Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765669

COMPARATIVE ANALYSIS OF CLASSIFICATION AND REGRESSION MODELS FOR AUTISM

[1]Yugandhar Bokka,[2]R.N.V. Jagan Mohan,[3]M. Chandra Naik
[1]Research Scholar, Gandhi Institute of Engineering and Technology (GIET) University-Gunupur, Odisha- 765022.
[2]Associate Professor, Sagi Rama Krishnam Raju Engineering College, Bhimavaram-534 204.
[3]Professor, GIET University-Gunupur,Odisha- 765022.
e-mail: bokka.yugandhar@giet.edu[1], mohanrnvj@gmail.com[2], srichandra2007@gmail.com[3]

Abstract:
An established subset of neurological disorders is known as autism spectrum disorder (ASD). Communication and social interaction ability may suffer a lifetime effect as a result of ASD. Autism symptoms begin appearing in children as early as three years old, and they continue to intensify as they get older, into adolescence and adulthood. ASD recovery is made possible by earlier diagnosis and prediction. In many medical and healthcare systems, machine learning algorithms are used widely. In this study, a variety of machine learning classification and regression models are applied to the electronic health records of pregnant women who had children with ASD in order to identify the disease at an early stage. The aim of this research is to make early ASD predictions.

Keywords: *Autism, classification, Regression, Electronic Health Records, Gestational Period.*

1. INTRODUCTION

The prevalence of autism spectrum disorder (ASD) has recently been increasing rapidly in people of all ages and genders[1,2]. ASD is problematic because there are too many illnesses with symptoms that resemble those of ASD [3]. Researchers want to continue being watchful in their efforts to understand the etiology of this disease even though a range of factors may show to be both cause and quasi [4,5]. Beginning in the third year of life and continuing throughout one's entire lifespan is an indication of ASD. No experimental therapy currently in use can completely cure this illness.

The few studies that have been done have assessed young autistic children. Several investigators have found structural abnormalities in the brain earlier in the disease symptoms of autism as a result of cohort study of brain development. Environmental factors also contribute to the emergence of ASD in people [6,7]. In India, pregnant women are disproportionately affected by obesity and diabetes. Women who are obese during pregnancy run the risk of developing a number of serious health issues, including gestational diabetes, preeclampsia, and sleep apnea [8]. Diabetes and obesity reach epidemic levels around the same time. Nowadays, which was before the diabetes, obesity, and gestational risks affect more than one-third of women of reproductive age [9]. The influence of maternal obesity before she became pregnant and maternal diabetes on the risk of autism spectrum disorder (ASD), either separately or jointly, have not yet been studied.

Maternal diabetes has now been linked to a child's incidence of autistic spectrum disorder. There has been conflicting research on the link between ASD and maternal pre-pregnancy obesity. The combined effects of these two conditions have not yet been investigated in any studies. Obesity and diabetic in mother were linked to a more severe risk of ASD than obesity or diabetic alone. It is possible that early diagnosis of some diseases may be aided by using models, which give a prediction for a patient's health. These studies of gestational period of mothers who have a child with ASD and Non-ASD help to minimize growth rate of ASD in infants.

The following outlines the topics covered in this article: Introduction is covered in Section 1. The literature review is in Section 2. The related work is discussed in Section 3, and the proposed technique for this study is discussed in Section 4. Section 5 discusses the results, and finally conclusion in Section 6.

2. LITERATURE REVIEW

The proposed that predictive performance [10] be used to predict symptoms in a variety of people. It is compared to a statistical technique called logistic-regression. Predictive performance, according to the results, has higher classification accuracy than logistic regression. A survey was used to gather data for the Random-Forest classifier [11], this model used to predict the stress levels and depression among workers in various locations. The Random Forest classifier is tested with four other models in order to determine the accurate algorithm to estimate the psychiatric issues. Study found that the compared classifier performs poorly at classifying and predicting psychological issues. The differentiation is implemented using different techniques such as artificial neural network, Support vector machine, logistic regression, Random Forest and gradient boosting [12]. Findings showed that logistic regression performed better at discriminating between anxiety and other psychiatric conditions. It was proposed to use genome-wide logistic regression [13] to identify the most significant mutations that are linked to the response to duloxetine.

a classification techniques study on ASD. This paper's main goals were to detect the levels of autism and the issues with it. SVM, Fuzzy, and WEKA tools are utilised in this neural network to study the social interaction and behavior [14]. This research has proposed method for a minimum number of characteristics for autism detection. In this, the clinical assessment of ASD was analyzed using a machine learning approach [15].

3. RELATED WORK

This research into the link between motherly obesity and the development of ASD has produced conflicting results. The results show that maternal prenatal diabetes raises the risk of autism spectrum disorder (ASD) in children. Sometimes have study attempts to identify the individual and combined effects of obesity and diabetes in ASD, however the fact that the two conditions are often comorbid. A method for efficiently extraction of features from huge amounts of data and able to produce specific diagnostic predictions is provided by machine learning [16-20].

In this paper, we study the electronic health records of mothers in her gestational period who have ASD and Non-ASD child. We've listed the dataset's characteristics in Table-1 so that you can understand it easily. This Table-1 contain nine variables, namely, age, Obese, Thyroid, Diabetic, Blood Pleasure, TC, LDL, HDL and TG. The women at high risk for pregnancy included those who were obese and over 25, as well as those who had one or more of the risk factors HDL less than 35 mg/dL, TG more than 250 mg/dL, TC more than 240 mg/dL, and LDL more than 190 mg/dL. We analyzed these data using various classification and regression models in machine learning to predict the ASD in infants.

Table-1: List of Attributes in the dataset

S.NO	AGE	OBESE	THYROID	DIABETIC	BP	TC	LDL	HDL	TG	TARGET
1	38	28	100	70	121	195	95	65	100	1
2	24	29	90	100	110	100	80	68	100	1
3	28	29	95	72	126	186	65	77	123	1
4	26	31	158	72	120	237	127	53	179	1
5	30	34	149	98	118	212	200	51	168	1
6	34	31	98	65	117	208	129	54	250	1
7	37	33	171	72	117	204	118	56	246	1
8	38	28	100	70	121	195	95	65	100	0
9	24	29	90	100	110	100	80	68	100	0
10	28	29	95	72	126	188	71	63	127	0
11	26	31	102	70	116	215	125	51	170	0
12	30	35	161	97	120	201	139	57	195	0
13	34	32	185	91	124	208	149	49	207	0
14	37	31	175	73	119	231	110	39	355	0

4. PROPOSED METHOD

The most of the studies involves machine learning techniques, which limits its performance. In order to accomplish this objective, this research compares the performance of various regression and classification models. The proposed workflow, which includes data preprocessing, assessment of results, and ASD prediction, is shown in Figure-1 below.

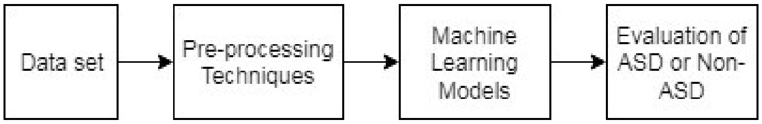

Figure-1. Steps in the proposed ASD detection

Pre-processing involves converting large amount data into a meaningful and understandable. In real world, the data is inconsistent and incomplete because it contains missing values and noise. Data that has been properly processed always yields good results. To deal with incomplete and inconsistent data, a variety of pre-processing techniques are used. The entire dataset is divided into two partition, with one portion being used for training and the other for testing, with a ratio of 80:20 each.

Classification is a supervised learning technique, which is used to predict the class labels whose class label is unknown.

Machine learning techniques:

- *Naïve bayes classifier*: This technique is based on bayes theorem. This model easy to build for large datasets.
- *Support vector machine (SVM):* SVM is widely used for pattern recognition and classification. It is also used to classify the images.
- *K- Nearest neighbor (KNN):* K-Nearest neighbor is a simple learning technique, its classifies based on similarity measures.
- *Decision trees*: It is a tree representation to solve the problem. Class labels are represented at leaf nodes and attributes at internal nodes.
- *Random forest*: Random Forest is a classification algorithm; it is used to randomly create decision trees and merges multiple decision trees into one forest.

Regression is used to estimate continuous values. Plot the line between the data that best fits the data is the main goal of the regression method, it shown figure-2 regression plot. There are several types of regression models, including polynomial, logistic, ridge, and lasso regression.

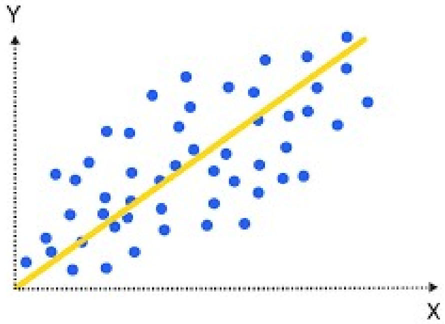

Figure-2. Regression plot

Regression analysis is a very important in the field of machine learning. Because both the input and output labels are trained to the algorithm, it is classified as supervised learning. It helps create a link between the variables by analyzing how one affects the other.

Lasso regression, which means Least Absolute Shrinkage and Selection Operator, is a significant study in this work. Lasso regression is also known as penalized regression technique. The selection of the subset of variables generally makes use of this method in machine learning. Its prediction accuracy is higher than that of other regression models. Un-important features in a dataset are identified by the lasso regression, those values are set to zero, removing them. A dataset with large-scale datasets and correlation is perfect for lasso regression. The given formal is

$$\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{P}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{P}\left|\beta_j\right|$$

Where λ shrinkage factor, βj regression coefficient.

5. RESULT AND DISCUSSION

In this study using various classification models, the random forest classifier model is comparing to other classification models, this model accuracy is better than each other classification models. As compared to other models, the random forest model's accuracy is 0.58. The classification models' accuracy values are shown in Table-2. The current analysis shows that the Lasso regressor model has done better than every other regression model. As compared to other models, the Lasso model's RMSE is calculated as values that are 0.5 smaller. The regression models' RMSE values are shown in Table-3.These analyses using correlation showed that several variables were highly correlated to each other, as shown in Figure-3.

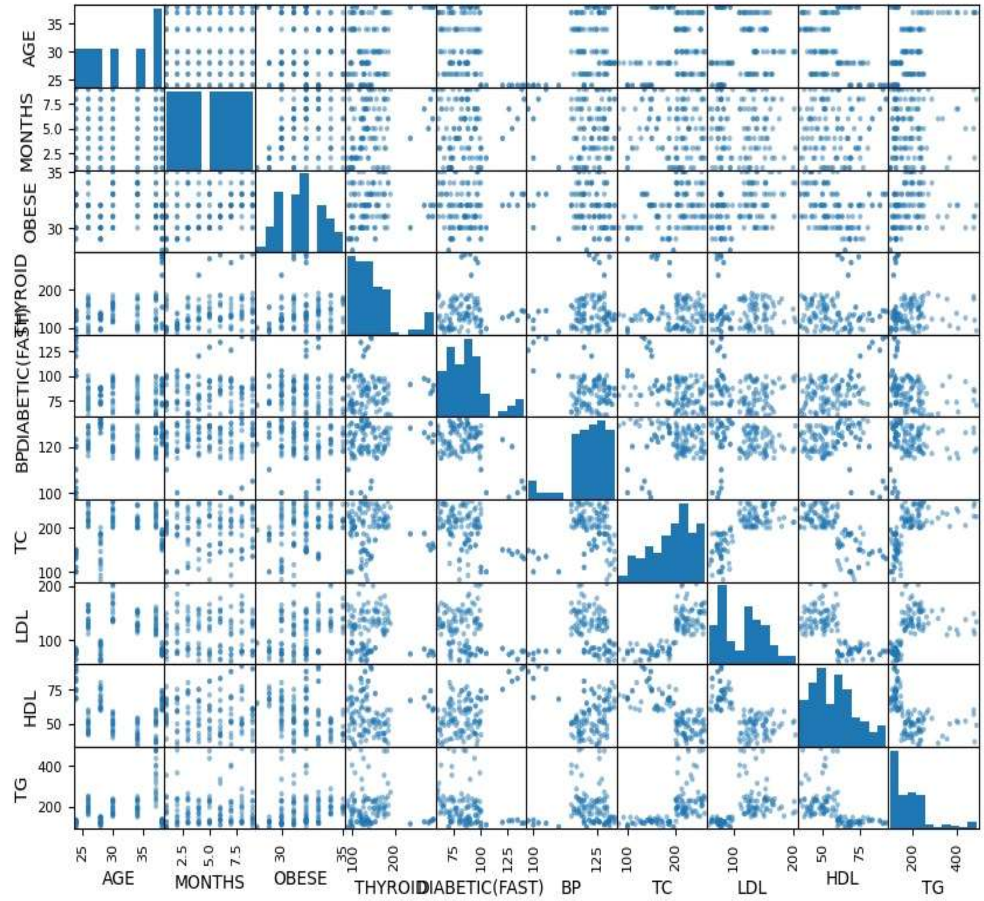

Figure-3 Correlation coefficients between all the variables

Evaluation Metrics:

Confusion Matrix: confusion matrix is a table; it contains a classification model performs on a set of test data.

		Actual values	
		Postive	Negative
Predicted	Postive	TP	FP
	Negative	FN	TN

Accuracy: accuracy counts how frequently the classifier predicts correctly. Accuracy is determined by dividing the total number of predictions by the number of true predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision is calculated as the ratio of actual positive values to predicted positive values.

$$\Pr ecision = \frac{TP}{TP + FP}$$

Recall (Sensitivity): Recall is calculated by dividing the total number of positive outcomes by the number of true positives.

$$\mathrm{Re}\,call = \frac{TP}{TP + FN}$$

F1 Score: F1-Score is a combined of Precision and Recall.

$$F1 - score = \frac{\mathrm{Pr}\,ecision * \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call}$$

Table-2: The model performance measures of Various classification techniques

Model	Accuracy	Balanced_Accuracy	ROC_AUC	F1-Score	Time_Taken
RidgeClassifierCV	0.42	0.44	0.44	0.42	0.01
DecisionTreeClassifier	0.42	0.44	0.44	0.42	0.02
AdaBoostClassifier	0.42	0.43	0.43	0.43	0.04
BernoulliNB	0.42	0.43	0.43	0.43	0
GaussianNB	0.35	0.36	0.36	0.34	0.01
KNeighborsClassifier	0.35	0.35	0.35	0.35	0
RandomForestClassifier	0.58	0.58	0.58	0.58	0.06

Mean Squared Error: A common error metric for regression models is mean squared error. Least squares minimization refers to reducing the mean squared error between predictions and expected values. The MSE is calculated using the average of the squared discrepancies between the expected and target values in a dataset.

$$MSE = \frac{1}{N} * \sum_{i=1}^{N} \left(x_i - y_i\right)^2$$

Where x_i is the expected value in the dataset and y_i is the predicted value.

Root Mean Squared Error: The root mean squared error is a type of the mean squared error. Finding the square root of the error is important since it shows that the goal value being predicted and the RMSE's units are the same. As a result, it would be typical to train a regression predictive model using MSE loss and then evaluate and report its performance using RMSE.

The RMSE can be calculated as follows:

$$RMSE = sqrt\left(\frac{1}{N} * \sum_{i=1}^{N} \left(x_i - y_i\right)^2\right)$$

Where x_i is the expected value and y_i is the predicted value.

Table-3: The model performance measures of various regression techniques

Model	Adjusted_R-Squared	R-Squared	RMSE	Time_Taken
LassoLarsCV	-0.73	-0.04	0.5	0.01
LarsCV	-0.73	-0.04	0.5	0.02
DummyRegressor	-0.73	-0.04	0.5	0
ElasticNet	-0.73	-0.04	0.5	0.01
ElasticNetCV	-0.73	-0.04	0.5	0.02
LassoLarsIC	-0.73	-0.04	0.5	0.01
LassoLars	-0.73	-0.04	0.5	0
LassoCV	-0.73	-0.04	0.5	0.05
Lasso	-0.73	-0.04	0.5	0.01
BayesianRidge	-0.74	-0.05	0.51	0
NuSVR	-0.85	-0.11	0.52	0
PoissonRegressor	-0.88	-0.13	0.52	0.02
TweedieRegressor	-0.94	-0.17	0.53	0
OrthogonalMatchingPursuit	-1.05	-0.23	0.55	0.02
RidgeCV	-1.11	-0.27	0.56	0
AdaBoostRegressor	-1.15	-0.29	0.56	0.07
SGDRegressor	-1.15	-0.29	0.56	0
Ridge	-1.19	-0.32	0.57	0
OrthogonalMatchingPursuitCV	-1.2	-0.32	0.57	0
Lars	-1.21	-0.32	0.57	0.01
TransformedTargetRegressor	-1.21	-0.32	0.57	0
LinearRegression	-1.21	-0.32	0.57	0
HuberRegressor	-1.39	-0.43	0.59	0.02
KNeighborsRegressor	-1.47	-0.48	0.6	0.01
LGBMRegressor	-1.51	-0.5	0.61	0.03
HistGradientBoostingRegressor	-1.51	-0.51	0.61	0.11
RANSACRegressor	-1.52	-0.51	0.61	0.02
RandomForestRegressor	-1.57	-0.54	0.61	0.07
LinearSVR	-1.82	-0.69	0.64	0
BaggingRegressor	-1.86	-0.72	0.65	0.01
SVR	-2.08	-0.85	0.67	0.02
GradientBoostingRegressor	-2.2	-0.92	0.68	0.02
MLPRegressor	-2.31	-0.99	0.7	0.08
ExtraTreesRegressor	-2.33	-1	0.7	0.05
PassiveAggressiveRegressor	-2.46	-1.08	0.71	0
DecisionTreeRegressor	-2.48	-1.09	0.71	0
XGBRegressor	-2.54	-1.12	0.72	0.06

QuantileRegressor	-2.94	-1.36	0.76	0.03
GaussianProcessRegressor	-2.94	-1.37	0.76	0
ExtraTreeRegressor	-3.07	-1.44	0.77	0.02
KernelRidge	-3.56	-1.74	0.82	0.01

The attributes of Obesity, HDL, LDL, TG, BP, and Diabetes are used in regression models in this study. The ideal scaling regression analysis between age and gestational obesity was conducted using the testing data set. The chance of ASD slowly increases with age, but the rise is not linear; beyond age 38, the risk of ASD rises more quickly with age. So far, early in pregnancy, pregnant women had HDL and LDL levels above normal.

6. CONCLUSION AND FUTURE WORK

This study to determine whether classification and regression models used with electronic health records could correctly predict ASD in children. The most important measures for minimizing ASD symptoms and to improve quality of life are early diagnosis and treatment.

In this study, machine learning classification and regression models were used to detect autism spectrum disorder. The model performances are used for ASD detection on Electronic health data was examined using a variety of performance evaluation metrics. When comparing the output of the various classification and regression models. The random forest classifier shows the highest accuracy for detecting of Autism Spectrum Disorder. Lasso regression models in these studies show the dataset's lowest RMSE. These results strongly indicate that, in place of the various machine learning models described in earlier work, the detection of Autism Spectrum Disorder can be accomplished using the Lasso regression model and the random forest model. Compared to all other models for creating models, these models were able to simulate data the best. The dataset has few features. For future development, if the dataset has more features, then the model is providing best accuracy and optimize the error.

REFERENCES

[1] Raj, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science*, *167*, 994-1004.

[2] BLESSIE, B. G. D. E. C. (2021). Autism spectrum disorder prediction using robust kalman filtering based neural network. *Journal of Theoretical and Applied Information Technology*, *99*(11).

[3] Thabtah, F. F. (2017). Autistic spectrum disorder screening data for children data set. *UCI machine learning repository*.

[4] Epifânio, J. C., & Da Silva, L. F. (2020). Scrutinizing reviews on computer science technologies for autism: Issues and challenges. *IEEE Access*, *8*, 32802-32815.

[5] Sidhu, G. (2019). Locally linear embedding and fMRI feature selection in psychiatric classification. *IEEE journal of translational engineering in health and medicine*, *7*, 1-11.

[6] Akter, T., Satu, M. S., Khan, M. I., Ali, M. H., Uddin, S., Lio, P., ... & Moni, M. A. (2019). Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access*, *7*, 166509-166527.

[7] Sun, J. W., Fan, R., Wang, Q., Wang, Q. Q., Jia, X. Z., & Ma, H. B. (2021). Identify abnormal functional connectivity of resting state networks in Autism spectrum disorder and apply to machine learning-based classification. *Brain Research*, *1757*, 147299.

[8] Ama Moor, V. J., Ndongo Amougou, S., Ombotto, S., Ntone, F., Wouamba, D. E., & Ngo Nonga, B. (2017). Dyslipidemia in patients with a cardiovascular risk and disease at the University Teaching Hospital of Yaoundé, Cameroon. *International journal of vascular medicine*, *2017*.

[9] Morales, D. R., Slattery, J., Evans, S., & Kurz, X. (2018). Antidepressant use during pregnancy and risk of autism spectrum disorder and attention deficit hyperactivity disorder: systematic review of observational studies and methodological considerations. *BMC medicine*, *16*, 1-14.

[10] Hatton, C. M., Paton, L. W., McMillan, D., Cussens, J., Gilbody, S., & Tiffin, P. A. (2019). Predicting persistent depressive symptoms in older adults: a machine learning approach to personalised mental healthcare. *Journal of affective disorders*, *246*, 857-860.

[11] Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science*, *167*, 1258-1267.

[12] Na, K. S., Cho, S. E., & Cho, S. J. (2021). Machine learning-based discrimination of panic disorder from other anxiety disorders. *Journal of Affective Disorders*, *278*, 1-4.

[13] Maciukiewicz, M., Marshe, V. S., Hauschild, A. C., Foster, J. A., Rotzinger, S., Kennedy, J. L., ... & Geraci, J. (2018). GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. *Journal of psychiatric research*, *99*, 62-68..

[14] Mythili, M. S., & Shanavas, A. M. (2014). A study on Autism spectrum disorders using classification techniques. *International Journal of Soft Computing and Engineering*, *4*(5), 88-91.

[15] Kosmicki, J. A., Sochat, V., Duda, M., & Wall, D. P. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational psychiatry*, *5*(2), e514-e514.

[16] Sahni, N., Simon, G., & Arora, R. (2018). Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: a proof-of-concept study. *Journal of general internal medicine*, *33*, 921-928.

[17] Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H., & Luscombe, N. M. (2018). Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PloS one*, *13*(8), e0202344.

[18] Devika Varshini, G., & Chinnaiyan, R. (2020). Optimized machine learning classification approaches for prediction of autism spectrum disorder. *Ann Autism Dev Disord. 2020; 1 (1)*, *1001*.

[19] Cheng, W., Fang, L., Yang, L., Zhao, H., Wang, P., & Yan, J. (2014, December). Varying coefficient models for analyzing the effects of risk factors on pregnant women's blood pressure. In *2014 13th International Conference on Machine Learning and Applications* (pp. 55-60). IEEE.

[20] Alwidian, J., Elhassan, A., & Ghnemat, R. (2020). Predicting autism spectrum disorder using machine learning technique. *International Journal of Recent Technology and Engineering*, *8*(5), 4139-4143.