Week4 Assessment:
Conceptual Understanding

Website: www.analytixlabs.co.in
Email: info@analytixlabsl.co.in

Disclaimer: This material is protected under copyright act AnalytixLabs©, 2011-2018. Unauthorized use and/ or duplication of this material or any part of this material including data, in any form without explicit and written permission from AnalytixLabs is strictly prohibited. Any violation of this copyright will attract legal actions.

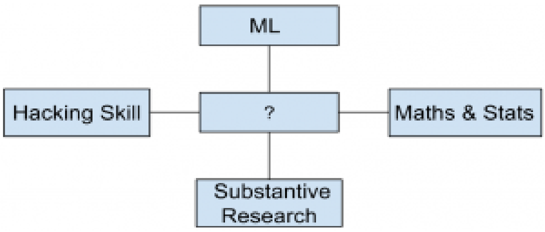
Marks weightage for different type of questions:

True/False Questions: 1 Mark
Objective Type Questions: 2 Marks
Descriptive Type Questions: 4 Marks

Note: More than one option can be correct for objective type questions. You need to provide all possible correct options if applicable.

QUESTIONS

Q1. Which of the following would be more appropriate to be replaced with question mark in the following figure?



a) Data Analysis
b) Data Science
c) Descriptive Analytics
d) None of the mentioned

Q2. Which of the following is performed by Data Scientist?
a) Define the question
b) Create reproducible code
c) Challenge results
d) All of the Mentioned

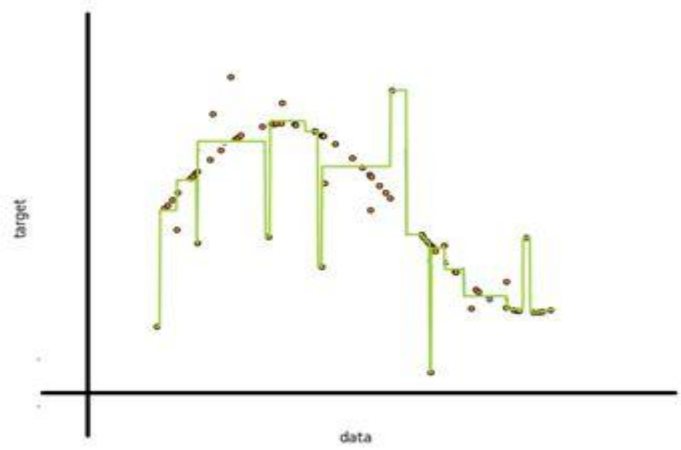
Q3. The data scientists at "Walmart" have collected 2014 sales data for 3000 products across 100 stores in different cities. Also, certain attributes of each product based on these attributes and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store during a defined period.
Which learning problem does this belong to?

a) Supervised learning
b) Unsupervised learning
c) Reinforcement learning
d) None

Q4. Point out the wrong statement:
a) Merging concerns combining datasets on the same observations to produce a result with more variables
b) Data visualization is the organization of information according to preset specifications
c) Subsetting can be used to select and exclude variables and observations
d) All of the Mentioned

Q5. Which of the following approach should be used to ask Data Analysis question?
a) Find only one solution for particular problem
b) Find out the question which is to be answered
c) Find out answer from dataset without asking question
d) None of the mentioned

Q6. Look at the below image: The red dots represent original data input, while the green line is the resultant model.



How do you propose to make this model better while working with decision tree?

A. Let it be. The model is general enough
B. Set the number of nodes in the tree beforehand so that it does not overdo its task
C. Build a decision tree model, use cross validation method to tune tree parameters
D. Both B and C
E. All A, B and C
F. None of these

Q7. Which Algorithms are used to do a Binary classification (Provide at least 5 algorithms)?

Q8. Random Forest has 1000 trees, Training error: 0.0 and validation error is 20.00.What is the issue here?

Q9. What is logit function?

Q10. Which Algorithms are used to do a multinomial classification (provide at least 5 algorithms?

Q11. What are all the Error metrics for Regression problem (provide at least 5 metrics)?

Q12. Give three ways to identify Outliers?

Q13. Give three ways handle outliers?

Q14. What are the key assumption for Naive Bayes?

Q15. Let's say you're building the recommended music engine at Spotify to recommend people music based on past listening history. How would you approach?

Q16. What is the curse of dimensionality?

Q17. How can you ensure that you don't analyze something that ends up meaningless?

Q18. How can you determine which features are the most important in your model (provide at least 10 methods)?

Q19. Let’s say you’re given an unfeasible amount of predictors in a predictive modeling task. What are some ways to make the prediction more feasible?

Q20. You have data on all purchases of customers at a grocery store. Describe to me how you would program an algorithm that would cluster the customers into groups. How would you determine the appropriate number of clusters include?

Q21. What is the difference between population and sample in data?

Q22. You have several variables that are positively correlated with your response, and you think combining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the possible issues?

Q23. What is the difference between univariate, bivariate and multivariate analysis other than the use of number variables used as part of analysis?

Q24. What is confusion matrix?

Q25. How will you detect the presence of overfitting?

Q26. How do you determine the number of clusters in k-means clustering (Provide at least 3 methods)?

Q27. What are the different ways of performing aggregation in python using pandas (Provide at least 3 methods)?

Q28. How to solve overfitting (provide at least 3 methods)?

Q29. Name few libraries that is used in python for data analysis/data science (provide at least 20 libraries)?

Q30. How do you deal with some of your predictors being missing (Provide at least 5 Methods)?

Q31. What is Imbalanced Data Set and how to handle them? Name Few Examples (at least 3 examples)?

Q32. Provide at least 5 algorithms for Supervised/Unsupervised learning algorithm?

Q33. When does multicolinierity problem occur and how to handle it? (Provide at least 3 methods to find multicolinierity presence in the data)?

Q34. Provide at least 5 Examples of Parametric machine learning algorithm and non-parametric machine learning algorithm

Q35. In Google if you type “How are “it gives you the recommendation as “How are you “/”How do you do”, this is based on what?

Q36. There are 24 predictors in a dataset. You build 2 models on the dataset:
1. Bagged decision trees and
2. Random forest
Let the number of predictors used at a single split in bagged decision tree is A and Random Forest is B. Which of the following statement is correct?

A. A >= B
B. A < B
C. A >> B
D. Cannot be said since different iterations use different numbers of predictors

Q37. Why do we prefer information gain over accuracy when splitting while building decision tree ?

A. Decision Tree is prone to overfit and accuracy doesn't help to generalize
B. Information gain is more stable as compared to accuracy
C. Information gain chooses more impactful features closer to root
D. All of these

Q38. Random forests (While solving a regression problem) have the higher variance of predicted result in comparison to Boosted Trees (Assumption: both Random Forest and Boosted Tree are fully optimized).

A. True
B. False
C. Cannot be determined

Q39. Assume everything else remains same, which of the following is the right statement about the predictions from decision tree in comparison with predictions from Random Forest?

A. Lower Variance, Lower Bias
B. Lower Variance, Higher Bias
C. Higher Variance, Higher Bias
D. Lower Bias, Higher Variance

Q40. What are the different activation functions in neural network (provide at least 5 activation functions)?

Q41. You are given a imbalanced data set on fraud detection. Classification model achieved accuracy of 95%.Is it good? (if answer is 'No", Provide at least 5 alternative methods to check model is good or not)

Q42. What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?

Q43. What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

Q44. Which Algorithm Suits for Text Classification Problem (provide at least 5 algorithms)?

Q45. Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?Q46. Which of the following tree based algorithm uses some parallel (full or partial) implementation?

a) Random Forest
b) Gradient Boosted Trees
c) XGBOOST
d) Both A and C
e) A, B and C

Q47. Point out the wrong statement:
a) ROC curve stands for receiver operating characteristic
b) Foretime series, data must be in chunks
c) Random sampling must be done with replacement
d) None of the Mentioned

Q48. Which of the following is a categorical outcome?
a) RMSE
b) RSquared
c) Accuracy
d) All of the Mentioned

Q49. Which of the following input can be accepted by Pandas DataFrame?
a) Structured ndarray
b) Series
c) DataFrame
d) All of the Mentioned

Q50. Point out the wrong statement:
a) Training and testing data must be processed in different way
b) Test transformation would mostly be imperfect
c) The first goal is statistical and second is data compression in PCA
d) All of the Mentioned

Q51. Point out the correct statement:
a) Combining classifiers improves interpretability
b) Combining classifiers reduces accuracy
c) Combining classifiers improves accuracy
d) All of the Mentioned

Q52. Predictive analytics is same as forecasting.
a) True
b) False

Q53. Which of the following is correct with respect to residuals?
a) Positive residuals are above the line, negative residuals are below
b) Positive residuals are below the line, negative residuals are above
c) Positive residuals and negative residuals are below the line
d) All of the Mentioned

Q54. Which of the following shows correct relative order of importance?
a) question->features->data->algorithms
b) question->data->features->algorithms
c) algorithms->data->features->question
d) none of the Mentioned

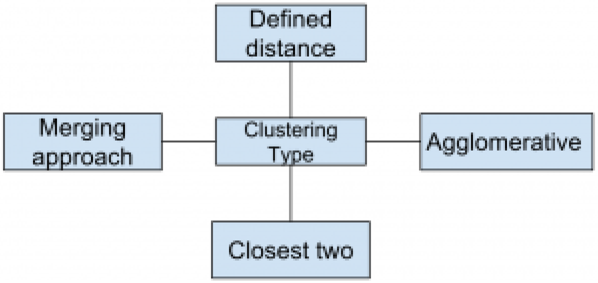
Q55. Which of the following is characteristic of best machine learning method?
a) Fast
b) Accuracy
c) Scalable
d) All of the Mentioned

Q56. Which of the following trade-off occurs during prediction?
a) Speed vs Accuracy
b) Simplicity vs Accuracy
c) Scalability vs Accuracy
d) None of the Mentioned

Q57. Which of the following random variable that take on only a countable number of possibilities?
a) Discrete
b) Non Discrete
c) Continuous
d) All of the Mentioned

Q58. Statistical inference is the process of drawing formal conclusions from data.
a) True
b) False

Q59. Which of the following clustering type has characteristic shown in the below figure?



a) Partitional
b) Hierarchical
c) Naive Bayes
d) None of the Mentioned

Q60. Merge function is used for merging data frames.
a) True
b) False

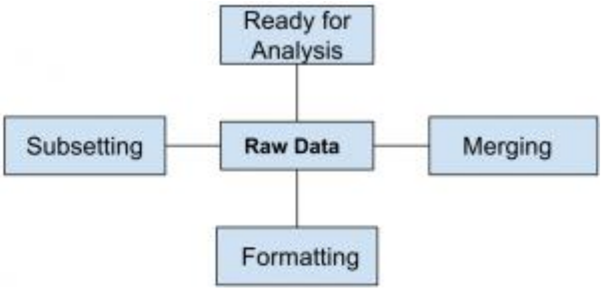
Q61. Regular expressions can be thought of as combination of literals and metacharacters.
a) True
b) False

Q62. Which of the following is used for machine learning in python?
a) scikit-learn
b) seaborn-learn
c) stats-learn
d) none of the Mentioned

Q63. Point out the correct statement:
a) Statsmodels provides powerful statistics, econometrics, analysis and modeling functionality that is out of pandas' scope
b) Vintage leverages pandas objects as the underlying data container for computation
c) Bokeh is a Python interactive visualization library for small datasets
d) All of the Mentioned

Q64. What is the role of processing code in the research pipeline?
a) Transforms the analytical results into figures and tables
b) Transforms the analytic data into measured data
c) Transforms the measured data into analytic data
d) All of the Mentioned

Q65. Which of the following block information is odd man out?

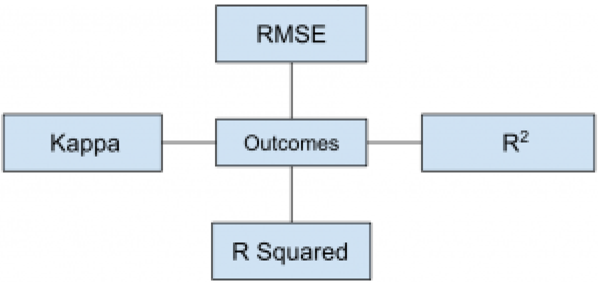


a) Subsetting
b) Raw Data
c) Ready for Analysis
d) None of the mentioned

Q66. Which of the following is required by K-means clustering?
a) defined distance metric
b) number of clusters
c) initial guess as to cluster centroids
d) all of the Mentioned

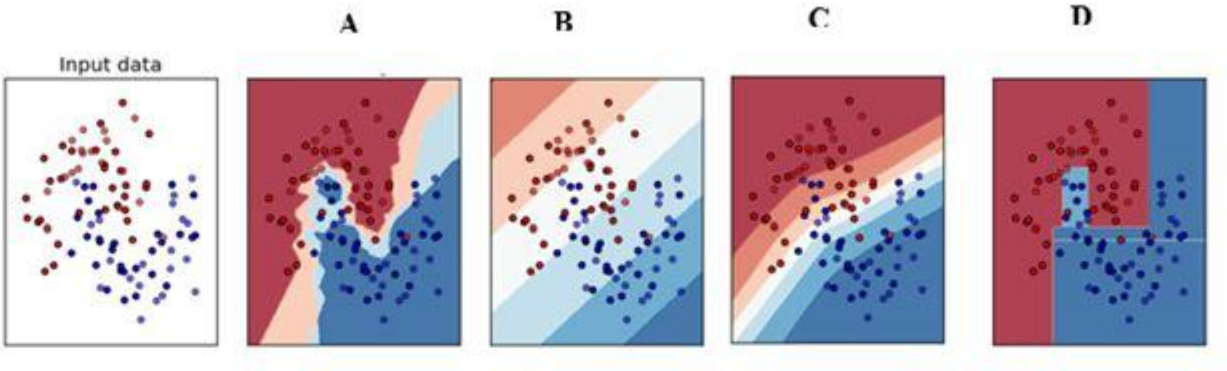
Q67. Residuals are useful for investigating best model fit.
a) True
b) False

Q68. Which of the following outcome is odd man out in the below figure?



a) R Squared
b) Kappa
c) RMSE
d) All of the Mentioned

Q69. Which of the following is a decision boundary of Decision Tree?



a. B
b. A
c. D
d. C
e. Can't Say

Q70. Let's say we have m numbers of estimators (trees) in a boosted tree. Now, how many intermediate trees will work on modified version (OR weighted) of data set?

a. 1
b. m-1
c. m
d. Can't say
e. None of the above

Q71. Given 1000 observations, Minimum observation required to split a node equals to 200 and minimum leaf size equals to 300 then what could be the maximum depth of a decision tree?

a. 1
b. 2
c. 3
d. 4
e. 5

Q72. Generally, in terms of prediction performance which of the following arrangements are correct:

a. Bagging>Boosting>Random Forest>Single Tree
b. Boosting>Random Forest>Single Tree>Bagging
c. Boosting>Random Forest>Bagging>Single Tree
d. Boosting >Bagging>Random Forest>Single Tree

Q73. In which of the following application(s), a tree based algorithm can be applied successfully?

a. Recognizing moving hand gestures in real time
b. Predicting next move in a chess game
c. Predicting sales values of a company based on their past sales
d. A and B
e. A, B, and C

Q74. Suppose we have missing values in our data. Which of the following method(s) can help us to deal with missing values while building a decision tree?

a. Let it be. Decision Trees are not affected by missing values
b. Fill dummy value in place of missing, such as -1
c. Impute missing value with mean/median
d. All of these

Q75. To reduce under fitting of a Random Forest model, which of the following method can be used?

a. Increase minimum sample leaf value
b. increase depth of trees
c. Increase the value of minimum samples to split
d. None of these

Q76. While creating a Decision Tree, can we reuse a feature to split a node?

a. Yes
b. No

Q77. Decision Trees are not affected by multicolinierity in features:

a. TRUE
b. FALSE

Q78. For parameter tuning in a boosting algorithm, which of the following search strategies may give best tuned model:

a. Random Search.
b. Grid Search.
c. A or B
d. Can't say

Q79. In Random Forest, which of the following is randomly selected?

a. Number of decision trees
b. features to be taken into account when building a tree
c. samples to be given to train individual tree in a forest
d. B and C
e. A, B and C

Q80. Which of the following are the disadvantage of Decision Tree algorithm?

a. Decision tree is not easy to interpret
b. Decision tree is not a very stable algorithm
c. Decision Tree will over fit the data easily if it perfectly memorizes it
d. Both B and C

Q81. Boosting is a general approach that can be applied to many statistical learning methods for regression or classification.

a. True
b. False

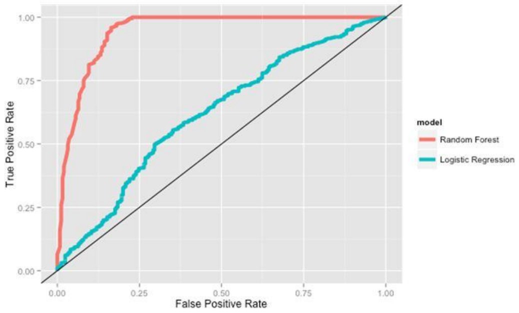
Q82. Predictions of individual trees of bagged decision trees have lower correlation in comparison to individual trees of random forest.

a. TRUE
b. FALSE

Q83. Which of the following is example of raw data?
a) original swath files generated from a sonar system
b) initial time-series file of temperature values
c) a real-time GPS-encoded navigation file
d) All of the Mentioned

Q84. Which of the following algorithm would you take into the consideration in your final model building on the basis of performance?

Suppose you have given the following graph which shows the ROC curve for two different classification algorithms such as Random Forest(Red) and Logistic Regression(Blue)
A) Random Forest
B) Logistic Regression
C) Both of the above
D) None of these



Q85. In random forest or gradient boosting algorithms, features can be of any type. For example, it can be a continuous feature or a categorical feature. Which of the following option is true when you consider these types of features?

A) Only Random forest algorithm handles real valued attributes by discretizing them

B) Only Gradient boosting algorithm handles real valued attributes by discretizing them

C) Both algorithms can handle real valued attributes by discretizing them

D) None of these

Q86. Which of the following algorithm are not an example of ensemble learning algorithm?

A) Random Forest

B) Adaboost

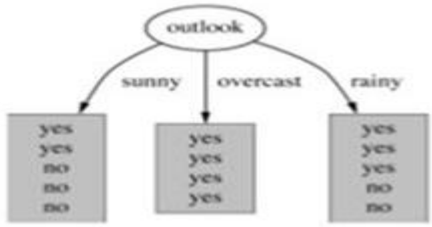
C) Extra Trees

D) Gradient Boosting

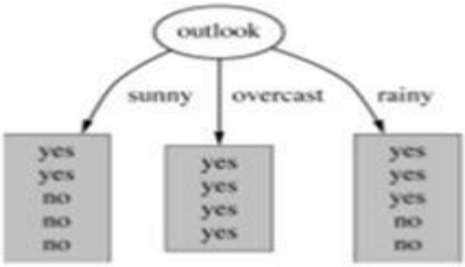
E) Decision Trees

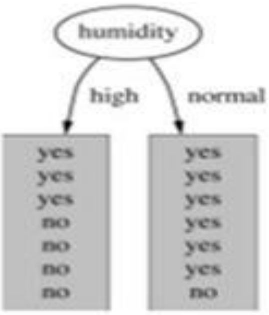
Q87. How is kNN different from kmeans clustering?

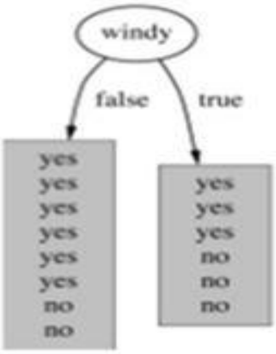
Q88. Calculate Gini, Entropy and Information gain for below tree at first level?

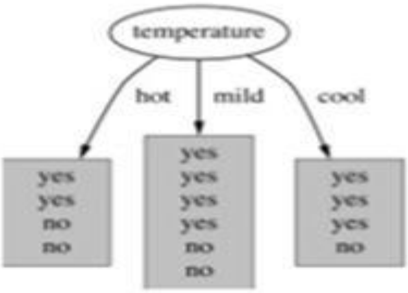


Q89. Suppose you are building random forest model, which split a node on the attribute that has highest information gain. In the below image, select the attribute which has the highest information gain?







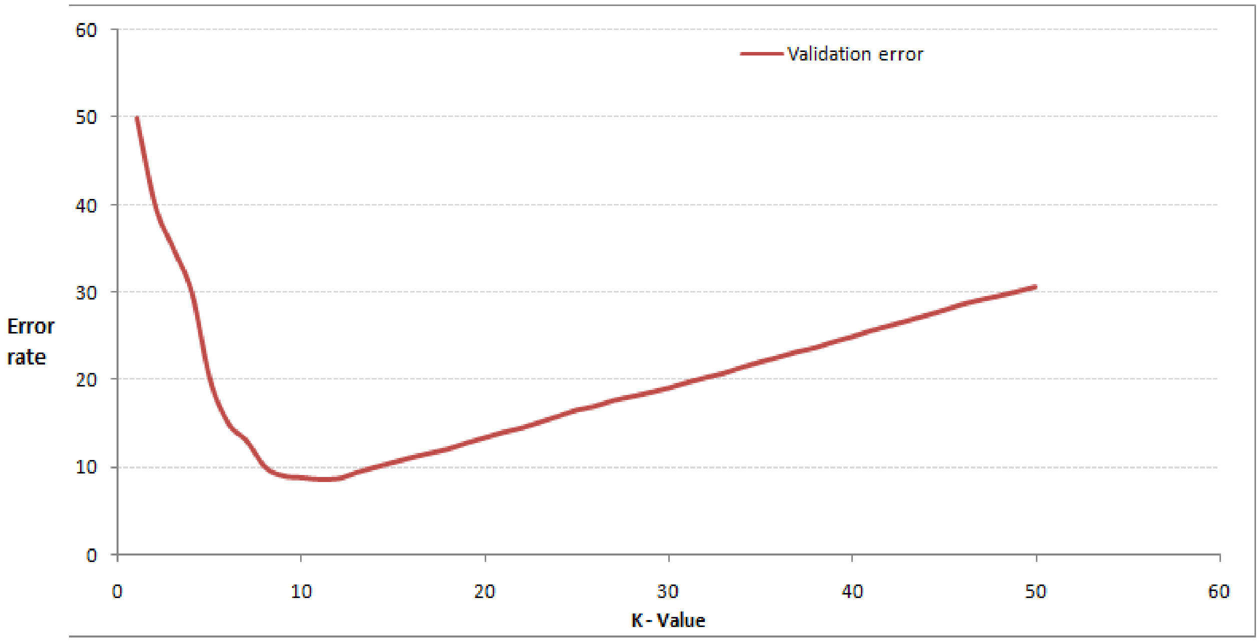


A) Outlook
B) Humidity
C) Windy
D) Temperature

Q90. The bagging is suitable for high variance low bias models?
A) TRUE
B) FALSE

Q91. k-NN algorithm does more computation on test time rather than train time.
A) TRUE
B) FALSE

Q92. In the image below, which would be the best value for k assuming that the algorithm you are using is k-Nearest Neighbor.



A) 3
B) 10
C) 20
D 50

Q93. Which of the following distance metric can not be used in k-NN?
A) Manhattan
B) Minkowski
C) Tanimoto
D) Jaccard
E) Mahalanobis
F) All can be used

Q94. Which of the following option is true about k-NN algorithm?
A) It can be used for classification
B) It can be used for regression
C) It can be used in both classification and regression

Q95. Which of the following statement is true about k-NN algorithm?

1. k-NN performs much better if all of the data have the same scale
2. k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large
3. k-NN makes no assumptions about the functional form of the problem being solved

A) 1 and 2
B) 1 and 3
C) Only 1
D) All of the above

Q96. Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?
A) K-NN
B) Linear Regression
C) Logistic Regression

Q97. Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?
A) 1
B) 2
C) 4
D) 8

Q98. Which of the following will be true about k in k-NN in terms of Bias?
A) When you increase the k the bias will be increases
B) When you decrease the k the bias will be increases
C) Can't say
D) None of these

Q99. Provide at least 10 packages related to text mining in Python

Q100. Provide at least 10 different types of analysis in text Mining