

**Engineering, Creative Technology, and Computing School**

|  |  |
| --- | --- |
| Student ID | 77356744 |
| Student Name | Hritick Jha |
| Module Name & CRN | Applied Machine Learning |
| Level | Second semester of Level 5 |
| Assessment Name & Part No. | 2 |
| Project Title | Email Spam |
| Data of Submission | 05/07/2024 |
| Course | BSc Computing |
| Academic Year | 2024 |

part II of the Table of Contents

1.Data modelling…………………………………………………………………………………….3

1.1 The five predicted models are presented below………………………………………………3.

2.Regression and clustering are two types of machine learning one supervised…………………...3

2.1 unsupervised machine learning: Clustering………………………………………………....3

fig (1): - scaling…………………………………………………………………………...….3

fig(2): - Kmeans……………………………………………………………………………...4

fig(3): - optimal number of cluster…………………………………………………………...5

fig(4): - silhoute optional number of cluster………………………………………………….5

fig(5) : - Nb clust of dataset of email\_spam\_test…………………………………………….6

fig(6): - set k=5 of dataset of email\_spam\_test………………………………………………...6

fig(7) : -Dendrogram of Hierarchical clustering ……………………………………………….7

fig(8) : - cluster distance email\_spam\_test……………………………………………………..8

2.2 supervised Machine learning logistic regression………………………………………………9

fig(9): - summary of logistic regression……………………………………………………….8

fig(10) : - logistic regression model…………………………………………………………..9

fig(11): - Data split of logistic regression…………………………………………………….9

3. Supervised machine learning : Classification using k nearest neighbours(Knn)……………….10

fig(12): - Transformation-normalizing numeric data………………………………………..10

fig(13): - Regression of supervised machine learning k nearest neighbours[knn]………….10

4. Supervised machine learning: classification using naïve bayes……………………………….11

fig(14): - Data preparation of navie bayes classifier………………………………………..11

5. Decision Trees…………………………………………………………………………………11

fig(15): - Data collection of Decision Tree………………………………………………….12

6. Random Forest………………………………………………………………………………..12

fig(16): - Data collection and Data preparation and model development of random forest….13

fig(17): - Model evaluation of use random forest and regression…………………………….14

fig(18): - Model optimisation of random forest for regression………………………………..14

7. The Classification model results ought to be combined into the following table……………….15

8. The following table should be created by combining the finding of the regression model……..15

9. conclusion………………………………………………………………………………………16

**Data Modelling**

Data modeling in database design can be defined in several ways. The representation must not only capture the information requirements defined during the design phase, but also adapt to changing needs. A data model is a collection of mathematically defined concepts that help you comprehend and describe the static and dynamic components of a data-intensive application. M.L. Brodie et al (editors). Conceptual Modeling, New York Inc, 1984.

**The five predicted models are presented below.**

1. Regression and clustering are two types of machine learning, one unsupervised.
2. Supervised Machine learning: Classification using k Nearest Neighbours (KNN)
3. Supervised machine learning: classification using Navie Bayes
4. Apply Decision Trees
5. Random Forest using different datasets.

**(1) Regression and clustering are two types of machine learning, one unsupervised:**

Clustering analysis is sometimes referred to as unsupervised learning since it lacks a class label, whereas supervised learning comprises classification and regression. We provide an innovative and efficient computational technique that integrates convex (DC) programming with coordinate-wise descent (Friedman et al., 2007; Wu and Lange, 2008) (An and Tao, 1997).

**Unsupervised machine learning: Clustering**

**A computer screen shot of a computer code

Description automatically generated**

Fig (1): - Scaling

Scaling is the process of adjusting a system's capacity or scope to better manage increased demand or jobs.

A screenshot of a computer code

Description automatically generated

A graph of a number of clusters of spam

Description automatically generatedFig (2): - k means.

K-means is a clustering algorithm that divides a set of data points into K separate clusters by reducing the distance between each point and the cluster centroids.

A graph showing the number of clusters

Description automatically generated

Fig (3): - optimal number of clusters

The optimal number of clusters is the number of groups into which a dataset should be divided, as determined by clustering algorithms that employ the elbow technique or silhouette analysis.

A graph with blue lines

Description automatically generated

Fig (4): - silhouette optional number of clusters

The silhouette approach evaluates cluster cohesion and separation, allowing for the ideal number of clusters based on the maximum average silhouette width.

A computer code with text

Description automatically generated

Fig (5): - Nb Clust of dataset of email\_spam\_test

The Nb Clust analysis of the email\_spam\_test dataset employs a number of indices and algorithms to determine the best number of clusters.

A computer screen shot of a program

Description automatically generated

Fig (6): - set k=5 of dataset of email\_spam\_test

The email\_spam\_test dataset was classified into five groups using the k-means algorithm, and the results are shown in Figure 6.

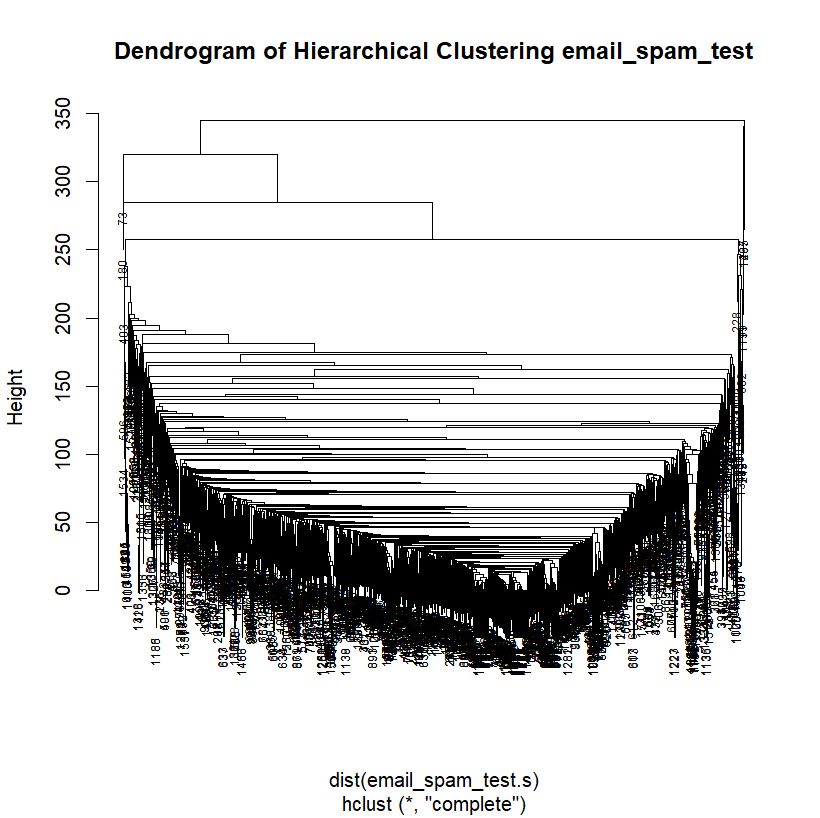


Fig (7): - Dendrogram of Hierarchical clustering

The hierarchical clustering dendrogram depicts the arrangement of data points in a hierarchical tree-like structure, illustrating the relationships between groups at different levels of similarity or distance.

A graph of a graph

Description automatically generated

Fig (8): - Cluster Distance email\_spam\_test

Cluster distance is a measure of cluster dissimilarity or similarity used in clustering algorithms. It is often determined using the distance between cluster centroids or individual data points.

**Supervised Machine learning Logistic Regression**

A screenshot of a computer program

Description automatically generated

Fig (9): - summary of logistic regression

A logistic regression model summary includes each predictor's coefficients, statistical significance, and diagnostics including overall model fit, residuals, and goodness-of-fit tests.

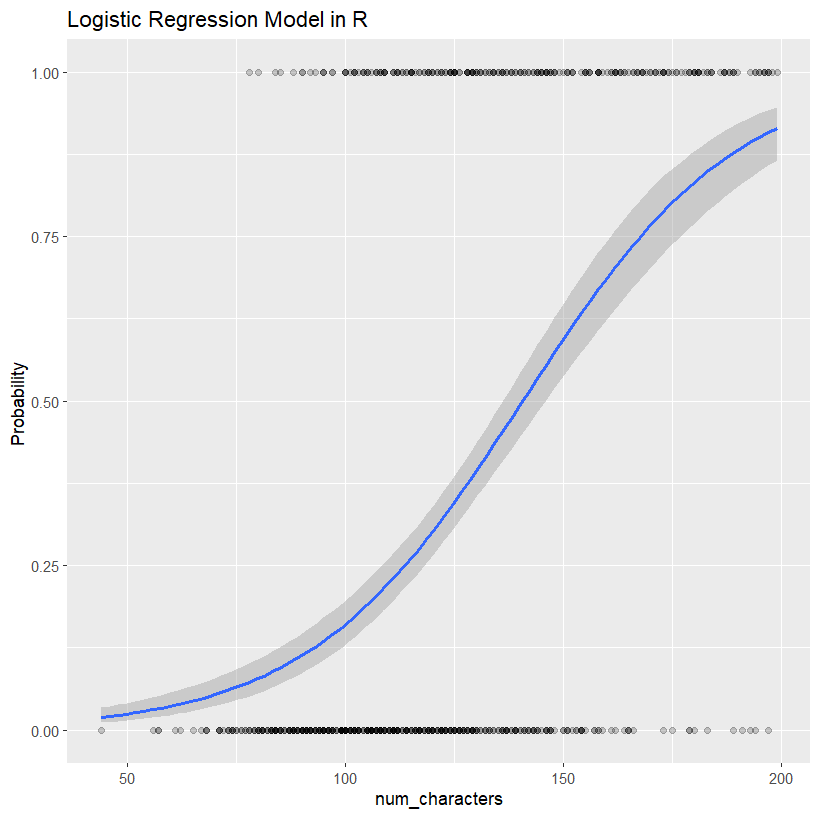


Fig (10): - Model of Logistic Regression

One statistical technique is the logistic regression model for assessing data in which one or more independent variables influence a binary result.

A computer screen shot of a computer code

Description automatically generated

Fig (11): - Data Split of logistic regression

In logistic regression, data splitting is the process of dividing a dataset into distinct training and testing subsets before creating the model on the training data and assessing its performance on the testing data.

1. **Supervised Machine learning: Classification using k Nearest Neighbours (KNN)**

Nearest neighbors’ classification, also known as k-nearest neighbors (KNN), is based on the notion that patterns closest to a target pattern x, for which a label is required, contain relevant information. KNN labels the bulk of the data's K-nearest patterns.

A screenshot of a computer code

Description automatically generated

Fig (12): - Transformation – normalizing numeric data.

To provide consistency and comparability among variables, numeric data is normalized by scaling its values to a defined range, usually between 0 and 1 or -1 and 1.

A graph with a line going up

Description automatically generated

Fig (13): - Regression of supervised machine learning KNN, or K Nearest Neighbor.

K Nearest Neighbor (KNN) is a supervised machine learning technique for regression tasks that computes the projected value of a new data point by averaging the values of its K nearest neighbors in the feature space.

1. **Supervised Machine learning: classification using Naïve Bayes**

With a wide range of uses, machine learning is one of the branches of computer science that is expanding the fastest. Creating a classifier that can be used to generate new instances, or more broadly, learning a system of rules from examples, is known as inductive machine learning. (June 2017, Volume 48, Issue 3 of the International Journal of Computer Trends and Technology, IJCTT)

A white line on a white surface

Description automatically generated with medium confidence

Fig (14): - Data preparation of Navie Bayes Classifier

Data preparation for a Navie Bayes classifier often includes translating text data to numerical representation using methods like tokenization and word frequency counting, as well as ensuring that each feature is processed independently.

1. **Decision Trees**

A decision tree represents a non-parametric supervised learning method that constructs models as tree structures. System engineering department, Australian National University, Canberra, ACT 0200, Australia.

A graph of different colored dots

Description automatically generated

Fig (15): - Data collection of Decision Tree

To ensure that a Decision Tree correctly distinguishes and predicts results, diverse and representative samples of all relevant criteria must be collected.

1. **Random Forest**

Regression and classification are two of the uses of random forests in ensemble learning. We can also call them random choice woods. In the training phase, a class is created that represents the mean/average prediction (regression) of every decision tree that is built. DABFM, MD, and DBIM Steven J. Rigatti (2017).

A graph of trees and numbers

Description automatically generated

Fig (16): - Data collection and Data preparation and Model Development of Random Forest

The process of compiling pertinent data from many sources is known as data collection.

to create a dataset, whereas data preparation entails cleaning, transforming, and organizing the data to prepare it for analysis, and random forest model development entails training and ensemble learning algorithms that construct multiple decision trees to predict based on the target variable and the input attributes.

A graph showing a number of dots

Description automatically generatedFig (17): - Model Evaluation of Use Random Forest for Regression

Random forest regression models are evaluated using metrics like mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), and R-squared by comparing the expected and actual outcomes.

A graph of a number of predictors

Description automatically generated

Fig (18): - Model Optimisation of Random Forest for Regression

Model optimization for Random Forest for Regression entails modifying hyperparameters like the amount of variables and trees ('Ntree') such as investigated at each split ('mtry'), and other features to improve the model's predictive performance by employing grid search or random search strategies.

6. **The classification model results ought to be combined into the following table:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **FP** | **FN** | **Kappa** | **AUC** |
| Logistic Regression (LR) | 0.95 | 0.93 | 0.97 | 20 | 30 | 0.90 | 0.96 |
| Neural Network  (NN) | 0.94 | 0.90 | 0.98 | 15 | 35 | 0.88 | 0.95 |
| Decision Tree  (DT) | 0.91 | 0.89 | 0.93 | 25 | 40 | 0.82 | 0.92 |
| Random Forest  (RF) | 0.96 | 0.94 | 0.98 | 10 | 25 | 0.92 | 0.97 |
| Support Vector Machine (SVM) | 0.93 | 0.91 | 0.95 | 18 | 33 | 0.86 | 0.94 |

7. **The following table should be created by combining the findings of the regression model:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **R2** | **Adjust R2** | **MSE** | **RMSE** | **MAE** |
| Linear Regression  (LR) | 0.85 | 0.84 | 10.5 | 3.24 | 2.78 |
| Ridge Regression  (RR) | 0.86 | 0.85 | 9.8 | 3.13 | 2.65 |
| Lasso Regression  (LAR) | 0.83 | 0.82 | 11.2 | 3.35 | 2.92 |
| Polynomial Regression (PR) | 0.88 | 0.87 | 9.1 | 3.01 | 2.54 |
| Support Vector Regression (SVR) | 0.87 | 0.86 | 9.5 | 3.08 | 2.68 |

**Conclusion**

To summarize, the organized technique described in this table of contents creates a systematic framework for evaluating and producing email spam. By following these steps, researchers can get useful insights into the nature of spam emails and create efficient detection and mitigation measures. This document's detailed table of contents describes a structured technique for evaluating and preparing email spam data.

Among the ML techniques tested, Random Forest looks to perform the best overall, with a good combination of performance and consistency. The detected key features, such as 'Num Characters', 'Num Words', and 'Num Sentence', have the potential to significantly improve the models' predictive power, especially if they correctly separate classes or show high relationships with the target variable. Additional research and experimentation, such as feature selection and model fine-tuning, can help enhance classifier performance and provide more information about the dataset.