Ilkogretim Online - Elementary Education Online, 2020; Vol 19 (Issue 4): pp. 3400-3413
http://ilkogretim-online.org
doi: : 10.17051/ilkonline.2020.04.764728

Evaluating Efficacy of Classification Algorithms on Personality Prediction Dataset

P William
Research Scholar Department of Computer Science and Engineering School of Engineering and Information Technology, MATS University, Raipur, India
Email ID: william160891@gmail.com
Dr. Abhishek Badholia
Associate Professor Department of Computer Science and Engineering School of Engineering and Information Technology, MATS University, Raipur, India
Email ID: abhibad@gmail.com

Abstract -Classification is a term that denotes to the process of labelling given input data pieces into predefined groups in machine learning. There are numerous classification algorithms focused on bayes, trees, functions, or laws that are commonly used. For a long time, the competence of these algorithmic approaches has become a big concern, attracting the attention of a sizable study group. The aim of this article is to examine the efficacy of various classification algorithms. There are several machine learning algorithms for classifying rules and performing various functions. This article compares classification algorithms used to identify and forecast personality using the HEXACO Model dataset and backs it up with implementation performance.
Keywords— *HEXACO Model, Personalityprediction, Classification, Comparative Analysis*

I. INTRODUCTION

Data mining is the method of identifying expressive, novel, and exciting correlations, trends, and patterns by the examination of vast quantities of data utilizing pattern recognition technology as well as computational and mathematical techniques [2-3]. Data mining covers more than data collection and manipulation; it also includes data prediction and interpretation. When people attempt to analyze or create associations between different features, errors often occur, rendering it difficult to find answers to specific problems. Machine learning may be beneficial in resolving these issues by improving the effectiveness of processes and the designs of computers. This article focuses on the classification issues that arise when coping with prediction [1]. A classification problem in machine learning can be thought of as an algorithmic practice for categorizing given data. A Classifier is a classification algorithm. Consider the input data as a case, and the types as classes. A vector of characteristics can be used to denote the characteristics of the instances. These attributes may take the form of ordinals, nominals, actuals, or integers..

Classification and clustering are also instances of problems involving general pattern recognition in which certain output values are allocated to specified input values. In machine learning, classification schemes obtained from observable data are first

classified according to their predictive accuracy. In reality, the clarity of a classifier is often often important. As a consequence, rule-based classifiers are more prevalent, as laws are generally simple to comprehend for humans. A person's personality pervades every part of existence. It denotes to the collection of opinions, moods, and behaviors that predict and characterize an individual's behavior and therefore has an effect on everyday life behaviors such as desires, preferences, motivations, and health [1].

Current work on recognizing personalities from social network text relies on controlled machine learning strategies applied to benchmark datasets [6], [7], and [8]. However, the primary problem is the datasets' skewness, i.e. the existence of imbalanced groups with various personality features. This concern is primarily responsible for the degraded efficiency of personality recognition systems. To resolve this problem, various techniques for reducing the skewness of the dataset are possible. When applied to imbalanced datasets from various domains, these techniques have demonstrated promising performance in terms of enhanced accuracy, recall, precision, and F1-score [9-10].
Theprimary objective is to investigate the efficiency of classification algorithms for personality prediction when the HEXACO Model is used. The remaining of the paper is divided into seven parts. Section II discusses the classification algorithms that were used in the research work. The following segment III provides an outline of the HEXACO Personality Prediction Model. Section IV outlines the various efficiency metrics for dataset used in the work. Sections V and VI discuss the comparative and analytical findings. The conclusions in Section VII are taken from the experimental findings, which are supplemented by references in the following section.

II. CLASSIFICATION ALGORITHMS

Detailed study has been done on all the machine learning classifiers before evaluation of the ease of use and to get better optimum result. Summary of the detailed study given below:

Logistic Regression : logistic regression is a data-science methodology Single-trial logistic algorithm It is true when applied to binary results, but says that all predictors are uncorrelated with one another [13]. It has been used to research a correlation between a variety of factors and a particular criterion. Independent variables yield a binary result. Since all variables may be categorical or numerical, the dependent variable must stay nominal. Exactly as published:

P(Y=1|X) or P(Y=0|X)

It calculates whether or not the dependent variable Y is more likely to arise. have a positive or bad connotation (0, 1, or on a scale between). Alternatively, to identify a person in a picture (a tree, a flower, etc.), any object has a likelihood between 0 and 1

Decision Tree: It is used to categorize data using attributes and classifications. Since it is challenging to categorize, and decision trees may lead to conflicting outcomes because of the various data, we must treat small variance as important. numerical and categorical details[16].

Since it takes advantage of trained rules, a decision tree is an ideal for classification problems. it breaks data points into "tree trunks, and then branches/leaves Resulting in non-hierarch, facilitating the classification without human control

Random Forest: The forest-based estimator suits a range of decision trees to a variety of subsamples and prevents model overfitting by using the average. The subsample is almost always the same as the initial, except by replacement It is a difficult and time-consuming project[17]. Reduced overfitting and the usage of a random forest are generally more accurate in certain instances.

The random forest is a variant on the decision tree algorithm, in that it forms some kind of real-world decision trees from your training dataset and then matches the new results.It sets the distance on the data scale to connect to the closest tree. Random forests are successful since they do not force data points into groupings arbitrarily

Support Vector Machine:A description of a machine learning (ML) machine data in the form of points in space is a help vector. Following that, the new instances are ranked based on their presence in the same room [15]. Five-fold time-consuming cross validation is not included in the algorithm, so in high-dimensional spaces it makes use of a subset of training points in the judgment. it utilizes algorithms to teach and categorize data in terms of polarity degrees. To display, we'll use two tags: red and blue along with two data features: X and Y.

Then, the SVM finds the most-separating hyperplane This is only two-dimensional. Whoever is on one side of the red line gets double the blue highlighting. In addition, there are both "happy" and "bad" feelings.

The machine learning plane must have the largest distance between each tag. when databases get more sophisticated, even Complexity correlates with precision. Consider what's in 3D and apply a Z-axis to get a 2D circle Multidimensional SVMs (MLMs) make for more precise learning.

K-Nearest Neighbours: Classification using thissubset-based learning is a subset-based method since it does not aim to construct a comprehensive model. Regimentation is determined by a clear majority of the k nearest points. To do this, there is a substantial computational expense in order to compute the distance between each instance and the entire training set. The algorithm is easy to apply, can handle noisy data, and does a good job with big volumes of data.

KNN uses training data to identify the locations of relatives as a trend It can locate data next to its neighbors, as it is a classification algorithm It will be provided to the class with the highest chance of being 1. Cautious decisions are made by majority rule

Naïve Bayes: Centered on Bayes' theorem, a Naïve Bayes is a probability model that infers independence for each function as useful real-world applications, such as document recognition and spam filtering, Naïve Bayes models shine [14]. Estimation of parameters needs just a small amount of details. naïve classifiers are really fast

Naïve Bayes is used to assess whether or not a data point corresponds to a certain group. Words and phrases may be identified as being part of a name in textual processing (classification).

III. HEXACO MODEL

HEXACO is a six-dimensional depiction of the basis of the human personality. Six factors comprise the HEXACO Model: Honesty-Humility (H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O) [22]. Each aspect consists of two end points, where the characteristics can be fully defined by their extremes. In other trait taxonomy models, the HEXACO was established in a similar manner to other traits. This model shares certain traits with other trait models, leading to an expansion of trait theory in the field In the other hand, the HEXACO paradigm stands out in that it combines the honesty and integrity concepts. Figure 1 shows the HEXACO Model of personality traits.

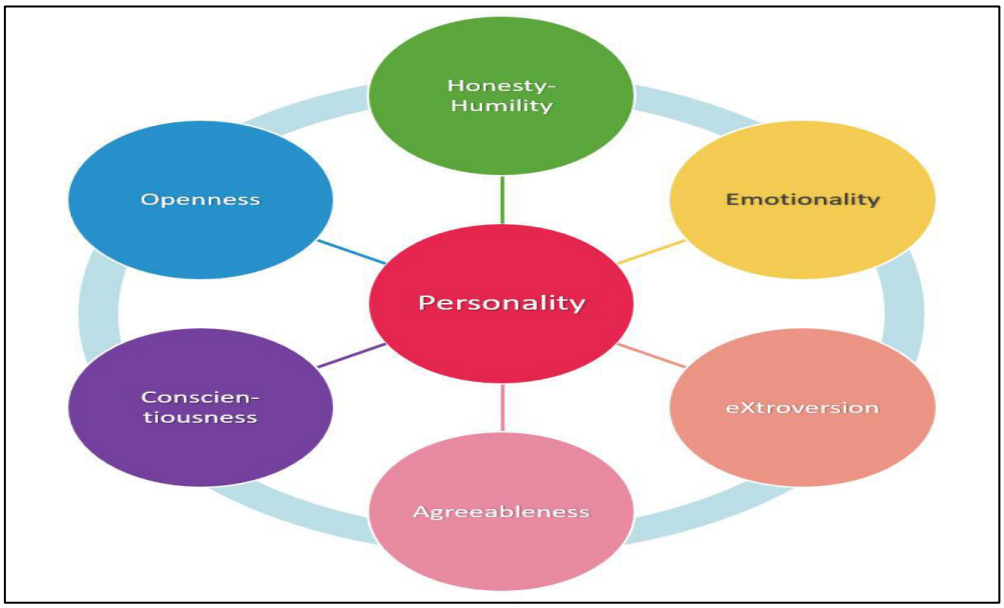


Figure 1: HEXACO - Model of Personality Trait

A brief look at six about the HEXACO personality models breaks the current rigid stereotypes of individual identity and gives us a whole new perspective. the HEXACO employs the results of several lexical studies were combined to create the model language models are also employed in the process of creating models of personality type which also been popularized by taxonomies On the basis of the research that humans use words like this, the word list process takes a broad approach to being expansive. The Factor Analytic Regression Model attempts to measure a few different aspects of a person's personality.

Quite often, self-analysis and analyst inventory testing are used to evaluate the complexity of the person's personality [20-21]. There are six main indicators that assist in the classification of an individual's language competence, each of which is measured by a set of questions. In addition, this have extra dimensions that make the study of interpersonal communications easier for researchers and can be used in a layman setting (HEXACO-PI-R). a robust HEXACO model, each of which has six aspects. there is a 25th supplementary moral facet, and this adds elements of Altruism which bring to the sum Honesty and Empathy [23-24]. Below are the six specific, subjective, and generalized stimuli that are present in each personality description, along with the adjective types that are characteristic of them:

Honesty-Humility (H):
Facets: Sincerity, Fairness, Avoidance of Greed, and Modesty
Adjectives: Sincere, trustworthy, devoted, respectful, modest/unassuming vs sneaky, deceptive, selfish, pretentious, hypocritical, boastful, pompous

Emotionality (E):
Facets: Fearfulness, Anxiety, Dependence, and Sentimentality
Adjectives: Emotional, nostalgic, scared, nervous, and fragile versus courageous, strong, independent, self-assured, and secure

Extraversion (X):
Facets: Self-esteem in social situations, social boldness, sociability, and liveliness
Adjectives: Many who are outgoing, vibrant, extraverted, sociable, talkative, enthusiastic, and active contrast with those who are shy, silent, distant, introverted, calm, and reserved

Agreeableness (A):
Facets: Forgiveness, gentility, adaptability, and patience
Adjectives: Tempered, quarrelsome, defiant, choleric versus patient, tolerant, calm, reasonable, accommodating, lenient, gentle

Conscientiousness (C):

Facets: Organization, diligence, perfectionism, and prudence are some of the facets
Adjectives: Sloppy, negligent, careless, lax, irresponsible, absent-minded versus ordered, disciplined, attentive, cautious, comprehensive, precise

Openness to Experience (O):
Facets: Appreciation of the aesthetic, inquisitiveness, creativity, and unconventionality
Adjectives: Intellectual, inventive, unconventional, and satirical vs superficial, unimaginative, and traditional

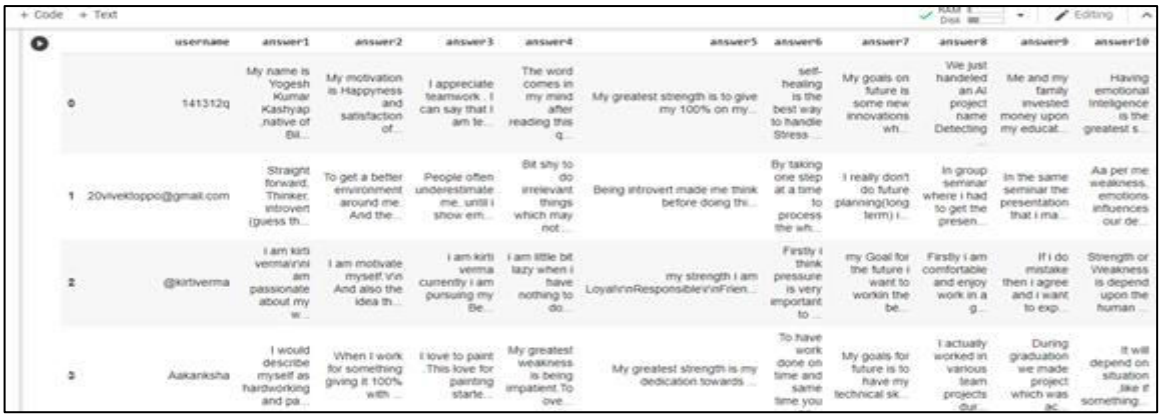
The HEXACO model developed as a consequence of researchers' ability to characterize individuals. Though not directly related to this quest, the HEXACO paradigm became well-known as a consequence of this quest and decades of effort [4-5]. Due to the difficulties of assessing personality, a formal technique was found to be appropriate, and factor analysis was selected as the solution. This introduced a new problem, as influential which characteristics to use in a factor analysis was contentious. This puzzle was solved using the lexical theory. Simply stated, this hypothesis postulates that as expressions are used to describe both high and low levels of significant personality traits in a group, words are used to characterize both high and low levels of these traits.

IV. DATA SET

The data collection used in this study is entirely dependent on the HEXACO Model personality prediction dataset obtained from a survey. The dataset consists of responses to a HEXACO Personality Test collected from a pool of career seekers or interviewees utilizing the HEXACO Factor Markers[11-12]. The HEXACO personality evaluation consists of a series of personality-related questions.

4.1 Data collection

Dataset used here for assessment is the interview answers got from student's survey interview. Datasets are the interview answers (In the form of text) got from the student's interview. 5000 student answers are recorded and processed as a dataset for personality prediction. Every student answered 12 Questions, which is considered as a dataset. Questions are completely open-ended and based on personality. Figure 2 shows the dataset processing in simulation environment, which is further combined to make a corpus.



	username	answer1	answer2	answer3	answer4	answer5	answer6	answer7	answer8	answer9	answer10
0	141312q	My name is Yogesh Kumar Kashyap ,native of Bil...	My motivation is Happyness and satisfaction of...	I appreciate teamwork . I can say that I am te...	The word comes in my mind after reading this q...	My greatest strength is to give my 100% on my...	self-healing is the best way to handle Stress ...	My goals on future is some new innovations wh...	We just handeled an AI project name Detecting ...	Me and my family invested money upon my educat...	Having emotional intelligence is the greatest s...
1	20vivekdoppo@gmail.com	Straight forward, Thinker, introvert (guess th...	To get a better environment around me. And the...	People often underestimate me. until i show em...	Bit shy so do irrelevant things which may not ...	Being introvert made me think before doing thi...	By taking one step at a time to process the wh...	I really don't do future planning(long term) i...	In group seminar where i had to get the presen...	In the same seminar the presentation that i ma...	Aa per me weakness, emotions influences our de...
2	@kirtiverma	I am kirti verma\r\n I am passionate about my w...	I am motivate myself.\r\n And also the idea th...	I am kirti verma currently i am pursuing my Be...	I am little bit lazy when i have nothing to do...	my strength I am Loyal\r\nResponsible\r\nFrien...	Firstly i think pressure is very important to ...	my Goal for the future i want to workin the be...	Firstly i am comfortable and enjoy work in a g...	If i do mistake then i agree and i want to exp...	Strength or Weakness is depend upon the human ...
3	Aakanksha	I would describe myself as hardworking and pa...	When I work for something giving it 100% with ...	I love to paint .This love for painting starte...	My greatest weakness is being impatient.To ove...	My greatest strength is my dedication towards ...	To have work done on time and same time you ...	My goals for future is to have my technical sk...	I actually worked in various team projects dur...	During graduation we made project which was ac...	It will depend on situation ,like if something...

Figure 2: Dataset Processing in the simulation environment

4.2 Data Pre-Processing

As the text is unstructured, it is very important to refine the data in such a way by which easily machine learning can be done. The corpus of textual data consists of many absent values, digits and stop words[18-19]. For filtering the data, the text is being cleaned by eliminating the useless text and all the processed data is combined in to a data frame. Various pre-processing tasks were performed with an intention to achieve better accuracy. Some of them are Tokenization, Stop Words and Lemmatization. After pre processing all the answers are combined to make a corpus. Figure 3 shows the combined answers which is considered as a dataset corpus.

	username	answer
0	141312q	My name is Yogesh Kumar Kashyap ,native of Bil...
1	20vivekdoppo@gmail.com	Straight forward, Thinker, introvert (guess th...
2	@kirtiverma	I am kirti verma\r\n I am passionate about my w...
3	Aakanksha	I would describe myself as hardworking and pa...
4	aasthachouhan99	I would like to describe myself as Ambitious a...
...	...	...
101	Vaibhav9898	I am ambitious and driven. I thrive on challen...
102	Venkatesh101299	Iam helpful and ambitious.I always look for op...
103	Vizi18	I am ambitious and driven. I thrive on challe...
104	yashish	A lazy person who try to stay away from any ki...
105	yashvanhvaishnav	I am in always go to Right way \r\nSelf depend...

106 rows × 2 columns

Figure 3: Combined Dataset Corpus

4.3 Rule Based Decision Table

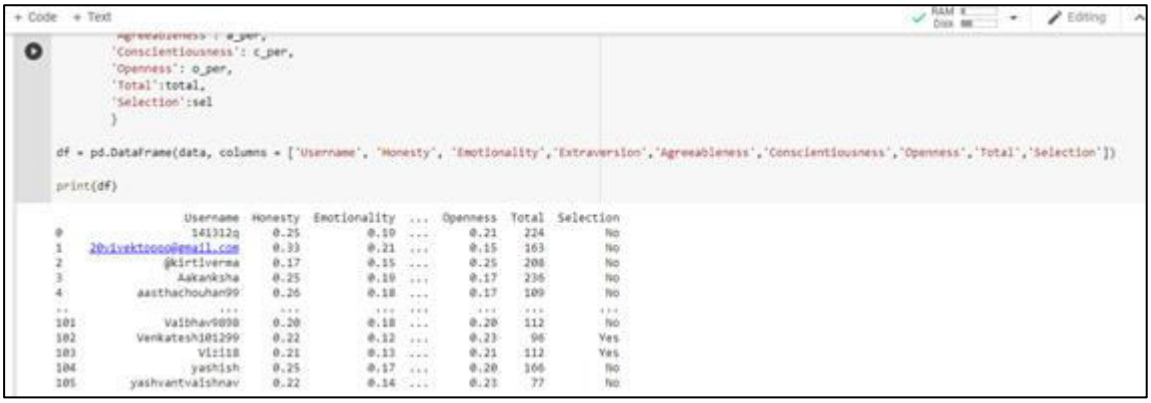
A decision table is made up of two parts:

(i) A set of attributes referred to as a schema.

(ii) The body is a set of labelled instances.

Each attribute, as well as the name, in the schema has a corresponding sense. A cell is a set of instances that all have the same meaning for a particular schema attribute. The decision table is structured similarly to a relational table, with each row representing

the mean of all documents with any possible combination of the attributes. Following the loading of the judgement table into memory, a hierarchy of tables is constructed, with each subsequent table being one level higher in the hierarchy and containing two fewer characteristics than the previous table. Finally, the top-level table features a single row containing the whole set of data. Along with columns for and attribute, a column for the total number of records is used, as is a column for a likelihood vector. Figure 4 shows the calculation of HEXACO Score where the Decision Table is applied to get the optimum result.



```
'Conscientiousness': c_per,
'Openness': o_per,
'Total':total,
'Selection':sel
}

df = pd.DataFrame(data, columns = ['Username', 'Honesty', 'Emotionality','Extraversion','Agreeableness','Conscientiousness','Openness','Total','Selection'])

print(df)
```

	Username	Honesty	Emotionality	...	Openness	Total	Selection
0	141312q	0.25	0.10	...	0.21	224	No
1	20vivektopco@gmail.com	0.33	0.21	...	0.15	163	No
2	@kirtiverma	0.17	0.15	...	0.25	208	No
3	Aakanksha	0.25	0.10	...	0.17	236	No
4	aasthachouhan99	0.26	0.18	...	0.17	109	No
..	...	...	...	...	...	...	...
101	Vaibhav9898	0.20	0.18	...	0.20	112	No
102	Venkatesh101299	0.22	0.12	...	0.23	96	Yes
103	Vi1i18	0.21	0.13	...	0.21	112	Yes
104	yashish	0.25	0.17	...	0.20	166	No
105	yashvantvaishnav	0.22	0.14	...	0.23	77	No

Figure 4: Calculation of HEXACO Score from the dataset corpus

It is used to find general trends in the dataset with the same amount of data used attributes are then added to the Decision classifier [is given the same amount of used attributes for it to find general patterns] by searching the table in the opposite direction to find the data, thereby giving it a new meaning to new data which are not already provided in the original line of the table. The wrapper approach is used to decide which attributes to use in the table's decision class. choosing outdents that only marginally increase the accuracy reduces overfitting and reduces the amount of data required for constructing the decision table, thereby producing a smaller and more straightforward decision model Either the top to bottom or top to bottom or the following of attributes is done to effectively grab the market share of the target market. Each stage of a top-to-bottom system increments the number of attributes. Additionally, forward picking is referred to as this. Starting from an exhaustive list of resources, a bottom to top technique is used to erase them one by one. As a consequence, this technique is often referred to as reverse elimination.

V. COMPARISON: PERFORMANCE MEASURES

The aim of this research is to compile the six most commonly used classification algorithms in Python, along with the Python code for each of them: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, and Nave Bayes. Classification is a technique that may be used on standardized or unstructured files. Classification is a process that divides data into a predetermined number of

classes. The primary objective of a classification issue is to ascertain the category/class under which new data will fall.

Below outlinedisthe process of developing a classification model:
Initialize the classifier to be used.
Train the classifier: All classifiers in scikit-learn uses a fit(X, y) method to fit themodel(training) for the given train data X and train label y.
Predict the target: Given an unlabeled observation X, the predict(X) returns the predicted label y.
Evaluate the classifier model

Classifiers' efficiency may be evaluated using a variety of metrics, including accuracy, precision, recall, F1 score, and error rate. where Accuracy refers to the model's ability to accurately predict the class mark. Precision: How often is a forecast positive value correct? Suggest: How reliable is the prediction when the actual value is positive? Both classification algorithms use the F1-Score, which is a weighted average of Precision and Recall values. As a consequence, this rating takes both false positives and negatives into consideration. F1-scores are often more valuable than precision, particularly if your class distribution is asymmetrical.The result is determined by the experimental outcome.

Machine learning classification was achieved by creating asimulation environment using multiple libraries and packages such as NumPy, NLTK, Pandas,Gensim, Seaborn, and others in Google Colab.In order to improve the accuracy of allmachine learning algorithms various libraries used based on the requirement to get improved result.Various in-depth observations about the data were gained after processing the mathematicalestimation. We gathered interview responses from 5000 students and divided them into groups.Following that, the data is classified and processed using various machine learningalgorithms by providing features extracted during the feature engineering phase. To examine the generalization of our proven model from training data to concealed data, we divided the initial dataset into training and test subsets separately.

A comparison of all of the machine learning algorithms used to complete the challenge can be found in the table below:

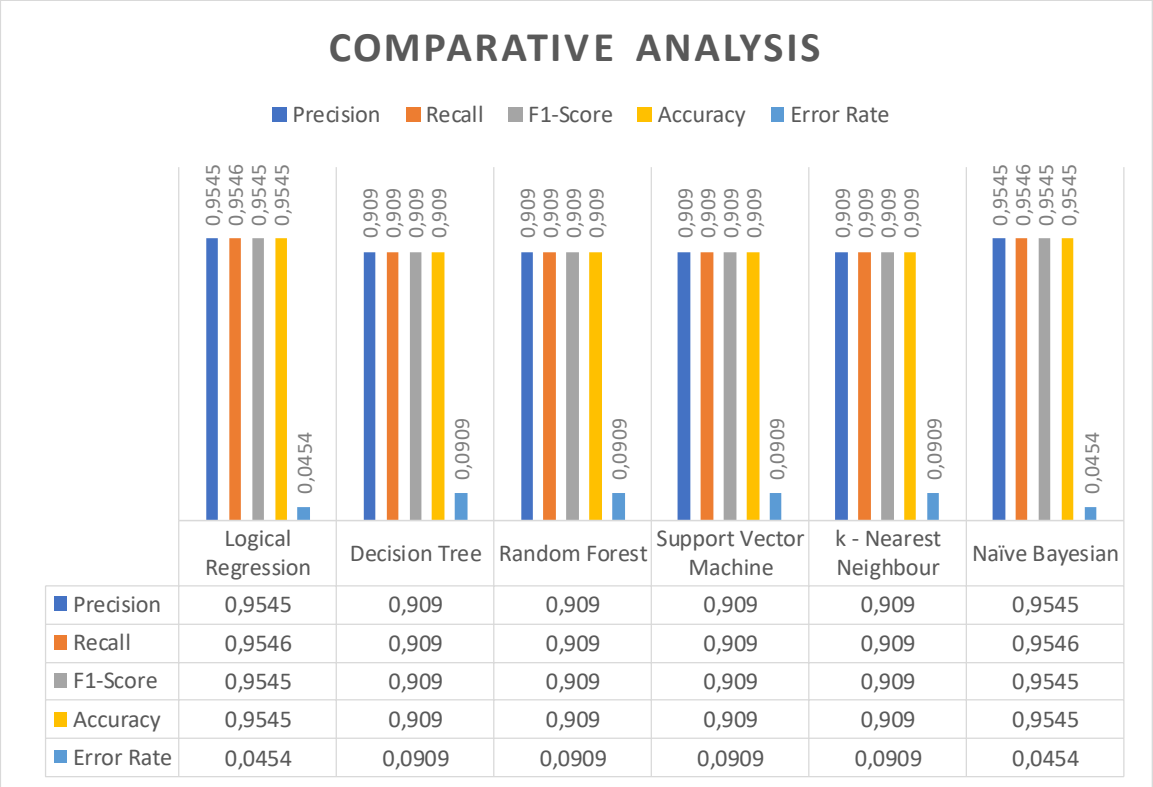
	Logical Regression	**Decision Tree**	**Random Forest**	**SVM**	**KNN**	**Naïve Bayes**
Precision	0.9545	0.909	0.909	0.909	0.909	0.9545
Recall	0.9546	0.909	0.909	0.909	0.909	0.9546
F1_Score	0.9545	0.909	0.909	0.909	0.909	0.9545

Accuracy	0.9545	0.909	0.909	0.909	0.909	0.9545
Error Rate	0.0454	0.0909	0.0909	0.0909	0.0909	0.0454

Table 1 Comparison of machine learning algorithms with Classification report

The result shows that Logical Regression and Naïve Bayesian Algorithm gives finer result than all other algorithms with precision 0.9545, recall 0.9546, F1 score 0.9545 and accuracy 95.45%. But when we compare the same with error rate,Logical Regression algorithm and Naïve Bayesian algorithm gives the best accuracy with lesserror rate as compared to other algorithms.

The below given Figure 5 is a schematic comparison of all of the machine learning algorithms used in our research :



	Logical Regression	Decision Tree	Random Forest	Support Vector Machine	k - Nearest Neighbour	Naïve Bayesian
Precision	0,9545	0,909	0,909	0,909	0,909	0,9545
Recall	0,9546	0,909	0,909	0,909	0,909	0,9546
F1-Score	0,9545	0,909	0,909	0,909	0,909	0,9545
Accuracy	0,9545	0,909	0,909	0,909	0,909	0,9545
Error Rate	0,0454	0,0909	0,0909	0,0909	0,0909	0,0454

Figure 5: Comparative Analysis of Machine Learning Algorithms Performance

VI. EXPERIMENTAL RESULTS

Each of the machine learning algorithms is evaluated against the above criteria for each query response in the dataset. The sum of the ten-question data for each of the six personality attribute categories – Honesty-Humility (H), Emotionality (E), Extroversion (X), Agreeableness (A), Conscientiousness (C), and Openness (O) – is calculated for each of the five success indicators to evaluate them. The results are shown below in the

context of a confusion matrix, along with a graph depicting the relationship between the real and expected labels.

Accuracy: (True Positive + True Negative) / Total Population

The accuracy of a prediction is described as the ratio of correctly predicted observations to total observations. Accuracy is the most intuitive metric of efficiency.

True Positive: The sum of correctly predicted occurrences that are positive.
True Negative: The percentage of correctly predicted occurrences that are negative.

Figure 6. clearly shows the output of evaluation of all machine learning algorithms used for personality prediction with the help of Confusion Matrix. In this graph is plotted between True Positive and True Negative to find out the accuracy. Logical Regression and Naïve Bayes algorithm given the best accuracy of 95.45% with less error rate. These two classifiers are of data science methodology, so it can execute fast with given dataset produced for personality prediction based on HEXACO Model.

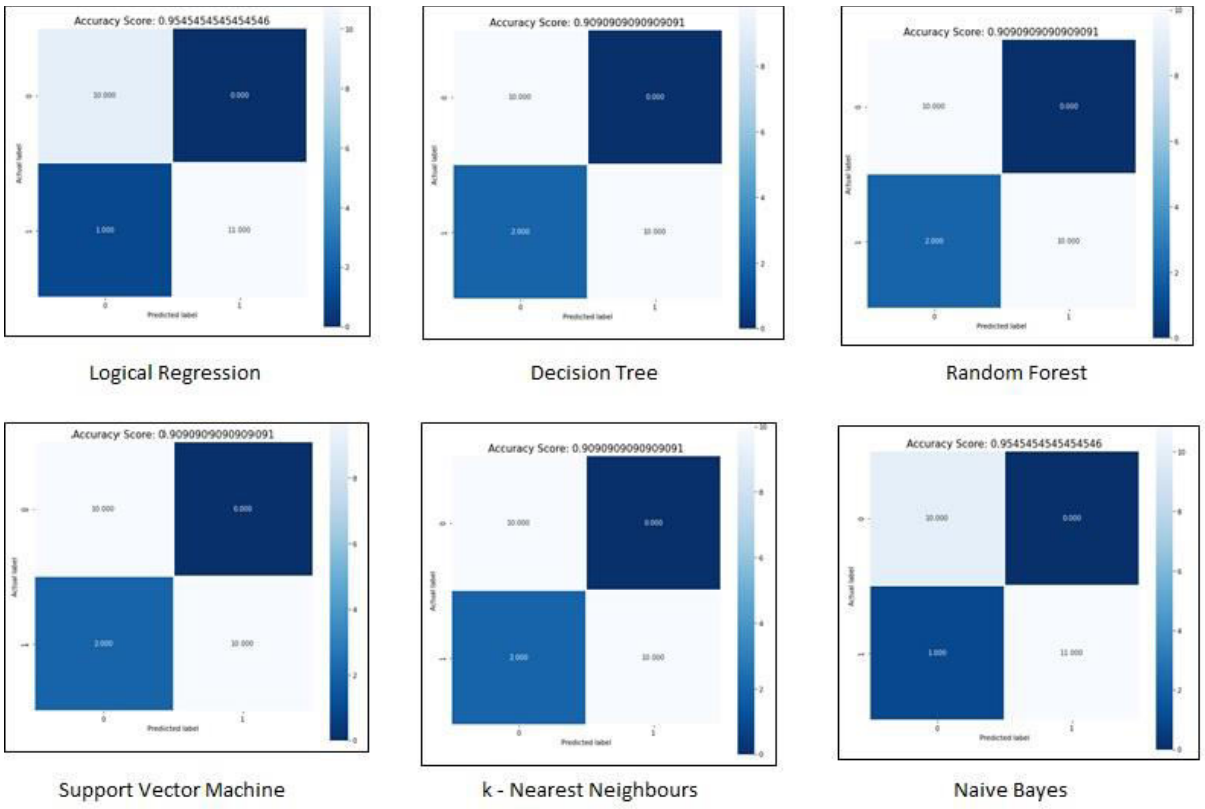


Figure6: Confusion Matrix for all Machine Learning Algorithms used for Prediction

VII. CONCLUSION

The primary objective of this analysis is to test and examine classification algorithms and personality predictability using the survey results. To evaluate the efficiency of the chosen classification algorithms, the HEXACO Model Personality data collection is used. Candidate responses may aid in predicting a person's characteristics using a variety of personality models. Previously, questionnaires were used, which was a costly and time-consuming operation. The aim of this work is to predict a person's personality based on their responses to interview questions. The research demonstrates the different methods and templates that were used. The algorithm with the lowest mean absolute error, which is typically associated with a higher accuracy score, is selected as the highest. The experimental research demonstrates that, although each algorithm exhibits a different accuracy rate for the various personality traits in the data collection, Logical Regression and Naïve Bayes algorithm gives the best accuracy contrasting different parameters with less error rate. Time can also be added as one of the parameter while evaluating the performance of classification algorithms.

REFERENCES

[1] Kumar, Raj, and Rajesh Verma. "Classification algorithms for datamining: A survey."International Journal of Innovations in Engineeringand Technology (IJIET) 1, no. 2 (2012): 7-14.

[2] Mehta, Y.; Majumder, N.; Gelbukh, A.; Cambria, E. Recent Trends in Deep Learning Based Personality Detection. Artif. Intell. Review **2020**, 53, 2313–2339. [CrossRef]
[3] Yang, H.-C.; Huang, Z.-R. Mining personality traits from social messages for game recommender systems. Knowl.-Based Syst. **2019**, 165, 157–168. [CrossRef]

[4] Huang, H.-C.; Cheng, T.C.E.; Huang, W.; Teng, C.I. Impact of online gamers' personality traits on interdependence, network convergence, and continuance intention: Perspective of social exchange theory. Int J. Inf. Manag. **2018**, 38, 232–242. [CrossRef]

[5] Anglim, J.; Sojo, V.; Ashford, L.J.; Newman, A.; Marty, A. Predicting employee attitudes to workplace diversity from personality, values, and cognitive ability. J. Res. Personal. **2019**, 83, 103865. [CrossRef]

[6] S. Bharadwaj, S. Sridhar, R. Choudhary and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1076-1082.

[7] M. Gjurković and J. Šnajder, "Reddit: A Gold Mine for Personality Prediction," In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media , pp. 87-97, 2018.

[8] B. Plank, and D. Hovy, "Personality traits on twitter—or—how to get 1,500 personality tests in a week." In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 92-98, 2015.

[9] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa and M. García-Borroto, "Study of the impact of resampling methods for contrast pattern-based classifiers in imbalanced databases," Neurocomputing, 175, pp. 935-947, 2016.

[10] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," 2016, arXiv preprint arXiv:1608.06048

[11] J. Levashina, C. J. Hartwell, F. P. Morgeson, and M. A. Campion, ``The structured employment interview: Narrative and quantitative review of the research literature,'' *Personnel Psychol.*, vol. 67, no. 1, pp. 241_293, Mar. 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/ 10.1111/peps.12052

[12] M. Mcdaniel, D. Whetzel, F. Schmidt, and S. Maurer, ``The validity of employment interviews: A comprehensive review and meta-analysis,'' *J. Appl. Psychol.*, vol. 79, pp. 599_616, Aug. 1994

[13] Description of Logistic Regression Algorithm. https://machine learningmastery.com/logistic-regression-for- machine-learning/.Accessed 15 May 2019

[14] Description of Multinomial Naı¨ve Bayes Algorithm https://www. 3pillarglobal.com/insights/document- classification-using-multi nomial-naive-bayes-classifier. Accessed 15 May 2019

[15] Khanday AMUD, Khan QR, Rabani ST. SVMBPI: support vector machine based propaganda identification. SN Appl. Sci. (accepted)

[16]Description of Decision Tree Algorithm: https://dataspirant.com/ 2017/01/30/how_decision_tree_algorithm_works/. Accessed 10July 2019

[17] Katuwal R, Suganthan PN (2018) Enhancing Multi-Class Classification of Random Forest using Random Vector Functional Neural Network and Oblique Decision Surfaces, Arxiv:1802.01240v1

[18] Kumar A, Dabas V, Hooda P (2018) Text classification algorithms for mining unstructured data: a SWOT analysis. Int J Inf Technol. https://doi.org/10.1007/s41870-017-0072-1

[19] Verma P, Khanday AMUD, Rabani ST, Mir MH, Jamwal S (2019) Twitter Sentiment Analysis on Indian Government Project using R. Int J Recent Tech Eng. https://doi.org/10.35940/ijrte. C6612.098319

[20] K. Lee and M. C. Ashton, ``Psychometric properties of the HEXACO- 100,'' *Assessment*, vol. 25, no. 5, pp. 543_556, Jul. 2018, doi: 10.1177/1073191116659134.

[21] N. Anderson, J. F. Salgado, and U. R. Hülsheger, ``Applicant reactions in selection: Comprehensive meta- analysis into reaction generalization versus situational specificity,'' *Int. J. Selection Assessment*, vol. 18, no. 3, pp. 291_304, Aug. 2010.

[22] M. C. Ashton and K. Lee, ``Empirical, theoretical, and practical advantages of theHEXACO model of personality structure,'' *Personality Social Psychol. Rev.*, vol. 11, no. 2, pp.150_166, May 2007

[23] K. Lee and M. Ashton, *The H Factor of Personality: Why Some People are Manipulative,Self-Entitled, Materialistic, and Exploitive_And Why It Matters for Everyone*.Waterloo, ON,Canada:Wilfrid Laurier Univ. Press, 2013.

[24] J. L. Pletzer, M. Bentvelzen, J. K. Oostrom, and R. E. de Vries, ``A meta-analysis of therelations between personality and workplace deviance: Big Five versus HEXACO,'' *J.Vocational Behav.*, vol. 112, pp. 369_383, Jun. 2019. [Online].