International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Accuracy Improvement using Machine Learning by Objects Count Based Feature selection method on Biological Data of Human Ancestors

Maradana Durga Venkata Prasad[1], Dr. Srikanth T.[2], Dr. N. Srivani[3], Balakrishna Gudla[4]

Submitted: 02/05/2024 **Revised**: 15/06/2024 **Accepted**: 22/06/2024

Abstract: In the present era huge data is generated by the IOT devices, mobiles, laptops and systems and stored either in data basesor files. Technique of Clustering is a used to extract the data from the data baseor files. Improving clustering accuracy always depends on the feature selection method. So feature selection always depends on choose of best feature selection method like wrapper, filter, embedded and hybrid.The original data set or data source's redundant, unnecessary, and noisy features can also be eliminated using feature selection techniques. Feature selection methods are used to reduce the computational costs, increases the accuracy, dimensionality is reduced and model is predictable.

Keywords *- Feature Selection,dimensionality reduction, Filter Method, Wrapper Method, Embedded Method, Hybrid feature selection.*

I. Introduction

A machine learning model's accuracy serves as a gauge for its effectiveness. It expresses the proportion of accurate classifications the model made [1]. It is represented as a value between 0 and.Accuracyof Machine learning model's performance depends on the feature selection methodalways [2].

Computing Model Accuracy

The accuracy is calculated by dividing the total number of predictions made across all classes by the number of accurate guesses [3]. Terms of Accuracy model are depicted in Table 1.

$$\textbf{Accuracy (ACC)} = (TP + TN) / (TP + TN + FP + FN)$$

Table 1: Terms of Accuracy model

	Model Predicted		
Data (True) Actual		Positives	Negative
	Positives	True Positives(TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

AccuracyScale

Either 0 (the model consistently predicts the incorrect label) or 1 (the model predicts the correct label) can be the accuracy. [4].

Connection to the Confusion Matrix

The counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), which are used to compute accuracy, are contained in the confusion matrix. The model's predictions are tabulated in the confusion matrix. [5].

Model accuracyStatistical Significance

It is used understand the model's performanceand even forecast future events or outcomes [6].

Importance of Models Accuracy

1. **Simplicity and understandability:**The accuracy metric is simple to use and comprehendIt displays the percentage of precise forecasts a model achieves. The simplicity of the model allows stakeholders who are neither technical nor non-technical to understand its performance.

2. **Error Rate complement**

Accuracy is treated like anerror ratecomplement.For Example; if accuracy is 1 then error rate is 1 minus accuracy.Therefore, the accuracy measure is used to determine how well a model predicts errors.

[1] *Research Scholar, Department of Computer Science and Engineering, Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh, India.*
powersamudra@gmail.com
[2] *Associate Professor, Department of Computer Science and Engineering, Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh, India.*
sthota@gitam.edu
[3] *Associate Professor, Department of Computer Science and Engineering, CMR Institute of Technology, Hyderabad, Telangana, India.*
vani.medipally@gmail.com
[4] *Associate Professor, Department of Computer Science and Engineering, CMR Institute of Technology, Hyderabad, Telangana, India.*
gudla.balakrishna@gmail.com

3. **Effectiveness and Efficiency:** Accuracy metric is used to evaluate a model performance.

4. **Common Research Metric:**In machine learning research, accuracy is employed when the datasets are well-balanced and clean.

ML Model Accuracy Improving Methods

The cycle of developing a model involves several steps, from gathering data to creating the model itself. To understand the relationships between the variables in the data, formulate hypotheses.

ML Model Accuracy Improving Methods

1. Add More Data.

2. Normalization/Standardization.

3. Treat Missing and Outlier Values.

4. Feature Engineering.

5. Feature Selection.

6. Choosing the Right Number of Clusters.

7. Selecting the Right Clustering Algorithm.

8. Multiple Algorithms.

9. TuningAlgorithm.

10. Ensemble Methods.

11. Cross Validation.

12. Iterative Refinement.

Add More Data

Adding more data is better to a data set consists less data because more data more correlations, less data less correlations between the attributes. So it is better to have more data which will gives use better relations for calculating the accuracy relaying on better attributes of the data set [8].

Normalization/Standardization:

To avoid features with bigger scales dominating the distance calculations, make sure all features are on the same scale.Min Max scaling / z score standardization are the general Normalization techniques [9].

Treat Missing and Outlier Values

Generally, in any data set unwanted, missing and outlier values data will be present. Due to this trained model accuracy will reduce intern affects the predictions (gives wrong predictions).This is the outcome of our flawed behavior analysis and correlation analysis with other variables. Therefore, for a machine learning model that is naturally improved and more dependable, it is crucial to handle missing and outlier variables carefully [10].

Methods to Handlethe Missing & Outlier Values:

1. Missing: For continuous variables missing values are replaced with mean, median, or mode. For categorical variables missing values treated variables as a separate class.

2. Outlier: The below three methods are used to deal with outliers

a. cutting weight, or lowering the weights of outliers
b. modifying the values of outliers through imputation, trimming, and windsorization
c. making use of reliable estimate methods (M-estimation).

Feature Engineering

• Select relevant features that capture the underlying patterns in the data.

• To more accurately depict the structure of the data, transform or engineer features.

• Eliminate superfluous or unnecessary elements that could cause noise [11].

By taking this step, more information can be gleaned from the available data. In terms of new features, fresh information is extracted. It's possible that these characteristics can better explain the variance in the training set. resulting in increased model accuracy. Generation of hypotheses has a great influence on feature engineering. Good features come from good theories. For this reason, I always advise devoting some time to the process of developing hypotheses. The feature engineering process can be divided into two steps:

Transformation of Features

It refers to the practice of changing or transforming input features in a dataset in machine learning in order to enhance a machine learning model's performance. To improve their fit for the learning process, the characteristics are subjected to statistical or mathematical manipulations [12].

Feature Creation

Feature generation is the process of creating a new variable or variables from an existing one. It facilitates the discovery of a data set's hidden relationships. Each time you develop a new feature. This may potentially result in the trained model performing less well or with less precision. Therefore, you need to consider how a new feature will impact the training process each time it is created by looking at its feature relevance.

Feature Selection

The process of determining which collection of attributes best describes how independent variables relate to the target variable is known as feature selection [13].

Features selection Based on Metrics like:

1. **Knowledgeof Domain:**We select a feature or characteristics based on domain expertise and knowledge that we believe could have a bigger impact on the target variable.

2. **Visualization:**

• To better comprehend the clustering findings, visualize the silhouette scores and clusters.

• Interpret the clusters to ensure they align with domain knowledge and expectations.

It facilitates the visualization of the relationship between variables, as the name implies, which eases the process of choosing variables.

3. **Statistical Parameters:**To select the optimal characteristicsP-values, information values, and other statistical measures are also considered.4. **Dimensionality Reduction:**

• Employ methods such as PCA(Principal Component Analysis) or t-SNE to decrease the number of dimensions in the data while maintaining its organization.

• Better silhouette scores and cluster separation can result from reduced dimensionality. Even while it helps to translate training data into lower dimensional regions, the data's intrinsic relationships are still characterized. It's a kind of method for reducing dimensionality. Numerous methods, such as factor analysis, low variance, higher correlation, backward and forward feature selection, and others, can be used to minimize the dimensions (features) of training data [14].

Choosing the Right Number of Clusters

• Try out several cluster counts and choose the one that optimizes the silhouette score.

• The ideal number of clusters can be established with the aid of methods such as silhouette analysis itself or the elbow method [15].

Selecting the Right Clustering Algorithm

• Various cluster geometries and data types are better suited for different clustering techniques. Try out several techniques, such as hierarchical clustering, DBSCAN, and K-means.

• If there are uneven forms and changing densities in the clusters, take into consideration density-based clustering techniques such as DBSCAN [16].

Multiple Algorithms

Although there are many different machine learning algorithms, choosing the appropriate one is the best way to increase accuracy. However, it is not as simple as it seems. It takes practice and experience to develop this intuition. Certain types of data sets are more suited for certain algorithms than for others. Therefore, we should use all pertinent models and evaluate the results. [17].

Algorithm Tuning Parameters

• To make the clustering algorithm better fit your data, change its parameters.

• Try varying the initializations, iteration count, or distance metrics, for instance, when using K-means [18].

Hyperparameters govern machine learning algorithms. The results of the learning process are significantly influenced by these hyperparameters. Hyperparameter tuning involves determining the ideal value for each hyperparameter in order to improve the accuracy of the model. You need to be well-versed in these definitions and how each one affects the model in order to properly adjust these hyperparameters. This is a procedure that can be repeated with several effective models.Example: The hyperparameters of the random forest classification algorithm include max_features, number_trees, random_state, oob_score, and others. Better and more accurate models will be produced by intuitively optimizing these parameter values.

Ensemble Methods

• To increase resilience, it aggregates the output of several clustering methods.

• Better results are obtained by combining the output of several weak models [19].You can achieve by the following ways:

1. Bootstrap Aggregating / Bagging.

2. Boosting.

Bagging/ Bootstrap aggregation

To lessen variance within a noisy data collection, ensemble learning is frequently employed. In bagging, a random sample of data from a training set is chosen with replacement, enabling each data point to be chosen more than once. [20].

Boosting

Boosting is a machine learning strategy that reduces errors in the processing of predicted data. Data scientists use labeled data to train software called machine learning models to make predictions about unlabeled data. [21].

Cross Validation

The efficiency of the model is validated using a technique called cross-validation, which involves training the model on a subset of input data and testing it on a subset of input data that hasn't been seen before. It can also be thought of as a technique for assessing how effectively a statistical model generalizes to an alternative dataset. [22]. Figure 1 depict the Test and Train split of a data set. Figure 2 shows the cross validation method.

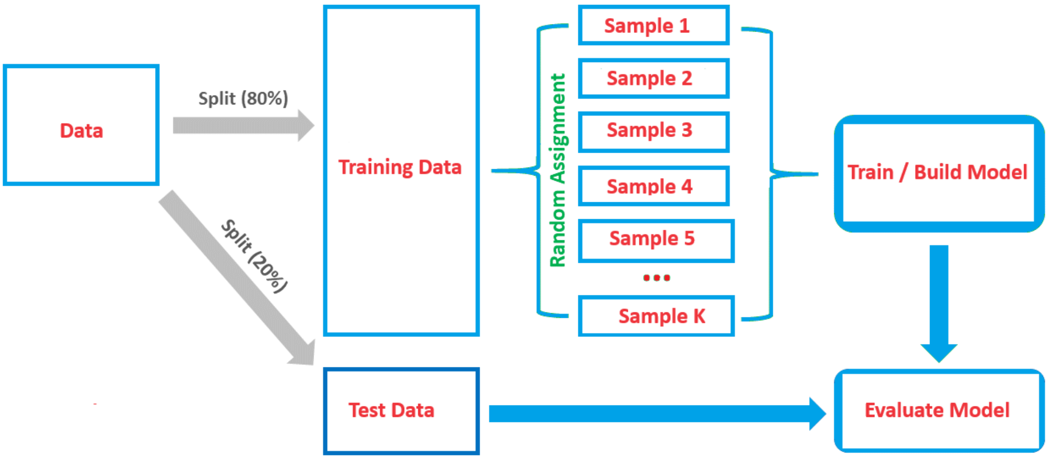


Fig 1 : Test and Train split of a data set

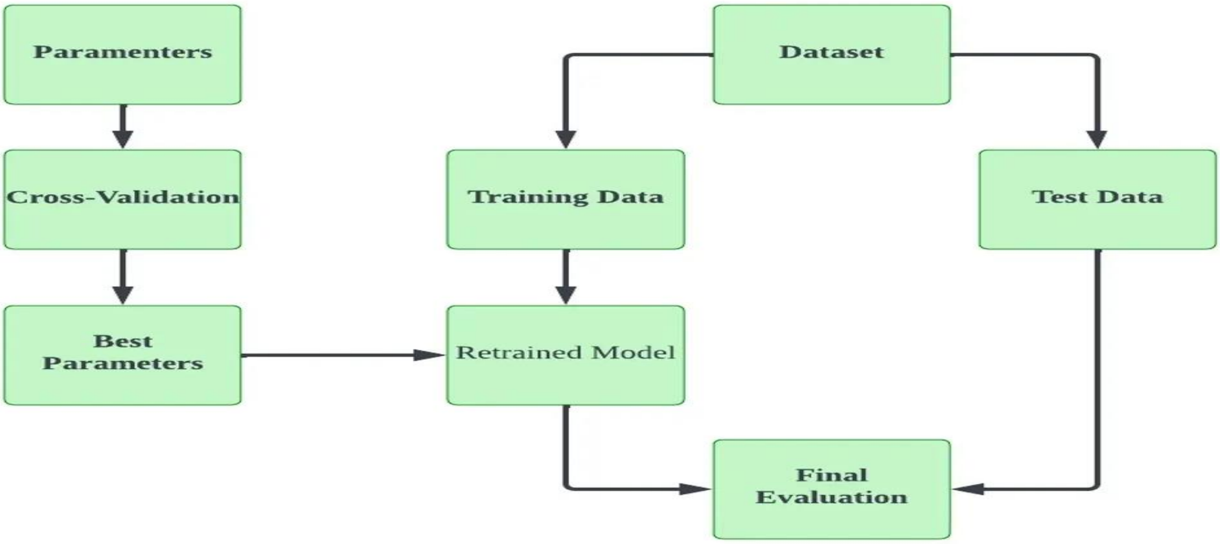


Fig 2: Cross Validation Method

Iterative Refinement:

• It used for refining the process based on the results obtained in previous iterations.

• Utilize domain experts' input to direct the process of refining [23].

II. Literature Survey

This survey consists of various Methods of feature selection methods and its Comparison. The right selection of the features is important to improve the accuracy and efficiency [23]. Table 2 depicts the feature selection methods and its sub methods. Table 3 show feature selection methods and its sub methodsdetails.

Table 2: Methods of Feature selection and its sub methods

Method of Feature Selection	**Sub Selection Method**	**Sub feature selection**
Filter Method	Information Gain	
	Chi Square	
	Fisher's Score	
	Correlation & Coefficient	
Wrapper Method	Sequential Search	SFS
		SBS

	Heuristic Search	Genetic algorithm
Embedded Method	Embedded	Regularization
Hybrid Method	Using Filter & Wrapper	
	Using Embedded& Wrapper	

Table 3: Details of feature selection methods and its sub methods

Method	**Techniques**	**Summary**	**Advantages**	**ResearchGaps**
Filter	InformationGain [24]	It assesses each variable's knowledge gain. It divides the dataset into clusters for categorization and lowers the dataset's entropy.	Fast, simplemethod,nodependencythe classifier.	This method's disadvantage is that individual aspects are assessed independently, and the dependencies among the features are not taken into account.
	ChiSquare [25]	By examining the correlation between the variables in the data set and the target variable, find the optimal chi-square score.		
	Fisher'sScore [26]	The ranks, which are determined by the fisher's score, are used to choose the variables.		
	Correlation&Coefficient [27]	Finds the variable which performs well and highly correlated with the target variable.		
Wrapper	SFS(Sequential Search) [28]	If the variable's outcome proves to be more accurate than the previous one, it is added permanently to the empty set.	When engaging with the classifier, it takes into account feature dependencies and is not as computationally demanding..	While this approach has a higher chance of overfitting and yields superior results than the filter method, it takes longer.
	SBS(Sequential Search) [29]	It begins with the entire collection and eliminates the variables that are less useful for the desired outcome.		
	Genetic	It uses an		

	algorithm (Heuristic Search) [30]	approximation technique in which a set of possible answers evolves through natural selection.		
Embedded	Regularization [31]	It shortens the wrapper method's computation time.	It carries out additional computationally demanding tasks and communicates with the classifier.	Overfitting models results in an increase in their weight. The characteristics are chosen based on the classifier.
Hybrid	Using Filter & Wrapper [32]	It uses the filter and wrapper method and has a faster computing time.		Overfitting models results in an increase in their weight. The characteristics are chosen based on the classifier.
	Using Embedded & Wrapper [33]	It is superior to the wrapper and embedded methods.		When compared to the filter and wrapper technique, the hybrid method's complexity has grown.

III Proposed Algorithm

The proposed algorithm is used to find the best featureswhich intern improves the accuracy of the any clustering algorithm. The algorithm name was Objects Based Feature selection method.

Algorithm steps

1. Count the Number of records (X).

2. Calculate the different objects count for a feature.

Example:Gender (Male (70), Female (30)).

3. Take the average of Number of Objects Count (ANOC) of all the attributes.

Average ofNumber of Objects Count (ANOC)= Sum of Total number of Unique objects per feature / Total Number of Records(X)

4. Compute the individual percentages of different objects count for a feature.

Example: Gender (Male (70), Female (30)).

5. Consider only max percentage of individual percentages of different objects count for a feature.

6. Sort the records in Descending Order based on max percentage of individual percentages of different objects count for a feature.

7. Choose /Select the max percentage of individual percentages of different objects count for a feature along with average of Number of Objects Count (ANOC) always less than the individual object count for running any clustering or classification algorithm.

8. Features with high individual percentages of different objects along with along with average of Number of Objects Count (ANOC) will give more accuracy compared to others

Example:

1. Sample data from biological data of human ancestor's data sets consisting of 5000 records

2. Individual count of objects of a feature are known using distribution graphs.

3. Calculation of Individual Feature sub percentages. Table 4 depicts the sample calculation of Individual Feature sub percentages. Table 5 shows Individual Percentage Calculation along with Objects Count.Table 6 depicts the Before and After sorting of max value of percentages of objects count of features.

Table 4: Calculation of Individual Feature sub percentages

Total Number of Records	**Percentage**
5000	100
217	X

X = (217*100)/5000 = 4.34

Table 5: Individual percentageCalculation along with Objects Count

Attributes/ Features	**Objects**	**Count**	**Individual percentage**	**Max**	**Objects Count**
	homininoOrrorintugenencin	217	4.34		
	homininoArdipithecusramidus / kabadda	211	4.22		
	Australopithecus Afarensis	208	4.16		
	Australopithecus Anamensis	222	4.44		
	Australopithecus Africanus	193	3.84		
	Homo Rodhesiensis	188	3.76		
	homininoSahelanthropustchadensis	217	4.34		
	Homo Neanderthalensis	207	4.14		
	ParanthropusAethiopicus	205	4.1		
	Homo Erectus	207	4.14		
	Homo Naledi	211	4.22		
	Homo Floresiensis	208	4.16		
	ParanthropusBoisei	211	4.22		
	Homo Rudolfensis	232	4.64		
	Homo Habilis	202	4.04		
	Homo Sapiens	195	3.9		
	Homo Antecesor	193	3.86		
	Homo Ergaster	201	4.02		
	Australopithecus Sediba	209	4.18		
	Homo Georgicus	219	4.38		
	Australopithecus Bahrelghazali	210	4.2		
	Australopithecus Garhi	190	3.8		
	ParanthropusRobustus	223	4.46		
Genus_&_Specie	Homo Heidelbergensis	220	4.4	4.46	23
	Africa	3744	74.88	74.88	3
	Asia	625	12.5		
Location	Europa	630	12.6		
	Central	855	17.1	53.36	4
	Oriental	2668	53.36		
	South	841	16.82		
Zone	west	635	12.7		
	Ethiopia	1005	20.1	29.12	8
	Georgia	211	4.22		
	Germany	413	8.26		
	Indonesia	414	8.28		
	Kenya	1456	29.12		
Current_Country	Republic of chad	442	8.84		

	South Africa	841	16.82		
	spain	217	4.34		
Habitat	Cold forest	413	8.26	32.74	8
	Forest	442	8.84		
	Forest-gallery	419	8.38		
	Forest-savanna	209	4.18		
	Jungle	442	8.84		
	Mixed	1230	24.6		
	Peninsular	207	4.14		
	savannah	1637	32.74		
Incisor_Size	Big	1247	24.94	41.46	5
	Medium large	428	8.56		
	Megadony	638	12.76		
	Small	2073	41.46		
	Very small	613	12.26		
Jaw_Shape	U shape	2491	49.82	49.82	4
	V shape	638	12.76		
	Conical	844	16.88		
	modern	1026	20.52		
Torus_Supraorbital	Flat	201	4.02	45.76	5
	Less protruding	388	7.76		
	Little protruding	1484	29.68		
	Ultra protruding	638	12.76		
	Very protruding	2288	45.76		
Prognathism	Absent	201	4.02	32.82	6
	High	1058	21.16		
	Medium	636	12.72		
	Medium-high	1641	32.82		
	Reduced	825	16.5		
	Very high	638	12.76		
Foramen_MÃ¡gnum_Position	Anterior	1851	37.02	37.54	4
	Modern	1877	37.54		
	Posterior	443	8.86		
	Semi-anterior	828	16.56		2
Canine Size	Big	2306	46.12	53.86	
	small	2693	53.86		
Canines_Shape	Canicalls	2067	41.34	58.64	2
	incisform	2932	58.64		
Tooth_Enamel	Medium-thick	412	8.24	37.88	7
	Medium-thin	1251	25.02		
	Thick	1894	37.88		
	Thick- Medium	190	3.8		
	Thin	413	8.26		
	Very thick	638	12.76		
	Very thin	201	4.02		
Tecno	Likely	193	3.86	54.84	3
	No	2742	54.84		
	Yes	2064	41.28		
Tecno_type	Mode 1	1050	21	54.84	6
	Mode 2	611	12.22		
	Mode 3	202	4.04		
	Mode 4	201	4.02		

	No	2742	54.84		
	primitive	193	3.86		
Biped	High probability	190	3.8	45.92	4
	Low probability	443	8.86		
	Modern	2296	45.92		
	yes	2070	41.4		
Arms	Climbing	3142	62.84	62.84	3
	Manipulate	1054	21.08		
	Manipulate with precision	803	16.06		
Foots	Climbing	1912	38.24	61.74	2
	walk	3087	61.74		
Diet	Carnivorous	614	12.28	37.5	5
	Dry fruits	1301	26.02		
	Hard fruits	831	16.62		
	Omnivore	1875	37.5		
	Soft fruits	378	7.56		
Sexual_Dimorphism	High	2513	50.26	50.26	3
	Medium-high	2285	45.7		
	Reduced	201	4.02		
Hip	Modern	1048	20.96	42.58	4
	slim	1208	24.16		
	Very modern	614	12.28		
	Wide	2129	42.58		
Vertical_Front	modern	1454	29.08	58.62	3
	no	2931	58.62		
	yes	614	12.28		
Anatomy	Mixed	1251	25.02	46.04	4
	modern	831	16.62		
	Old	2302	46.04		
	Very modern	614	12.28		
Migrated	No	3750	75	75	2
	Yes	1249	24.98		
Skeleton	Light	2311	46.22	46.22	3
	refined	611	12.22		
	robust	2077	41.54		

Table 6:Before and After sorting of max value of percentages of objects count of features.

Before Sorting			After Sorting	
Attributes	Max		Attributes	Max
Genus_&_Specie	4.46		Migrated	75
Location	74.88		Location	74.88
Zone	53.36		Arms	62.84
Current_Country	29.12		Foots	61.74
Habitat	32.74		Canines_Shape	58.64
Incisor_Size	41.46		Vertical_Front	58.62
Jaw_Shape	49.82		Tecno_type	54.84
Torus_Supraorbital	45.76		Tecno	54.84
Prognathism	32.82		Canine Size	53.86
Foramen_MÃ¡gnum_Position	37.54		Zone	53.36

Canine Size	53.86		Sexual_Dimorphism	50.26
Canines_Shape	58.64		Jaw_Shape	49.82
Tooth_Enamel	37.88		Skeleton	46.22
Tecno	54.84		Anatomy	46.04
Tecno_type	54.84		Biped	45.92
Biped	45.92		Torus_Supraorbital	45.76
Arms	62.84		Hip	42.58
Foots	61.74		Incisor_Size	41.46
Diet	37.5		Tooth_Enamel	37.88
Sexual_Dimorphism	50.26		Foramen_MÃ¡gnum_Position	37.54
Hip	42.58		Diet	37.5
Vertical_Front	58.62		Prognathism	32.82
Anatomy	46.04		Habitat	32.74
Migrated	75		Current_Country	29.12
Skeleton	46.22		Genus_&_Specie	4.46

Average of Number of Objects Count (ANOC) Of All the Attributes = (2 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 3 + 3 + 4 + 4 + 4 + 4 + 4 + 4 + 5 + 5 + 5 + 6 + 6 + 7 + 8 + 8 + 23) / 25 = 123 /25 =4.92

IV. Results

The results are generated using orange Tool of data mining. The Figure 3 depicts the mappings of required for comparison of different algorithms like Naïve Bayes, logistic regression and SVM for accuracy. Figure 4,5 and 6 depicted the use of different features for the improvement of accuracy using the Object Count Based Feature selection method on different Algorithms like Naïve Bayes, logistic regression and SVM.

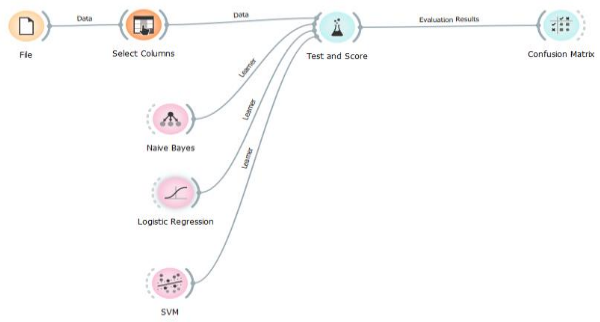


Fig 3: Comparison Naïve Bayes, logistic regression and SVM for accuracy.

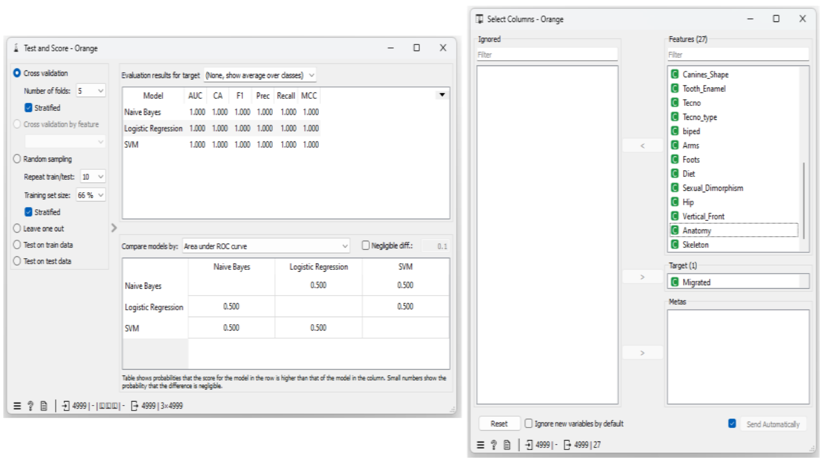


Fig 4: Mitraged feature is used for Comparison Naïve Bayes, logistic regression and SVM algorithmsfor accuracy.

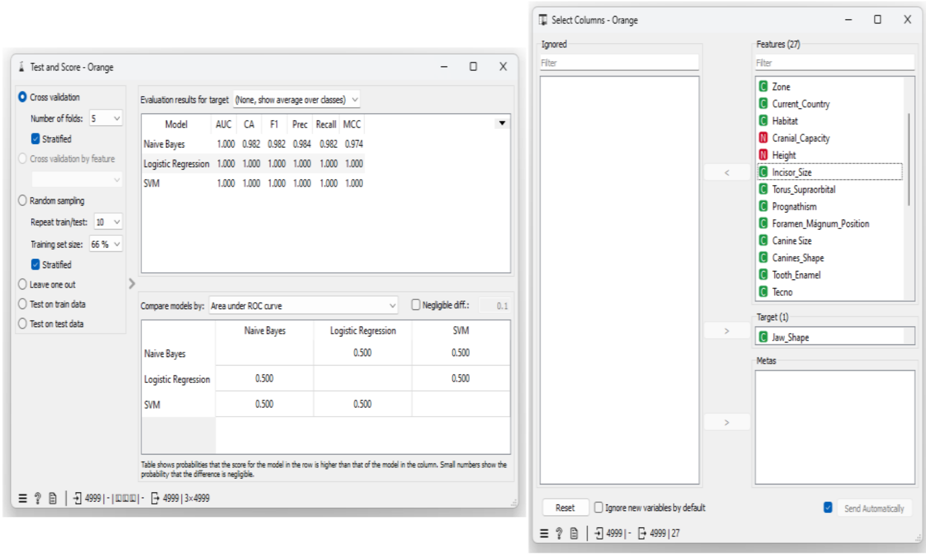


Fig 5: Jaw_Shape feature is used for Comparison Naïve Bayes, logistic regression and SVM algorithmsfor accuracy.

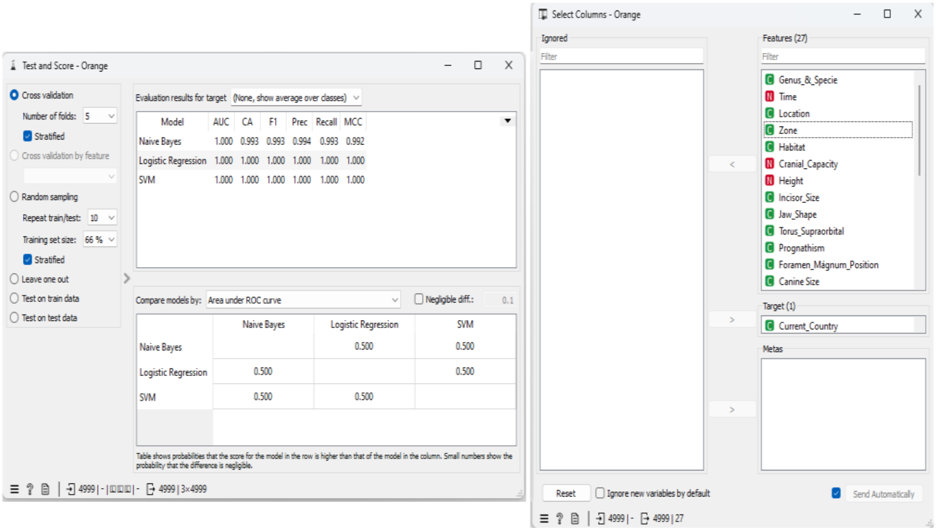


Fig 6: Current_country feature is used for Comparison Naïve Bayes, logistic regression and SVM algorithmsfor accuracy.

V. Conclusions

Clustering is used to group the data. But the accuracy of the clustering always depends on the feature section method. Object count based feature selection method is used for clustering to improve its accuracy.

Author Contributions

Conceptualization, MDVP, ST; methodology, MDVP, ST; software, MDVP; validation, MDVP; formal analysis, ST; investigation, MDVP, ST; resources, MDVP, ST; data curation, MDVP; writing—original draft, MDVP; writing—review and editing, MDVP; visualization, MDVP, ST; supervision, ST; project administration, ST; funding acquisition, MDVP, ST. All authors have read and agreed to the published version of the manuscript.

Ethical considerations

Not applicable

Funding

No External or Internal Funding for this project.

Informed Consent Statement

There is no research with human subjects included in this article.

Data Availability Statement

No data sets were used or generated in this article.

Conflicts of Interest

The authors certify that they have no competing interests with relation to the work they have submitted.

References

[1] V. K and G. Dayalan, "Framework for Improving the Accuracy of the Machine Learning Model in Predicting Future Values," 2023 IEEE 8th International Conference for Convergence in

Technology (I2CT), Lonavla, India, 2023, pp. 1-8, DOI: 10.1109/I2CT57861.2023.10126236. [2]. A. Benkessirat and N. Benblidia, "Fundamentals of Feature Selection: An Overview and Comparison," 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 2019, pp. 1-6, DOI: 10.1109/AICCSA47632.2019.9035281.

[2] A. V. Veligosha, D. I. Kaplun, D. M. Klionskiy, V. V. Gulvanskiy, D. V. Bogaevskiy and I. I. Kanatov, "Model of computation accuracy in modular digital filters," 2017 XX IEEE International Conference on Soft Computing and Measurements (SCM), St. Petersburg, Russia, 2017, pp. 275-277, DOI: 10.1109/SCM.2017.7970559.

[3] S. Sun et al., "Investigation of Prediction Accuracy, Sensitivity, and Parameter Stability of Large-Scale Propagation Path Loss Models for 5G Wireless Communications," in IEEE Transactions on Vehicular Technology, vol. 65, no. 5, pp. 2843-2860, May 2016, DOI: 10.1109/TVT.2016.2543139.

[4] B. P. Salmon, W. Kleynhans, C. P. Schwegmann and J. C. Olivier, "Proper comparison among methods using a confusion matrix," 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 2015, pp. 3057-3060, DOI: 10.1109/IGARSS.2015.7326461.

[5] S. Gupta, W. Zhang and F. Wang, "Model Accuracy and Runtime Tradeoff in Distributed Deep Learning: A Systematic Study," 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 2016, pp. 171-180, DOI: 10.1109/ICDM.2016.0028.

[6] W. Wang and L. Bi, "Research on strategies to improve model accuracy based on incomplete time series data," 2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT), Haikou, China, 2021, pp. 45-52, DOI: 10.1109/ACAIT53529.2021.9731336.

[7] S. Tabassum, M. B. Sampa, R. Islam, F. Yokota, N. Nakashima and A. Ahmed, "A Data Enhancement Approach to Improve Machine Learning Performance for Predicting Health Status Using Remote Healthcare Data," 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), Dhaka, Bangladesh, 2020, pp. 308-312, DOI: 10.1109/ICAICT51780.2020.9333506.

[8] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi and V. Padma, "Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 729-735, DOI: 10.1109/ICSSIT48917.2020.9214160.

[9] M. Seliem, Muhammad. (2022). Handling Outlier Data as Missing Values by Imputation Methods: Application of Machine Learning Algorithms. Turkish Journal of Computer and Mathematics Education (TURCOMAT). 13. 273-286.

[10] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," SoutheastCon 2016, Norfolk, VA, USA, 2016, pp. 1-6, DOI: 10.1109/SECON.2016.7506650. [12]. A. Kusiak, "Feature transformation methods in data mining," in IEEE Transactions on Electronics Packaging Manufacturing, vol. 24, no. 3, pp. 214-221, July 2001, DOI: 10.1109/6104.956807.

[11] Aparna U.R. and S. Paul, "Feature selection and extraction in data mining," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, India, 2016, pp. 1-3, DOI: 10.1109/GET.2016.7916845.

[12] N. Sharma and K. Saroha, "Study of dimension reduction methodologies in data mining," International Conference on Computing, Communication & Automation, Greater Noida, India, 2015, pp. 133-137, DOI: 10.1109/CCAA.2015.7148359.

[13] K. M. A. Patel and P. Thakral, "The best clustering algorithms in data mining," 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India, 2016, pp. 2042-2046, DOI: 10.1109/ICCSP.2016.7754534.

[14] T. T. Chikohora and E. Chikohora, "An Algorithm for Selecting a Data Mining Technique," 2021 3rd International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Windhoek, Namibia, 2021, pp. 1-7, DOI:. 10.1109/IMITEC52926.2021.9714525.

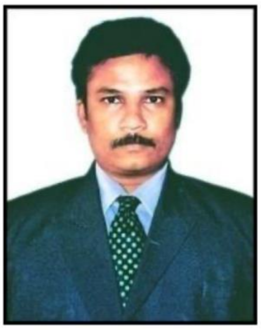
[15] Z. M. Fadhil and R. A. Jaleel, "Multiple Efficient Data Mining Algorithms with Genetic Selection for Prediction of SARS-CoV2," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 2016-2020, DOI: 10.1109/ICACITE53722.2022.9823757.

[16] F. Arden and C. Safitri, "Hyperparameter Tuning Algorithm Comparison with Machine Learning Algorithms," 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2022, pp. 183-188, DOI: 10.1109/ICITISEE57756.2022.10057630.

[17] T. N. Rincy and R. Gupta, "Ensemble Learning Techniques and its Efficiency in Machine Learning: A Survey," 2nd International Conference on Data,

Engineering and Applications (IDEA), Bhopal, India, 2020, pp. 1-6, DOI: 10.1109/IDEA49133.2020.9170675.
[18] D. P. Gaikwad and R. C. Thool, "Intrusion Detection System Using Bagging Ensemble Method of Machine Learning," 2015 International Conference on Computing Communication Control and Automation, Pune, India, 2015, pp. 291-295, DOI: 10.1109/ICCUBEA.2015.61.
[19] L. Cherif and A. Kortebi, "On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification," 2019 Wireless Days (WD), Manchester, UK, 2019, pp. 1-6, DOI: 10.1109/WD.2019.8734193.
[20] T. Gunasegaran and Y. -N. Cheah, "Evolutionary cross validation," 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, 2017, pp. 89-95, DOI: 10.1109/ICITECH.2017.8079960.
[21] T. R. N and R. Gupta, "Feature Selection Techniques and its Importance in Machine Learning: A Survey," 2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS), Bhopal, India, 2020, pp. 1-6, DOI:. 10.1109/SCEECS48394.2020.189.
[22] M. S. De Sousa, C. E. L. Veiga, R. D. O. Albuquerque and W. F. Giozza, "Information Gain applied to reduce model-building time in decision-tree-based intrusion detection system," 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), Madrid, Spain, 2022, pp. 1-6, DOI: 10.23919/CISTI54924.2022.9820579.
[23] S. Rosidin, Muljono, G. FajarShidik, A. ZainulFanani, F. Al Zami and Purwanto, "Improvement with Chi Square Selection Feature using Supervised Machine Learning Approach on Covid-19 Data," 2021 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarangin, Indonesia, 2021, pp. 32-36, DOI: 10.1109/iSemantic52711.2021.9573196.
[24] Zhongzhi Shi, Youping Huang and Sulan Zhang, "Fisher Score Based Naive Bayesian Classifier," 2005 International Conference on Neural Networks and Brain, Beijing, China, 2005, pp. 1616-1621, DOI: 10.1109/ICNNB.2005.1614941.
[25] -h. Yang, G. -l. Shan and L. -l. Zhao, "Correlation Coefficient Method for Support Vector Machine Input Samples," 2006 International Conference on Machine Learning and Cybernetics, Dalian, China, 2006, pp. 2857-2861, DOI: 10.1109/ICMLC.2006.259069.
[26] M. Peña, M. Cerrada, D. Cabrera and R. -V. Sánchez, "Fast feature selection based on cluster validity index applied on data-driven bearing fault detection," 2020 IEEE ANDESCON, Quito, Ecuador, 2020, pp. 1-6, DOI: 10.1109/ANDESCON50619.2020.9272146.
[27] U. Haq, J. Li, M. H. Memon, M. HunainMemon, J. Khan and S. M. Marium, "Heart Disease Prediction System Using Model of Machine Learning and Sequential Backward Selection Algorithm for Features Selection," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-4, DOI: 10.1109/I2CT45611.2019.9033683.
[28] Raj, A. Kumar, V. Sharma, S. Rani, A. K. Shanu and T. Singh, "Applications of Genetic Algorithm with Integrated Machine Learning," 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), Uttar Pradesh, India, 2023, pp. 1-6, DOI: 10.1109/ICIPTM57143.2023.10118328.
[29] H. Osman, M. Ghafari and O. Nierstrasz, "Automatic feature selection by regularization to improve bug prediction accuracy," 2017 IEEE Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE), Klagenfurt, Austria, 2017, pp. 27-32, DOI: 10.1109/MALTESQUE.2017.7882013.
[30] Suto, S. Oniga and P. P. Sitar, "Comparison of wrapper and filter feature selection algorithms on human activity recognition," 2016 6th International Conference on Computers Communications and Control (ICCCC), Oradea, Romania, 2016, pp. 124-129, DOI: 10.1109/ICCCC.2016.7496749.
[31] T. Hamed, R. Dara and S. C. Kremer, "An Accurate, Fast Embedded Feature Selection for SVMs," 2014 13th International Conference on Machine Learning and Applications, Detroit, MI, USA, 2014, pp. 135-140, DOI: 10.1109/ICMLA.2014.104.

AUTHOR DETAILS:

	Dr. SrikanthThota received his Ph.D in Computer Science Engineering for his research work in Collaborative Filtering based Recommender Systems from J.N.T.U, Kakinada. He received M.Tech. Degree in Computer Science and Technology from Andhra University. He is presently working as an Associate Professor in the department of Computer Science and Engineering, School of Technology, GITAM University, Visakhapatnam, Andhra Pradesh, India. His areas of interest include Machine learning, Artificial intelligence, Data Mining, Recommender Systems, Soft computing.
	Mr. MaradanaDurgaVenkata Prasad received his B. TECH (Computer Science and Information Technology) in 2008 from JNTU, Hyderabad and M.Tech. (Software Engineering) in 2010 from Jawaharlal Nehru Technological University, Kakinada, He is a Research Scholar with Regd No: 1260316406 in the department of Computer Science and Engineering, Gandhi Institute of Technology and Management (GITAM) Visakhapatnam, Andhra Pradesh, INDIA. His Research interests include Clustering in Data Mining, Big Data Analytics, and Artificial Intelligence. He is currently working as an Assistant Professor in Department of Computer Science Engineering, CMR Institute of Technology, Ranga Reddy, India.
	Dr. N Srivani received her Ph.D (Computer Science and Engineering) in 2023 from JJTU, Rajasthan and M.Tech. (CSE) in 2013 from Anurag University, Hyderabad. Her Research interests include Deep learning networks, Big Data Analytics, and Artificial Intelligence. She is currently working as an Assistant Professor in Department of Computer Science Engineering, CMR Institute of Technology, Ranga Reddy, India.
	Dr. Balakrishna Gudla have obtained his Ph.D degree from National Institute of Technology, Surathkal, Karnataka. His Ph.D. research was in the area of Graph Coloring. I have obtained her M.E. from TIET, Patiala, Punjab. I have more than 5 years of teaching and research experience. My previous associations were with Mewar University- Chittorgarh, AIGS- Bengaluru and Dayananda Sagar University-Begaluru. I have published ten research articles, one patent and one conference paper. My research interest lies in the areas of Graph Coloring, Graph Algorithms, Big Data and Data science analytics.