JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DESIGNING A WEB BLOG POSTER UTILIZING RANDOM FOREST ALGORITHM AND DECISION TREES

1st Dr. T. Amalraj Victorie,2nd M. Vasuki, 3rd Moniksha. M

[1]Associate Professor, Department of Master of Computer Application, Sri Manakula Vinayagar Engineering College Puducherry-605 107, India.

[2]Associate Professor, Department of Master of Computer Application, Sri Manakula Vinayagar Engineering College Puducherry-605 107, India.

[3]PG Student, Department of Master of Computer Application, Sri Manakula Vinayagar Engineering College Puducherry-605 107, India

ABSTRACT:

With the widespread proliferation of blogs, there arises a crucial necessity to extract pertinent information for addressing a spectrum of issues spanning social, political, criminal, and other domains. It can extracting meaningful information from blogs is crucial . The goal of blogs is to classify the bloggers based on their writing style, content , or other features. It can be analysis the data from the Random Forest algorithms to classify bloggers whether they are professional or otherwise from the Kozhikode and Boyer Ahmad province in Iran. In this algorithm , it can analysis the data through the tree-based ensemble learning technique in machine learning .In this algorithm, it can be process through the ensemble learning methodology that analysis the data for predictions of these tree ,it can predicts accuracy value. As a means of comparison, this paper suggests employing the Random Forest algorithm for categorizing blogs, utilizing the identical dataset . The findings indicated that employing ensemble classification using the Random Forest algorithm resulted in greater accuracy compared to previous research utilizing the same algorithm.

KEYWORD: Random forest , decision tree process , ensemble learning method, blog poster prediction, Python 3.6, Machine Learning, Data Analysis, Training Data ,Testing Data

1.INTRODUCTION:

A blog is an online journal or informational website run by an individual, group, or corporation. It offers regularly updated content in the form of blog posts about a specific topic such as such as social, political, criminal and others . Blogs have had a fascinating evolution over the years. They started as personal online journals or diaries and gradually transformed into platforms for sharing expertise, opinions, and news on various topics. The shift from single-author blogs to multi-author ones reflects the growing diversity of voices and perspectives on the internet. With the rise of microblogging platforms like Twitter, the landscape has become even more dynamic, with blogs often serving as sources for more in-depth analysis and discussion. The versatility of blogs as both noun and verb showcases their adaptability in the digital age.

Python is the most popular programming language in the field of machine Learning and has been used more widely. Python 3.6 stands out as a powerful and widely adopted programming language, suitable for projects of various scales,

including those in the field of predictions. Python strengths lie in its clean syntax, readability, and extensive library ecosystem. Its diverse array of algorithms and rich data structures make it well-equipped for handling a wide range of tasks, including predictive modeling. Moreover, Python's ability to integrate seamlessly with other languages further enhances its utility, allowing developers to leverage existing codebases and technologies . In the landscape of machine learning systems, Mahout shines as a distributed machine learning library, favored by major corporations like Yahoo, Twitter, and LinkedIn. Its distributed nature enables efficient processing of large datasets, making it a popular choice for real-world applications . Prediction analysis, a critical aspect of machine learning, often involves selecting the most appropriate algorithm and model for a given problem. In the context of random forest algorithms, Python offers robust implementations through libraries like scikit-learn, enabling practitioners to tackle prediction tasks effectively.

The paper aims to establish a random forest algorithm and decision tree model of one variable with the help of Python3.6 to predict the effect of the multiple decision trees over different parts of the same training dataset . The model will define the fitting relations between,random samples from the training set through
Build a decision tree for each training data point and then, it can be analysis the new dataset to be find and finally go back to the assumed equation to predict the variation tendency of the trained dataset based on the typical decision tree model of predicting the future accuracy data of web blog poster according to the assigning the new data points to the category that wins the majority votes or average.

LITERATURE SURVEY:

1.Title: "Random Forests"
Authors: Leo Breiman
Journal: Machine Learning
This seminal paper by Leo Breiman introduces the Random Forest algorithm, explaining its principles and advantages. It covers topics such as ensemble learning, decision trees, and the construction of Random Forests. It also discusses applications of Random Forests in classification and regression tasks.

Title: "Random Forests for Classification in Ecology"
Authors: Andy Cutler, Douglas R. Edwards, Kathryn H. Beard, Anne Cutler, Tom Hess, John Gibson, and James J. Lawler
Journal: Ecology
This paper explores the application of Random Forests in ecology, specifically for classification tasks. It discusses how Random Forests can handle complex ecological datasets and compares their performance with other classification methods. The paper provides insights into using Random Forests for species distribution modeling and other ecological studies

Here's a journal paper that provides a comprehensive overview of the Decision Tree algorithm:

2.Title: "Decision Trees: A Comprehensive Review from a Statistical Perspective"**

- Authors: Lior Rokach, Oded Maimon

- Journal: Knowledge and Information Systems

This paper offers an in-depth examination of Decision Trees from a statistical perspective. It covers various aspects of Decision Trees, including different algorithms (e.g., ID3, C4.5, CART), pruning techniques, handling missing values, and dealing with continuous attributes. The paper also discusses the strengths and weaknesses of Decision Trees compared to other machine learning algorithms. Additionally, it provides insights into practical considerations for applying Decision Trees in real-world scenarios.

This journal paper should serve as a valuable resource for understanding the Decision Tree algorithm and its implications in statistical modeling and machine learning. You can access the full text of the paper through the provided DOI link.

2. ABOUT BLOG :

Blogs are dynamic websites regularly updated with content that offers in-depth analysis and perspective on specific subjects. The word blog is a combined version of the words “web” and “log.” At their inception, blogs were simply an online diary where people could keep a log about their daily lives on the web. They have since morphed into an essential forum for individuals and businesses alike to share information and updates. In fact, many people even make money blogging as professional full-time bloggers.

A blog consists of a series of articles or posts. While the look and feel of your blog may differ depending on the platform and design preferences you, as the blogger, select, here are some typical components you might encounter in a typical blog and include in your own, keeping in mind the importance of user experience design and web design:

2.1 HEADER SIDE:

The top section of a blog includes the title of your blog or its logo, accompanied by a navigation menu facilitating visitors in navigating various sections or categories of your blog. You may opt to organize blogs on similar topics together, catering to returning readers seeking posts within their area of interest.

2.2CONTENT OF THE BODY:

This is where the content of your blog posts is displayed. Every post typically features a title, the author's name, publication date, and the primary content, which may encompass text, images, videos, or other multimedia elements.

2.3 SLIDEBAR OF THE ARTICLE:

A blog may have a sidebar on one or both sides of the main content area. The sidebar frequently hosts supplementary details or functionalities like a search bar, recent and popular posts, categories, tags, social media links, an about section, and advertisements. This section of your blog aids in enhancing navigation and site organization for both users and search engines.

2.4 COMMENTS OF THE POSTER:

Many blogs allow readers to leave comments on their posts. The comments section usually shows up beneath the primary content of every post and might offer readers the option to reply to comments or give them an upvote. Prior to enabling comments on your blog, ensure that you have the necessary time and resources to effectively manage them.

2.5 FOOTER OF POSTER:

The bottom section of your blog usually contains copyright information, links to your privacy policy and terms of service, additional navigation links, and sometimes widgets like a subscription form, social media icons (social share buttons), or related posts.

Blog designs can vary greatly depending on the theme, customization options, and personal preferences that you chose. These components offer a basic glimpse into the appearance of a blog, but blogs can boast unique layouts or extra features depending on the platform you've chosen and your design preferences. Typically, to establish and maintain a blog, you'll require a blogging platform, along with a domain name and web hosting service such as wix hosting.

2.2 TYPES OF BLOG:

Personal blog: This type of blog usually works like an online diary where the blogger shares opinions, often not aiming to reach a target audience or sell an item Personal blogs have the flexibility to delve into diverse topics, ranging from family gatherings and introspective musings to professional endeavors and creative projects.

Niche blog: Provides information on a particular topic, usually related to the blogger's passions, skills, and knowledge. Instances of this type of blog encompass book blogs, culinary blogs, and lifestyle blogs.

Multimedia blog: It uses a blog format but publishes multimedia content, like videos and podcasts, instead of written posts. It also usually includes the video or podcast's summary, table of contents, and essential quotes.

New Blog: A news blog concentrates on delivering timely updates and fresh releases within a particular industry or field of interest**.** Unlike other blogs, news blogs typically do not usually include opinions or personal content.

Business blog: A business blog is primarily dedicated to sharing content pertinent to a company's industry or informing the target audience about any developments within the business. It can exist either as a section on the company's website or as a standalone site.

Affiliate blog: An affiliate blog operates on affiliate marketing principles, where the blogger promotes products and services from third-party vendors. Blog owners earn a commission for each purchase made through their customized affiliate links.

Reverse blog: A reverse blog, also referred to as a group blog, involves multiple authors collaborating to create posts on interconnected subjects. The blog owner oversees the process, proofreads, and ultimately publishes the content.

3. DECISION TREE ALGORITHM:

The application of the decision tree algorithm can be observed in various fields. Text classification and text extraction are the field where they are used. Additionally, in libraries, books can be categorized into various genres using the Decision Tree algorithm. Similarly, companies, hospitals, schools, colleges, and universities employ it to manage their records efficiently. Decision Tree algorithms are effective in that they provide human-readable rules of classification. Beside this it has some drawbacks, one of which is the sorting of all numerical attributes when the tree decides to split a node. Such split on sorting all numerical attributes becomes costly (e. g) efficiency or running time and memory size, especially if Decision Trees are set on data the size of which is large i.e. it has more number of instances.

3.1 Decision Tree Terminologies:

1.Root Node: It is the topmost node in the tree, which represents the complete dataset. The root node serves as the initial step in the decision-making journey..

2.Decision/Internal Node: A decision or internal node represents a decision point concerning an input feature, with connections branching out to leaf nodes or other internal nodes.

3.Leaf/Terminal Node: A node without any child nodes that indicates a class label or a numerical value.

4.Splitting: The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.

5.Branch/Sub-Tree: A subsection of the decision tree starts at an internal node and ends at the leaf nodes.

6.Parent Node: The parent node is the starting point that branches out into one or more child nodes.

7.Child Node: The nodes that emerge when a parent node is split.

8.Impurity: A measurement of how consistent or uniform the target variable is within a specific subset of data. The Gini index and entropy are commonly utilized metrics to gauge impurity in decision trees for classification tasks.

9.Variance: Variance assesses the extent of variation between the predicted and target variables across different samples within a dataset. It is used for regression problems in decision trees. Mean squared error, Mean Absolute Error, Friedmans , or Half Poisson deviance are used to measure the variance for the regression tasks in the decision tree.

10.Information Gain: The splitting criterion is chosen based on the feature that yields the highest information gain, aiming to identify the most informative feature for splitting at each node. This process aims to generate subsets that are as pure as possible.

11.Pruning: Pruning involves eliminating branches from the tree that contribute little or no additional information, or that may cause overfitting.

3.2 STRUCTURE OF DECISION TREE:

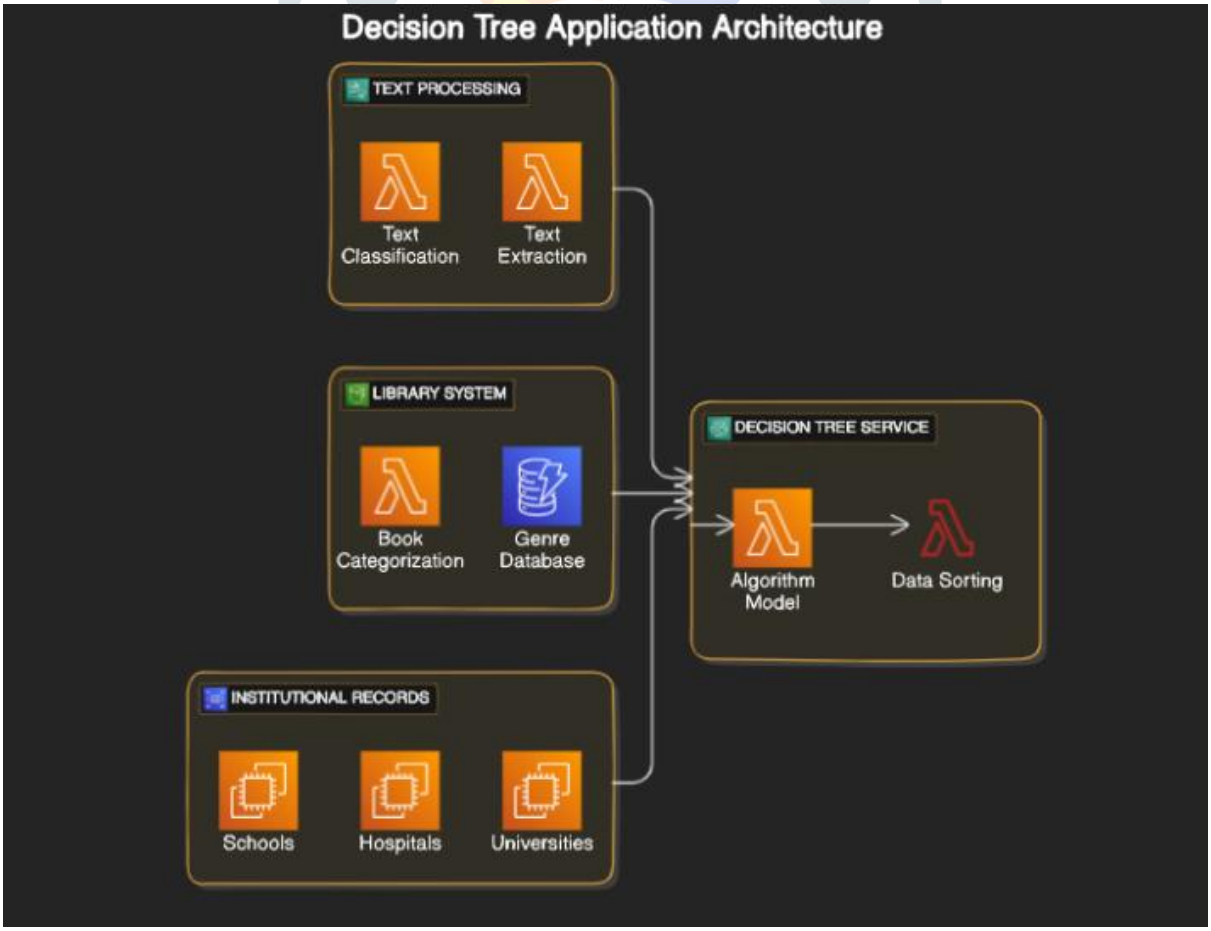


FIG 1.1 Architecture diagram for decision tree

3.3 ADVANTAGES OF THE DECISION TREE:

1.It is simple to understand as it follows the same process which a human follow while making any decision in real-life.

2.It can be very useful for solving decision-related problems.

3.It helps to think about all the possible outcomes for a problem.

4.There is less requirement of data cleaning compared to other algorithms.

3.4 DISADVANYAGES OF THE DECISION TREE:

1.The decision tree contains lots of layers, which makes it complex.

2.It may have an overfitting issue, which can be resolved using the Random Forest algorithm.

3.For more class labels, the computational complexity of the decision tree may increase.

4. RANDOM FOREST ALGORITHM:

The concept of Random Forests has demonstrated superior performance compared to other classifiers such as Support Vector Machines, Neural Networks, and Discriminant Analysis, while effectively mitigating the issue of overfitting. Approaches like Bagging and Random Subspaces, which involve ensembles of diverse classifiers and employ randomization to enhance diversity, have proven highly effective. Random Forests have garnered significant attention in machine learning due to their efficient discriminative classification.

Each Decision Tree within a Random Forest is constructed by randomly selecting data from the available dataset. For example a Random Forest for each Decision Tree can be built by randomly sampling a feature subset, and/or by the random sampling of a training data subset for each Decision Tree. In a Random Forest, the features are randomly selected in each decision split. The correlation between trees is reduces by randomly selecting the features which improves the prediction power and results in higher efficiency.

1. Accurate predictions results for a variety of applications
2. Through model training, the importance of each feature can be measured
3. Trained model can measure the pair-wise proximity between the samples

Random Forest are not only keeps the benefits achieved by the Decision Trees but through the use of bagging through its utilization of sample-based voting and random selection of variable subsets, Random Forests typically outperform individual Decision Trees, yielding superior results.

4.1 STRUCTURE OF RANDOM FOREST ALGORITHM:

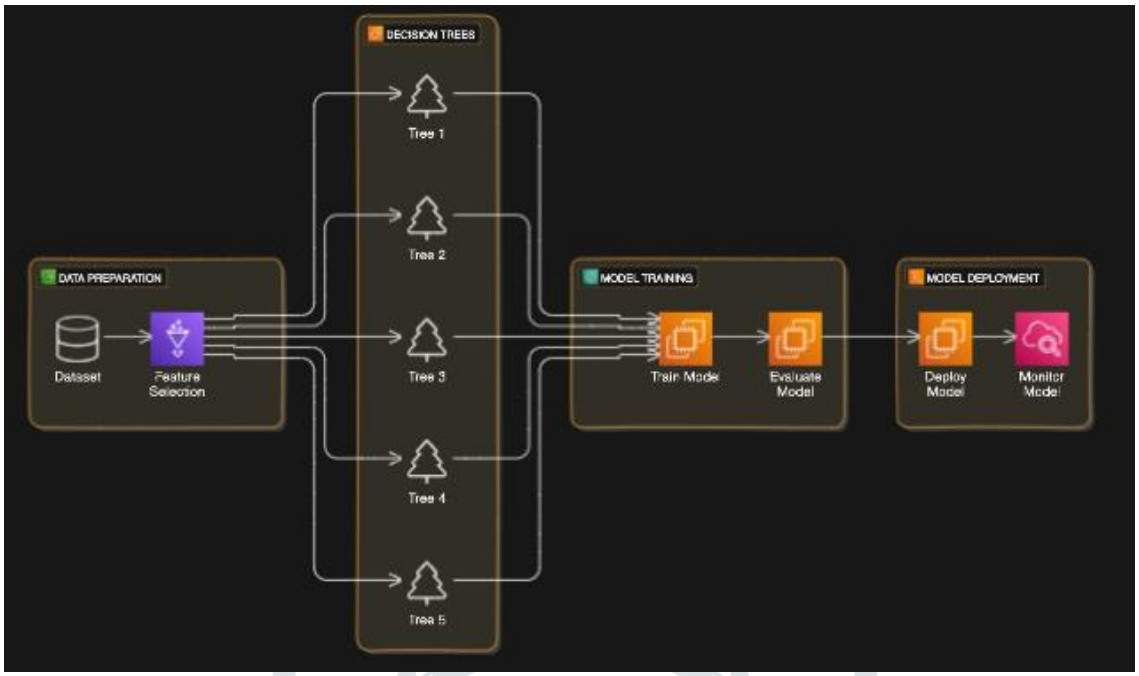


Fig 1.2 : Architecture of preparing the data for trained set through the Random Forest Algorithm

4.2 Advantages of Random Forest:

1.Random Forest is capable of performing both Classification and Regression tasks.

2.It is capable of handling large datasets with high dimensionality.

3.It enhances the accuracy of the model and prevents the overfitting issue.

4.3 Disadvantages of Random Forest:

1.Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

5. RESULT AND ANALYSIS:

5.1 GATHERING THE DATA:

Every blog creater can collect the data for the current content of the blog poster. They can first, start to focus on there content and take the keywords for that . And then, they can related the keywords for the subtopics and content of the blogs. Then, analyze the questions for there related subtopics for the extra content . Then , Anlyze the result through the process of historical data. They can explore the extra features for the blog through analysis of the

old data . Get the input for the community or organisation for complete the blog poster.

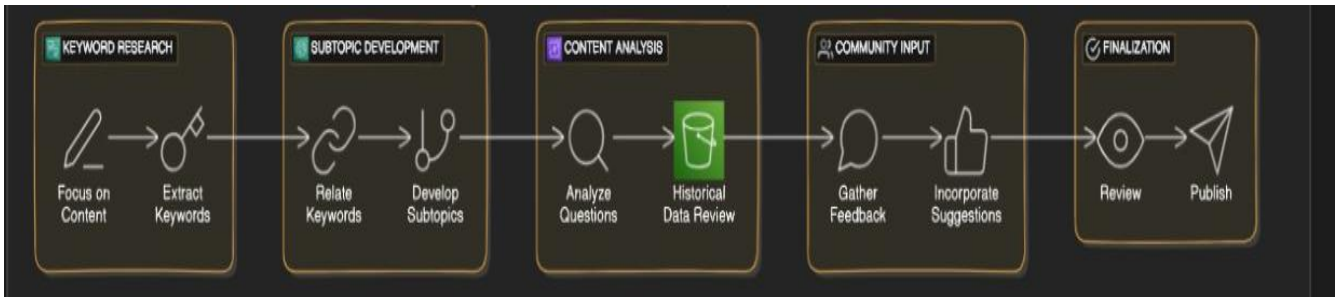


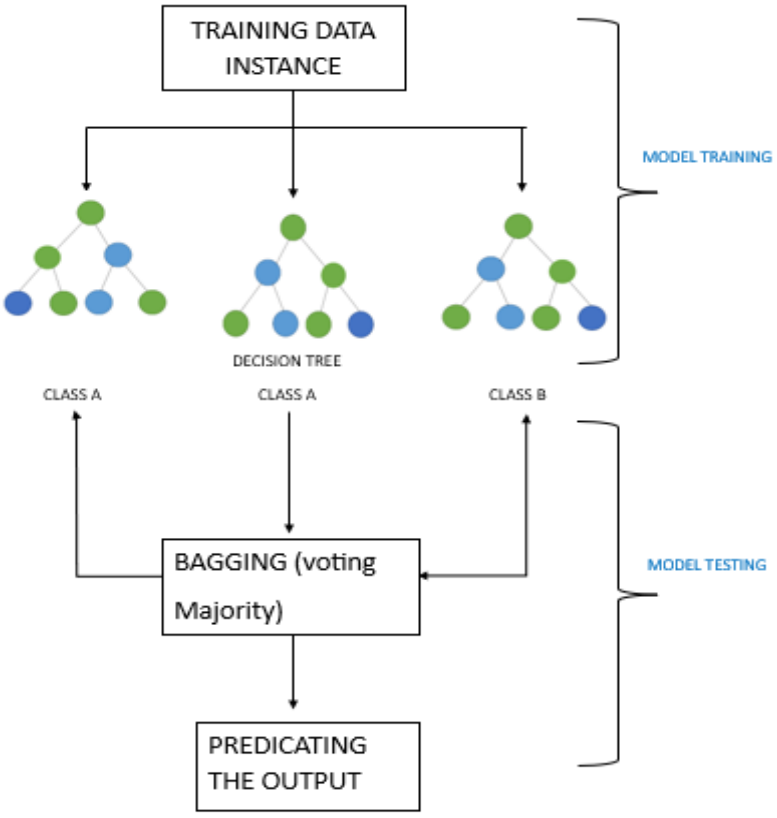
Fig 1.3 blog poster can analysis the content for there topics

5.2 DATA ANALYSIS:

Data analysis is the process of inspecting, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facts and approaches, encompassing diverse techniques under a variety of names, and is used in different social, political, criminal and others domains. In today's busy world, data analysis plays a role in making decisions more scientific and helping for creating the content for the bloggers.

5.3 TRAINED DATA:

Training data is the dataset for use to train your algorithm or model so it can accurately predict your outcome. Validation data is used to access and inform the choice of algorithm and parameters of the model that are building. Test data is used to measure the accuracy and efficiency of the algorithm used to train the machine to see how well it can predict new answers based on its training.



Fid 1.4 It can analysis the trained and tested data through algorithms

5.4 TESTED DATA:

Test data refers to the process of creating and maintaining values for testing with the intention of using the model for testing purposes. It consists of creating artificially or representative data to validate the functionality, performance, security, and various other aspects of the software. There are several reasons why creating test data is crucial it should match the test environment and be relevant to the testing at present. Not all data is suitable for every type of testing. Therefore, it's essential to generate data that is relevant and beneficial for the specific testing objectives.

By using test data creation automation tools
Migrating existing data from production to the testing environment

5.5 PREDICTION FOR THE BLOG:

The output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome, such as whether or product dissatisfaction. The algorithm predicts likely values for unknown variables in each record of the new data, enabling the model builder to determine the most probable outcome.

6.CONCLUSION:

The mentioned paper outlines a machine learning algorithm and model aimed at analyzing and predicting blog poster. The authors employed a random forest algorithm and decision-tree model, specifically utilizing Python 3.6, to establish a random forest algorithm and decision-tree model. In this model, that can analysis the historical dataset after that trained the supervised dataset and tested that for the prediction . The paper discusses the application of the analysis the tested data through the voting in random forest algorithm and decision- tree model techniques for data cleansing, for training the ensemble learning model also used . The simultaneous analysis of historical and current data enables the prediction of future values. The ultimate outcome of the study successfully guided the company in dynamically adjusting the poster and content creating strategies for a blog based on future value predictions. This not only provides significant commercial value for the company but also establishes a crucial theoretical foundation that can benefit other companies or organisation, communities.

REFERENCE:

1. Random Forest:
- Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
- The original paper by Leo Breiman introducing the Random Forest algorithm, which demonstrates its effectiveness in classification and regression tasks.

Scikitlearn Documentation:

[RandomForests}(https://scikitlearn.org/stable/modules/ensemble.html#)forest Scikit-learn is a popular Python library for machine learning. Their documentation provides detailed information on how to implement and use Random Forests for classification and regression tasks.

- "Random Forests" Chapter in "Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman.

- This book is a comprehensive resource for various machine learning algorithms. The chapter on Random Forests offers in-depth insights into the algorithm's theory and applications.

2. Decision Tree:
- Quinlan, J. R. (1986). "Induction of decision trees." Machine learning, 1(1), 81-106.
- This seminal paper by J. R. Quinlan introduces the decision tree algorithm and discusses its application in machine learning for classification and regression tasks.

- Scikit-learn Documentation: [Decision Trees](https://scikit-learn.org/stable/modules/tree.html)
- Similar to Random Forests, the Scikit-learn documentation provides a comprehensive guide on implementing and using Decision Trees for various machine learning tasks.

- "Decision Trees" Chapter in "Data Mining: Concepts and Techniques" by Jiawei Han, Micheline Kamber, and Jian Pei.
- This book offers a thorough overview of data mining techniques, including Decision Trees. The chapter provides both theoretical understanding and practical applications of Decision Trees in data analysis.

3. Ensemble Learning:
- "Ensemble Methods in Machine Learning" by Thomas G. Dietterich. This paper provides an overview of various ensemble methods, their theoretical foundations, and practical applications.

- "An Introduction to Statistical Learning" by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Chapter 8 of this book covers ensemble methods, providing both theoretical understanding and practical examples using R.

- Scikit-learn Documentation: Ensemble Methods. The documentation of the Scikit-learn library offers detailed explanations and examples of various ensemble methods implemented in Python.

- "Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. This book offers a comprehensive treatment of ensemble methods, including bagging, boosting, and random forests, along with theoretical insights and practical examples.