iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41354

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

Finding Feature Selection for Business Dataset using PCA on Big Data Environment using Spark

Shubham Tiwari[1], Rohit Bhansali[2], Gagan Kumar[3], Chandrashekar D K[4]
[1, 2, 3]*Graduate,* [4]*Associate Professor, SJBIT Bangalore*

Abstract: Business is important in development of any country as it contributes into the country's GDP. India has 29% of GDP and 28% of employment. Its nominal output ranks 16th in the world and the service industry ranks 15th. This project is example of business process improvement or re-engineering. The proposed project applies machine learning concepts and determines the appropriate change in tread for any business firm. Using Principle Component Analysis as a dimensionality reduction technique we have reduced number of feature to minimum. This feature reduction technique helps the model to execute efficiently and provide better results.

Index terms: Big Data, Spark, PCA, machine learning and data visualization.

I. INTRODUCTION

Data is a collection of raw facts and statistics together for reference or analysis. Data and information are often used interchangeably. Graphs, pictures, and other tools can be used to compute, aggregate, and describe data. Large amount of data is accumulated by organizations, institutions, governments etc. And data can be in a variety of formats, including text, numbers, and multimedia. Managing large amounts of data has become a time-consuming task for businesses, which can be alleviated by integrating big data. Big data is a term used for accumulation of huge and complex data sets, processing this huge amount of data is difficult by using normal Data Management technologies. Data collection, storage, search, exchange, transmission, analysis and visualization are all issues when dealing with large data.

Big data can be explained with many examples and one such example is everyday millions of data is been uploaded and in last two years 90% of the worlds data is been generated. Big data is like data but in large in size, which contain both structured and unstructured data. The 3V's in Big data refers to variety, volume and velocity. Variety describes the size of data and velocity describes the speed of data. Presently data is increased rapidly because if the increasing large number of mobile devices, aerial, camera etc. by increasing accuracy efficiency can be increased thereby reducing the cost and risk. Big data can process huge amount of data and its data size increases in peta byte range. MapReduce is a compiled language which allows parallel and distributed processing of large amounts of data. MapReduce code has lower level abstraction which makes it a complex programming model. MapReduce has more lines of codes as it has two functions in it, both map and reduce function works together which makes it technically complex. The different steps in MapReduce program is splitting of data then mapping of data after mapping the data sets are shuffled then it is reduced. After reducing the result is obtained which gives the list of data and its count. Code performance of MapReduce code is high but is complex to use since it has more functions. MapReduce is suitable for complex business data and logic. It can be used for both structured and unstructured data. MapReduce software framework used for large datasets. MapReduce is named from the two phases it consists of: the map phase and the reduce phase. In each part of MapReduce, both the input and output are in the form of a key. The input will be separated into key pair values before being sent to mapreduce. Whenever a data is passed MapReduce will generate the value of the most recent key pair. Every different key value is subjected to the reduction technique. The reduce part generates one key value pair for each different key. The final output is a key value pair. MapReduce will process data in the same way that an input file is processed.

Apache Spark is a universal, distributed cluster computing framework that is open sourced. It's the Provan interface, which allows you to program large clusters with implicit data parallelism and fault tolerance. It was built on a Resilient Distributed Data architectural foundation to address issues with MapReduce cluster computing. Spark necessitates the use of a cluster manager as well as a distributed storage system. The aim of spark is for extensibility, speed and interactive analytics. Spark is usually used in huge dataset where the general problem for other platforms is execution speed but, we can also use spark for pseudo-distributed mode. By using spark application framework, it simplifies the analysis.

A. Advantages of Apache Spark:

1) *Speed:* Dealing with speed dependably matters when it comes to Big Data. Because of its speed, Apache Spark is very popular among data analysts. For gigantic extension data planning, Spark is 100x quicker than Hadoop. Spark stores data in an in-memory enumeration structure, whereas Hadoop stores data in a nearby memory region. Spark is capable of handling many petabytes of gathered data from over 8000 centre points in real time.
2) *Multilingual*: Spark has a wide range of usability and can be used in any programming applications such as Java, R, Python, Scala, MySQL. It even provides the interactivity between any of the shells.
3) *Ease of Use*: Apache Spark passes on easy to-use APIs for dealing with tremendous datasets. It offers in excess of 80 verifiable level executives that simplify it to manufacture equivalent applications.

Principal Component Analysis is one of the dimension reduction techniques. It is the process of finding the key components in the dataset. Every Data Scientist will prefer huge datasets as it will provide the better accuracy for the model but, with huge dataset the complexity of the program increases and hence resulting in the drop of efficiency. In these huge dataset not all components will be useful for the model and only few components are required to get the results. The best method to avoid these unnecessary components is by applying a Dimension Reduction Technique. PCA is one of the reduction techniques. The purpose of PCA is to find and discover relationships between variables. If there is a significant amount of correlation and its found then you could reduce the dimensionality. The PCA algorithm's major functions are to normalize data, get eigenvectors and eigenvalues, and then sort the eigenvalues in descending order. To transform the original dataset, construct the projection matrix W from the selected K eigenvectors. PCA is similar to fitting a p-dimensional ellipsoid to data, with each axis representing a primary component. When one of the ellipsoid's axes is small, the variation along that axis is minimal as well. To find the ellipsoid's axes,
we ought to first dispose of the mean of every variable from the dataset, that allows you to center the information across the foundation. The Eigen values and accompanying Eigen vectors of the facts covariance grid are then calculated. To convert the orthogonal Eigen vectors to unit vectors, we ought to first normalize them. After that, each of the unit Eigen vectors may be read as an axis of the ellipsoid suited for the facts. Our covariance matrix will be diagonalised as a result of this choice of basis, with the diagonal entries indicating each axis' variance. By differentiating the Eigen value corresponding to that eigen vector by the sum of all Eigen values, you can calculate the proportion of variance that each eigenvector represents. PCA is an issue of unsupervised learning. The entire procedure for extracting primary components from any dataset may be broken down into six parts. To make our new dataset d dimensional, select the entire dataset that has d+1 dimensions and ignore the label. Calculate the mean for each dataset dimension. Calculating the dataset's covariance matrix. Calculate the eigenvector and its eigenvalues.

$$\text{Cov(X,Y)} = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - x)(Y_i - y)$$

The eigenvalues can be calculated using the below formula:

$$\det(A - \vartheta I) = 0$$

Feature selection, additionally referred to as variable choice, characteristic choice, or variable subset choice, is the method of selecting a subset of crucial traits (variables, predictors) to be used in model production. techniques for feature selection are employed for a spread of motives:

1) Model simplification to lead them to less complicated for researchers/users to interpret.
2) Quicker training instances.
3) To keep away from the dimensionality curse.
4) To make data more compatible with a learning model class.

A characteristic choice method accommodates of a seek approach for recommending new function subsets and a scoring metric for the numerous characteristic subsets. The most effective basic set of rules involves assessing every feasible subset of functions and selecting the only with the lowest blunders charge. All however the tiniest characteristic sets are computationally intractable in an exhaustive seek of the gap. The selection of assessment metric has a extensive impact on the technique, and it's far those assessment metrics that distinguish among the three fundamental categories of capabilities, particularly wrappers, filters, and embedding techniques. When it comes to dealing with massive datasets and running sophisticated models, Google Colab is a lifesaver for data scientists. Google's Colab is built on Jupyter Notebook, a dynamic application that takes advantage of Google Docs functionalities. We don't need to install anything locally because it runs on Google's server, whether those are Spark or a deep learning model. The free GPU and TPU support is one of Colab's most appealing features. Because the GPU support runs on Google's own server, it is really quicker than some other alternatives.

Cloud computing refers back to the on-call for availability of computer device sources, such as statistics storage (cloud storage) and computational strength, with out the consumer having to actively manipulate them. Large clouds commonly divide functions across multiple locations, each of which serves as a data center. Cloud computing relies on sharing resources to achieve coherence and scale economies. The purpose of computing on cloud is to enable consumers to take benefit of all of those technology with out requiring tremendous understanding or enjoy in every one. The cloud pursuits to cut charges via allowing clients to pay attention on their core business in preference to being hampered with the aid of IT problems. Cloud computing is primarily enabled through virtualization. Virtualization software splits a physical computer into one or more "virtual" computers that can be used and managed separately to fulfill computing tasks. Virtualization at the operating system level can be utilized to better allocate and use idle computing resources, which essentially creates a scalable system of several separate Computing machines. Virtualization delivers the flexibility needed to accelerate IT processes while also lowering costs by Maximizing infrastructure use.

II. LITERATURE SURVEY

Swati A Sonawale [1] proposed a model with a selection algorithm for reduction elimination. In this model she compressed high dimensional data to low level to get accurate results. Tarun Kumar Gupta proposed a model which uses Meta Heuristics for dimensionality reduction by this technique he reduces space and increases performance for better classification. K Sudhakar uses Data Mining techniques and Neural Networks to predict heart disorder in his model. In the present model feature extraction and feature selection are used to get related data. Data cleaning is done to remove outliers then we extract critical features which have the most relevant features.

Saraswathi, V. et al.[2] derived the best solution for tumor detection using RF-PCA. Brain Tumor is one of the most dangerous diseases in the world at this moment, so identification of brain tumor at an earlier stage is very important as it can save many lives. There are generally three types of tumors meningioma, glioma, and pituitary. Meningioma is one of the most common tumors and is present 20% but they are considered as benign because of their slow architecture. Glioma is a primary brain tumor and it can be cured easily due to their lack of locomotion and spreading in the brain.The Feature Extraction is an important step as it is used to classify these different types of dataset and are able to successfully predict the types of tumor. Selvaraj et al. uses 5 features and have targeted an accuracy of 98% for normal and 96% for abnormal ROI.

Saima Farhan [3] proposed a model using Dimension Reduction, PCA. In this model complexity of high dimensions descriptors is reduced and image segmentation is done. P Aljbar, ST Roweis and M Belkin derived a nonlinear method where the images are arranged according to graphs, where the vertex of each graph represents weighted edges, then the graph is used to give a relationship between images after segmentation. The coordinates of results can be viewed as descriptors. The above models were not efficient as they were not optimized. The present model is optimized and is derived from linear and nonlinear approaches. Brain Atrophy is developed using a combination of PCA and Manifold and later is evaluated for T-test cases, this model helps to differentiate among various age groups to classify Brain Atrophy. It is highly optimized and can be used in practical situations.

David Enke [4] derived a Machine Learning algorithm where he gives the relation between variables using Neural Network. Dase RK derived a model using Neural Network, as it can be used to get large info from a large database to predict the Stock Market. Halbert White derived a combination of Neural Network and Nonlinear model to predict the Stock Market. Debashish Das et al derived a model using the Neural Network and Data Mining approach. The model can efficiently predict the Stock Market. Phichhang ou et al proposed a model with 10 different DM techniques to predict the Hong Kong Stock Market. In the present model PCA is used to reduce the dimension of datasets/databases. Drawback of this is it can affect the prediction of the Stock Market.The model has been tested on 3 different Stock Market Datasets and is found to predict the Stock Market efficiently.

D Ravi [5] proposed a model with a combination of two features: Shallow and Learnt. He used IOT and wearable devices to keep track of computation time which reduced the data collection effort. S Samarah proposed a model to track human activities by using Spatio Temporal Mining technique. Z Yang proposed a model where the machine is able to adapt to human interaction by their body language with the help of high level semantics. D Tao, L Jin proposed a model where he uses spectral-geometry for faster recognition of human activity. This model helps to implement assistive smart homes. In the present model we use a combination of PCA and ANN which boosts the activity recognition and is more efficient.

Qamar A et al. [6] proposed a model using SML algorithm and mining techniques for classification and analysis of tweets. R M Dumairi proposed a model for sentiment analysis using SVM, SV, BM, Naive Bayes to get accurate results. Assiri A proposed a model for sentiment analysis using NLA by using Twitter dataset on Saudi dialect. In the present model sentiment analysis is used to classify tweets. Mining techniques are used to classify tweets and these tweets are classified depending on the reviews given by customers on social media/any other platform. R programming language and ML algorithm are used to extract and classify tweets.

III. METHODOLOGY

The initial step in each suggested model is to gather appropriate datasets. The website kaggle.com was used to gather open-source datasets.

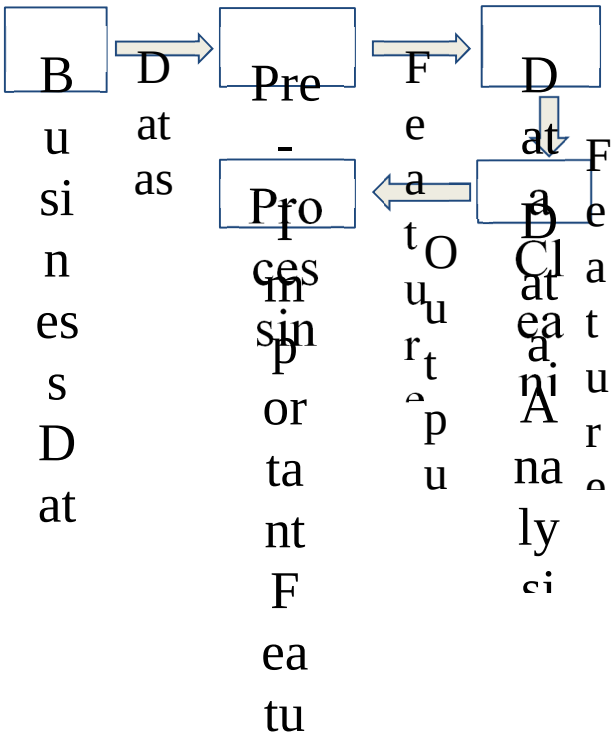


Fig. Methodology used for prediction

Preprocessing facts is an information mining method for changing raw statistics right into a layout this is both useable and efficient. The expression "junk in, rubbish out" is especially relevant in information mining and system learning tasks. Statistics gathering techniques are frequently unregulated, resulting in out-of-variety numbers, impossible facts combinations, and missing records, to name a few examples. Reading data that hasn't been fully scrutinized for such flaws can lead to incorrect results. As an end result, data representation and quality must come first before any analysis. Cleaning up the data many sections of the data may be irrelevant or missing. To deal with this, data cleaning is carried out. Features that have been processed. We can use dimension reduction or another learning model to process and classify the features.

We are employing dimensional reduction techniques such as principal component analysis in our project. Analyzing data It's a method of examining, purifying, data manipulation and modelling with the goal of unearthing beneficial information, forming conclusions, and helping choice-making. Data analysis encompasses a wide range of approaches and methodologies, as well as a variety of titles, and is utilized in a variety of business, technology, and social technological know-how regions.

A. Map Reduce PCA

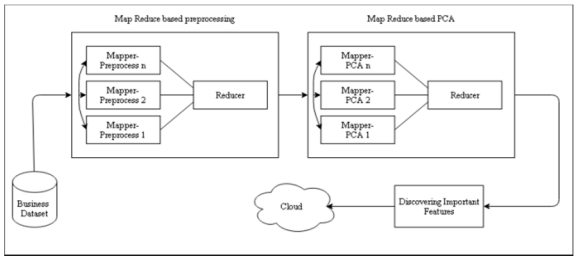


Fig. Architecture of Map Reduce PCA

Map-reduce PCA is divided into two parts: preprocessing-based map reduction and PCA-based map reduction.

B. Map Reduce based Preprocessing:

Consider a business dataset, which is often large and contains a huge number of attributes. The number of attributes in any dataset will vary, and identifying only the most significant features might be tough. Because there are *n* features, *n* mappers preprocess are created. Each mapper will accept input from the business dataset and perform parallel preprocessing. Many approaches are used in the preprocessing, including string indexing, vector assembler, and one hot encoder. The preprocessing is carried out in parallel by *n* mappers in Spark. The output of all *n* mappers will be combined into a single file by the one reducer.

C. Map Reduce based PCA:

The reducer of the preprocessing section provides input to the map reduce pca. There are five significant features in this, hence five mappers are produced. In our technique, we employ a single mapper to locate the pca feature of a given feature. We have five significant features in our example, thus we'll need five mappers to discover each one. Each mapper will be in charge of a specific feature. These mappers will work in tandem to generate pca features of key features. The next reducer receives each attribute value and checks to see if the feature values are correct. Reducer will now aggregate all of the features into a single feature, which will be constructed as a single important feature because our model has five essential features, thus five reducers will be built for each of them. Algorithm 1 Map Reduce PCA

1: Procedure: Finding important features from business dataset
2: Input: T : Business Dataset
3: Output: Important Features: This section contains the most important features that help you get better outcomes.
4: Begin
5: 1. Using pyspark, create a spark context and run.
6: 2. Map Reduce Preprocessing
7: MapperPreprocess () //input: B: Business dataset
8: { //output: Important Features
9: for each item bi ∈ B do
10. Seek each 'bi' and do the preprocessing
11. end for
12. }
13. Reducer () input: B: Mappers output
14. { //Output: Preprocessed data
15. for each item bi ∈ B do
16. P = P ∪ Bi
17. end for
18. }
19. 3. Map Reduce PCA on the Pre-Processed Data.
20: PCA () //input: P: Pre-Processed Data
21: { //output: Accumulate all important features.
22. for each item pi ∈ P do
23. pca =Seek each 'pi' and finds the correlation between each important features
24. end for
25. }
26. Reducer () // input: MG: Gained value of attributes
27. { //output: RG: PCA value
28: for each item mi ∈ MG do
29: Combine important features as a single feature, state it as pca features.
30: end for
31: }
32: 4. Discovering important features using pca.
33: 5. Generated features can be used for further execution performed on cloud.
34: END
35: End Procedure

IV. IMPLEMENTATION

For Business Dataset we have considered an open Dataset which has been referred from kaggle. The dataset which we have considered is Iowa liquor sales. According to the Iowa Department of Commerce, every business selling bottled alcohol for off-premises consumption requires a class "E" liquor licence. All alcoholic purchases made by stores that have registered with the Iowa Department of Commerce are entered into department's system, which is then made public by the state.

The brand name, kind of beverage, retail price, quantity, and address of alcoholic beverage, sales of individual containers or bundles of boxes are all included in this dataset. This dataset is quite easy, but this Gist provides further details on the data's contents.

Next step we have performed is data cleaning, You can investigate your data using a variety of statistical analysis and data visualization tools to uncover data cleaning activities you might want to execute. Before moving on to the more advanced methodologies, you should definitely undertake some fundamental data cleaning activities on any project based on machine learning. These are so basic that even seasoned machine learning practitioners forget about them. However, they are so critical that if they are overlooked, models may break or generate too optimistic performance results.

In our Project we have implemented Vector Assembler is a transformer that creates a single vector column from a list of columns. By merging raw features and features created by different feature transformers into a single feature vector, it can be used in various models like regression and trees. Vector Assembler accepts all integer types, boolean types, and vector kinds as input column types. The input values will be incorporated into a vector. Next is StringIndexer converts a column of labels in a string to a column of label indices. Multiple columns can be encoded using StringIndexer. The indices are in the format [0, numLabels], and there are four ordering choices available: “frequencyDesc” is a descending order based on label_frequency (the most common label is 0), whereas “frequency_Asc” is an ascending order based on label frequency, “alphabet_Desc” refers to descending alphabetical order, whereas “alphabet_Asc” refers to ascending alphabetical order (the default is “frequency_Desc”). The strings are further sorted by alphabet in the case of equal frequency when using "frequencyDesc"/"frequencyAsc." And Next is Large datasets are becoming more common, yet deciphering is complicated and difficult. PCA is a technique for reducing the dimensionality of such datasets and retaining as much value as possible. PCA is a data analysis technique because it makes discovering new components, or principle components, more easier when addressing an Eigen value/Eigen vector issue. It's also pliant in different manner: multiple variants of the approach has been developed for various data types and architectures. The first section of this essay will cover the essential notions of PCA, as well as what it can and cannot do. After that, it will go over a few different forms of PCA and how to use them.

V. CONCLUSION

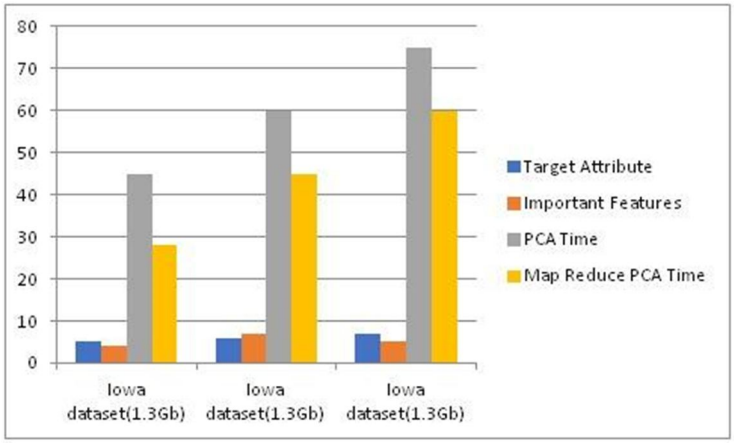
The outlook for machine learning in feature selection is very promising. In this project, the proposed systems help in data visualization and data predictions using many functions. In this model by using PCA we will be able to get the most powerful features which are necessary for the model to run more efficiently. These solutions provide a more efficient approach to analytics and forecasting. These systems are more accurate and save a lot of time.

Table 1. Performing proposed model on different sizes of datasets.

Business Dataset	PCA (Minutes)	Map-Reduce PCA (Minutes)
11MB	2	2
500 MB	20	15
1.3 GB	45	30
4.3 GB	125	106

Table 2. Selecting different important features for various target attribute.

Business Dataset	Target Attribute	Important Features	PCA (Minutes)	Map-reduce PCA (Minutes)
1.3 GB	7	5	40	30
1.3 GB	6	8	37	28
1.3 GB	8	4	45	34
1.3 GB	5	5	32	25



REFERENCES

[1] Kale, A. P., & Sonavane, S. (2018). PF-FELM: A robust PCA feature selection for fuzzy extreme learning machine. IEEE Journal of Selected Topics in Signal Processing, 12(6), 1303-1312.Konstantinos G. Liakos, Patrizia Busato , Dimitrios Moshou Simon Pearson ID and Dionysis Bochtis," Machine Learning in Agriculture: A Review", Lincoln Institute for Agri- food Technology (LIAT), University of Lincoln, Brayford Way, Brayford Pool, Lincoln LN6 7TS, UK, spearson@lincoln.ac.uk, pg4,2018

[2] Saraswathi, V., & Gupta, D. (2019, January). Classification of Brain Tumor using PCA-RF in MR Neurological Images. In 2019 11th International Conference on Communication Systems & Networks (COMSNETS) (pp. 440-443). IEEE.Amrutha, A., Lekha, R., & Sreedevi, A. (2016, December). Automatic soil nutrient detection and fertilizer dispensary system. In 2016 International Conference on Robotics: Current Trends and Future Challenges (RCTFC) (pp. 1- 5). IEEE.

[3] Alam, Saadia Binte, Ryosuke Nakano, Syoji Kobashi, and Naotake Kamiura. "Feature selection of manifold learning using principal component analysis in brain MR image." In 2015 International Conference on Informatics, Electronics & Vision (ICIEV), pp. 1-5. IEEE, 2015.

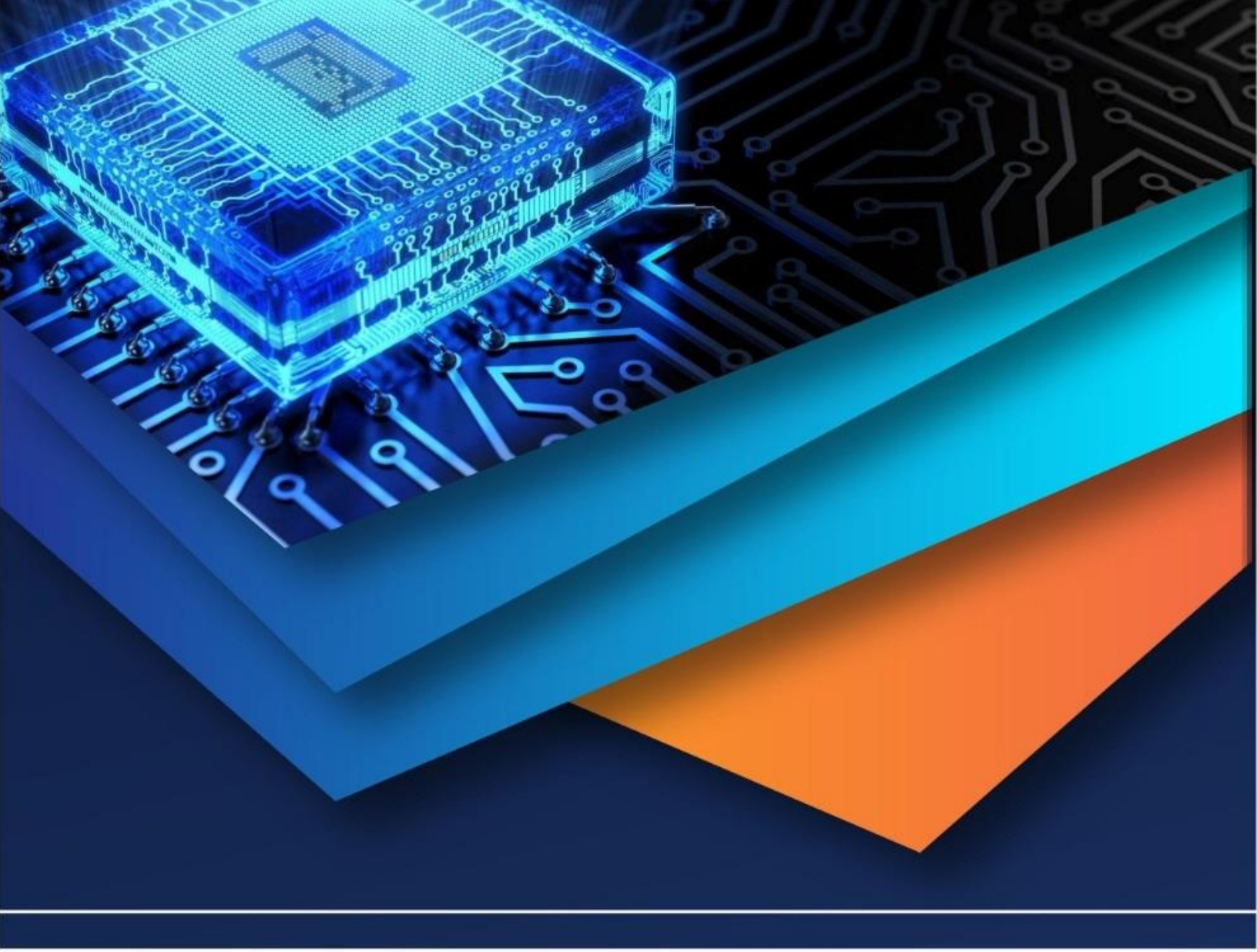
[4] Waqar, Muhammad, Hassan Dawood, Ping Guo, Muhammad Bilal Shahnawaz, and Mustansar Ali Ghazanfar. "Prediction of stock market by principal component analysis." In 2017 13th International Conference on Computational Intelligence and Security (CIS), pp. 599-602. IEEE, 2017.

[5] Kishore, Swapnil, Sayandeep Bhattacharjee, and Aleena Swetapadma. "A hybrid method for activity monitoring using principal component analysis and back-propagation neural network." In 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon), pp. 885-889. IEEE, 2017.

[6] Joy, Asif Ahmmed, and Md Al Mehedi Hasan. "A Hybrid Approach of Feature Selection and Feature Extraction for Hyperspectral Image classification." In the 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), pp. 1-4. IEEE, 2019.

[7] Mandloi, Lokesh, and Ruchi Patel. "Twitter Sentiments Analysis Using Machine Learning Methods." In 2020 International Conference for Emerging Technology (INCET), pp. 1-5. IEEE, 2020.

[8] MondherBouazizi and TomoakiOhtsuki, "A Pattern Based Approach for Multi-Class Sentiment Analysis in Twitter"-Digital Object Identifier 10.1109/ACCESS.2017.2740982, Volume 5, 2017, August 18,2017, Page 20617-20639.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY