International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

A Novel Approach for Detecting Outliers by Using Isolation Forest with Reducing Under Fitting Issue

Mr.Prashant M Goad[1]
Electronic & Telecommunicatoin
RCPIT Shirpur
Shirpur Maharashtra
prashantgoad@gmail.com

Dr. Pramod J Deore[2]
Electronic & Telecommunicatoin
RCPIT Shirpur
Shirpur Maharashtra
pjdeore@yahoo.com

Abstract— The effectiveness of machine learning for a particular activity depends on a variety of parameters. The incident database's description and validity come first and primary. Information retrieval even during the training cycle is more challenging if there is a lot of repetitious, unimportant information or incomplete information available. It is good knowledge that running time for ML tasks is significantly impacted by conditions as follows and sorting stages. To increase the accuracy of any model data cleansing is essential. Without sufficient data scrubbing, no predictive model accuracy can begin. EDA, or exploratory data analysis, is the name of this procedure. In this study, we discussed outlier identification, one of many EDA processes for complete perfect data. In this research, we attempted to use the isolation forest approach to calculate the outlier factor. Then a model known as an outlier finding model is created. The problem of outlier detection leads to a collection of connected supervised learning for binary classification. We carry out in-depth tests on various datasets and demonstrate that in our latest outlier finding technique compare with the old way. Our approach yields superior outcomes in terms of accuracy, precision, recall & F-1 score. Additionally, we successfully lowered the machine learning algorithms' under fitting issue.

Keywords- Machine learning, Data analysis, Outlier detector, Isolation forest, Threshold selection.

I. INTRODUCTION *(HEADING 1)*

Heart disease describes a condition of heart. The heart disease which include blood vessel diseases, such as coronary artery disease , heart bits problems (arrhythmias) and heart defects at the time of born(congenital heart defects), among others.

The Today's world has a much better growth of information. Human data processing is not that simple. Systems for data analysis enable even greater insight. With its 'English' inputs and simple syntax, the scripting language Python provides an incredibly potent accessible alternative to established methods and tools.

Organizations can use business intelligence to analyze their productiveness, which in turn enables them to make good decisions. An online store, for instance, would be interested in evaluating consumer profiles to show focused adverts in increasing revenue. If a person is familiar with the technologies able to handle material, findings may be applied to practically any area of an organization. E-commerce organizations are employing the right approach of visualization to analyze consumer reviews. Exploratory statistical analysis (EDA) is a method for condensing knowledge by emphasizing its key features and representing it appropriately. EDA seems to be more strictly focused on addressing incomplete data, transforming parameters as necessary, and validating the hypotheses necessary for pattern building and proposition validation. IDA is a part of EDA [34] [35] [36].

The input sets' amount of rows and columns, incomplete data, info kinds, etc. display are briefly described by EDA [6]. Correct incorrect numbers, resolve incomplete data [7], and mop up contaminated data. EDA uses graphical representations and box plots to visualize statistical parameters [8] [9]. Make relationships visible & calculable.

Exploratory research is a strategy to determine whatever the facts can tell us outside of formal modeling or hypotheses tasks. EDA [5] assists in the investigation of the data sets to identify distinctive qualities, focusing on four key elements: How is data transmitted, exactly? Are there outliers in the data? Are there any missing values? And how do the various attributes relate to one another.

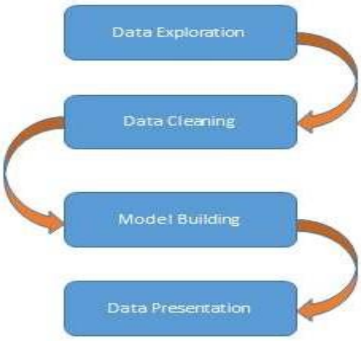


Fig 1. Steps of pre-processing

a) Data Exploration

Users examine and comprehend their data using statistical and graphical techniques during knowledge discovery, sometimes referred to as exploratory data analysis (EDA)[26]-[28]. Choosing a framework or method to employ in the basic sequence, and also spotting issues and trends in the database, are all aided by it process. The following methods are adopted for data exploration.

b) Finding of outliers

The outcomes of a study may be significantly or predominately impacted by outliers [14] or data that are sizable or tiny in value compared to the overall data. Although there may be different ways to cope with issues, it is crucial to be aware of their existence. Tools for identifying outliers include boxplots and Cleveland dot plots [29]-[33].

c) Uniformity of the variance

In qualifications, linear extrapolation modelling, and way anova, unity in variance is a key presumption. It implies that the features factors' variances must be comparable. By investigating the model's residuals, i.e., by plotting multicollinearity vs. predicted value, it can be verified. The variance of the error terms should be comparable in the plots.

d) Distributed data

Many statistical methods, including t-tests and linear regressions, presuppose normalcy. Data variances can be displayed using histograms.

e) Data with a missing value

The investigation becomes more challenging if there are no or null data. Given that they are all exactly zero, they can be wrongly classified as being similar [1] [2].

f) Covariate collinearity

One is likely to produce a perplexing statistical study where nothing is meaningful if a correlation is disregarded. Nevertheless, if one correlational attribute is eliminated, the others may start to matter. Calculating variance inflation factors, paired scatter plots evaluating covariates, or biplots applied to all variables are all methods for spotting collinearity.

g) Relationships between the variables

When two variables interact, their connection will change depending on the value of the second variable. When using regression analysis, this kind of information can be discovered by looking at the weight of the attributes.

h) Impartiality of the data set

A dataset's data points ought to be generated separately. Any geographical time correlations can be modeled for analysis, or information can be nested in a rigid hierarchy.

The effectiveness of an ML algorithm's ability to generalize is significantly impacted by concerns with information pre-processing. The most crucial tasks in preprocessing are to find outliers and reduce the number of missing values. In this research, a novel machine learning-based approach to outlier detection is brought proposed. Here, we take into account a supervised data set, where some data are utilized for training and the rest data are used for testing. The mixture of unsupervised learning with supervised learning increases the accuracy of outlier detection and the usability of methods for detecting outliers by merging the unsupervised learning with the labeled data.

II. OVERVIEW LITERATURE SURVEY

Kenji Yamanishi et al [3] (2004) proposed the theoretical underpinnings of SmartSifter are presented in this publication, along with actual evidence of its efficacy. Through continuous unsupervised of a prediction model of a source of data, SmartSifter may identify outliers in an online application. SmartSifter uses an active discounted learning method to learn the probability model every time a data is entered. SmartSifter's innovative qualities proposed following features (1) its ability to adjust to non-stationary datasets; (2) its clear statistical and details significance of a score; (3) its low processing cost; and (4) its capability to handle either both categorical & continuous variables. The Smart Sifter programmer for online unsupervised outliers has been suggested in this research. Through studies, they provided a probabilistic reasoning for Smart Sifter and showed its effectiveness as a function of correctness and computing time. Aindrila Ghosh et al [5] (2018) examined a set of new exploratory needs for huge datasets while being in analyzing an industry columnar dataset. Furthermore, they give a thorough overview of a recently developed discipline of exploratory data analysis. Authors assess 50 professional and quasi visual data exploration programs with regard to their usefulness inside the 6 phases of an exploratory data analysis process. They also look at how well these contemporary information extraction tools meet the extra requirements for studying enormous datasets. In this research ,researcher conclude that Only a few recent EDA solutions take into account tackling the issues of large analytics, while the majority of products support the basic steps of a EDA procedure. In spite of the tools examined, there is still a barter in between range of allowed variables and comprehensive data processing. Seyedamin Pouriyeh et al [6] (2017) studied the comparison and contrasts the efficacy of several data mining categorization systems for predicting heart problem using ensemble machine learning methods. Author also contributed an amount of information has indeed been increased using 10-Fold Cross-Validation.in this research author applied various machine learning algorithms. It has been determined how well each trained to handle on its own and when used in conjunction using the bagged, boost, and stack strategies. In average, after using the bagged, boost, and stack procedures, author have seen some gains whenever the boost technique was used, SVM fared better than each of the remaining techniques. Mohammad Shafenoor Amin et al [8] (2018) defines the important characteristics and data mines that can increase the precision of heart disease prediction. Seven classification methods and feature set combinations were used to create prediction models. The findings of the study reveal that the automated diagnostic model, which was created utilizing the major features shown to be key and the strongest data mining methodology, has a cardiac illness accuracy rate of 87.4%. Markus M. Breunig et al [9] (2000) suggest that it is greater appropriate to give each item a level of outlierdom for numerous circumstances. The local outlier factor (LOF) of an item is the measure of this extent. It is local in that the extent is determined by how remote an item is from its immediate surroundings. Author include a thorough rigorous analysis demonstrating that LOF has numerous advantageous characteristics. Author show that LOF is able to identify anomalies that seem to be important but cannot be discovered with current techniques using database.

Samir yadav et al [11] (2020) proposed that any model's performance will suffer when all features are taken into account at once when training it. Efficiency is calculated as the rolling sum of each non-diagonal column in a binary classification, where heavier weights are given to more extreme incorrect negative instances. The feature is derived through clinical methods with various costs for diagnosing heart disease. It is considered when choosing the optimum feature set to get the functionality that is both cost- and computationally efficient. Author was successful in developing a Classifier model with an F-score of 0.892 for the issue of identifying heart disease. It has been demonstrated that a filter performs better when its extracted features is smaller. Omar Alghushairy et al [12] (2020) gives a survey of the research on LOF methods for local feature extraction in dynamic & stream contexts. It compiles and classifies current local anomaly identification methods, then examines each one's features. The study also addresses the benefits and drawbacks of those techniques and suggests a number of prospective lines of research for the improvement of local background subtraction approaches for data stream. Author conclude that the techniques for detecting anomalies in stable & streaming situations are reviewed in his study. It analyses the many techniques of regional anomalies detection that have recently developed and much more especially discusses the difficulties of localized anomalies detection in systems. The LOF in system habitats has received additional consideration. The report additionally offers a way for boosting the LOF's effectiveness in stream situations depends on the findings of this research. Wen Jin et al [14] (2001) Describe a fresh technique for quickly locating those top n anomalies in sizable collections. To condense the information, the idea of "micro-cluster" is presented. Having this in mind, an effective segments and sub local anomaly mined technique is created. Author presented a significant trim approach for redundant data since the duplication in the micro-clusters can have a negative impact on their method. The written report and experimentation demonstrate this strategy may successfully locate the most notable anomalies. According to the author, their methodology is excellent at ranking the most improbable anomalies. Fei Tony Liu et al [16] (2008) proposed that according to data testing, iForest outperforms ORCA, a length technique with near-linear computation time, LOF, and Random Forests in terms of AUC and processing time, particularly for big data sets. Additionally, iForest performs well enough in rising issues with a lot of irrelevant qualities and when the training dataset is devoid of outliers. In contrast to typical instance profile, the design approach proposed in this research concentrates on anomalous separation. Lian Duan et al [20] (2008) report a novel concept of an anomaly is a cluster-based outlier. This is crucial and gives weight to the data behavior. Additionally describe how and where to identify outliers using the grouping LDBSCAN, that can locate groups and allocate LOF to specific points. Edwin M. Knorr et al [21] (1997) author examine an obvious concept of anomalies in this research. This writer's major contribution is to demonstrate how the concept of oddities suggested integrates or generalizes various existing notations of anomaly given by discordancy tests with such a single component identifying just the types of outlier presented. This writer's second analysis is the creation of a method for locating all anomalies in a data. Mauricio A Hernadez et al [22] (1997) built a method to complete this feature extraction operation and show how it can be used to cleanse lists of the names for potential clients in an operation similar to direct sales. While the information is processed numerous times utilizing secret key for categorizing on each subsequent pass, they findings for randomly produced data are demonstrated to be precise and efficient. Utilizing reflexive completion to combine the outcomes of multiple rounds over the separate findings yields much more reliable data at a reduced cost. The software has a rules design component that really is simple to set up and successful at locating duplication, particularly in settings where there is a lot of information. M.R. Brito et al [23] (1996) examine the connection among two reciprocal k-NN graphs, as well as the existence of grouping and anomalies, in the information. A method to identify grouping organization and anomalies is suggested, as well as the effectiveness of the analysis is assessed using model output. Edwin M. Knorr et al [25] (1998) introduce easily separate methods, each with a sophistication of O (k N'), where k is the dataset's fractal dimension and N represents the total number of artefacts. These methods easily accommodate sets of data with numerous additional attributes. Furthermore, author present an improved fibroblast method, whose sophistication is sequential in terms of N but incremental in terms of k. Third, they present a new iteration of the cell-based method for data sources which are primarily disk resident, which ensures a maximum of 3 needs to pass through the set of data.

III. OUTLIERS IDENTIFICATION

In particular, filter and wrapper instance choice algorithms are differentiated [23], [24]. Data reduction is the sole factor taken into consideration while evaluating filters; operations are not considered. The ML component is clearly emphasized in wrapper techniques, which also use the ML algorithm's instance selection mechanism to evaluate outcomes [21] [22]. Samples of how this metadata can be used to identify various potential issues with data integrity are provided in Table 1. Additionally, several publications, including [25], focused on the issue of duplicated occurrence detection and deletion. [3], [4],[12]

Table 1

Problems	**Commons**	**Cases**
Inappropriate Values	Gender	Value f gender > 2 showsa problem
	Maximum , Minimum	The range must be inside the values
	Deviation	The deviation is statically value must be less than the threshold value
Incorrect values	Feature	When filtering numbers, incorrect values frequently appear adjacentto the right ones.

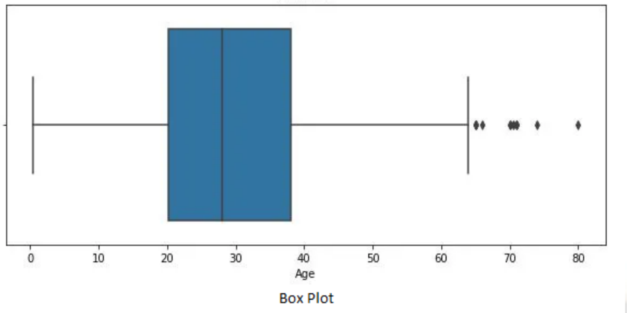
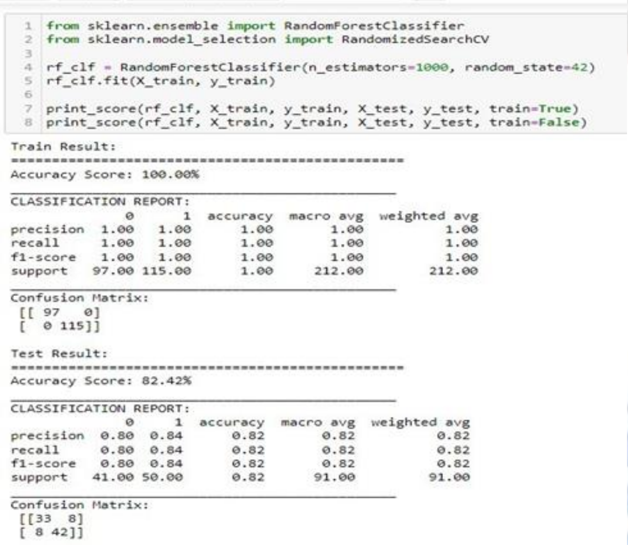


Fig 2. Distribution plot before Log transformation



```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import RandomizedSearchCV

rf_clf = RandomForestClassifier(n_estimators=1000, random_state=42)
rf_clf.fit(X_train, y_train)

print_score(rf_clf, X_train, y_train, X_test, y_test, train=True)
print_score(rf_clf, X_train, y_train, X_test, y_test, train=False)
```

```
Train Result:
================================================
Accuracy Score: 100.00%
_______________________________________________
CLASSIFICATION REPORT:
              0      1  accuracy  macro avg  weighted avg
precision  1.00   1.00      1.00       1.00          1.00
recall     1.00   1.00      1.00       1.00          1.00
f1-score   1.00   1.00      1.00       1.00          1.00
support   97.00 115.00      1.00     212.00        212.00
_______________________________________________
Confusion Matrix:
 [[ 97   0]
 [  0 115]]

Test Result:
================================================
Accuracy Score: 82.42%
_______________________________________________
CLASSIFICATION REPORT:
              0     1  accuracy  macro avg  weighted avg
precision  0.80  0.84      0.82       0.82          0.82
recall     0.80  0.84      0.82       0.82          0.82
f1-score   0.80  0.84      0.82       0.82          0.82
support   41.00 50.00      0.82      91.00         91.00
_______________________________________________
Confusion Matrix:
 [[33  8]
 [ 8 42]]
```

Fig 3. Box plot before Log transformation

In Fig 2, a slight line shows in the range of 60 to 80, which shows the outlier. In Fig 3, dots are shown after the age of 60, that's dot indicates outlier.

IV. Related Proposed Method

A. *Abbreviations and AcronymsDetect Outliers*

A concept of machine learning to create the exclusionary framework of outliers is the foundation for the anomaly detection technique provided in this research. A 2-class model that primarily consists of two components outlier element creation and outlying element threshold selection is created from the anomaly detection model. Construct a discriminatory function.

$$(y) = \ Y_i \geq threshold(t) \qquad (1)$$

Whenever values of (y) are 1 then it considers to an Anomaly, if its value is 0 then it considers correct data. Here Y_i is the outlier variable determined from every dataset pattern's anomaly model. $thresh\ (t)$ is indeed the outlying element threshold, which needs to be learned as a parameter estimate.

B. *Creation of outlier factor*

This article's approach for generating the anomaly value is Isolated Forest [17, 18]. Contrasting density-based and distance-based Isolation Forest specifically eliminates anomalies in its procedures. Instead of usual instances being profiled. In isolation, forest splits are produced by choosing a random characteristic, followed by a divided number between both the feature's quantitative relationships. Primarily requires taking benefit of two characteristics of anomalies: rare and distinct, which makes them more vulnerable than usual to isolation scores. You can efficiently build a binary tree to separate each instance. Due to their vulnerability Anomalies are separated nearer to the tree's base due to isolation, while the bottom end of both the factor are isolated trees. Consider "T'" as a tree node. This node has one test and two daughter nodes, however, it also might not have any offspring [19] [20]. At node point T, there are two tests and an attribute q available. To create a ranking that accurately reflects the level of outlier is the aim of the creation of the outlier factor. When data points are categorized using iTrees as per their mean distances, outliers are those that appear at the forefront of the list. A local outlier factor approach compares the relative concentrations of a point's entire surroundings to determine how dense a point is. A place is potentially an outlier if it is significantly less populated than its surroundings. We get a high LOF [10] [13] that denotes an outlier if the proportion of the densities of the neighbors towards the concentration of the spot is just too large.

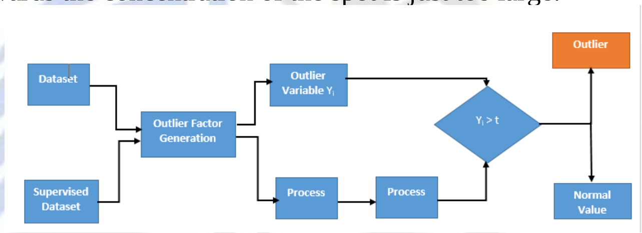


Fig 4. The framework of outlier detection

Here, we used an outlier factor generator using a dataset. Outlier factor Y (i) and the outlier factor produced from the supervised dataset are both subjected to the output of the outlier factor generator. The supervised data set's outlier factor is sent to the threshold values (t). We will now contrast the threshold and the outlier factors Y(i) & (t). If Y(i) exceeds t, it will be regarded as an outlier; otherwise, it will be treated as a normal value.

C. *Units Selection of outlier factor threshold*

The outlier factor threshold is typically chosen by selected by the algorithm's user based on their prior knowledge as well as knowledge of the methodology. This section makes use of some attributes, via the training labels with steadily increasing complexity random searching to choose the Outliers factor threshold.

V. Experimental Analysis

Following experimental analysis, we carried out specific algorithms. With the aforesaid technique, we were able to prevent outliers and enhance the accuracy scores. Here, we used the random forest algorithm to depict two different results: the accuracy score with an orthodox method to calculate outlier detection was 82.42%, and the accuracy score after applying the proposed technique of outlier detection was 84.62%. We also succeeded in reducing the underfitting problem.

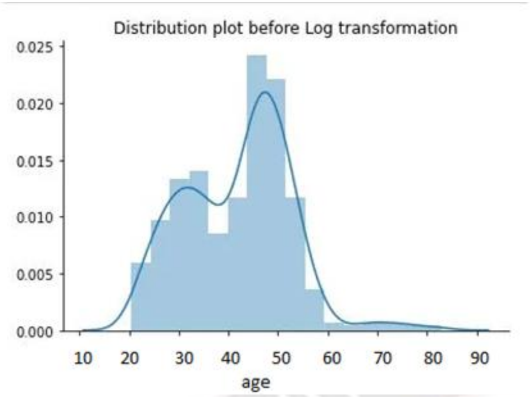
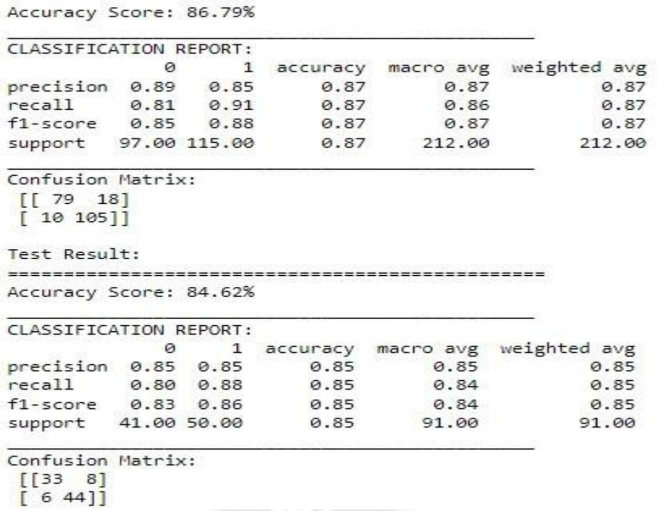


Fig 5. Analysis with orthodox method of outlier detection

With the orthodox method [11] [15] [16] of outlier detection, as shown in the analysis above, the training accuracy score was 100%, but the testing accuracy score was only up to 82.42%. Fig 4 shows after analyzing the results of the outlier identification approach, we discovered that training accuracy was 86.79%, but testing accuracy was only up to 84.62%.



```
Accuracy Score: 86.79%
_______________________________________________
CLASSIFICATION REPORT:
                 0       1  accuracy  macro avg  weighted avg
precision     0.89    0.85      0.87       0.87          0.87
recall        0.81    0.91      0.87       0.86          0.87
f1-score      0.85    0.88      0.87       0.87          0.87
support      97.00  115.00      0.87     212.00        212.00
_______________________________________________
Confusion Matrix:
 [[ 79  18]
 [ 10 105]]

Test Result:
================================================
Accuracy Score: 84.62%
_______________________________________________
CLASSIFICATION REPORT:
                 0      1  accuracy  macro avg  weighted avg
precision     0.85   0.85      0.85       0.85          0.85
recall        0.80   0.88      0.85       0.84          0.85
f1-score      0.83   0.86      0.85       0.84          0.85
support      41.00  50.00      0.85      91.00         91.00
_______________________________________________
Confusion Matrix:
 [[33  8]
 [ 6 44]]
```

Fig 6.Analysis after the outlier detection method

Table 2

Classification Report			Precision	Recall	F1-Score	Accuracy
Analysis with Orthodox method of outlier detection	0	Training	0.1	0.1	0.1	100%
		Testing	0.80	0.80	0.80	82.40%
Analysis after the outlier detection method	1	Training	0.89	0.81	0.85	86.79%
		Testing	0.85	0.80	0.83	84.62%

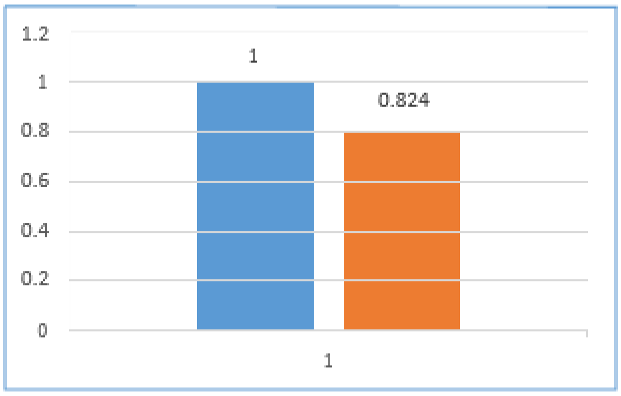


Fig 7. Accuracy of a model after using an orthodox method for outlier detection

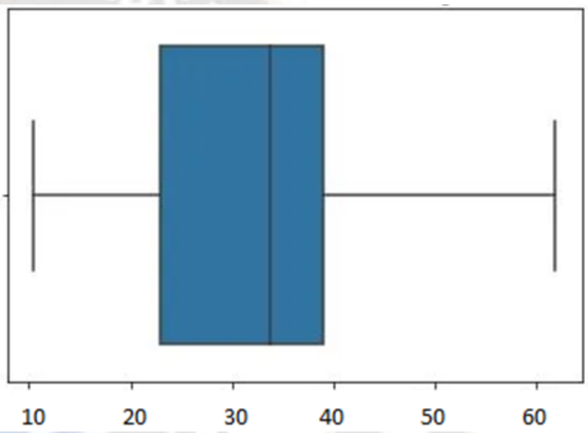


Fig 8. Accuracy of a model after using a proposed method of outlierdetection

In Fig 5, when using the conventional approach of outlier detection, the blue block in Fig. 5 depicts training accuracy as 100% and the orange block, testing accuracy as 82.40 percent. The blue block in Fig. 6 depicts training accuracy as being 86.79%, while the orange block depicts testing accuracy as being 84.62%. So here under fitting issue is also reduced. We were able to successfully eliminate the machine learning using algorithm under fitting issues.

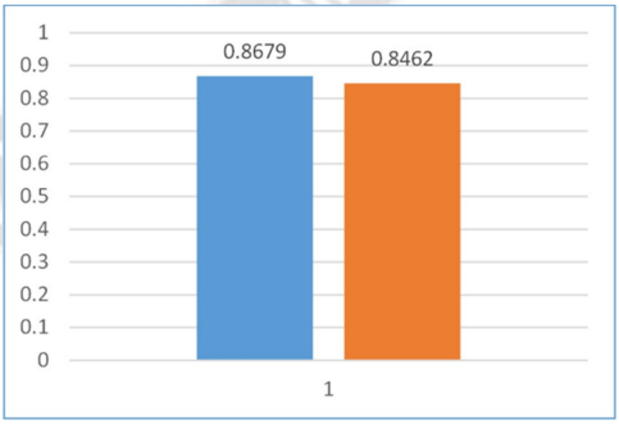


Fig 9. Box Plot after removing Outliers

VI. CONCLUSION

Specifically addressing issues with the application such as, the insufficient ability to gauge the threshold and the method's applicability, a novel outlier's identification technique using machine learning is created using the few data labels available, along with supervised learning. A new threshold method of selection is suggested. The formula is further combined by using collective learning to enhance detection performance. The experimental results support the finding. Effect of the novel outlier detection technique, enhance the original detection algorithm's detection precision, and the method's applicability, and address the issue of choosing a threshold based on experience. The underlying technical problems need more research in the future such as we need to take into account all parameters for EDA, such as deleting null values, outliers, and correlation which may increase the accuracy of any model. The problem of under fitting in machine learning algorithms was successfully resolved.

ACKNOWLEDGMENT

This paper has been posted as a preprint on Research Square. Preprint assigned a Digital Object Identifier (DOI) . Here I am attaching DOI link
https://doi.org/10.21203/rs.3.rs-2376758/v1

REFERENCES

[1] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer-aided decision making for heart disease detection using the hybrid neural network-genetic algorithm," Comput. Methods Programs Biomed., vol. 141, pp. 1926, Apr. 2017.

[2] H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, "A multilayer perceptron-based medical decision support system for heart disease diagnosis," Expert Syst. Appl., vol. 30, no. 2, pp. 272-281, 2006.

[3] Yamanishi K, Takeuchi J I, Williams G, et al. "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms", [J]. Data Mining & Knowledge Discovery, 2004, 8(3):275-300.

[4] Han Jiawei, Micheline Kamber, Pei Jian, et al. "Data Mining Concept and Technology ", [M]. China Machine Press, 2012.

[5] [5] Aindrila Ghosh, Mona Nashaat, James Miller, Shaikh Quader, and Chad Marston, "A Comprehensive Review of Tools for Exploratory Analysis of Tabular Industrial Datasets," Visual Informatics, Volume 2, Issue 4, December 2018, pp. 235-253.

[6] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in Proc. IEEE Symp. Comput. Commun. (ISCC), Jul. 2017, pp. 204-207.

[7] [7] V. Bernett and T.Lewis, "Outlier in statical data", john Velly & Sos,1994.

[8] [8] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease", Telematics Informat., vol. 36, pp. 82-93, Mar. 2019.

[9] [9] Breunig M M, Kriegel HP, Ng R T, et al. "LOF: identifying densitybased local outliers". Acm Sigmod Record, 2000, 29(2):93- 104.

[10] [10] Yadav, S.S., Jadhav, "S.M.: Machine learning algorithms for disease prediction using iot environment", International Journal of Engineering and Advanced Technology 8(6), 4303{4307 (2019).

[11] Yadav Samir & Jadhav Shivajirao, "Machine Learning-Based Cardiovascular Disease Diagnosis Using Feature Selection." 10.13140/RG.2.2.15329.33127,2020.

[12] Raed Alsini et al. "A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams", Article in Big Data and Cognitive Computing • December 2020

[13] M. Ashraf, S. M. Ahmad, N. A. Ganai, R. A. Shah, M. Zaman, S. A. Khan, and A. A. Shah, "Prediction of Cardiovascular Disease Through Cutting-Edge Deep Learning Technologies: An Empirical Study Based on TensorFlow, PYTORCH and KERAS", Singapore: Springer, 2021, pp. 239-255.

[14] W. Jin, A.K. Tung, J. Han, "Mining top-n local outliers in large databases, in Proceedings", of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001.

[15] Liu F T, Kai M T, Zhou Z H. "Isolation-Based Anomaly Detection[J]. Acm Transactions on Knowledge Discovery from Data", 2012, 6(1):1- 39.

[16] Liu F T, Kai M T, Zhou Z H. "Isolation Forest". Eighth IEEE International Conference on Data Mining. Pisa, Italy, 2008:413-422.

[17] Trunk, G. V. (July 1979). "A Problem of Dimensionality: A Simple Example", IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (3): 306307.

[18] F. Andreotti, F. S. Heldt, B. Abu-Jamous, M. Li, A. Javer, O. Carr,S. Jovanovic, N. Lipunova, B. Irving, R. T. Khan, R. Dürichen, "Prediction of the onset of cardiovascular diseases from electronic health records using multi-task gated recurrent units," 2020, arXiv:2007.08491. https://arxiv.org/abs/2007.08491

[19] L. Duan, et al., "Cluster-based outlier detection", Ann. Oper. Res. 168 (1) (2009) 151–168.

[20] Marek Grochowski, Norbert Jankowski "Comparison of Instance Selection Algorithms II. Results and Comments", ICAISC 2004a: 580- 585.

[21] Knorr E. M., Ng R. T "A Unified Notion of Outliers: Properties and Computation", Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD'97), Newport Beach, CA, 1997, pp. 219-222.

[22] [22] Hernandez, M.A.; Stolfo, S.J "Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem. Data Mining and Knowledge Discovery", 2(1):9-37, 1998.

[23] M.R. Brito, E.L. Chavez, A.J. Quiroz, et al., "Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection", Stat. Probab. Lett. 35 (1) (1997) 33–42.

[24] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander. "Loft Identifying density-based local outliers". In Proc. 2000 ACM- SIGMOD Int. Conf. Management of Data (SIGMOD'O0), Dallas, Texas, 2.

[25] E. Knorr and R. Ng. "Algorithms for mining distance-based outliers in large datasets", In Proc. 1998 Int. Conf. Very Large Data Bases (VLDB'98), pages 392-403, New York, NY, Aug. 1998.

[26] Jingke Xi, "Outlier Detection Algorithms in Data Mining", Second International Symposium on Intelligent Information Technology Application, 12/2008.

[27] Gupta, Manish, Jing Gao, Charu C. Aggarwal, and Jiawei Han, "Outlier Detection for Temporal Data: A Survey", IEEE Transactions on Knowledge and Data Engineering, 2014.

[28] Christy, A., G. Meera Gandhi, and S.Vaithyasubramanian, "Cluster Based Outlier Detection Algorithm for Healthcare Data", Procedia Computer Science, 2015.

[29] Bin Wang, "Distance-Based Outlier Detection on Uncertain Data", 2009 Ninth IEEE International Conference on Computer and Information Technology, 10/2009.

[30] Seung Kim, Nam Wook Cho, Bokyoung Kang, and Suk-Ho Kang, "Fast Outlier detection for very large log data," Expert system with an application, 2011.

[31] V.Kathiresan, Dr.N.A.Vasanthi, "A Survey on Outlier Detection Techniques Useful for Financial Card Fraud Detection" 2015, International Journal of Innovations in Engineering and Technology (IJIET).

[32] C.C.Agrawal, Philip S. Yu, "Outlier detection with uncertain data," Proceedings of the 2008 siam international conference on data mining, 2008.

[33] Peng Yang, Qingsheng Zhu, "Finding Key Attributes subset in a dataset for outlier detection," Knowledge-Based System, Vol: 24, No: 2, Pp. 269- 274,2011.

[34] HG. Kim, "Environmental Sound Event Detection in a Wireless Acoustic Sensor Networks for Home Telemonitoring," China Communications, vol. 14, no. 9, 2017, pp. 1-10.

[35] D.M. Hawkins, "Identification of outliers [M]," London: Chapman and Hall, 1980.

[36] J. Huang, Q. Zhu, L. Yang, et al., "A novel outlier cluster detection algorithm without top-n parameter," Knowledge-Based Systems, vol. 121, 2017, pp. 32-40.

[37] B. Tang, H. He, "A local density-based approach for outlier detection," Neurocomputing, vol. 241, 2017, pp. 171-180.

[38] C.S. Hemalatha, V. Vaidehi, R. Lakshmi, "Minimal infrequent pattern based approach for mining outliers in data streams," Expert Systems with Applications, vol. 42, no. 4, 2015, pp. 1998- 2012.