**AP EAPCET Students’ Qualification Prediction Using Machine Learning Algorithms**

1 Dr. Siddique Ibrahim Peer Mohammad

Assistant Professor,

School of Computer Science

And Engineering,

VIT-AP University,

Amaravathi, A.P, India

[siddique.ibrahim@vitap.ac.in](mailto:Siddique.ibrahim@vitap.ac.in)

2 Malladi Niharika 3 Mokara Hemanth Kumar

UG Student UG Student

School of Computer Science School of Computer Science

and Engineering, and Engineering,

VIT-AP University, VIT-AP University,

Amaravati, A.P, India Amaravati, A.P, India

[niharika.21bce8053@vitapstudent.ac.in](mailto:niharika.21bce8053@vitapstudent.ac.in) hemanthkumar.21bce9109@vitapstudent.ac.in

4 Karumuri Leela Vaishnavi 5 Pisipati Srirama Sai Bhavya Mani

UG Student UG Student

School of Computer Science School of Computer Science

and Engineering, and Engineering,

VIT-AP University, VIT-AP University,

Amaravati, A.P, India Amaravati, A.P, India

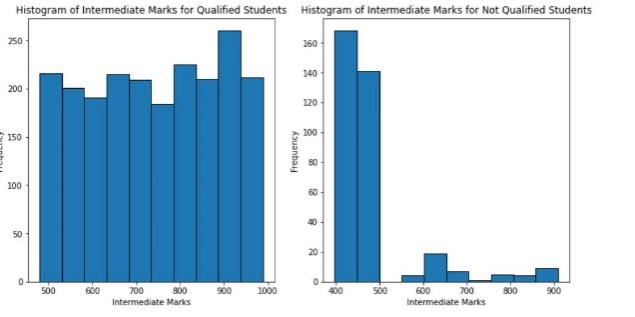
[vaishnavi.21bce9075@vitapstudent.ac.in](mailto:vaishnavi.21bce9075@vitapstudent.ac.in) [bhavya.21bce7628@vitapstudent.ac.in](mailto:bhavya.21bce7628@vitapstudent.ac.in)

***Abstract*—This research paper presents a mini project focused on predicting whether the student qualifies in AP EAMCET Examination or not based on their region, district and intermediate marks. In order to predict the qualification of the student we use two Machine Learning algorithms which are SVM and Decision tree classification. In which Decision tree classification algorithm gives the result more effectively for large data and classify the candidates based on their performance. The dataset is split into training and testing datasets to fit into the classifier (or)model in order to continue with the process. These results in qualification prediction with a high-level accuracy. This qualification helps us to get accurately either the student qualifies or not which will help the candidates to feel better without any tensions.**

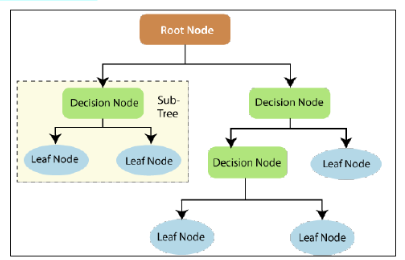
Keywords— EAMCET prediction, SVM, Decision Tree

# **Introduction**

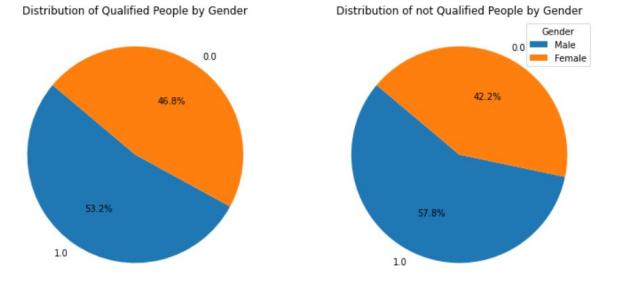
**T**he EAPCET exam plays a crucial role in any intermediate candidate’s path decision-making as it is the Entrance examination for getting the admission into any Engineering colleges in Andhra Pradesh. We can help them by using this model to predict their qualification based on their performance in Intermediate and they can manage well for their bright future. As students have a constant passion for knowledge and desire to make a meaningful impact on world through research.



The primary objective is to help the students by predicting their Qualification through their performance in Intermediate. This research report is important since it can help pupils recognize their advantage over the competition. We worked on the dataset which consists of 2020,2021,2022 EAPCET student’s details to check the accuracy.

The dataset that we used was live data. It consists of 2482 tuples with 14 characteristics. And Name, gender, hall ticket number, state, district, region, Intermediate marks, EAMCET year, status, EAMCET rank, allotted college, branch, coed are the properties. We employed district, region, and Intermediate marks among these attributes to predict the Qualification through this machine learning model. The target column in this dataset is “*status*” to know whether they are qualified or not.The target commands was used to remove all of the remaining attributes .To perform an algorithm on data, it must only include integers. So, *Label Encoder* function is taken from the *sklearn* module and this will change the strings to binary (0&1’s format) for machine understanding purpose.

Now to proceed further we used SVM algorithm and Decision tree classification algorithm as the dataset is large and we aimed for high accuracy of the prediction. Attributes named region\_n, district\_n and Intermediate marks\_n are created after encoding is done. The column ‘status’ contains the qualification class label as Qualified and Not Qualified and these class labels are normalized as 1 and 0 respectively to change the categorical data into numerical data.



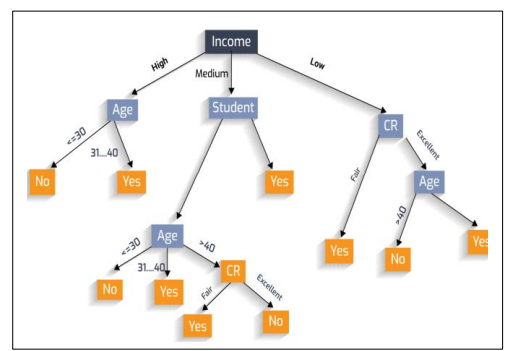
# **PROPOSED SYSTEM**

1. **Decision Tree Algorithm:**

Decision trees are a most used machine learning, statistics, data mining and for decision making in technique that falls under the genre of supervised learning. They are basic yet powerful models that are ideal for both classification and regression tasks. While they are susceptible to overfitting, ensemble approaches and regularization techniques help to improve their performance, making them invaluable tools in the data scientist's toolbox.

Method that is used to create classifiers are prominent data mining technique. In Data mining where sorting of large datasets and identifying relationship between variables and observing the patterns here classification algorithms like decision tree algorithm can handle huge amounts of data. These algorithms can be used to generate hypotheses about class names of different categories, sort and shuffle the data using data set and class names or labels, and used to add the new data point under the class label. A variety of algorithms are used in machine learning classification techniques, and this research focuses on the decision tree approach in general.

Decision trees are the algorithms which are the most effectual process in a different area of study and work and, which includes machine learning, relation between variables identifying and pattern finding. A Decision Tree is a successive or serial mode. Is largely used for categorizing. In addition, Decision Tree is a popular categorization model in Data Mining. Due to their ease of decision trees have proven to be a reliable tool for analyzing and predicting many types of data.



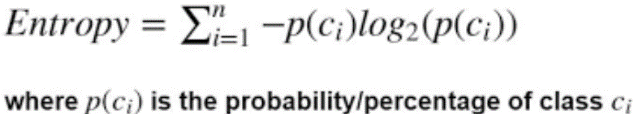
**A. Decision Tree Algorithm Types**

Decision Tree algorithms are classified into numerous types, like ID3, Inheritor of ID3 (C4.5), and its Classification.

* Chi-squared Automatic Regressive Tree (CART)
* Multivariate Adaptive Interactive Detector (CHAID)
* Generalized, Unbiased Regressive Splines (MARS)
* Conditional Interaction Detection and Estimation (GUIDE),
* CTREE Inference Trees, Classification Rule with
* CRUISE stands for Unbiased Interactive Selection and Estimation.
* QUEST stands for Quick, Unbiased, and Efficient Statistical Tree.

**B. Entropy and Information Gain**

The randomness or impurity in the dataset is called Entropy. Its range is always between 0 to 1. It’s values always nearer to 0. The equations is:



**C. Decision tree advantages and disadvantages**

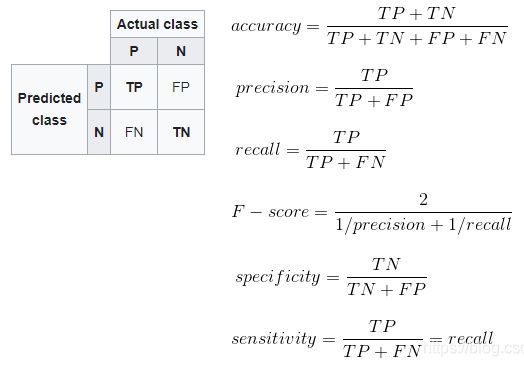
The Decision Tree algorithm is a part of supervised ML Alogrithms. The primary aim is to construct a model that can be trained to predict the type through the application of inferred decision tree rules from the sample data set. Regression and classification problems can be solved using the Decision Tree method, however the outcome may be impacted by benefits and drawbacks.

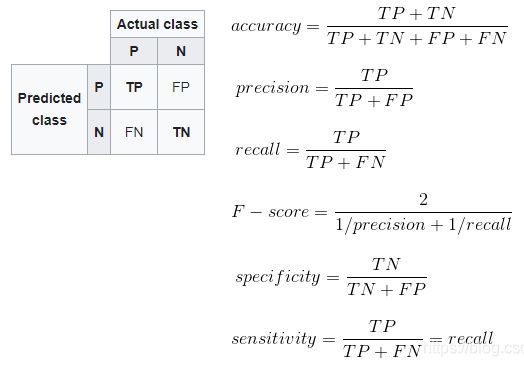
The decision tree algorithm has a structure similar to a flowchart, with each point standing for an attribute. By using the actual class and the predicted class, we can build the confusion matrix or it is also known as error matrix and it is used to determine accuracy, precision, sensitivity, and specificity.

Confusion matrix

A table known as a confusion matrix or error matrix which was used to show us type of classification system performed by the model  and they are,

* TP = True Positive (our prediction is correct and the actual class is also yes)
* TN = True Negative (our prediction is incorrect and the actual class is no)
* FP = False Positive (our prediction is correct and the actual class is no)
* FN = False Negative (our prediction is incorrect and the actual class is yes)





With the use of training data, we constructed the model using the Decision Tree classifier function. By providing the test data, we were able to determine how accurate the algorithm was for this data. With this approach, we obtained a 99% accuracy rate.

**ii) SVM ALGORITHM:**

A Support vector machine (SVM) is a novel method of researching which originated from Vapnik et al. based on small sample learning to achieve the optimal classification solution in training text samples. This is the general framework of the statistical learning theory.

A Support Vector Machines (SVM) is well known for classification work and it has since established itself. The aim is to develop an accurate predictive model that can assist educational institutions in identifying whether a student is qualified or not qualified based on certain input features. The study discusses the SVM algorithm's theoretical foundation, preprocessing steps, feature selection, model training, and evaluation methods. The outcome of the algorithm when applied on training dataset will prove the efficiency of the SVM algorithm in EAPCET Qualification prediction. The ability to predict whether a student will be qualified or not qualified is significantly important for Academies as well. It can aid in early intervention strategies and personalized support, ultimately Strengthening the quality of education and student outcomes. Support Vector Machines, which is a class of supervised machine learning algorithms, have shown promise in various classification problems due to their capacity to know optimal hyperplanes that disconnect the nearest planes or classes to the hyperplane.

Support Vector Machines spin off a hyperplane to fully differentiate classes in characteristic space. The process consists of four main steps: The algorithm transforms the data into a multi-dimensional space. A hyperplane with the largest margin between classes is sought. The margin itself is defined as the distance between the hyperplane and the nearest data points of each class to find a hyperplane that separates the classes while still allowing a good fit for new, unseen data points

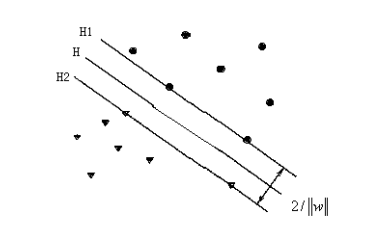
Selecting relevant features is essential to increase the SVM model's performance and generalization ability. Domain knowledge and feature analysis can guide the selection process. Irrelevant or redundant features can lead to overfitting, hindering the model's ability to predict accurately.

In SVM, the key concept is to observe the hyperplane that increases the margin while minimizing the classification error. This is achieved through optimization techniques, where the algorithm seeks to minimize a cost function while satisfying constraints related to the margin and correct classification. Kernel functions play a critical role by implicitly mapping data into a higher-dimensional space, allowing SVMs to handle the data which cannot be separated in linear way.

The operations of an SVM model can be measured in several different ways, as examples of which can be found among accuracy rate, precision, recall or even the F1-score. Moreover, you can evaluate the single grid test if it's anything like something along those lines. The generalized performance mean of such a model can in practice be roughly estimated by cross-validation methods, like k-fold cross-validation. Moreover, by procedures such as grid search, hyperparameters like kernel type and regularization parameter can be tuned to achieve better predictions from the model.

Assessing the performance of SVM model includes a variety of metrics that are mentioned above can give us a rough estimate of a machine learning model's generality. Moreover, by making hyperparameters such as the kernel type and regularized constant of support vector machine training proper through method like grid search, we can make our prediction more accurate.

The support vector machine is a new type of learning machine based on the sinker structural risk principle which can replace traditional transforming methods in many applications.It is like a linear line which can be viewed in two dimensional ways in a plane surface or any ideal surface. As per the given figure below, H is a linear line on the surface. H1 and H2 are the lines which passes through the nearest points of the H line. This kind of separating of H line’s nearest points with H1 and H2 lines is known as class interval. The purpose of an ideal separating surface, also known as the biggest class interval, is to guarantee the two classes of samples without any error.



The above hyper plane equation can be classified as

(ω.x) + b = 0 can be represented as the linear discriminate function in pattern recognition: h(x) = ω.x + b. The discriminating function h(x) was made to work in the linear seperable plane surfaces. So now, h(x) >= 1 training samples are met, even when they are not in the categorization of the sample |h(x)|=1, The class interval is comparable to 2/||w||. Making interval on ||w|| or ||w||2.

Equivalent categorization of the surface of All samples must be appropriately classified, and they must also:

Yi[(ω.Xi) + b] - 1 ≥ 0, i=1,2,3,…,n.

As this SVM model seems to be suitable for our dataset, we have trained the model with part of our dataset. The training dataset will only consist of the relevant features for the prediction. The relevant futures considered are District, region, Intermediate marks of the student. The target feature or predicted feature should be whether the student will be Qualified or Not Qualified for the EAPCET Exam.

After training the model with the training data, we have achieved the accuracy of approx. 96% in EAPCET Qualification Prediction.

**iii) Linear Regression:**

In statistics, linear regression is a type of regression analysis for examining the relationship between a continuous dependent variable and one or more explanatory variables. As the name suggests, simple linear regression deals with one explanatory variable and multiple linear regression has more than one. This phrase is in contrast to multivariate linear regression, which predicts multiple correlated dependent variables rather than a single scalar variable.

A linear predictor function is used to model the relationship and the unknown model parameters are inferred from data with a linear regression, Such models are called as linear models. That is, the expectation of the answer conditional on the values of the explanatory variables (or predictors) is usually assumed to be an affine function of those values.

Linear regression is a very useful application. This is how most applications fall under one of these two large groups:

A predictive model may be developed from an observed data set of values of the response and explanatory variables if a purpose of fitting a linear regression to a model is to minimize the difference between future forecasts and what has actually been observed. Then it is possible to predict the response if additional values of the predictor variables were collected, which is called the fitted model.

Linear regression analysis is a statistical technique that enables estimation of the strength of the relationship between the response, and the explanatory factors if the goal is to explain variability in the response variable that is a result of variability in the explanatory variables.

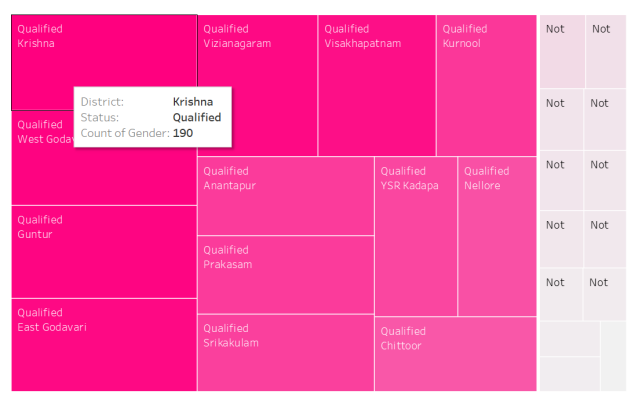
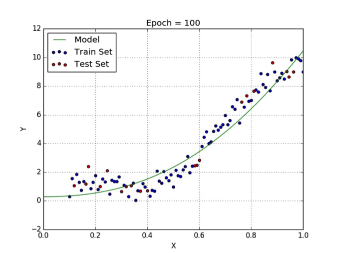
Additionally, the power of the analysis is increased, even if the independent variables are normally distributed and linearly-related to the dependent variable. We explore these assumptions in the context of the present technique, a poor fit to the data, which can make a classical regression toolkit approach to a lot of clinically relevant scenarios cumbersome to apply. In Linear Regression it simply calculates the prediction of the probability that a case will have certain variable values.

At first, we are taking the Gender and status columns for performing the Algorithm. And also, we are removing the remaining all attributes. The missing values are filled by 0 and we are encoding ‘male’ to 1 and ‘female’ to 0. The status attribute also encoded into 1’s and 0’s for Not Qualified and Qualified respectively. To proceed further we also removed the rows which are not the primary attributes for prediction. Now, the data is divided into training and testing datasets The model is trained with the training dataset. The confusion matrix is formed to get the count of the Not Qualified students in the upcoming year.

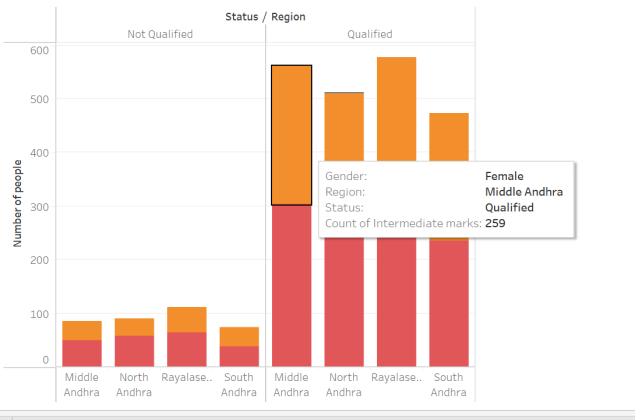
Linear regression is used for counting of disqualified candidates in APEAMCET. The status 0 (Not qualified) is considered and using count we calculated the disqualified candidates.

We used Linear regression for predicting the count of not qualified people. The best linear equation is formed to get rid of maximum errors.

# **Result and discussion**

District wise Qualified and Not Qualified Ratio:

Region wise Qualified and not Qualified students count along with the gender:



# **CONCLUSION**

SVM algorithms, decision tree algorithms, and machine learning approaches are used in this situation. The Decision Tree methods ID3 C4.5 were applied on the dataset. Decision trees perform better than other methods in terms of precision, accuracy, and time. It mainly depends on the recommendation algorithm to find interesting websites.

The purpose of a regression algorithm is to plot the best-fit line or curve between two sets of data. The trained regression model is evaluated using three factors: variance, bias, and error. When variance is great, overfitting occurs, and when bias is high, underfitting occurs.

Finally, a thorough examination of the decision tree method reveals that the approach for this dataset is the most precise and accurate among all. The Linear regression provides the count of Not qualified candidates with their total information and it may help the officials to improve the regions or particular districts candidates.

We achieved 97% accuracy by using the SVM method, and when we represented the decision tree, we reached 99% accuracy. In order to provide pupils with accurate information for a bright future, we need to be more accurate. This makes sense for all young people to enter more, and 99% is a better number to know whether they can perform well or not. Although intermediate marks and EAMCET rank may change, young people can gain greater confidence and have future conversations with their parents.

##### References

[1] Scholarly articles for Vapnik, V.N.: Statistical learning theory. Springer, New York (1995)

[2] <http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf> (BARBER BOOK)

[3] <http://noiselab.ucsd.edu/ECE228/Murphy_Machine_Learning.pdf> (KEVIN MURPHY)

[4] Agresti, A. (2002). Categorical Data Analysis. 2nd ed., Wiley-Interscience Google Scholar

[5] <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf> (MITCHELL)

[6] <https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf>

[7] Amemiya, T. (1985). Advanced Econometrics. Harvard University Press [Google Scholar](http://scholar.google.com/scholar?hl=en&q=+%0AAmemiya%2C+%0AT.+%281985%29.%0AAdvanced+Econometrics.%0A%0AHarvard+University+Press+)

[8]<https://www.analyicsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners> (ANSHUL SAINI)

[9]<https://www.cambridge.org/core/books/abs/understanding-machine-learning/decision-trees/F1C6923D0974184C4F161B406AD1393C> (CAMBRIDGE UNIVERSITY PRESS)

[10]<https://books.google.co.in/books?id=GlKIIR78OxkC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false>