2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-1451-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/Confluence51648.2021.9377162

Comparison study of Regression Models for the prediction of post-Graduation admissions using Machine Learning Techniques

Naveen S.Sapare
Dept of Engineering Mathematics
K. L. E. Institute of Technology
Hubballi-580027, Karnataka, India
naveen.sapare@gmail.com

Sahana M.Beelagi
Dept of Computer Science
Jain College of Engineering and Technology
Hubballi-580032, Karnataka, India
sahana.saa@gmail.com

***Abstract*-In today's technological world, a student's graduate performance plays a vital role in building either a net worthy career or opting different university for master's studies. One simple wrong decision made while Shortlisting University without the knowledge of university ranking by a student can ruin an entire year of hard work and success. A poor university choice may conflict with the student's inner gift and talent, wasting invested time and can cause confusion in choosing the right path and directions. Especially the student who is opting for a master's degree based on their GRE/TOFEL score face real difficulty in choosing different research-based universities that need a high score in these exams. This study can help the student to take a measurable step towards selecting. Our study focuses on using analytics to propose a model for predicting the chance of admissions. This paper attempts to define different regression models and predict students' chances of getting admissions. Prediction is performed using regression models, namely linear regression, ridge regressions, lasso, KNN, and elastic net regression, support vector machine, and few other regression models based on the available data. Further, our study explores two different sampling methods naming random forest and cross-validation sampling. All these regression methods are used to predict students' chance of admitting to the University of their Interest based on their graduate performance. The best regression method is used to predict new unseen data.**

Keywords—machine learning, regression, academic performance, validation

I. INTRODUCTION

Machine Learning (ML) is a promising application in information science and technology and has wide applications. In the education domain, machine learning plays an essential role in predicting student performance and improves performance. With its enormous ability to predict the future of student performance, machine learning is helping to redefine and shape the educational system. With the help of machine learning and intelligent technology, we can take our education to a new height and help the student choose their carrier smartly. A student is opting for a master's degree face difficulty selecting universities, mainly research-based universities that require a high score in GRE/TOFEL exams, which students may be unaware of. Students need to meet specific requirements while taking admission and score minimum marks to fulfill the eligibility criteria for individual universities' admission. Wrong decisions while applying, may spoil a student's entire year effort. In this case, our machine learning model developed helps to predict the admission based on the student performance parameters suggests ideas about the selection criteria**.** In today's modern day's value-based education is gaining more popularity. Education is the main backbone of our nation, and it is necessary to sustain quality. Much work is carried out in performing regression methods for measuring student performance. The data is usually collected from students. For this collected student data, different machine learning and data mining technology are applied to find the hidden patterns and relationship that is further used in predicting many unseen data and extracting valuable information[1].[2]The more significant work is done to predict admission using multiple linear regression and classification[3]. Here the author explains to predict student performance on a specific group of courses. The model proposed is tested by CAS's dataset [4]. Different prediction models using regression over is performed in the higher school education domain to process the academic (structured data) with a career [5]. A study proposed by the author [6] focuses on identifying predictor variables to learn the intrinsic relationship between the identified features and students' academic grades. In [7] the author presents a way of predicting student grades by making students into different clusters. Early prediction of grades is predicted mainly using two methods linear regression and clustering. [8] The author presents a machine learning technique to improve the hyper parameters three different techniques of optimization, namely Adaptive Moment Estimation, Limited-memory Broyden-Fletcher-Goldfarb-Shanno, and Stochastic Gradient Descent. Multi-Layer-Perceptron (MLP) is mainly used for data set analysis. Data mining and machine learning techniques is used to predict the performance of student which predicts their grades. Linear regression and Markov network is used in this process [9]. Several ML algorithms were compared in [10] to predict student results mainly using mainly different predictive analytics techniques of linear regression and neural networks. Work presented in [11] mainly used different machine learning models namely logistic regression, Support vector machines, Step regression and support vector Regression. Total of 1995 student's data is used for prediction of results and later compared with actual results obtained. Support vector machine model is chosen over others with accuracy of 77%.Academic analytics is often used in education domain which helps to build strong research in institutions and improving decision making and performance measuring capability[12][13].Many research is being carried out in education to understand the student learning patterns and also to predict different results based on the input data. In our study we focus mainly to study the importance of predicting the result of graduate aspirant which intern helps student to find the gap in their study and prepare for the competitive exams namely GRE and TOFEL.The study of predicting the result is focused in our work. Our work includes,

- Collection of data and performing normalization.
- Applying different regression models namely linear regression, lasso, elastic net, Decision tree, KNN, Random Forest, Ridge regression and support vector study.
- Applying validation method namely random sampling and cross validation sampling.
- Comparing the performance of these different regression models.
- Using the best method to predict the unseen data

II.METHODOLOGY

A. Information about the data set

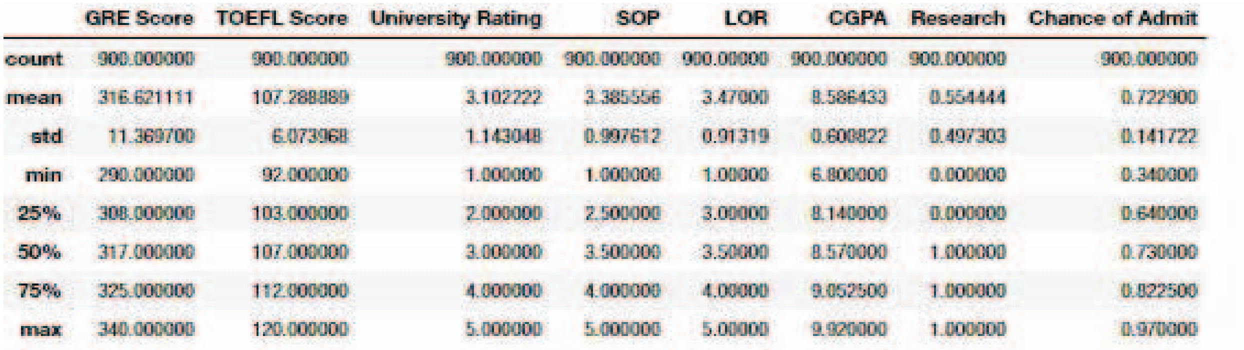
The data used to perform our study is collected from www.kaggle.com[24].The data consisting of 900 student's observations and attributes. The details of the dataset are given below,

Table: I Attributes present in Dataset

Features	Serial No	University Ranking	GRE Score	TOEFL Score	SOP (Statement of Purpose)	LOR (Letter of Recommendation)	CGPA (Academic Performance)	Research	Chance of Admit
Valid Count	**900**	**900**	**900**	**900**	**900**	**900**	**900**	**900**	**900**
Selected		✓	✓	✓	✓	✓	✓	✓	

The above Table shows information about the attributes of our data set. The data set was found to have more than900 instances. The dataset contains various information about student performance. This data set contains information of about over 900 Students with the different attributes as described in Table I.

The dataset downloaded was in its original CSV form and contained the above-described feature. Fig 1 shows the data set sample used for our study.



	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	900.000000	900.000000	900.000000	900.000000	900.00000	900.000000	900.000000	900.000000
mean	316.621111	107.288889	3.102222	3.385556	3.47000	8.586433	0.554444	0.722900
std	11.369700	6.073968	1.143048	0.997612	0.91319	0.600822	0.497303	0.141722
min	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.340000
25%	308.000000	103.000000	2.000000	2.500000	3.00000	8.140000	0.000000	0.640000
50%	317.000000	107.000000	3.000000	3.500000	3.50000	8.570000	1.000000	0.730000
75%	325.000000	112.000000	4.000000	4.000000	4.00000	9.052500	1.000000	0.822500
max	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.000000	0.970000

Figure 1: Sample of data used for our study

B. Data Preprocessing and Transformation

This section describes the data preprocessing. The collected data is loaded, and preprocessing is performed to remove any incomplete and inconsistent data[14]. The data normalization is performed. In our study, we compare the different regression methods for predicting student graduation admission and also choosing the best method.

The relationship between attribute features distribution is shown in Fig 2

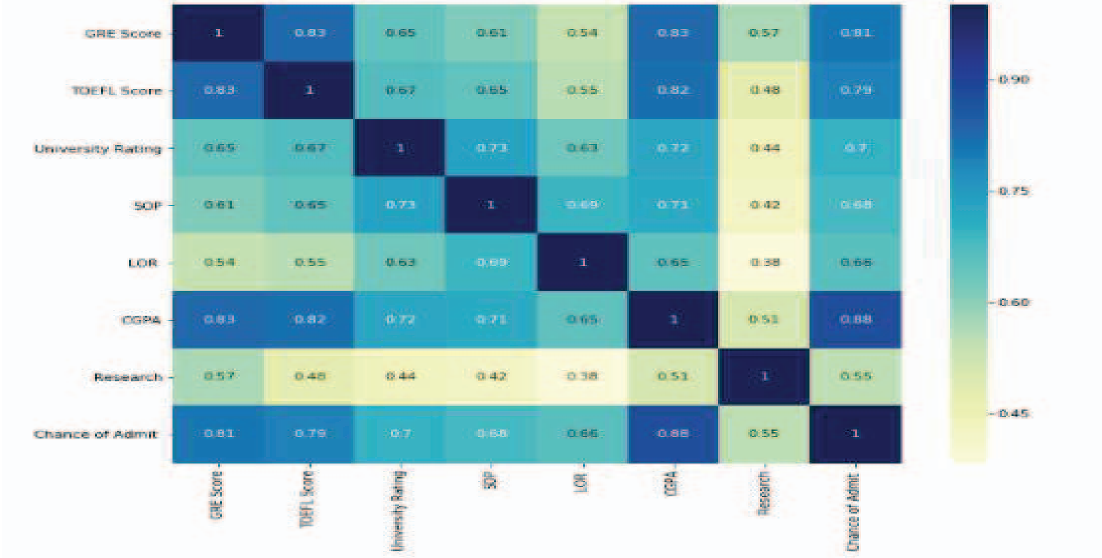


Figure 2: correlation of different attribute

In our study, we have used different regression methods for implementation, and further comparison study is carried out to find the best performance model. The reason for using the regression method is because of its supervised nature. It is observed that the independent variables like GRE, TOFEL, and university ranking are in a high relationship with the dependent variable chance of admission. The metrics used for evaluation are Root Mean Square Error, Mean Absolute Error [15], and R2.

A. Root Mean Square Error

It is defined as the square root of differences between predicted and observed values. And is given by the formula

RMSE=$\sqrt{\frac{\sum_1^n (X_0-X_p)^2}{n}}$, where X0—Observed Values, Xp ----Predicted values

B. Mean absolute Error

Mean absolute Error of the given data set is the mean of absolute deviations from a center point of the data set. And is given by the formula,

MAE=$\sum_1^n \frac{|X_i-\bar{X}|}{n}$, n-number of data values, Xi- Input data values, $\bar{X}$- mean of data value

C. R-Squared

This is also called a coefficient of determination. The metric expresses how well the model fits the available data. It shows how near the regression line is to the actual data values. The R-squared range is 0, and 1.where 0 says the model does not fit the available data and 1 say the model fits perfectly to the available data.

III. FOR THE COMPARISON, THE FOLLOWING MACHINE LEARNING REGRESSION MODELS ARE USED

A. Multiple linear regressions

Multiple linear regressions is the most common type of statistical method used to predict the outcome. The main use of multiple linear regressions is to define the relationship between predictors and predicted variables. The main purpose of multiple linear regressions is to model the relationship between the predictor (independent) variables and predicted (dependent) variables [16]. Every time the value of the predictor variable x depends on the value of the predicted variable y.

Common form of multiple regressions is represented as below

$y = b_1x_1 + b_2x_2 + ... + b_nx_n + c$ …………(1)

In Equation (1)the coefficients of regressions are bi's (i=1, 2…n) In which y(dependent) variable predicts values depending on the values of x(Independent) .The value of the dependent variable chance of admission is predicted using the other attribute features listed in Table I.

B. Ridge Regression

The variation in linear regressions that is used is said to be Ridge regression. When the data set is having the nature of multi co linearity.[17]Ridge regression never differentiates between variables like weather data variable is more important or less important. It uses the following formula to estimate coefficients. Here X is a variable feature that is independent, β is beta Coefficient.

$$\beta^{ridge} = argmin\|y - XB\|_2^2 + \lambda\|B\|_2^2$$

C. Elastic Net

The selection of the Elastic Net variable is unstable, and it depends much on data[18]. To get the best results, we combine the penalties of both ridge and lasso regressions. In Elastic Net, it concentrates on minimizing the loss function:

$$L_{enet}(\beta) = \frac{\sum_{i=1}^{n}(y_i - x_i\beta)^2}{2n} + \lambda\left[\frac{1-\alpha}{2}\sum_{j=1}^{m}\beta_j^2 + \alpha\sum_{j=1}^{m}|\beta_j|\right]$$

In ridge (α = 0) and lasso (α = 1) the mixing parameter is given as α

D. Lasso regression

Lasso means for the Least Absolute Shrinkage and Selection Operator.[19] The penalty applied for L2 is equal to the absolute value of the magnitude of the coefficients: Model specification shown in the equation below,

$$L_{lasso}(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\sum_{j=1}^{m}|\beta_j|$$

E. K nearest Neighbors Regression

K nearest neighbors or KNN is a simple algorithm used to predict the numerical values for a given set of data. KNN is used in the estimation of statistics and pattern recognition. In our model, we use this regression method along with other regression models to find the best fit model.[20]

F. SVM

Support vector regression supports both linear and nonlinear type of regression[21]. More detailed information can be found in [22] .

IV. EXPERIMENTAL SETUP AND RESULTS

In our study, we carried out an experimental analysis of the above-defined data. The coding is implemented by using python high-level language on Jupiter notebook. Intel Core i3 with 4GB of RAM. The details of the dataset are represented in Table-I. The evaluation is performed using all the regression methods explained in the above section III. The evaluation method is carried out using different sampling methods, namely random sampling and cross-validation sampling. Table IV and Table V show the performance of different regression methods using two defined validation technique.

A. Cross-validation Sampling

For the cross-validation sampling, we have experimented by dividing the data into many folds. We have used the k-Fold method [23]. The experiment is conducted by dividing data into namely 2,5,10, and 20 folds. From the results, we have found that the best value was achieved for RMSE, MAE, and r2 for the folds 10 and 20. Results are displayed in Table II. The value of RMSE for Linear and ridge regression can be seen very close to each other. It also observed that random regression performs well with the least RMSE for 1o fold. The results confirm that Random forest might be an efficient model than the other.

Table II: Cross-validation Performance of different models

Number of folds	Evaluation Metrics	Model						
		Linear Regression	*Ridge*	*KNN*	*Elastic net*	*SVM*	*Lasso*	*Random Forest*
2 fold	**MAE**	-0.04410096	-0.0441129	-0.0488533	-0.0744617	-0.066291611	-0.0968635	-0.02631
	R2	0.8105424	0.81047635	0.7801428	0.55316687	0.699322197	0.2647158	0.9079491
	RMSE	-0.003814226	-0.00381570	-0.0044162	-0.0089775	-0.006037537	-0.0147756	-0.0018609
5 fold	**MAE**	-0.0457845	-0.04579909	-0.0500511	-0.0759930	-0.06487804	-0.0983046	-0.0219455
	R2	0.799418491	0.801801292	0.76503192	0.53449041	0.702208994	0.23988309	0.94125737
	RMSE	-0.0040949	-0.00409772	-0.0045312	-0.0093672	-0.005444119	-0.0152497	-0.0011491
10 fold	**MAE**	-0.04507609	-0.04507666	-0.0499955	-0.0754062	-0.0648881	-0.0977541	-0.0220755
	R2	0.8018218	0.80180129	0.75600734	0.5271123	0.687441342	0.21832525	0.93914463
	RMSE	-0.00396794	-0.0039689	-0.0046111	-0.0092014	-0.005441792	-0.0150599	-0.0011251
20 fold	**MAE**	-0.04479843	-0.0448014	-0.0497555	-0.0750731	-0.064826881	-0.0974018	-0.0224277
	R2	0.79943651	0.79947413	0.74953890	0.52639045	0.670870668	0.20794263	0.93510466
	RMSE	-0.0.0392373	-0.0039242	-0.0045643	-0.0091242	-0.005439057	-0.0149630	-0.0011789

B. *Random Sampling*

For our study we have considered dividing into different Training set namely 5%, 10%, 20%, 30%, 40%50%, 60% And 70%.The obtained results are presented in Table III. The results found that all regularization was found to give the least Error on 60% training set for linear regression and 70% in ridge regression modules. Our study shows that Random Forest gives on 60% of the training set the best performance among all the other models.

Table III: Random sampling Performance of different models

Training Set	Evaluation Metrics	Model						
		Linear Regression	*Ridge*	*KNN*	*Elastic net*	*SVM*	*Lasso*	*Random Forest*
5%	**MAE**	0.05268711	0.05274607	0.0567111	0.0852421	0.0612097	0.1050083	0.026357
	R2	0.7440316	0.74361265	0.7405355	0.4788384	0.8006980	0.2100745	0.942204
	RMSE	0.07923263	0.07929745	0.0797718	0.1130569	0.0699144	0.1391887	0.037649
10%	**MAE**	0.05064065	0.0507059148	0.0497111	0.0823370	0.0657080	0.1022016	0.025717
	R2	0.7585860	0.75807917	0.7600908	0.4977862	0.7507397	0.2325904	0.931228
	RMSE	0.07369810	0.07377543	0.0734680	0.1062966	0.0748861	0.1313980	0.039335
20%	**MAE**	0.04708954	0.04710260	0.0475888	0.0789327	0.0662026	0.1001678	0.026788
	R2	0.79752546	0.797203040	0.7967719	0.532833	0.7356596	0.2460510	0.928324
	RMSE	0.06604417	0.066096	0.0661669	0.1003194	0.0754624	0.1274442	0.039294

30%	**MAE**	0.0471585	0.047194	0.0510666	0.0770109	0.0681513	0.0989617	0.030397
	R2	0.7945325	0.794031	0.752660	0.5357021	0.6930071	0.2480879	0.899546
	RMSE	0.0656091	0.065689	0.0719845	0.098625	0.080196	0.1255094	0.045874
40%	**MAE**	0.04667727	0.04669885	0.0536388	0.0767412	0.0695998	0.0995314	0.034423
	R2	0.79037190	0.7901325	0.7334623	0.5372563	0.6759015	0.2328209	0.865754
	RMSE	0.06633852	0.06637638	0.0748032	0.0985623	0.0824858	0.1269081	0.053087
50%	**MAE**	0.04583334	0.04585263	0.0554977	0.0753477	0.0705630	0.0972723	0.035462
	R2	0.79670244	0.79639797	0.7211618	0.5401913	0.6560027	0.2475416	0.855487
	RMSE	0.06445228	0.06450053	0.0754828	0.0969305	0.0838397	0.1239976	0.054340
60%	**MAE**	0.0455663	0.04554719	0.0558814	0.0719778	0.0701611	0.0943195	0.038220
	R2	0.80030981	0.8002997	0.7098830	0.5649932	0.6503587	0.2658638	0.839361
	RMSE	0.06277433	0.062775923	0.0756641	0.0926513	0.0830644	0.1203627	0.056302
70%	**MAE**	0.04600700	0.045894131	0.0554571	0.0726351	0.0724619	0.0962518	0.042866
	R2	0.8001443	0.80101042	0.7271293	0.5555118	0.6235967	0.2430685	0.793784
	RMSE	0.06279800	0.06266180	0.0733780	0.0936521	0.0861815	0.1222125	0.063789

Table II Table III: Results of RMSE and MAE from different regression models using random Sampling and Cross validation Sampling (RMSE-Root Mean Square Error and MAE-Mean Absolute Error)

The performance evaluated considering the MAE and RMSE for evaluation of the accuracy of the model. The better model can be identified with minimized RMSE value [7].A smaller value of RMSE indicates the better fitting of the model. From the experiment we have conducted on our dataset by applying different regression model results, it is found that among all the models, random forest performs well when compared to other models. The results of multiple linear regressions are closely followed by ridge regression with very minimal difference.

After getting our best performing model, we have applied a few test instances to see the performance evaluation using a random forest module. The profile of test instances created is shown in below Table, along with the results obtained. After getting our best performing model, we have applied a few test instances to see the performance evaluation using ridge regression. The profile of test instances created is shown below Table along with the results obtained for various features given in the Table:

Table IV: Predicted value for the unseen data

Attribute	GRE score	TOEFL	University Rating	SOP	LORs	CGPA	Research	Predicted Value
Test Value1	315	105	4	4	4	9.5	1	0.8312
Test Value2	337	118	4	4.5	4.5	9.65	1	0.9212
Test Value3	330	120	5	4.5	5.0	9.56	0	0.9175

Table IV shows that the values predicted by our model closely resemble the actual chances of admission in the observed data.

V. CONCLUSION

The machine learning regression model is one of the majorly used concepts to predict values in many applications. Our study carried out different regression models to develop the best prediction model for students seeking postgraduate admission in different universities. The model's performance has been evaluated by using three evaluation metrics, namely MAE, R2, and RMSE, on a different distribution of data. The best-chosen regression method from the observed result is applied for new unknown data. Our study found random forest outperforms all other regression methods based on the observations made from our experiment. The same regression model is used to predict new unseen data. From the obtained results, it's observed that the predicted value closely resembles the actual value. Our work focuses on using different regression models and chooses the best among the entire regression model to advise students to plan and choose a career for them and join the University of their Choice. In the future, we plan to carry out our study by using different volumes of data and with more attributes, mainly considering neural networks.

REFERENCES:

[1] M. R. Mufid, A. Basofi, and M. Udin, "Design an MVC Model using Python for Flask Framework Development," no. Mvc, pp. 214–219, 2019.

[2] M. R. Rimadana, S. S. Kusumawardani, P. I. Santosa, and M. S. F. Erwianda, "Predicting Student Academic Performance using Machine Learning and Time Management Skill Data," 2019 2nd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2019, pp. 511–515, 2019, doi: 10.1109/ISRITI48646.2019.9034585.

[3] M. S. Acharya, A. Armaan, and A. S. Antony, "A comparison of regression models for prediction of graduate admissions," ICCIDS 2019 - 2nd Int. Conf. Comput. Intell. Data Sci. Proc., 2019, doi: 10.1109/ICCIDS.2019.8862140.

[4] R. S. Abdulwahhab and S. S. Abdulwahab, "INTEGRATING LEARNING ANALYTICS TO PREDICT STUDENT PERFORMANCE BEHAVIOR."

[5] R. Harimurti, Y. Yamasari, Ekohariadi, Munoto, and B. I. G. P. Asto, "Predicting student's psychomotor domain on the vocational senior high school using linear regression," 2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018, vol. 2018-Janua, pp. 448–453, 2018, doi: 10.1109/ICOIACT.2018.8350768.

[6] F. Aman, A. Rauf, R. Ali, F. Iqbal, and A. M. Khattak, "A Predictive Model for Predicting Students Academic Performance," 10th Int. Conf. Information, Intell. Syst. Appl. IISA 2019, pp. 2–5, 2019, doi: 10.1109/IISA.2019.8900760.

[7] V. K. Anand, A. R. S. K, E. Ben George, and A. S. Huda, "Recursive Clustering Technique for Students ' Performance Evaluation in Programming Courses," pp. 1–5, 2018.

[8] C. Verma, V. Stoffová, Z. Illés, S. Tanwar, and N. Kumar, "Machine Learning-based Student ' s Native Place Identification for Real-Time," vol. X, 2020, doi: 10.1109/ACCESS.2020.3008830.

[9] A. Slim, G. L. Heileman, J. Kozlick, and C. T. Abdallah, "Employing Markov Networks on Curriculum Graphs to Predict Student Performance," 2014, doi: 10.1109/ICMLA.2014.74.

[10] V. U. Kumar and A. Krishna, "Advanced Prediction of Performance of A Student in An University using Machine Learning Techniques," no. Icesc, pp. 121–126, 2020.

[11] T. Zhang, Y. Xu, and X. Li, "Predicting the Performance Fluctuation of Students Based on the Long-term and Short-term Data," pp. 126–127, 2017, doi: 10.1109/EITT.2017.38.

[12] J. L. Harvey and S. A. P. Kumar, "A Practical Model for Educators to Predict Student Performance in K-12 Education using Machine Learning," pp. 3004–3011, 2019.

[13] M. Lange et al., "Devito : Towards a generic Finite Difference DSL using Symbolic Python," 2016, doi: 10.1109/PyHPC.2016.9.

[14] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data Preprocessing for Supervised Leaning," no. June 2014, 2006.

[15] T. Chai and N. Oceanic, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," no. February, pp. 3–7, 2015, doi: 10.5194/gmd-7-1247-2014.

[16] N. S. Sapare and S. M. Beelagi, "Comparison Study of Python in Scientific Computing Using Mathematical Problems," vol. 7, no. 9, pp. 1–10, 2020.

[17] A. K. E. Saleh, "Introduction to Ridge Regression," pp. 1–13, 2019.

[18] T. Menu, "sklearn.linear_model," pp. 1–6, 2020.

[19] C. F. Account, "Regularization : Ridge , Lasso and Elastic Net," vol. 2020, pp. 1–24, 2020.

[20] L. E. Peterson, "K-nearest neighbor Characteristics of kNN," vol. 4, no. 2009, pp. 1–6, 2020, doi: 10.4249/scholarpedia.1883.

[21] S. Girgin, "Support Vector Regression in 6 Steps with Python," Medium, pp. 1–7, 2019, [Online]. Available: https://medium.com/pursuitnotes/support-vector-regression-in-6-steps-with-python-c4569acd062d.

[22] A. J. SMOLA and B. SCHOLKOPF, "A tutorial on support vector regression," in Statistics and Computing, vol. 14, 2004, pp. 199–222.

[23] "K-Fold Cross Validation." .

[24] "Mohan S," p. 2020, 2020. https://www.kaggle.com/mohansacharya/datasets