IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

DATA CATEGORIZATION IN CLOUD COMPUTING USING XGBOOST ALGORITHM

[1,2,3,4]Kayalvizhi S, Jaffer Ali Akbar Ali, Sudha Senthil Kumar, Gokul A
[1,4]Department of Computer Science and Engineering
[1]Excel Engineering College, Namakkal
[4]Paavai College of Engineering, Namakkal
[2,3]University of Technology and Applied Science, Sohar

Abstract:

In cloud computing, efficient data categorization is crucial for optimizing storage, retrieval, and security. Machine learning techniques, particularly ensemble methods like XGBoost (Extreme Gradient Boosting), have proven effective in handling large-scale data classification tasks. This study explores the application of the XGBoost algorithm for data categorization in cloud environments, leveraging its ability to handle high-dimensional data, manage missing values, and provide high accuracy with minimal computational cost.The proposed model preprocesses cloud-stored data by extracting relevant features, encoding categorical values, and handling imbalances. XGBoost is then trained and fine-tuned using hyperparameter optimization to classify data into predefined categories. The performance is evaluated using metrics such as accuracy, precision, recall, and F1-score, comparing it against other traditional machine learning classifiers.Results indicate that XGBoost outperforms conventional algorithms like Decision Trees, Random Forests, and Support Vector Machines (SVM) in terms of speed and accuracy. Additionally, its scalability makes it well-suited for cloud-based environments where real-time data processing is required.The proposed model preprocesses cloud-stored data by extracting relevant features, encoding categorical values, and handling imbalances. XGBoost is then trained and fine-tuned using hyperparameter optimization to classify data into predefined categories. The performance is evaluated using metrics such as accuracy, precision, recall, and F1-score, comparing it against other traditional machine learning classifiers.

Key Words: Cloud computing, data, categorization, XGBoost

1. Intoduction

Cloud computing (CC) is the most disturbing platform that involves of one-of-a-kind superior degree technological functions the place a man or woman can store, retrieve and securely save their non-public documents. Users can get right of entry to cloud services and functions the use of mobile gadgets such as mobile, laptop computer and android phones. As a long way as the protection of the archives is concern, cloud computing can be reflect onconsideration on as most dependable and impervious platform. Whereas, different storage gadgets like cell telephones and laptops may additionally unable to product such impenetrable platform for records storage due to lack of battery storage, overall performance and storage potential [1, 3]. Cloud storage offerings are regularly used to save and backup arbitrary records in a cost-effective, consumer pleasant and in quickly reachable manner [1]. They additionally supply the facility of information sharing and system synchronization in simpler way.

There is no fashionable set of attributes for cloud storage architecture, and countless cloud storage structure schemes exist throughout exceptional cloud storage systems. However, to grant cloud storage offerings to clients, cloud storage often consists of heaps of storage units clustered collectively and linked by way of a disbursed le system, a network, and different storage middleware.

Cloud storage consists of a storage useful resource pool, carrier stage agreements (SLAs), a allotted le system, and offerings interfaces. The most important intention of cloud storage designs is to furnish on-demand storage in a multi-tanned, particularly scalable manner. A frequent cloud storage diagram consists of a the front stop with an API for storage get admission .

Figure 1 gives a high-level picture of NIST's cloud computing standard architecture [3]. Five major actors Cloud customer, cloud supplier, cloud carrier, cloud auditor, and cloud broker de ned by the architecture. A individual or a company that engages in a transaction or process and/or ful ls responsibilities is referred to as an actor in CC. The actors identi ed in the NIST cloud computing standard architecture are summarized in Table 1 [3]. Security is a component of the Cloud Provider in this case.

Table 1
Actors in NIST Cloud Computing reference architecture adopted from [26]

Actor	Definition
Cloud Consumer	A person or organization that maintains a business relationship with, and uses service from, Cloud Provider.
Cloud Provider	A Person, organization, or entity responsible for making a service available to interested parties.
Cloud Auditor	A party that can conduct independent assessment of cloud services, information system operations, performance and security of the cloud implementation.
Cloud Broker	An entity that manages the use, performance and delivery of cloud services, and negotiates relationships between Cloud Providers and Cloud Consumers.
Cloud Carrier	An intermediary that provide connectivity and transport of cloud services from Cloud Providers to Cloud Consumers.

The information on the cloud is saved randomly. As the quantity of statistics increases, the person has extra for looking that data. Conversely, if the facts is saved in prepare form, the consumer will be in a position to effortlessly get right of entry to the required data. Therefore, there is a want to strengthen a mannequin that approves statistics to be effortlessly saved on cloud prepared form. The benefits of classication are stated below:

* Ensures correct classification with fewer false positives through using compound phrase search.

* Has an index, permitting you to search for touchy phrases barring having to re-crawl your facts storage.

* Comes witha taxonomy supervisor that lets in you to tailor your classi cation parameters.

* Offers strategies for automating duties like transferring touchy statistics from public sharing.

* Supports on-premises and cloud content material sources, as nicely as structured and unstructured data.

Users are hesitant to upload personal and con dential les to the internet storage because they are concerned that the service provider may misuse them. They are also concerned that their data may be hacked and compromised as a result of the widespread adoption of effective cloud storage attacks [2]. Existing cloud based architectures use the same key length to encrypt all data, which may be infeasible, and do not consider the data's level of secrecy. Treating low and high secret data the same creates extra overhead and slows processing.

As a result of the previous facts, this study concentrates on three key components of cloud computing: data sensitivity, automatic classi cation, and high-level accuracy. Before the transmission and storage activities, we provide an effective system. Which ensures automatic data classi cation by using machine learning algorithms to maintain con dentiality and integrity in cloud storage. However, on the fact that data con dentiality is particularly essential in the cloud environment. Moreover, this framework will reduce manual efforts for classi cation and will achieve high-level accuracy rate.

Following is the organization of remaining of this paper. Related work is mentioned in the second portion. The proposed work is presented in the third part. Experiments are presented in the fourth part. Results and discussions are discussed in fth and future direction are mentioned in the part sixth, the evidence has been achieved with coming analysis directions.

2 Related Works

In [4] digital signatures have been optimized to beautify security, whilst the RSA approach used to be used to invulnerable the con dentiality element of security. The encryption manner is carried out in ve steps. The rst step consists of key generation. A digital signature is carried out in step two After that encryption and decryption are carried out in steps three and four In step five signature verification is performed. [5] Proposed an structure that consists of digital signatures and exchanges Di e Hellman keys with the Advanced Encryption Standard (AES) encryption approach to tightly closed the con dentiality of information saved in the cloud. The Di e Hellman key trade facility renders the key in transit ineffective even if it is compromised due to the fact it is vain barring the user's non-public key, which is solely on hand to approved users. Hackers will have a tough time breaking into the protection mechanism due to the fact of the architecture's three-way approach, which protects facts saved in the cloud.

Using the creational set of tests de ned at each node or branch decision tree, the training dataset is recursively partitioned into smaller sub-divisions [20]. This feature has one value on each branch descending from the node, and each node of the tree represents a test of a feature from the training dataset. The dataset is categorized by starting with the root node and then testing each characteristic. Then, according to the value of the feature in the given dataset, going down the tree branch, and this method is repeated recursively [17].

3 Proposed Work

This lookup article analyzed the have an effect on of a variety of computing device gaining knowledge of records categorization methods such as the NB, RF, SVM, and KNN algorithms. Data categorization is carried out on sensitivity tiers for the cloud. Our proposed mannequin consisting of three classes: Basic Class, Con dential Class, and Highly Con dential Class, as proven in Figure 2.

3.1 Basic class

Our proposed model's simple type consists of a frequent kind of data, such as textual content documents, with a low degree of con dentiality. Basic data such as advertising, announcements, and notices can be determined in textual content documents. As a result, this degree offers a fundamental degree of facts security. The primary classification does now not require encryption on the patron side; nevertheless, when sent, it will be encoded on the server-side the use of the backup service's key.

3.2 Confidential Class

Personal les, such as personal accounts, internet accounts, and expert details, are protected in this class. Our con dential category is meant for statistics with a medium stage of con dentiality. Security is indispensable to defend our records due to the fact this category continues music of secret and personal information. At the con dential level, encryption strategies such as AES can be utilized for this purpose. In this class, encryption will be used on the client-side.

3.3 Dataset

We have collected the Reuters-21578 text categorization collection dataset from the UCI ML repository. We also collect con dential and highly con dential data from the CIA public library for text composition to test the recommended system. The compatible material like commercials, announcements, news articles data, accounts information documents (of the organization) and military information are compiled from the different mentioned international repository.

The datasets generated at some stage in and/or analyzed throughout the modern-day find out about are on hand in the [UCI, CIA] repositories (Access hyperlinks are supplied in Table 2).

Table 2
Dataset with corresponding repositories

S.NO	Dataset Type	Number of Dataset	Link document
1	Public and Confidential Data	4000	https://miguelmalvarez.com/2015/03/20/classifyingreuters-21578-collection-with-python-representing-the-data/
2	Highly Confidential Data	2010	https://www.archives.gov/research/intelligence/cia

3.4 Data Processing

Natural language processing is a technique of analyzing, manipulating, and extracting which means from human language in such a way that computer systems can apprehend it. Before the textual content enter is despatched to the algorithm, it is modified the usage of the NLTK library. The unstructured textual content statistics is as a result changed into a structured format. Many computer studying methods rely on processing as a signi cant component. It has a substantive affect on the classification system as nicely [15].

3.4.1 Tokenization

Tokenization is the technique of breaking down a personality association into components, every represents a phrase or phrase. In natural language processing, there are two kinds of tokenization: phrase tokenization and sentence tokenization. The listing of tokens, which can be a phrase or a phrase, is then utilized to technique the records [15]. Fig. 3 suggests the tokenization process.

3.4.2 Filtering

Filtering a text le is a common practice to remove some of the more inconsequential terms. A reciprocal ltering mechanism prevents the removal of words. Stop words are terms that regularly appear in text that lacks substantive information (for example relational words, conjunctions, and so on.).

Thus, words that appear frequently in the content are said to have insu cient information to distinguish between reports, whereas terms that appear infrequently may similarly be of low signicance and can be eliminated from the content document [17].

3.5 Feature Extraction

Before being fed into the classifer, the facts from the textual content record is represented in indexes. Words may want to be used to describe features. The Bag of Words technique, which represents the file as a series of words, is a many times used structure. We have to rst de ne various phrases and variables that will be typically used in the following to enable for formal or formal descriptions of characteristic extraction. If there are wonderful phrases or phrases in a set of archives D=d1,d2,..., dD, then V=w1,w2,...,wv exists, then V is regarded as the vocabulary [18]. fd(w) represents the prevalence of the phrase w V in the record dD, and fD represents the quantity of archives containing the phrase w. (w). tD = (fd(w1), fd(w2),..., fd(wv) represents the characteristic vector for file t.
Algorithm for developing a XG model
Step 1: Creating the XG Boost model
Step 2: word2count = {}
Step 3: for facts in dataset:
Step 4: phrases = nltk.word_tokenize(data) Step 5: for phrase in words:
Step 6: if phrase no longer in word2count.keys():
Step 7: word2count[word] = 1 Step 8: else:
Step 9: word2count[word] += 1

There are two general approaches for symbolizing a document using a list of features, namely the local dictionary technique and the global dictionary methodology [13, 18]. The international dictionary will be built using just relevant texts. As a result, if a term appears in the relevant document, it can be added to the lexicon as a feature. As far as the local dictionary technique is concerned, it can produce better results [19].

3.6 Feature Vector

Transforming files into numeric vectors is the most usual strategy to characterize them. The “Vector Space Model” is some other identify for this demonstration. Its structure, on the different hand, is easy and used to be designed with records retrieval (IR) and indexing in consideration. The vector house mannequin is widely used in more than a few textual content mining methods and IR classifications, and it permits for wise evaluation of a big quantity of archives [19]. Each phrase in VSM is identified by means of a numeric range that the word's weight or 'importance' in the document. The two fundamental function weight fashions is the Boolean model. If a function is existing in the document, it has a weight of 1; otherwise, it has a weight of zero if it is now not existing in the document.

The weight assigned to every phrase w D is derived as follows the usage of the TF weighting approach as an example:

$$f(w) = fd(w) * \frac{\log|D|}{fD(w)} \quad (1)$$

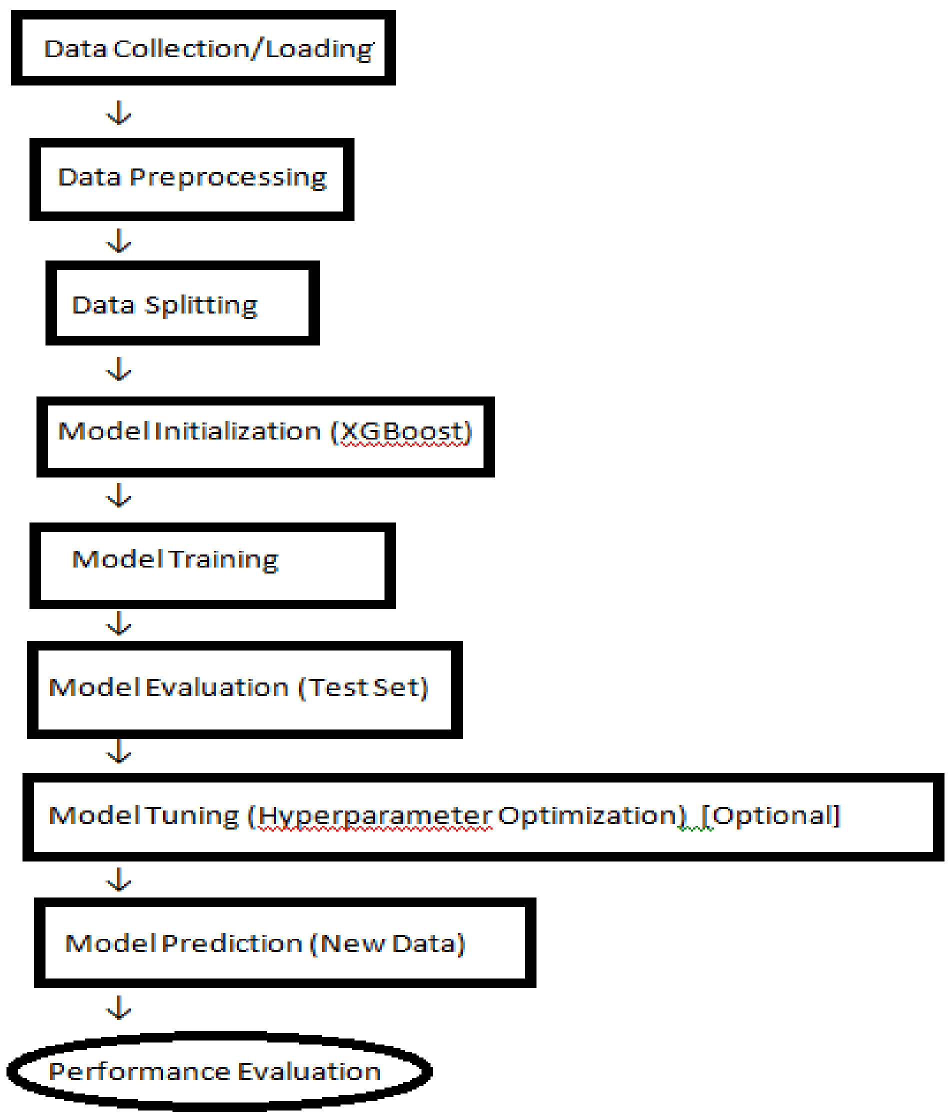
Algorithm for T df Vectorizer to calculate tf-idf score

```
import os
import csv
from nltk.tokenize import RegexpTokenizer
from nltk.stem import PorterStemmer
from sklearn.feature_extraction.text import
TfidfVectorizer

# Initialize tokenizer and stemmer
tokenizer = RegexpTokenizer(r'\w+')
stemmer = PorterStemmer()
# Path to dataset
path = "E:/gokul/dataset/"
# List to hold documents
docs = []
# Read all documents from the directory
for subdir, dirs, files in os.walk(path):
 for file in files:
file_path = os.path.join(subdir, file)  # Get full file path
        with open(file_path, encoding="latin-1") as f:
            text = f.read()
docs.append(text)
# Function for stemming tokens
def stem_tokens(tokens, stemmer):
    return [stemmer.stem(item) for item in tokens]
# Function for tokenizing and stemming text
def tokenize(text):
    tokens = tokenizer.tokenize(text)
    return stem_tokens(tokens, stemmer)
# Initialize the TF-IDF Vectorizer

vectorizer = TfidfVectorizer(tokenizer=tokenize,
stop_words='english')
# Fit the vectorizer and transform the documents into a
document-term matrix
DocumentVectorizerArray =
vectorizer.fit_transform(docs).toarray()
# Write the TF-IDF values to a CSV file
with open('E:/gokul/fade.csv', 'w', newline='') as f:
    writer = csv.writer(f)
    # Write header row (vocabulary terms)

writer.writerow(vectorizer.get_feature_names_out())

    # Write document vectors

    for doc_vector in DocumentVectorizerArray:

writer.writerow(doc_vector)
```

3.7 Sensitivity Base Classification

Classification is a technique of supervised learning. Which classifier ought to be used to predict the type label on new facts whilst additionally studying from the coaching data? It's used in a range of disciplines, together with clinical diagnosis, photograph processing, file management, and textual content classification. It is additionally taken into consideration in a range of communities, along with computer learning, database, IR, and records mining [23].The necessary purpose of classification is to assign specified classifications to textual content files [17]. The following is a clear definition of the classification problem. We want a D = {d1,d2,.. .,dn} education set of documents, so that every record di is related with a label ℓi from the series L = {ℓ1,ℓ2,. .,ℓ3

Figure 1: Control Sequence



3.8 Training Module

In the education section preprocessing is carried out the use of NLP and acquired facets automatically. So extracted facets are used for prediction by way of making use of exceptional classification algorithms. The anticipated output is matched with the enter occasion to instruct it. The cautioned methodology takes as enter textual content archives along with basic, sensitive, and especially con dential data, and at the quit of the education phase, a nal predictive mannequin is chosen to forecast type labels. We supply 6010 textual content archives to instruct our model. KNN, NB, RF, and SVM are the 4 classifiers used to analyze the data.The class labels are a set of outputs in this module that are used to train a prediction model by combining them with features (also known as variables). To allow a machine-learning system to predict class labels, use a training model. To train the classifier and test the efficiency of the trained model, crossvalidation is performed. After being trained, the model will be able to pr edict the class label for any new text documents based on the features they include. The suggested method uses the training dataset, which was manually labelled with class labels, to create a classier.

3.9 Testing Module

The educated mannequin is evaluated by means of the use of a new set of checking out information to see how efficiently our mannequin was once trained. A new dataset of textual content archives used to be used as enter in the trying out phase, and the new textual content archives have been pre-processed. For pre-processing, the studying objects have been tokenized into phrases primarily based on the homes they contained. The most modern instances have been loaded into the skilled mannequin at the end. Which expected the text's category label precisely to consider the classifiers, researchers employed about 2030 checking out datasets from quite a number publications. This is pretty beneficial to the trying out module.

3.10 Development of Prototyping

The proposed methodology's aim is to assemble the application's modules such that they can be validated. The system's lower back stop used to be constructed in Python 3.7, which is a effective language to work with for facts evaluation when mixed with the right equipment and modules.

The system's returned stop was once constructed in Python 3.7, which is a effective language to work with for statistics evaluation when mixed with the right equipment and modules. Python is a free and open-source programming language designed to be easy to study and powerful. We used a range of Python libraries, consisting of nltk, sklearn, numpy, os, csv, and scipy. These libraries are used to preprocess data, extract facets automatically, assemble function vectors, produce dataset csv les, and then supply the csv les to the classi er to instruct and take a look at the models.

4.Experiment

This part includes a range of experiments that are carried out to verify the proposed work on computerized facts classification. RF, NB, KNN, and SVM are 4 sorts of classifiers used in experiments to gain accuracy on a given dataset.We need to rst de ne quite a few phrases and variables that will be oftentimes used in the following to enable for formal or formal descriptions of function extraction. Given a set of archives D = {d1,d2,...,dD}, and the set of a variety of phrases or phrases in the set V= {w1,w2,...,wv},V is referred to as the vocabulary. fd(w) represents the prevalence of the phrase wV in the report dD, and fD represents the variety of files containing the phrase w. (w). tD = (fd(w1), fd(w2),..., fd(wv) represents the characteristic vector for record t.

Python is a free and open-source programming language designed to be easy to study and powerful. We used a range of Python libraries, consisting of nltk, sklearn, numpy, os, csv, and scipy. These libraries are used to preprocess data, extract facets automatically, assemble function vectors, produce dataset csv les, and then supply the csv les to the classi er to instruct and take a look at the models.

Table 3:Data Values

abort	accept	access	account	action	adapt	address	lable
0	0	0.23514	0.65652	0.14154	0.23241	0.02154	basic class
0	0.326235	0.14545	0.23265	0	0	0	basic class
0.16464	0	0	0	0	0	0	confidential
0	0	0.32323	0	0	0.68778	0	basic class
0	0	0.65855	0	0	0	0	confidential
0	0	0	0	0	0	0.21215	confidential
0.00124	0	0.32654	0	0	0	0.87542	basic class
0	0	0	0	0.65844	0.66874	0	confidential
0	0	0.21548	0	0	0	0	basic class

4.1 Model Evaluation Metrics

The goal is used to check the classification model's performance. For this purpose, we reserved a random quantity of the labeled files take a look at set. Classify the check set and join the anticipated labels with the actual labels, as nicely as investigate the quantitative performance, after the classier has been skilled with a labelled dataset. Accuracy is de ned as the percentage of precisely classified archives to the whole range of records. Precision, recall, and F-Measure are three wellknown goal assessment or qualitative measurements for classification. The recall is an evaluation of the system's Genius in classifying essential documents, whilst the precision is an evaluation of the system's brain in figuring out inappropriate documents. F-measure will additionally be used at some stage in contrast to overcome the bias difficulty in precision and recall.

4.2 Precision

Precision is the calculation of the wide variety of archives that are precisely classified through an complete wide variety of documents.

tp numberofdocuments

$$\text{precision(p)} = \frac{tp}{tp+fp} \quad \frac{\text{number of documents}}{\text{number of label documents}} =$$

4.3 Recall

The recall is the calculation of the wide variety of retrieved archives that are significant or precisely classified by means of quite a few relevant documents.

Tp number of label documents

$$\text{Recall(r)} = \frac{tp}{tp+fn} \quad \frac{\text{number of label documents}}{\text{number of documents}} =$$

4.4 F1-Measure

In F1 Measure the calculated accuracy and calling off is used to dig out the symphonic motive amongst them. The matchless documents are 1 of F1 measure when the accurateness and precision are best in sequence and lowest when the F1 measure is 0.

$$F1 - \text{Measure(f)} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

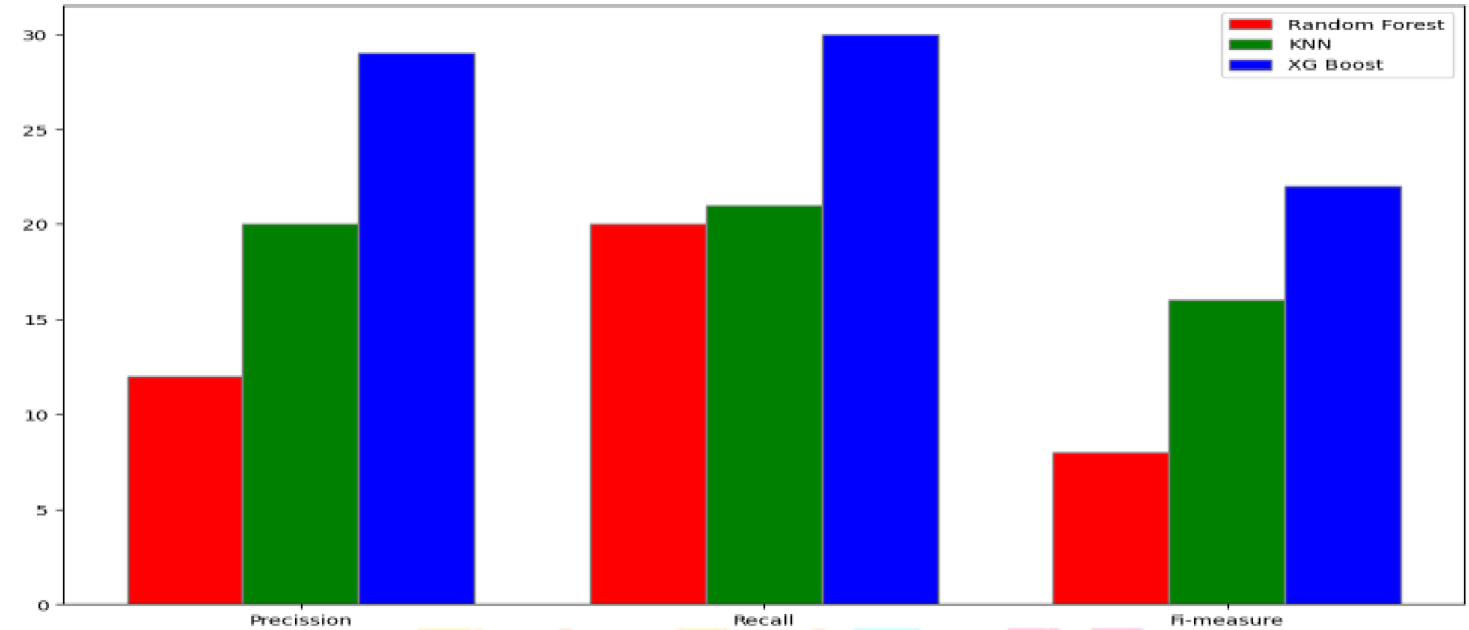


Figure 2:Feature Vector

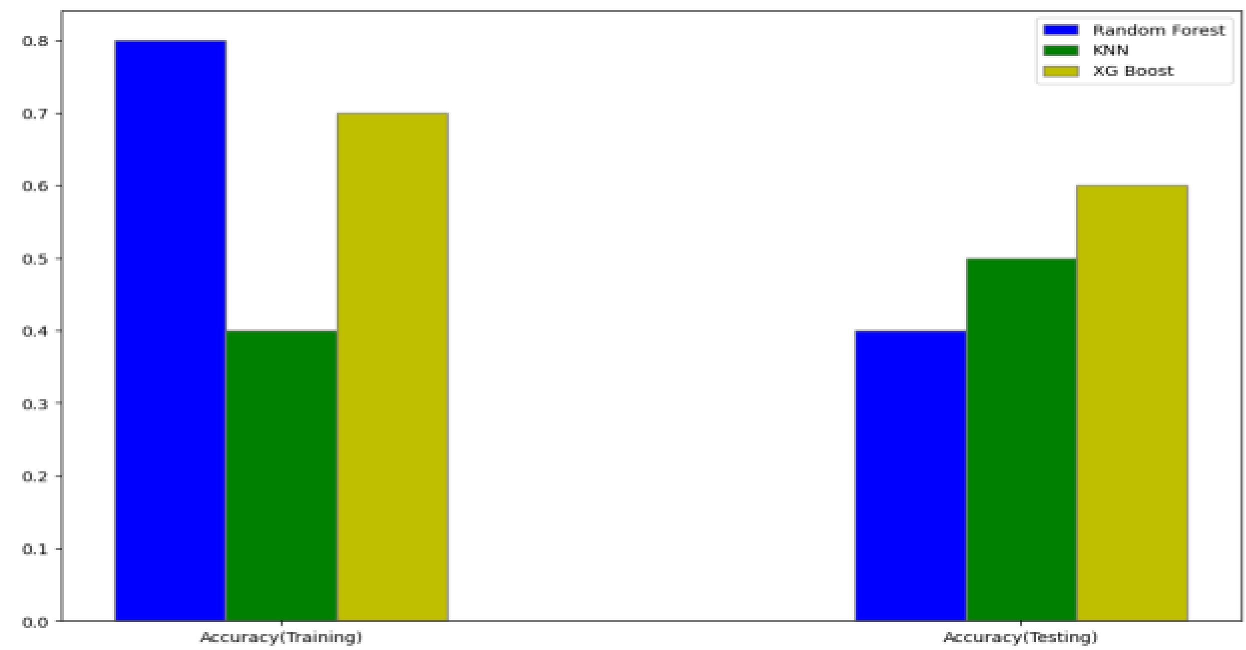


Figure 3:Perfomance Evaluation

5. Results And Discussion

The proposal's purpose is to validate the utility by means of growing modules for it. The use of Python 3.7 at the system's lower back cease is a high-quality instance of statistics evaluation with the right equipment and frameworks. The rst segment of record classification is training, observed via the 2nd segment of testing. Multiple components of the education section encompass NLP pre-processing, characteristic extraction, and function vector, whilst the prediction module consists of making use of specific classification algorithms for comparison. Figures 5 and 6 exhibit the RF, NB, KNN, and SVM algorithms the use of a number of methodologies. In the sketch above, the proposed strategy is in contrast with the current technique in phrases of performance.

6. Conclusion And Future Work

XGBoost algorithm is a powerful, flexible, and dependable laptop gaining knowledge of library for supervised and unsupervised desktop mastering tasks. It is an environment friendly implementation of the gradient boosting algorithm and can be used for both regression and classification problems. XGBoost is effortless to use and offers quite a few benefits over different laptop gaining knowledge of libraries such as quickly coaching speed, parallel computing capabilities, and superb overall performance with massive datasets. XGBoost algorithm is an magnificent desire for any computing device gaining knowledge of assignment and can be used to rapidly and precisely construct fashions that can be used in manufacturing systems.

References

1. Sun, X., Wang, Z., Wu, Y. et al. A price-aware congestion control protocol for cloud services. J Cloud Comp 10, 55 (2021). https://doi.org/10.1186/s13677-021-00271-5
2. Al-Said Ahmad, A., Andras, P. Scalability resilience framework using application-level fault injection for cloud-based software services. J Cloud Comp 11, 1 (2022). https://doi.org/10.1186/s13677-02100277-z.
3. Song, D., E. Shi, I. Fischer and U. Shankar, “Cloud data protection for the masses”, IEEE Computer. Soc., Vol. 45, Issue 1, pp.39-45, 2012.
4. Somani U, Lakhani K, Mundra M. Implementing digital signature with RSA encryption algorithm to enhance the Data Security of cloud in Cloud Computing. In Parallel Distributed and Grid Computing (PDGC), 2010 1st International Conference on 2010 Oct 28 (pp. 211-216). IEEE.
5. Rewagad P, Pawar Y. Use of digital signature with Di e Hellman key exchange and AES encryption algorithm to enhance data security in cloud computing. In Communication Systems and Network Technologies (CSNT), 2013 International Conference on 2013 Apr 6 (pp. 437-439). IEEE.
6. Sinha N, Khreisat L. Cloud computing security, data, and performance issues. In2014 23rd Wireless and Optical Communication Conference (WOCC) 2014 May 9 (pp. 1-6). IEEE.
7. Diwan V, Malhotra S, Jain R. Cloud security solutions: Comparison among various cryptographic algorithms. IJARCSSE, April. 2014 Apr.