**Report I239 Machine Learning**

**Name: Sakshi Agarwal**

**Student ID: 1710402**

***Name of the Dataset: Blogger.csv***

***Problem:*** There are five attributes in this dataset named Degree, Political Caprice, Topics, Local Media Turnover (LMT), and Local, Political and Social Space (LPSS). These attributes are the features of bloggers. We use them to predict the whether the blogger is Professional Blogger (Pb) or not.

***Data Set Characteristics:***

| **Data Set Characteristics** | **Multivariate** |
| --- | --- |
| Tasks | Classification |
| Number of Attributes | 6 |
| Attribute Type | Categorical |
| Missing Value | No |
| Number of Instances | 100 |

***About the Attributes:***

Degree: This attribute shows the education level of a blogger and has three values; -1: for less educated people, 0: for B.Sc. level, and 1: for M.Sc. and Ph.D.

Political Caprice (caprice): It shows political affiliation of a blogger.

It has also three values; -1: Bloggers affiliated with reformists party, 0: bloggers who have no political affiliations, and 1: bloggers affiliated with a traditional conservative party.

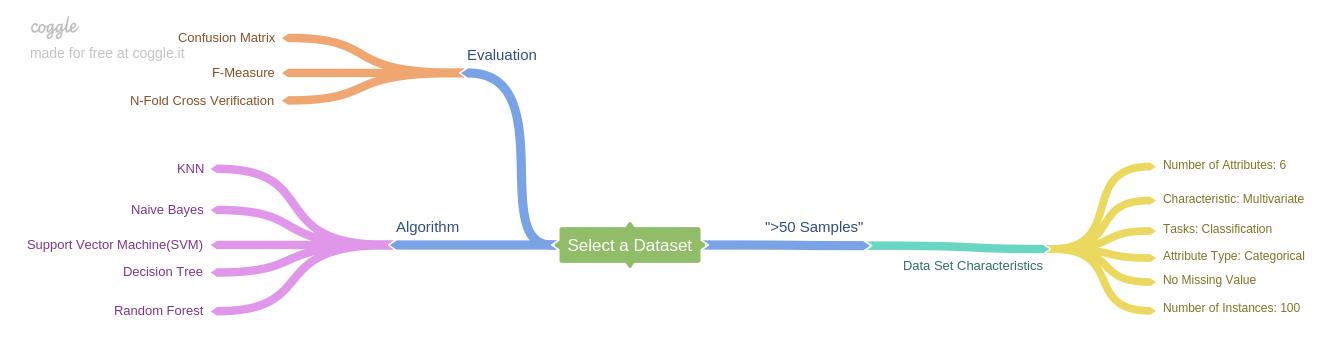
Topics: Bloggers are intended to blog in five areas named Impression (1), Political (3), Tourism (5), Scientific (4), and News (2).

LMT: This attribute has two possible values: Yes/ No. “Yes (1)” value is used for those bloggers who believe in local media effect on blogging and vice-versa (0).

LPSS: This attribute has also Yes/No values. “Yes (1)” value is used to denote those bloggers who believe that local, social and political conditions affect blogging and vice-versa (0).

Pb: It is target class with Yes/No values. “Yes (1)” value denotes a professional blogger, and “No (0)” value indicates a temporary blogger.

***MindMap:***



***Algorithms:***

We have run following algorithms on the dataset:

1. Naive Bayes
2. Decision Tree
3. SVM
4. KNN
5. Random Forest

Data distribution:

Train Data: 80%

Test Data: 20%

Random state = 42

***Evaluation Results:***

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms/Evaluation Measure | Confusion Matrix | F-Measure | Cross Value Score (Accuracy) |
| Naïve Bayes |  | 0.7142857142857142 | 0.68 (+/- 0.14) |
| Decision Tree |  | 0.8 | 0.88 (+/- 0.14) |
| SVM |  | 0.7000000000000001 | 0.75 (+/- 0.12) |
| KNN |  | 0.6 | 0.75 (+/- 0.12) |
| Random Forest |  | 0.58 | 0.78 (+/- 0.05) |

***Discussion and Conclusion:***

Decision Tree works best in this case as per all the three evaluation measures used. It is the fastest way to identify most significant variables and relation between two or more variables. Decision Trees are excellent tools for helping you to choose between several courses of action.