**Acknowledgment:** This is the **r**eport of the project completed as part of the online certificate course in "Machine learning in Python – 2nd batch from 01/06/2020 to 09/06/2020" conducted by ICFOSS, Trivandrum, Kerala, India. This project was completed by **Riyas M. J.**

---

# OSCAR AWARD PREDICTION USING MACHINE LEARNING TECHNIQUES IN PYTHON

## Abstract

*Oscar awards are the ultimate recognition in the film industry and the hardest to achieve. While a group of experts determines the Oscars awards, public platforms are available to rate a movie according to individual experience. However, the public-rating of movies is not directly correlated with awarding of Oscars because there are various parameters to be considered for Oscar awards apart from the user experiences. IMDb and Rotten Tomatoes being the largest film databases, meta-data of most of the movies are found in these databases. In this study, we are investigating the probability of a movie to get an Oscar award by analyzing various attributes in the meta-data of the movie. As the Oscar movie selection has a poor correlation with any of its single characteristics, we utilized diverse Machine learning techniques to produce predictive models to assess the Oscar potential of a movie. The obtained models are evaluated with real datasets to derive the accuracy of each model. Finally, a comparison has done between the models to decide which model to choose for the best results.*

## Objectives

1. Prepare a database of movies, containing movies with the highest rating and movies got Oscar awards.

2. Investigate the possibility of performing linear separations for predicting the Oscar awarding of movies.

3. Perform machine learning techniques including Logistic regression, Decision tree, and Support Vector Machine (SVM) to obtain different models by feeding the database of movies for predicting Oscar awards.

4. Test the model and compute the accuracy assessment of each model.

## Introduction

The Academy Awards, popularly known as the Oscars, are awards for artistic and technical merit in the film industry. Given annually by the Academy of Motion Picture Arts and Sciences (AMPAS), the awards are international recognition of excellence in cinematic achievements, as assessed by the Academy's voting membership. The various category winners are awarded a copy of a golden statuette, officially called the "Academy Award of Merit", although more commonly referred to by its nickname, the "Oscar". The statuette depicts a knight rendered in the Art Deco style.

George Stanley originally sculpted the award from a design sketch by Cedric Gibbons. AMPAS first presented it in 1929 at a private dinner hosted by Douglas Fairbanks in the Hollywood Roosevelt Hotel in what would become known as the 1st Academy Awards The Academy Awards ceremony was the first broadcast by radio in 1930 and was televised for the first time in 1953. It is the oldest worldwide entertainment awards ceremony and is now televised live worldwide. It is also the oldest of the four major annual American entertainment awards; its equivalents – the Emmy Awards for television, the Tony Awards for theater, and the Grammy Awards for music are modeled after the Academy Awards. They are widely cited as the most famous and prestigious competitive awards in the field of entertainment.

Since the Oscars are awarded to movies having the best film-characteristics, we assume the same will reflect in the public rating of movies. Unfortunately, this doesn't happen in many cases. For example, the movie entitled "The Shawshank Redemption" being the highest public rated movie with a score of 9.2 out of 10, didn't taste any Oscar awards. On the contrary, the movie entitled "The English Patient" achieved nine Oscar awards has a public rating of only 7.4 out of 10. Rather than considering the two entities as totally uncorrelated, it can be portrayed more accurately as public rating itself are unable to predict the Oscar potential of a movie. However, the public ratings of most of the Oscar winner movies are above average.

Even though the Oscar awards were started from 1929 onwards, the award predictions have always been challenging. However, professional critics are predicting the awards long before the award ceremonies, and many of them had been correct. Movie critics are different from other audiences as they look into complex details of the movies in addition to the audience experience. The predictions are not being an easy task because of the complexity involved in it. However, if we look into the history of Oscar awards, few facts are valid in most of the

Oscar awarding scenarios. For example, the award for best motion-picture has never assigned to a non-English language movie till last year.

Considering the possibility of predicting the Oscar awards by learning from the history of Oscar awards and by investigating the characteristics of each movie, we are trying to make an Oscar predicting model based on computation techniques. Due to the higher complexity of the attributes involved in this prediction, we chose machine learning techniques to tackle this challenge. Due to the unawareness of the inter-attribute relations, contribution of each attributes towards Oscar probability, and considering the complexities, we decided to perform various models, including logistic regression, decision tree, and SVM to find the most suitable method. Further, we are validating and estimating the accuracy of each model and identify the suitable model for Oscar prediction.

## Methodology

The methodology involved in this study has been categorized into four sections, such as data pre-processing, assessment of inter-attribute relationships, creating machine learning models, and finally, the model validation and accuracy assessments. Each of the four sections has described below in detail.

### *1. Data pre-processing*

A movie database containing the details of 260 movies has created for this study. The list comprises of best-rated movies in IMDb (IMDb best 250 movies), and movies got Oscars from 1990 onwards. Films having a rating higher than 8.0 out of 10 were considered for the study. Movies released before 1929 were ignored irrespective of the rating since the Oscar awarding has started in 1929. We have incorporated various attributes including year, Rotten Tomatoes meter, Rotten Tomatoes Audience score, IMDb rating, IMDb movie genre tags, Film certificate, Director's Oscar history, Oscar awarded to the film, language, and duration (Figure 1). After collecting the details of the movies from IMDb, ratings of corresponding movies from Rotten Tomatoes were collected separately and merged with the dataset.

The database was prepared in a spreadsheet and exported to UTF-8 CSV format. Further, the exported data was imported in Python as a Pandas DataFrame using the library *read_csv()*. Various columns are modified, added, and removed to make the attributes more meaningful. For example, the total duration of the movie segregated into two columns, such as total hours and total minutes were removed while creating a new column showing the total duration in

minutes. Similarly, there were unwanted columns that were created just to make the data entry easier, such as the name of the Director. The distribution of movies won the Oscar and movies did won the Oscar are shown in Figure 2.

| | Movie | Year | Tomato_meter | T_Audience | Imdb | Tag-1 | Tag-2 | Tag-3 | Certificate | Director got an oscar already? | Oscar | Language | Duration |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | The Shawshank Redemption | 1994 | 90 | 98 | 9.2 | Drama | NaN | NaN | R | N | N | English | 142 |
| 1 | The Godfather | 1972 | 98 | 98 | 9.1 | Crime | Drama | NaN | R | Y | Y | English | 175 |
| 2 | The Godfather 2 | 1974 | 97 | 97 | 9.0 | Crime | Drama | NaN | R | Y | Y | English | 202 |
| 3 | The Dark Knight | 2008 | 94 | 94 | 9.0 | Action | NaN | NaN | PG-13 | N | Y | English | 152 |
| 4 | 12 Angry Men | 1957 | 100 | 97 | 8.9 | Crime | Drama | NaN | Approved | N | N | English | 96 |
| 5 | Schindler's List | 1993 | 97 | 97 | 8.9 | Biography | Drama | History | R | Y | Y | English | 195 |
| 6 | The Lord of the Rings: The Return of the King | 2003 | 93 | 86 | 8.9 | Adventure | Drama | Fantasy | PG-13 | Y | Y | English | 201 |
| 7 | Pulp Fiction | 1994 | 91 | 96 | 8.9 | Crime | Drama | NaN | R | Y | Y | English | 154 |
| 8 | The Good, the Bad and the Ugly | 1966 | 97 | 97 | 8.8 | Western | NaN | NaN | R | N | Y | Foreign | 178 |
| 9 | The Lord of the Rings: The Fellowship of the Ring | 2001 | 91 | 95 | 8.8 | Action | Adventure | Drama | PG-13 | Y | Y | English | 178 |

Figure 1: Database sample

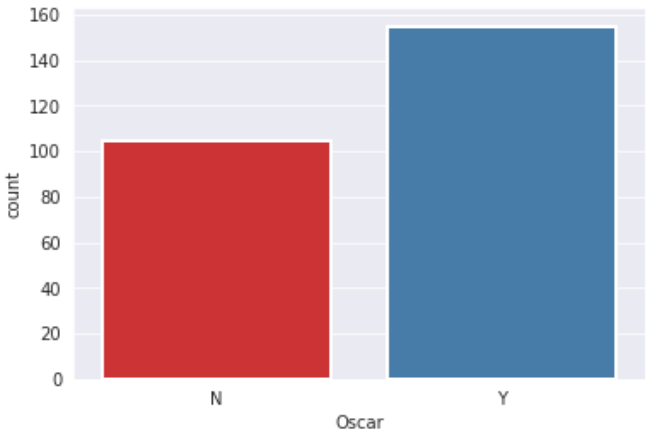


Figure 2: Count plot showing the number of samples having an Oscar award and not-having awarded an Oscar used in this study

## 2. *Assess of inter-attributes relationship*

Prior to executing machine-learning techniques, we have looked into the inter-attribute relationship to assess the possibility of implementing a linear regression that can separate the movies as those with Oscar potential and those without Oscar potential. Instead of looking into each pair, a pair plot was executed after defining the attributes to be considered.

### *3. Creating machine learning models*

We have incorporated three machine learning techniques, including Logistic Regression, Decision Tree, and SVM, for developing a prediction model to predict the Oscar probability of movies. *LogisticRegression(), DecisionTreeClassifier() and SVC()* functions are used in for the models respectively for each methods. Machine learning was done using 80% of the available datasets selected randomly (208 movie meta-data). *train_test_split()* function imported from sklearn.model library was used for this purpose with a random state of 100. The function separates the datasets for model training and model testing in a random manner.

Prior to the execution of machine learning techniques, the categorical data columns were converted to integer numbers starting from 0 to the number of unique string characters available in the column. *LabelEncoder()* function from sklearn.pre-processing library was used for this purpose. The conversion of strings to integers was necessary since the modeling techniques used in this study accept only numerical characters. Later on, the three columns representing the genre tags of movies were removed, and a new column was created for each of the unique tags. After identifying the unique tags, a new column was created in the Pandas DataFrame for each of the unique genre tags and assigned a value of 0 if the movie doesn't fall in that genre category and value 1 if the movie falls in that genre category. A new category entitled "X" was also created to represent column entries without any tags. After appending all the newly created genre columns, the column labeled as "X" was removed since it represents null values only.

### *4. Validation and accuracy assessment*

After computing the models using various techniques, all the models were validated using the test datasets, comprising 20% of the input movie dataset (52 movie meta-data). Various parameters, including sensitivity, specificity, precision, recall factor, F1 score, and error-matrix were derived for each model. The validation was done by comparing the original values of the test datasets with the model predicted values. All the parameters used for model accuracy assessment were described below.

**True Positive (TP)** is the number of correct predictions that an example is positive, which means positive class correctly identified as positive.

**False Negative (FN)** is the number of incorrect predictions that an example is negative, which means positive class incorrectly identified as negative.

**False positive (FP)** is the number of incorrect predictions that an example is positive, which means negative class incorrectly identified as positive.

**True Negative (TN)** is the number of correct predictions that an example is negative, which means negative class correctly identified as negative.

Unlike sensitivity, specificity, precision, and accuracy defined in Figure 3, **F1 score** is a weighted average of the sensitivity and precision. F1 score is a good choice while seeking a balance between precision and sensitivity. F1 score can be mathematically defined as shown below.

F1 Score = 2 * (Precision * Sensitivity/ Precision + Sensitivity)

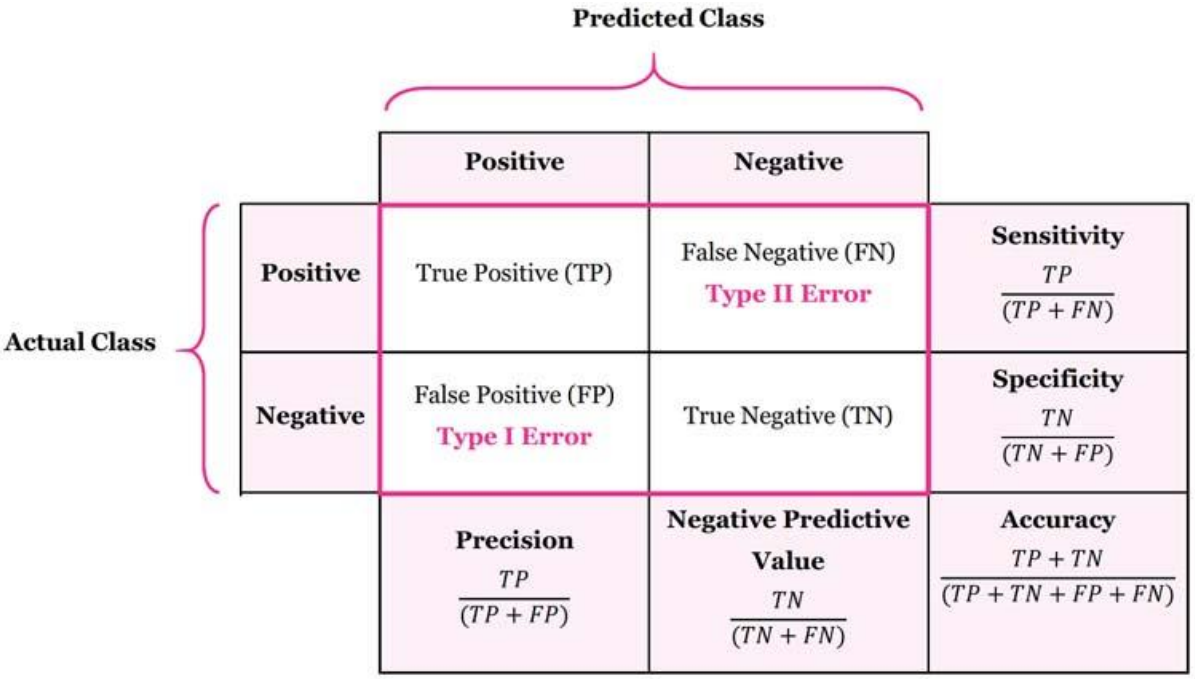


Figure 3: Defines various parameters used in estimating the usability of a predictive model in Machine Learning

## Results

The primary task was to identify possible relations between various parameters of a film in a one-to-one manner. A pair-plot was prepared using *pairplot()* function in the seaborn library. This function creates a simple scatterplot between all the parameters fed to the function. A hue value was also set to the function to discriminate sample points in plots as those awarded Oscar and those don't have.

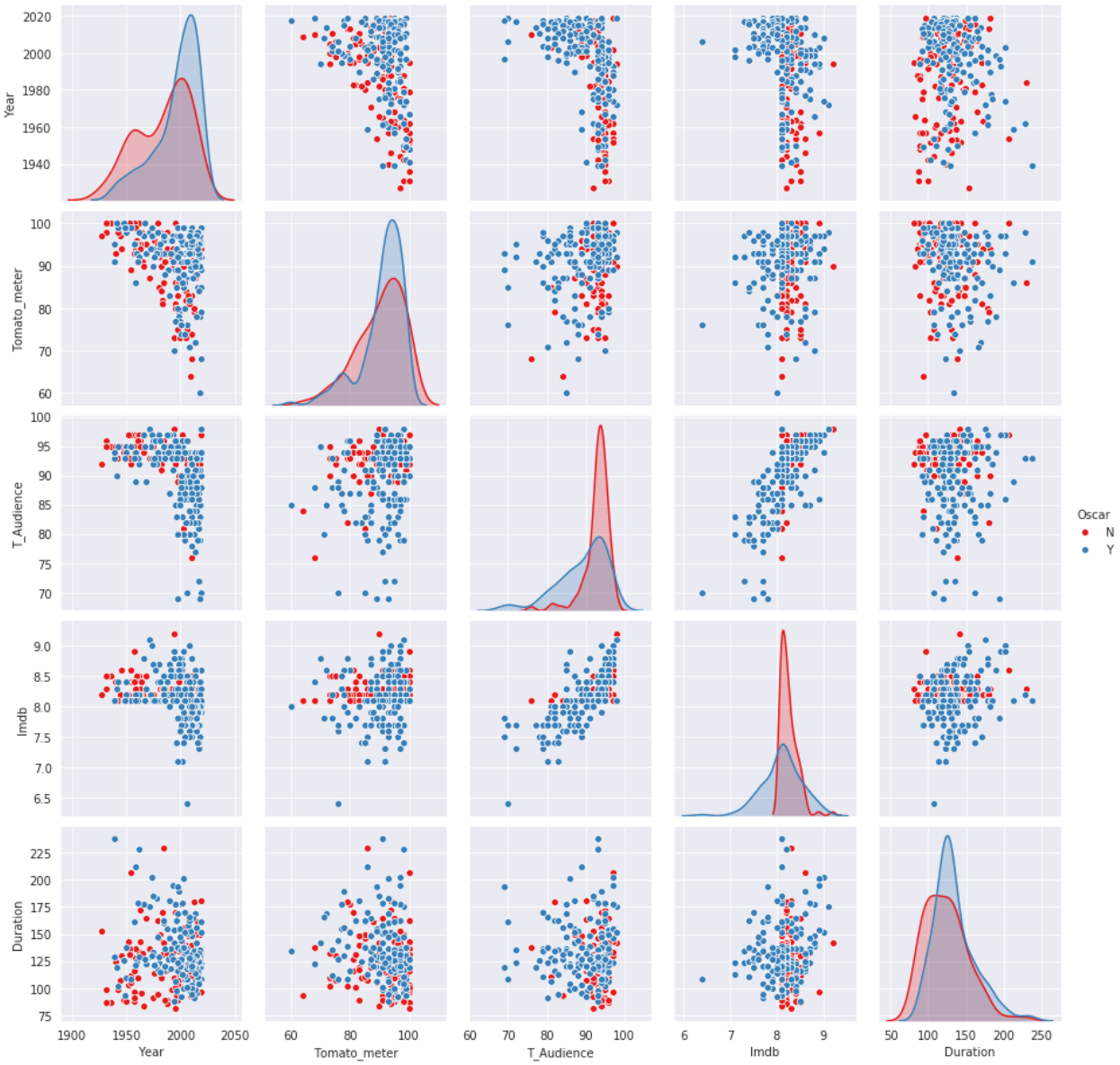


Figure 4: Pair-plot showing the scatterplot between each pair of attributes used in the study. The sample points in blue color as the movies got Oscar award and red points those movies which didn't get Oscar award.

The pair-plotting was done for various attributes, including the year of release, duration, IMDb rating, Rotten Tomatoes Meter, and Rotten Tomatoes audience meter. Other attributes were ignored for pair analyses since they are categorical data, and simple linear regressions are not recommended for categorical datasets. It is evident from the figure 4 that it is difficult to obtain any linear relationship between the attributes even while trying to relate various film rating scores. It was also noticed that movies won Oscar and non-Oscars movies exhibit a similar pattern in scattering plots, making it even difficult for identifying any linear relationship to discriminate the movies on the basis of Oscar possibility. However, significant peaks were observed in the self-plotting of IMDb rating as well as the Rotten Tomatoes audience rating. Unfortunately, these peaks are associated with the limitation of considering movies having IMDb rating above 8.1 only.

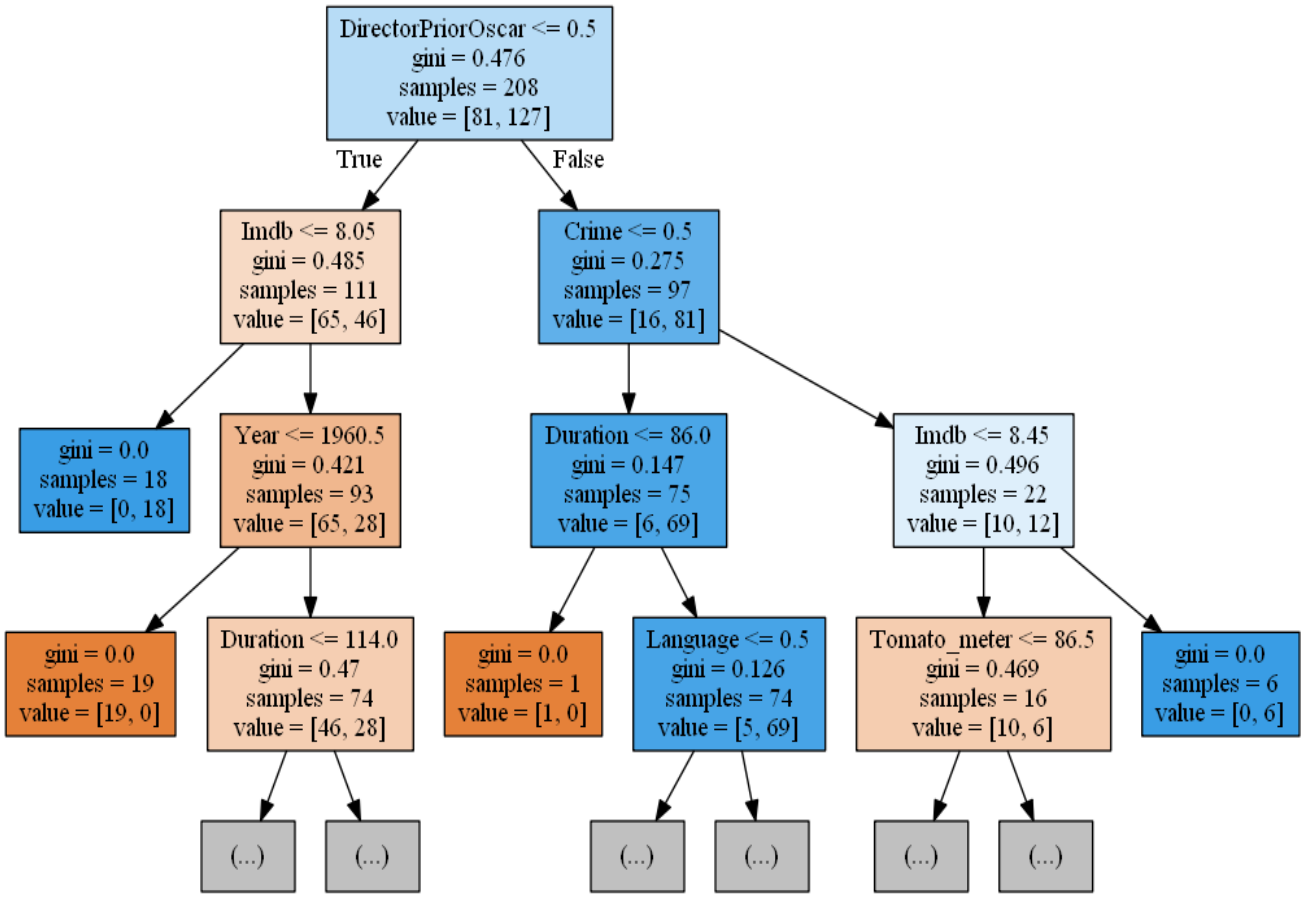


Figure 5: Decision tree used for the prediction modeling

The systematic criteria used in the decision tree method for deriving the prediction model has shown in figure 5. It signifies that the movies directed those already won an Oscar award have a higher chance of winning the Oscar award. Also, the movies in the crime genre and IMDb rating got a good influence on the classification criteria once they are categorized according to the Director's prior Oscar experience.

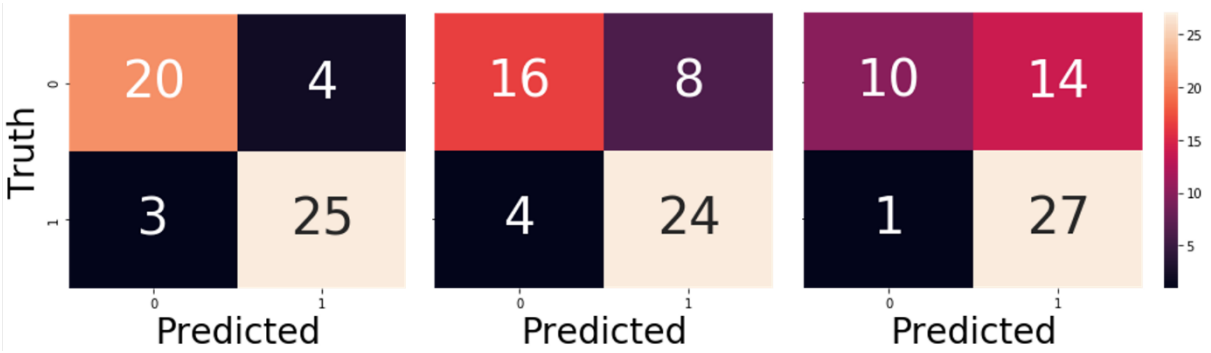


Figure 6: Heatmaps of error-matrices of Logistic regression, Decision Tree method, and SVM technique, respectively.

Heatmaps representing the error matrices of each model has shown in figure 6. The accuracy is highest in logistic regression, while SVM exhibits the least accuracy. In the Logistic regression-based model, while correctly predicting 20 non-oscar movies, it misclassified only three of the oscar winner movies as non-oscar potential movies. On the other hand, while correctly classifying 25 of the Oscar winner movies, only four non-oscar movies were misclassified as the Oscar potential movies. It is noticed that Type-2 errors are slightly high than Type-1 errors, which represents the misclassification of movies without the Oscar potential as movies with high probability to win Oscar. On the contrary, the system excelled in truthful classifying movies with Oscar potential.

Detailed accuracy assessment has completed by analyzing various error estimate related parameters of each model, including sensitivity, specificity, precision, F1 score, and accuracy, as shown in Table.1. As observed in error-matrices, the best performance of the predicted model developed using Logistic regression exhibits higher accuracy than its competitors. The only instances where the Logistic regression was slightly outrun are in case of type-1 error (False positive) and better classification of Oscar potential movies by the SVM model. False-positive represents the misclassification of Oscar potential movies as a movie without the

potential to win an Oscar. However, if we check the overall performance, the SVM model is the least robust among the models.

Table 1: Parameters used for model accuracy assessment

| ERROR ESTIMATION PARAMETERS | MACHINE LEARNING TECHNIQUE | | |
|---|---|---|---|
| | **Logistic Regression** | **Decision Tree** | **SVM** |
| **Sensitivity** | 0.89 | 0.86 | 0.96 |
| **Specificity** | 0.83 | 0.67 | 0.42 |
| **Precision** | 0.86 | 0.75 | 0.66 |
| **F1 Score** | 0.88 | 0.80 | 0.78 |
| **Negative Predictive Value** | 0.87 | 0.83 | 0.91 |
| **Accuracy** | 0.87 | 0.77 | 0.71 |

## Summary and Conclusion

Machine learning techniques are used to create models that can predict whether a movie is capable of achieving an Oscar award or not by analyzing the meta-data of the movie. Logistic regression, Decision tree method, and SVM technique are utilized for preparing Oscar prediction models. A prediction model was derived from each of the models, and all the models were validated using real movie datasets comprised of movies got Oscars, and those didn't get Oscar. It is identified that Logistic regression performed well among the models with an accuracy of 87%. The logistic regression model also demonstrates consistently better performance while assessing other error estimating parameters, including sensitivity, specificity, precision negative predictive value, and F1 score.

This study also shows the flexibility of using machine learning in various applications and the advantage of deriving the objectives in different methods. The model can be further enhanced by increasing the number of sample datasets that are used to train the models and validating them. It is also important to mention that, assigning proper weightage for each input parameter while training the model would also be able to produce better prediction accuracy. We are looking forward to implementing these shortcomings to enhance the current prediction model performance.