Curve on Bejaia Data (score-0.97).

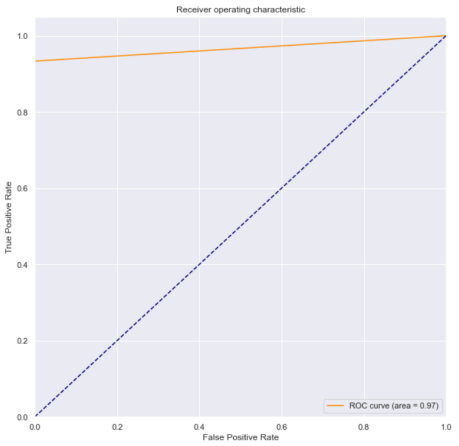

Fig. 2. ROC-AUC Curve on Abbes Data. (score-0.95)

IV. Conclusion

There were multiple inferences that were made from the dataset, one being that the wind speed attribute had no effect in the classification. This can be first observed in the correlation matrix where wind speed had very less correlation near to 0 with other attributes and can also be observed when the simultaneous confidence intervals were found, wind speed's interval remained almost same in both the case, that is, when there was a fire and when there was no fire. Also we have observed several attributes that had very high correlation with each other and hence we have removed such attributes to reduce redundancy. Simultaneous confidence interval and hypothesis testing gave us an idea of where the attribute values lie. Logistic Regression model performed well and provided good results in classification. High precision, recall and accuracy were observed in both regions both in training and testing phase removing the possibility of an overfit. Inference regarding two populations has made it clear that there is a very slim possibility(probability of 0.01%) that both region's data is from same population. And at 90% confidence we can reject the hypothesis that both the data is from same population.

References

[1] Faroudja ABID et al. , â€œPredicting Forest Fire in Algeria using Data Mining Techniques: Case Study of the Decision Tree Algorithmâ€., International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD 2019) , 08 - 11 July , 2019, Marrakech, Morocco.