**A report on**

**“Text Classification”**

**Mentor by:**

**Prof. Arya Rahgozar**

**Team Members:**

**Jayshil Patel**

**Josh Patel**

**Hrutik Champaneri**

**Pratik Gade**

**University of Ottawa**

**Electrical and Computer Engineering**

**Index:**

1. **SECTION 1: - Choosing the data sets……………………………3**
2. **SECTION 2: - Data Preprocessing……………………………….4**
3. **SECTION 3: - Machine Learning………………………………..6**
4. **SECTION 4: - Cross Validation Score………………………….12**
5. **SECTION 5: - Confusion Matrix………………………………..13**
6. **SECTION 6: - Accuracy comparison……………………………15**
7. **SECTION 7: - Hyperparameter tuning…………………………16**
8. **References…………………………………………………………17**

**Introduction: -** The purpose of the assignment was to perform the classification of the data. Particularly the data derived from random five books which are already uploaded onto the Gutenberg’s digital library. Firstly, the data sets are chosen from the digital library. Then after, the data is imposed on various machine learning models to find out which one has the highest accuracy. In addition to that some other parameters such as model’s precision, recall, f1 score and support is also displayed. Note that in this case, two types of data sets are used. The first one contains five books which are of random genres and the other one is used with the books that are of the same genre. The predictions and validations done by machine learning models are presented with the help of confusion matrix and various visuals. Furthermore, the parameter tunning of the model having the highest accuracy is also done at the end. The report is divided into seven different sections, each of them containing the information about a specific action taken in order to complete the assignment.

**SECTION 1: - Choosing the data sets.**

As aforementioned, two data sets have been used in this case. One is having the books that are of the same genre and the other one contains books from random genres. The reason behind doing this was to check if a particular algorithm only works well with a certain type of data set or is it suitable for almost all the types of data sets. Below are the two data sets which were taken into the consideration while doing this assignment.

Data set 1(Random): -

|  |  |  |
| --- | --- | --- |
| Books | Genre | Authors |
| Pride and prejudice | Romance | Jane Austen |
| Frankenstein | Horror Fiction | Mary Wollstonecraft Shelley |
| Moby-Dick | Adventure fiction | Herman Melville |
| The Origin of Species | Science | Charles Darwin |
| War and Peace | Historical | Leo Tolstoy |

Data set 2(Specific): -

|  |  |  |
| --- | --- | --- |
| Books | Genre | Authors |
| Middlemarch | Romance | George Eliot |
| Hints for Lovers | Romance | Arnold Haultain |
| Falling in Love | Romance | Grant Allen |
| Herein is Love | Romance | Reuel L. Howe |
| The Evolution of Love | Romance | Emil Lucka |

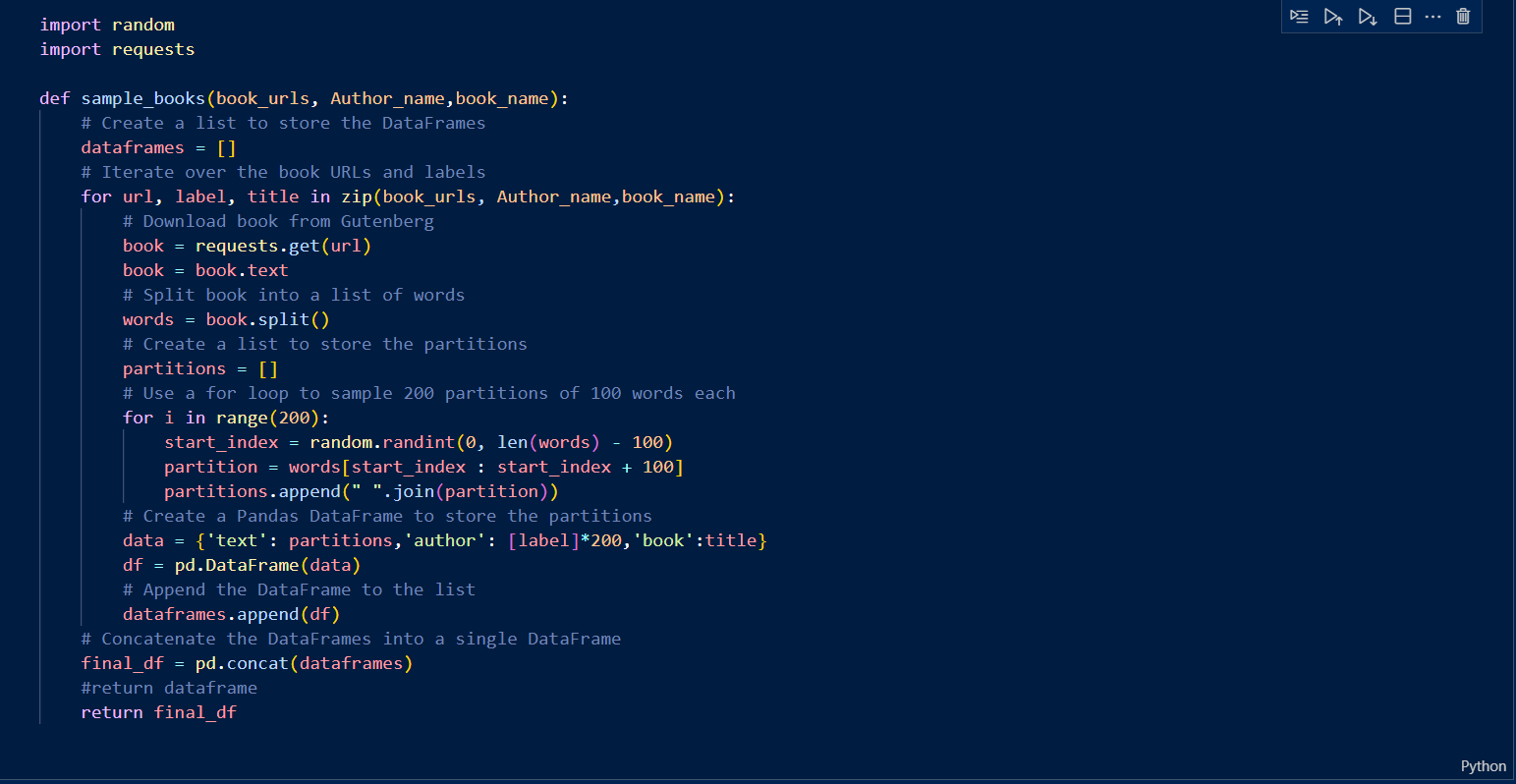
**SECTION 2: - Data Preprocessing.**

Data pre-processing is an essential phase in the data science process since it makes that the data is accurate, consistent, and fit for analysis. Data pre-processing is significant for the reasons listed below, among others:

* Handling missing values: Missing values are a common occurrence in real-world data, and they might affect how accurate the results that can be drawn from the data are. Data pre-processing aids in the management of missing values by either eliminating instances where missing values exist or imputes appropriate values.
* Getting rid of Outliers: Outliers can have a big impact on how data analysis and modelling turn out. To ensure that the outcomes are more accurate, data pre-processing aids in the identification and removal of outliers.
* Feature scaling: The magnitude of the features in the data must be uniform for many machine learning methods. The features are scaled by data pre-processing to make them appropriate for analysis.
* Encoding Categorical Variables: Categorical variables must be translated into numerical values before being encoded since some machine learning algorithms cannot handle them. Categorical variables are helped to be encoded during data pre-processing so they may be used in the analysis.

In conclusion, data pre-processing aids in enhancing data quality and preparing it for analysis, which in turn produces more accurate results and improves decision-making.

In our case the data pre-processing part was very simple as we just had to split the text of the books into the words and we had to get rid of the special characters and the punctuations which were there. The snippet of the part of the code which is doing that is mentioned below.



Text

Description automatically generated

After doing this, the next step was to transform the data in such a way that the data can be recognized by the Machine Learning model. So, we transformed the data into a bag-of-words representation.

Text

Description automatically generated

**SECTION 3: - Machine Learning**

Here we have to do the data classification and for the classification the data, supervised machine learning algorithms are used. So, here we have used 6 different types of supervised machine learning models for the prediction. The details of the models are mentioned below.

1. Decision Tree: - Decision tree is a ML approach for classification of the data. It is condition-based algorithm. The tree contains two types of nodes first leaf node and second decision node, Leaf nodes are the final outcomes and decision nodes are the choice between alternatives. The sample of the decision tree is shown below. Decision tree is very easy to understand because of its step-by-step approach moreover they reduce vagueness in the decision making.

Diagram

Description automatically generated

This type of algorithm has the components given below,

Root node: the node from which the system is starting.

Leaf node: they are the final nodes; the tree cannot be further divided.

Sub tree: the secondary tree in the system.

Parent/child node: the parent node is the root node all the other nodes are child nodes.

Splitting: the process of dividing decision node into other nodes.

Pruning: the process of cutting off unwanted branches.

Diagram

Description automatically generated

This system works by classifying the data into the most basic level hence leaf node. First the data set f (assuming the label of data set is f) is given to the root node, here the first level of tree is created. After that the data is further divided into sub sets which have the best attributes, this attributes are calculated by attribute selection measure(ASM). this generates decision nodes, the same steps are performed until the information can no longer be classified hence reaching the last level of the tree which is known as leaf node.

1. SVM (Support Vector Machine): - Support vector machine is the algorithm for the classification of data. This algorithm uses a hyperplane to distinguish the data. To separate two classes many possible that could be chosen. The main objective for us is to find the most suitable hyperplane. The hyperplane which completely separates two classes is the ideal hyperplane. Support vectors are the data points which are the closest to the hyperplane. Support vectors help us in achieving the best possible hyperplane for the given data set. If that data points are removed from the data set it will certainly change the coordinates of the hyperplane.

Chart, scatter chart

Description automatically generated

This algorithm is supervised learning algorithm. For example, in the above figure class 1 is of dogs and class 2 is of cats. So firstly, we will have to put a lot of images of dogs and cats with tags so that the algorithm can understand the characteristics of both the objects. Now when we put the un-tagged data the algorithm will try to identify whether it’s a cat or a dog by its previous learning examples. However, with the large data set SVM can consume a lot of time classifying the data and the memory may blow up too.

1. Regression: - Regression analysis is a mathematical method which reflects the relationship between two or more variables. In our day-to-day life how much food we consume has direct relation with our weight as our weight depends on our food consumption. If we were to put this relationship in form of graph, we will get the idea of what is our weight going to be in next 6 months or 12 months or so….

One can say that regression is the algorithm that guesses future outcome based on current data set. The formula for linear regression analysis is given below.

Y = f(x) + e

Here y is the prediction(outcome) f(x) is function which relates x and y and e is the margin of error. X is an independent value and y is dependent value as it depends on the value of x.

Chart, scatter chart

Description automatically generated

Suppose at the base of the graph above the experience of the employee is very less so he will be provided with lower salary, as experience increases the salary also increases. Linear regression algorithm as shown in the above equation will be able to predict the salary of the employee for the next years as it takes the value of x and predicts the outcome in terms of y. if there is only one input variable x then the system is called simple linear regression but if there are multiple values of x then it is called multiple linear regression.

1. K-NN (K-Nearest Neighbor): - k-nn is one of the most popular and important algorithms. This learning algorithm is easier and simpler than most algorithms. In this algorithm all the training instances are stored in master database. When new query is added it is compared with the existing data and the results are derived. In k-nn algorithm the distance between each data points are generally calculated by euclidean distance function. when it comes to practical applications it does not perform very well in each case mainly because this algorithm decides the data’s class by the same number of nearest neighbours each time it is applied.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Suppose there are two data set in the system already which are represented by circle and square in the figure, and another data point is added to the system. Now firstly the number of k is decided. In our case we can see the number of k is five. By using euclidean distance function the distance of five nearest neighbours is calculated. We can see that the three nearest neighbours are from class 2(circle class) and two nearest neighbours are from class 1(square) so the newly added data point will be classified as class 2 object.

1. Bayesian Learning: - Bayesian method is a purely mathematical or one can say that it is a statistical method. It gives the probability of an event to occur. Bayes’ theorem is being used in this algorithm to check the conditional probability of any event. The formula of the theorem is given below.

p(A|B) =[P(B|A) X P(A)]/P(B)

Here, p(A) is the probability of occurrence of event A. probability of occurrence of event B is denoted by p(B). p(A|B) represents the probability of occurrence of event A given event B and probability of occurrence of event B given A is denoted by p(B|A).

In this method firstly the given data set is converted into a frequency table. In the second step it will generate the table containing likelihood means the possibilities of the features to happen, and then using Bayes’ formula the output is calculated.

1. Random Forest: - Random Forest falls under the category of supervised learning. It is used for the classification of the data. This algorithm is very similar to the decision tree algorithm, but this algorithm contains multiple trees instead of just one tree thus making it more accurate. The principle behind this method is to generate multiple trees by selecting random features and random samples [15]. random forest eliminates the risk of over-fitting of the data, and it also eliminates the error because of the imbalanced data up to some extent.

Diagram

Description automatically generated

Above figure illustrates that there are multiple decision trees included in the random forest algorithm and multiple training data set are given to all the decision trees present in the system. The average of all the decision trees’ output is considered as prediction. The averaging of all the outcomes makes this algorithm more precise than the single decision tree algorithm.

Below are the implementations of all the above-mentioned models.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

**SECTION 4: - Cross Validation Score.**

The model's performance on the validation data is summed up by a statistic called the cross-validation score. Cross-validation scores come in a variety of forms, including:

Accuracy: This is the proportion of the model's total predictions that were correct forecasts that were made. It is a frequently employed metric in categorization issues.

F1-score: The F1-score represents the harmonic mean of recall and precision. It is a frequently used statistic for datasets that are unbalanced, meaning that some classes have much less samples than others.

Precision: This is the percentage of all positive forecasts that were correct. Remember that this represents the proportion of genuine positive forecasts to all actual positive events.

AUC-ROC: Area under the receiver operating characteristic curve, or AUC-ROC is a standard measurement for binary classification issues.

The cross-validation score aids in estimating how well the model performs on unobserved data and can be used to compare many models and choose the most appropriate one for the given issue.

Below is the graph indicating the data of the models used in this assignment.

Graphical user interface

Description automatically generated

**SECTION 5: - Confusion Matrix**

Machine learning uses a confusion matrix as a performance evaluation tool, especially for classification issues. It is a table that lists how many predictions a classifier made correctly and incorrectly. Typically, a confusion matrix is shown as a 2x2 table with the following entries:

* The number of incidents that were successfully identified as positive cases, or true positives (TP).
* The number of incidents that were mistakenly labelled as positive cases, or false positives (FP).
* The number of cases that were accurately identified as negatives, or true negatives (TN).
* The number of cases that were mistakenly labelled as negative is known as the False Negatives (FN) rate.

Accuracy, precision, recall, and F1-score are a few performance metrics that can be calculated from the numbers in the confusion matrix. These metrics provide a more comprehensive picture of the classifier's performance, including how well it can distinguish between positive and negative instances and how well it can prevent false positives and false negatives.

The confusion matrix is a crucial tool for assessing and enhancing a classifier's performance since it reveals its advantages and disadvantages and helps pinpoint areas for development.

Below are the snippets of the best and the worst performing models that were used in the project.

1. Decision tree (Worst): -

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

1. Naïve Bias (Best) : -

Graphical user interface, text

Description automatically generated

Graphical user interface, application

Description automatically generated

From the above two cases we can see that Naïve Bias algorithm has predicted the correct author for almost every single time, whereas, the decision tree algorithm was not so accurate regarding its predictions.

**SECTION 6: - Accuracy Comparison**

The comparison of accuracy of all the used models has been added by the help of a horizontal bar chart. The snippet of the same is mentioned below. One can see that the Naïve bias had the highest accuracy of all the trained models closely followed by the logistic regression model. The Random forest algorithm had almost same accuracy as the logistic regression. SVC also falls into the same category as above two. Two models having very less accuracy were the K-nn and Decision Tree.

Chart

Description automatically generated

**SECTION 7: - Hyperparameter tuning.**

The process of finding the ideal set of hyperparameters for a machine learning model is known as hyperparameter tuning. Hyperparameters are settings made before the model is trained and cannot be determined from the training set of data. The learning rate, the quantity of trees in a random forest, or the degree of regularisation in a linear regression are a few examples of hyperparameters. Finding the hyperparameters that provide the model the highest performance on a validation set is the goal of hyperparameter tweaking. This is accomplished by experimenting with various hyperparameter combinations and assessing the model's performance using a performance metric, such as accuracy, F1-score, or AUC. The performance of the model can be greatly impacted by hyperparameter tuning, making it a crucial stage in the machine learning pipeline. The generalisation to fresh data and prediction accuracy of the model can be enhanced by selecting the optimal hyperparameters.

Text

Description automatically generated

**REFERENCES: -**

1. *S. S. Gavaskar and S. D. Savarkar, "Eager decision tree," 2017 2nd International Conference for Convergence in Technology (I2CT), 2017, pp. 837-840, doi: 10.1109/I2CT.2017.8226246.*
2. *Xiaorong Cheng, Lin Sun and Ping Liu, "Application of regression analysis based on genetic particle swarm algorithm in financial analysis," 2010 International Conference On Computer Design and Applications, 2010, pp. V4-335-V4-338, doi: 10.1109/ICCDA.2010.5541106.*
3. D. C. Montgomery, E. A. Peck, G. G. Vining, Introduction to linear regression analysis, Vol. 821,John Wiley & Sons, 2012.
4. *Okfalisa, I. Gazalba, Mustakim and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2017, pp. 294-298, doi: 10.1109/ICITISEE.2017.8285514.*
5. *S. Sun and R. Huang, "An adaptive k-nearest neighbor algorithm," 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010, pp. 91-94, doi: 10.1109/FSKD.2010.5569740.*
6. F. V. Jensen, An introduction to Bayesian networks, Vol. 210, UCL press London, 1996.
7. [scikit-learn: machine learning in Python — scikit-learn 1.2.1 documentation](https://scikit-learn.org/stable/index.html)