

Machine Learning CA02

(10515513)

Submitted by-Prasad Tambe(10515513)

Submitted to- Abhishek

Contents

[Q1 Definations 3](#_Toc37891041)

[Q1 Define the following (4\*3 marks) ● Define Data pre-processing and its steps with examples. (200 words). 3](#_Toc37891042)

[● Define Decision tree, Information gain and Entropy. (200 words) 4](#_Toc37891043)

[● Define Chinese restaurant algorithm with example. (200 words) 4](#_Toc37891044)

[Report 5](#_Toc37891045)

[Regression: 6](#_Toc37891046)

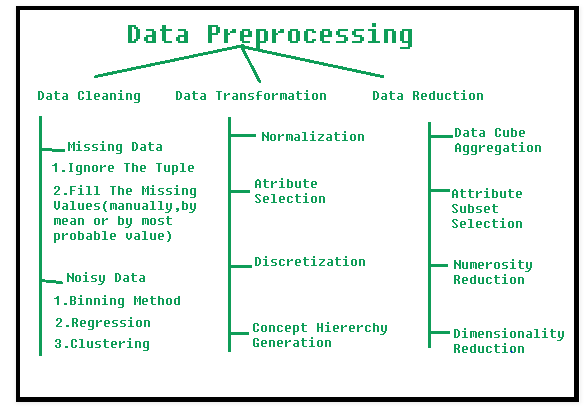
[**Classification:** 9](#_Toc37891047)

[Reference 11](#_Toc37891048)

# Q1 Definations

# Q1 Define the following (4\*3 marks) ● Define Data pre-processing and its steps with examples. (200 words).

Data Pre-processing: pre-processing of data is a method in data Machine Learning, used to turn raw data into an effective and usable format.



Data Pre-processing Steps Involved: 1. Cleaning of data: The data may contain several incomplete and unrelated components.

[a]. Lost data: If any details in the data are incomplete, this condition occurs. It can be handled in different ways.

[b]. Noisy Data: Noisy data is useless, not interpretable by computers. It may be made because of inaccurate data processing, mistakes in data analysis, etc. The following form should be handled:

 This system functions through ordered data to smooth it.

Binning Process: The data are divided into similarly large fragments and then the process is performed by various methods. Every is independently stored.

The regression can be linear (which has one independent variable) or multiple (having many independent variables). This regression may make the data smooth.

Clustering: The clustering of the related results. The outliers cannot be identified or dropped off the clusters.

2. Data transformation: This step is taken to convert data into appropriate types of mining. It covers the following ways:

Normalization: to scale data values within a given range (-1.0 to 1.0 or 0.0 to 1.0) Collection of attributes: new attributes are created from the set of attributes for the mining process.

3. Data reduction: Data mining is a technique used for processing vast volumes of data. In such cases, interpretation becomes more difficult when dealing with large quantities of data.

Reduction in dimension: this decreases the data size by encoding mechanisms. Unless the initial data can be recovered after the restoration of compressed files, the lossless reduction is called then the loss reduction.

# ● Define Decision tree, Information gain and Entropy. (200 words)

The decision tree for classification and prediction is the most effective and common method in the world. A Decision Tree is a flowchart like the tree structure in which every internal node is an attribute test, every branch is a test result, and each leaf node has the class symbol.

Information Gained: The sum of data can be defined as (1 – entropy) in a collection of data. If any findings about the data given are made, new information can be retrieved.

Decision trees can generate understandable rules. Decision trees can handle both continuous and categorical variables. Decision trees provide a clear indication of which fields are most important for prediction or classification. (T, 2019)

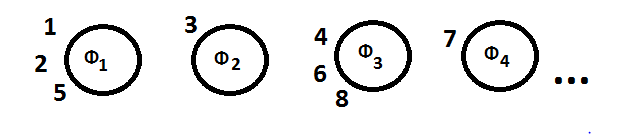
"Information gain" is the contrast between the two meanings. In other words, the probability change is the information the measurement generates. When you partition a T set into T1 and T0, based on certain data characteristics, this partition can be described as an information gain.

Entropy is a measure of the randomness of the information produced in relation to machine learning. The greater entropy, the more difficult it becomes to draw conclusions. Flipping a coin is an example of a measure which gives random information.

# ● Define Chinese restaurant algorithm with example. (200 words)

A easy way of describing the Chinese method is by classifying individuals according to a given set of possibilities used by the algorithm into an hypothetical Chinese restaurant with infinite tables. The algorithm then models the number and randomization or probabilistic dimension of the processes at Chinese restaurants can be seen mathematically at every table of which tables are the "partitions."

Customers 1, 3, 4 and 7 are sat on empty tables in the above case, customers 2, 5, 6 and 8 are sat at the table already in place. The restaurant's number of tables is limitless.



Consumer 1 will sit anywhere he wants. If consumer 1. Table 1: 1 (1) + (α) New Table (i.e. every empty seat): α (1 + α) The probability on which customer 9 will be seated shall be as follows: Seat 1: 3 / (8 + α) Table 2: 1 / (8 + α) Table 3:3 / (8 + α) Table 4: 1 / (8 + α) Table 1:

The numerator is the number of people already sat at a particular table, and the denominator is the number of customers in the restaurant (i – 1) plus α (a positive scalar [hyperparameter](https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)), which is set before the process starts).

* The probability of the i th customer sitting at an existing table is nk / (α + i – 1),
* The probability of the i th customer sitting at a new table is α / (α + i – 1).

As more people sit at a table, those tables increase in popularity, so new patrons are less likely to sit at empty tables. (Statistics How To, 2018)

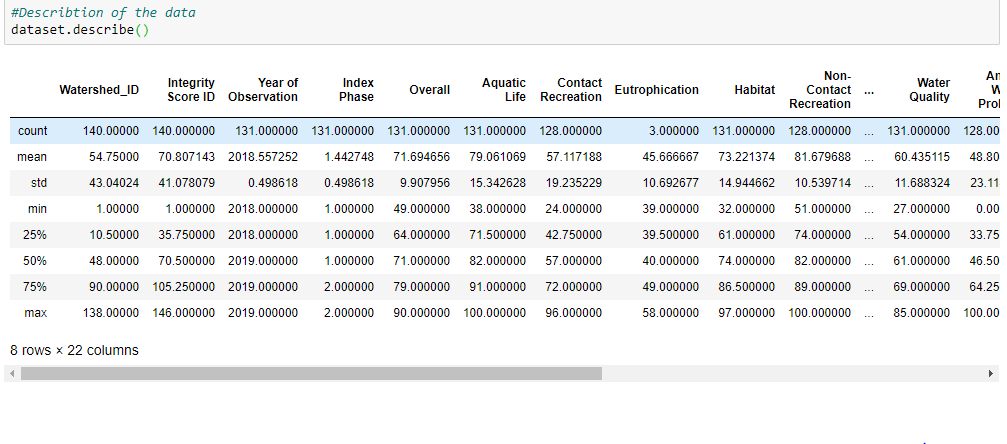
# Report

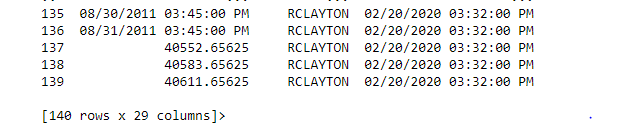
**Dataset and Explanation:**

We have chosen the data set from Data.gov which is related to Watershed Reach Index and Problem Scores which is been used in order to perform the regression model. The data is about the Track and assess environmental quality and degree of degradation in watersheds in Austin(USA).

I have chosen this dataset because it looks quite interesting how the Overall score can affect the different features i.e. Habitat, Aquatic life etc. of Austin.

Now, Understanding our dataset.





The dataset is having 140 rows and 29 columns in total with some missing values in it. So in order to clean the data which comes in the pre processing part we need to clean the data and remove all the null values/missing values in the dataset.

The dataset is having 29 features which decides the Track and assess environmental quality and degree of degradation in watersheds which can impact the over performance score in watersheds.

**Technique:**

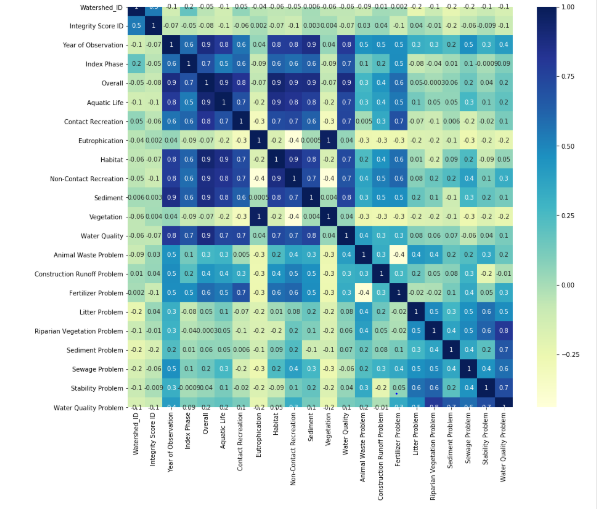
# Regression:

Regression model is used in this case because the Overall feature is playing an important role in deciding what factors can affect the other features depending in watersheds. That is the reason I chose to do a regression model on this dataset.

In order to predict the overall performance the overall feature is used in this case against the remaining 28 features.

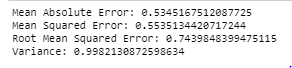
I have used 3 Different types of regression in this regression model.

1)Linear Regression 2) Gradient Boosted tree 3) Random Forest.

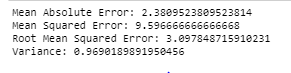


The heatmap in the above figure shows that the data is not highly correlated and gives the best result in the regression techniques which are used on the dataset. Because the data is giving different types of correlations which can be used to predict the overall performance in our case.

Here we can compare the results of the error rates and RMSE for each cases.

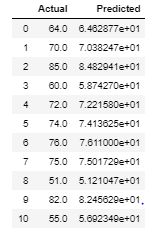
 Linear Regression Vs

Gradient Boosted Tree Regression Vs

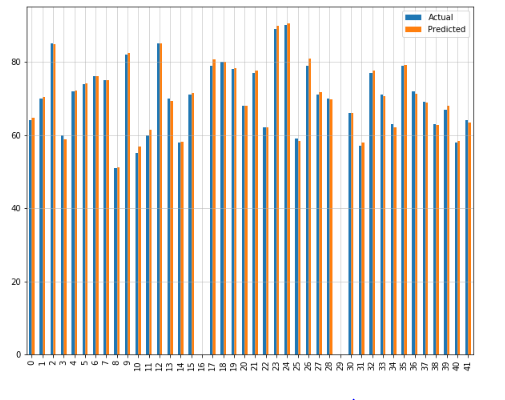
Random Forest Regression

Here we can see that Linear regression gave us the best results, but also the Gradient Boosted Tree Regression and Random Forest gave us the best performance but not as compared to the Linear Regression Technique.

Further we can see the Actual vs predicted table in the below figure.



The below graph shows the actual values vs the predicted values with respect to Overall Performance of the watersheds.

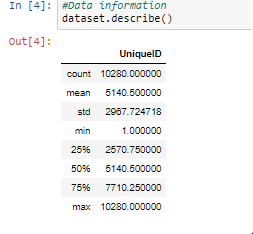


Here, we can conclude that the actual values and the predicted values are quite nearly similar because of the data which is not highly correlated, and the predictions can be seen in the graph above. The overall Performance can be dependent on the variables which we extracted from the feature’s selection technique. So, the linear regression is the best fit for the dataset.

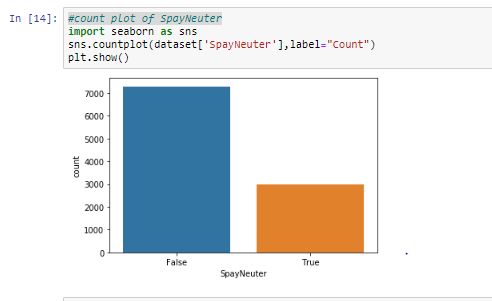
# **Classification:**

Here, we have used a new dataset to perform the classification technique with respect to dogs bitten as per area. The classification is carried out in the "Spay neuter" component as this defines the area or divisions of USA The Health Department is helped to assess that the biting dog is safe ten days after the human was bitten to discourage unwanted vaccinations of rabies being bitten. Every record is an incident involving a single dog Spay neuter. The Spay neuter records will predict the incident of an area with respect to the bite date on which the person was bitten.

In order to classify the data, we will understand the data.



The data is having a Unique ID for a particular bite and we will predict the Spay neuter is True or False on the basis of its feature and data.



The above Figure shows Spay Neuter’s count as indicates the True or False values.

I implemented 4 Classification techniques on the dataset

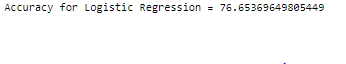
* Decision Tree
* KNN Classifier
* Logistic Regression
* Naive Bayes Classifier

Comparison of the Classification Techniques:

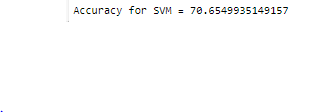
  
Accuracy for Decision Tree Vs



Accuracy for KNN Vs



Accuracy for Logistic Regression Vs



Accuracy For SVM Vs



Accuracy for Naïve Bayes

As seen in the table above, the Decision Tree Classifier showed 77.14% with the highest accuracy relative to other methods and the lowest 46.40% for SVM.

The data set is non-linear and therefore performs best as Decision Tree Classifier did not expect any data Decision Tree Classifier functions faster and easier with noisy data as we have seen from the regression output. In comparison, the SVM classification provides the lowest 46.40% accuracy which clearly demonstrates the overlaps between our data and target groups of noise.

From the above accuracy’s we can predict that Decision tree gave us the best performance with respect to the dependent variables.

Hence, we conclude that the Classifiers were consistent throughout the dataset excluding the SVM as it gave the lowest performance. We tried with the previous dataset to perform the classification, but none were relatable, so we used the new data set which was best suited, and the technique was performed.

# Reference

1)T, S. (2019). *Entropy: How Decision Trees Make Decisions - Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8 [Accessed 15 Apr. 2020].

2) Scikit-learn.org. (2020). *scikit-learn: machine learning in Python — scikit-learn 0.22.2 documentation*. [online] Available at: https://scikit-learn.org/stable/ [Accessed 15 Apr. 2020].

‌

3) Sidana, M. (2017). *Intro to types of classification algorithms in Machine Learning*. [online] Medium. Available at: https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14 [Accessed 15 Apr. 2020].

‌

4) How to fix: ’only integers, slices (`:`), ellipsis (`…`), numpy.newaxis (`None`) and integer or boolean arrays are valid indices (2019). *"How to fix: “only integers, slices (`:`), ellipsis (`…`), numpy.newaxis (`None`) and integer or boolean arrays are valid indices”? - Intellipaat Community*. [online] Intellipaat.com. Available at: https://intellipaat.com/community/17810/how-to-fix-only-integers-slices-ellipsis-numpy-newaxis-none-and-integer-or-boolean-arrays-are-valid-indices [Accessed 15 Apr. 2020].

5) Statistics How To. (2018). *Chinese Restaurant Process: Simple Definition & Example - Statistics How To*. [online] Available at: https://www.statisticshowto.com/chinese-restaurant-process/ [Accessed 15 Apr. 2020].

‌

‌

‌