

# College Prediction using ML

Prof.Amol bhilare , Om Soni , Shivam Pandagale , Digvijay Patil , Manish Patil

Department of Computer Engineering  
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

**Abstract** —Each year, the majority of students apply for graduate programs to further their education and gain deeper knowledge on an academic or professional path. Universities consider JEE Rank

to be selected from the candidate pool. In this paper, we examine and analyze these for our application, giving the probability of admission to this university a problem. Compare the models with the highest scores across time. In addition to the R2 score, other performance error metrics such as mean absolute error, mean squared error, and root means the squared error are calculated and displayed.

## I. INTRODUCTION

The admissions process is very hectic, requiring months of intensive preparation to develop a list of suitable options and select universities that offer relevant programs. Many students struggle to know their chances of getting into a particular university based on their educational background, and even if they spend a lot of time and energy searching and applying to many institutions, Some of them even give up on their plans, want to go to a lower ranked college, or even face mental health issues like depression and feelings of failure. There is. Here we use a machine learning approach to train on data with a large number of parameters. After training, new data is fed into the system to determine results, and

sample data is tested against the three different models mentioned above to gain a theoretical and visual understanding of each model's performance. This white paper takes into account all the parameters required for authorization purposes, such as rank, gender, and category.

The admissions process is very arduous and hectic, requiring months of intensive preparation to develop a list of suitable options and select universities that offer relevant programs. Many students struggle to know their chances of getting into a particular university based on their educational background, and even if they spend a lot of time and energy searching and applying to many institutions, Some of them even give up on their plans, want to go to a lower ranked college, or even face mental health issues like depression and feelings of failure. There is. Here we use a machine learning approach to train on data with a large number of parameters. After training, new data is fed into the system to determine results and sample data is tested against the three different models mentioned above to gain a theoretical and visual understanding of each model's performance. This white paper takes into account all the parameters required for authorization purposes, such as rank, gender, and category.

## II. LITERATURE REVIEW

The paper named An Automated Prediction Model For College Admission System, Proposes application benefits for the student admission community that accommodates the need of students to choose the best college and helps colleges too to recognize their stand in attracting students and finer prediction implies better results for the students. The authors of this paper are

**Dr. Arunakumari B. N, Vishnu Sastry H K, Sheetal Neera, and Shashidhar R**

The paper named Predictive Models of Student College Commitment Decisions Uses Machine Learning for prediction for colleges from this paper indicates that the logistic regression classifier performed best in modeling the student college commitment decision problem, i.e., predicting whether a student will accept an admission offer, with an AUC score of 79.6%. The authors of this paper are

**Kanadpriya Basu, Treena Basu, Ron Buckmire, and Nishu Lal**

The paper named The research of regression model in machine learning field calls the corresponding library function to predict the sale of iced products according to the variation of temperature, which will provide the foundation for the company to adjust its production each month, or even each week and each day. As a result, the situation of overproduction can be avoided. The final result correctly leads the company to adjust the production and sale of iced products flexibly according to the variation of temperature, which definitely provides great commercial value and offers a crucial theoretical foundation for the sale of other companies that produce iced products. The authors of this paper are **Shen Rong Zhang Bao-wen**

In The paper named College Admission Prediction the System design focuses on the detailed implementation of inputs (sources),

outputs (destinations), databases (data source), and procedures (data flows) all in a format that meets the user's requirements. The system works by defining the design specifications, which tell the programmers exactly what the candidate system must do. The programmers write the necessary programs that accept input from the user, perform necessary processing on accepted data through call and produce the required report on a hard copy or display it on the screen. The authors of this paper are **Annam Mallikharjuna Roa, Nagineni Dharani, A. Satya Raghava, J. Buvanambigai, K. Sathish**

## III. METHODOLOGY/EXPERIMENTAL

### A. Shuffling and Splitting:

The dataset is consistently shuffled and split into Training and Test Set i.e 80% will be in the Training Set and the rest 15% will be in Test Set.

### B. Linear Regression:

Linear Regression Linear regression is one of the simplest regression models for predicting outcomes. Its main purpose is to obtain the relationship between the independent variable (X) and the dependent variable (Y). Linear regression was the first type of regression analysis to be rigorously studied and widely used in practice. This is because a model that depends linearly on the unknown parameters is easier to fit than a model that is nonlinearly related to the parameters and is easier to determine the statistical properties of the resulting estimator.

Simple linear regression has one independent variable and one dependent variable. However, since this data set has more than independent variables and one dependent variable, we will use a multiple linear regression model.

Multiple linear regression attempts to establish relationships between two or more exploratory data sets and a response variable by fitting a linear equation to observed data.

### Linear Regression

Linear regression is one of the simplest regression models for predicting outcomes. Its main purpose is to obtain the relationship between the independent variable (X) and the dependent variable (Y). Linear regression was the first type of regression analysis thoroughly studied and widely used in practice. This is because a model that depends linearly on the unknown parameters is easier to fit than a model that is nonlinearly related to the parameters and is easier to determine the statistical properties of the resulting estimator. Simple linear regression has one independent variable and one dependent variable. However, our dataset has more than independent variables and 1 dependent variable, so we use a multiple linear regression model. Multiple linear regression attempts to establish relationships between two or more exploratory data sets and a response variable by fitting a linear equation to observed data.

Table 1: Evaluation Metrics of Decision Tree

Matrices	Values
Mean Squared Error (MSE)	0.06401272212028147
Root Mean Squared Error (RMSE)	0.2530073558620015
Mean Absolute Error	0.1818847231663019

### C Decision Tree Regression

A decision tree builds a regression or classification model in the form of a tree structure. We decompose the dataset into small subsets while stepwise evolving the associated decision trees. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing the value

of the attribute being tested. Leaf nodes represent decisions about numerical goals. The top decision node of the tree corresponding to the best predictor is called the root node. Decision trees can handle both categorical and numerical data.

Decision tree regression typically uses the mean squared error (MSE) to decide whether to split a node into two or more subnodes.

Here we evaluated decision trees using the maximum number of features that are standard in practice. Table 2 shows the metrics obtained.

Table 2: Evaluation Metrics of Decision Tree

Matrices	Values
Mean Squared Error (MSE)	0.05051294206569571
Root Mean Squared Error (RMSE)	0.22475084441597926
Mean Absolute Error	0.15671299839508196

### E. Support Vector Regression:

So we also used SVR and we came to know that it is best suited to regression for our project. We've learned that SVR is one convenient and flexible technique. We've seen how it is used for predictions based on data points. It is important to note that kernel selection plays an important role in SVR modeling. We've seen how kernel choices affect machines. We also learned how SVR works and confirmed that it outperforms other regression algorithms.

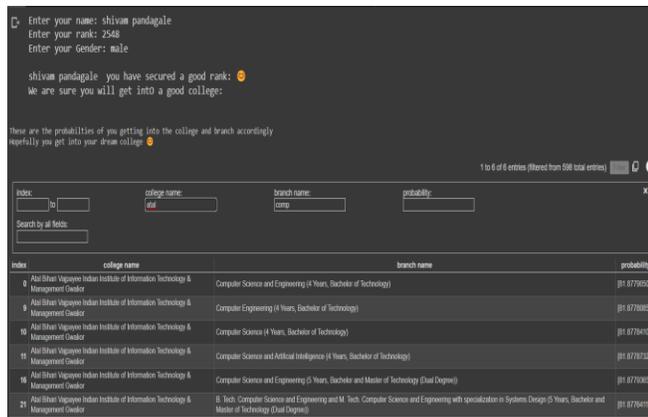
Table 4: Evaluation Metrics of Support vector regression

Matrices	Values
Mean Squared Error (MSE)	0.062280260849708495
Root Mean Squared Error (RMSE)	0.24956013473651695
Mean Absolute Error	0.169253954422593

#### IV. RESULTS AND DISCUSSIONS

These are the results we have got. Our model has been successful at predicting the probability of a certain student getting into the college and branch accordingly. All 4 regressions are running successfully and predicting the probability.

We can sort the results according to branch and college name.



#### V. FUTURE SCOPE

Our system now can only show the probability of getting into a certain college on the basis of JEE data. further, we can add other entrance exams such as NEET, UPSC, CET, etc. we can also add features to show probability on the basis of caste, and home university quota. In the future, more data related to additional universities and courses can be added to the system. Also, the system can be enhanced to a web-based application by making

changes to the code. Other classification algorithms can be evaluated to resolve the problem if they perform better than the current algorithm; the system can be easily updated to support the new algorithm by changing the code.

#### Conclusion

After evaluating all dataset models, model performances were compared on the basis of two important metrics Mean Squared Error(MSE) and R-Squared Score to find out which model is better. Table 4: Analysis of performance Now, it is pretty clear that Support Vector Regression performs best on our data set. Our Project predicts the probability of getting a college for an individual student based on three important factors which are college, rank, branch, and gender.

#### REFERENCE

- [2] Gupta, Narender, Aman Sawhney, Dan Roth. Will I get in? modeling the graduate admission process for American universities. In 16th International Conference on Data Mining Workshops (ICDMW) on, IEEE, 2016.
- [3] Roa, Annam Mallikharjuna. College Admission Predictor. In journal of Network Communications and Emerging Technologies (JNCET), 2018.
- [4] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Conference on AI (IJCAI), pages 1137–1145, San Mateo, CA, 1995. Morgan Kaufmann.
- [5] J. D. D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In Third annual symposium on document analysis and information retrieval, volume 33, pages 81–93, 1994.
- [6] B. Minaei-Bidgoli and W. F. Punch. Using genetic algorithms for data mining optimization in an educational web-based system. In Genetic and Evolutionary Computation—GECCO 2003, pages 2252–2263. Springer, 2003.
- [7] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. Classification and regression trees. In CRC press, 1984.
- [8] J. Demsar. Statistical comparisons of classifiers over multiple data sets. In The Journal of Machine Learning Research, volume 7, pages J. Zimmermann, K. H. Brodersen, J.-P. Pellet, E. August, and J. M. Buhmann. Predicting graduate-level performance from undergraduate achievements. In 4th International, Educational Data Mining Conference (EDM11), pages 357–358, 2011.
- [9] M. Young, *The Technical Writers Handbook*. Mill Valley, CA: University Science, 1989.
- [10] V. Vapnik. The nature of statistical learning theory. In Springer Science & Business Media, 2000.