**BOX OFFICE PREDICTION**

**USING MACHINE LEARNING ALGORITHM**

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE AND ENGINEERING**

**By**

**BANDI REETHIKA (2103A51447)**

**SANDI SRIJA (2103A51536)**

**Under the guidance of**

**Mr. D. Ramesh**

**Assistant Professor, Department of CSE.**



**Department of Computer Science and Artificial Intelligence**



**CERTIFICATE**

This is to certify that project entitled **“BOX OFFICE PREDICTION USING MACHINE LEARNING ALGORITHMS“** is the Bonafede work carried out b**y BANDI REETHIKA AND SANDI SRIJA.**

As a course project of the partial fulfillment to award the degree **BACHELOROF TECHNOLOGY in ARTIFICIAL INTELLIGENCE AND MACHINELEARNING** during the academic year **2022-2023**  under the guidance and supervision.

**Mr. D.Ramesh Dr. M.Sheshikala**

Asst**.**Professor, Asst.Prof .&HOD(CSE) ,

SR University, SR University,

Ananthasagar, Warangal.Ananthasagar, Warangal

**ACKNOWLEDGEMENT**

We express our thanks to Course co-coordinator **Mr. D. Ramesh, Asst. Prof** for guiding us from the beginning through the end of the course Project . We express our gratitude to Head of the department CS&AI, **Dr. M. Sheshikala** , **AssociateProfessor** for encouragement , support and insightful suggestions,. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean , School and Artificial Intelligence , **DrC.V.Guru Rao** , for his continuous support and guidance to complete this project in the institute.

**ABSTRACT**

Box office in movies refers to the total amount of revenue generated by a film during its theatrical release. The box office is the important measure of a film’s commercial success and is often used a benchmark for evaluation of a movie’s popularity and profitability. A Movie box office collections dataset is mainly focuses on each and every single movie to predict the financial success of a movie.

The dataset contains various features such as the movie name, year, score, IMDB rating, votes etc., which can be used to train machine learning models to predict the box office revenue for a new movie. By analyzing the historical box office data, we can able to identify movie performance, which can then be used to make predictions about the future financial success of a movie. Movie box office collections dataset can be useful in various fields, including the movie industry, market research, and investment analysis. For example, movie studios can use machine learning models to optimize their production strategies and make data-driven decisions about which movies to produce and how much to spend on them. Market researchers can also use the dataset to analyze trends and consumer behavior related to movie-watching, which can help businesses make informed decisions about marketing and promotion.

**TABLE OF CONTENTS**

***ABSTRACT***

**Chapter No. Title Page No.**

**1. INTRODUCTION**

1.1 Problem Statement

1.2 Existing System

1.3 Proposed System

1.4 Objectives

1.5 Architecture

**2. LITERATURE SURVEY**

2.1Analysis of the Survey

**3. DATA PRE-PROCESSING**

3.1Data Description

3.2Data Visualization

**4. METHODOLOGY**

4.1Procedure to solve

4.1.1Using KNN

4.1.2 Using SVM

4.1.3 Using Logistic Regression

4.1.4 Decision Tree

4.2 Software Description

4.2.1 Through KNN

4.2.2 ThroughSVM

4.2.3Through Logistic Regression

4.2.4 Through Decision Tree

4.2.5 Through Random Forest

**5. RESULTS**

**6. CONCLUSION**

**7. REFERENCES**

**CHAPTER-1**

**INTRODUCTION**

Box office prediction is the process of forecasting the potential earnings of a movie before it is released in theaters. The accuracy of these predictions can have a significant impact on the movie industry, as they inform investment decisions, marketing strategies, and release schedules.

Box office prediction has become increasingly important in recent years as the movie industry has become more competitive and costly. With millions of dollars at stake, accurate predictions can help studios optimize their investments and maximize their returns. On the other hand, inaccurate predictions can lead to financial losses, reputational damage, and even production shutdowns.

To predict box office earnings, analysts use a variety of methods and data sources, including historical box office data. Social media trends, audience demographics and critical reviews, machine learning algorithms and artificial intelligence are also being used to improve prediction accuracy by analyzing vast amounts of data and identifying patterns and correlations.

Despite advances in technology and data analysis, predicting box office earnings remains an inherently challenging task due to the unpredictable nature of audience preferences and behaviors. Nevertheless, accurate predictions can provide valuable insights into audience preferences and trends, and help studios make more informed decisions.

**1.1-PROBLEM STATEMENT**:

**Problem Statement:**

The goal of this project is to develop a machine learning model that can predict the success rate of the film whether it reached the aim or not. And also to develop accurate and reliable methods to forecast the potential earnings of the movie before releasing in theatres.

**T**he main motive is to prove the prediction accuracy using the different regression models and compare which model performs better regarding the problem.

**1.2-EXISTING SYSTEM:**

We have collected the data set from a website called Kaggle.com. Our data set contains imbd ratings and different movie names along with the actors and technicians.It also consists of years when movies were released and money collected. We removed all the strings involved in the data set for further progression.

**1.3-PROPOSED SYSTEM:**

With the assist of dataset obtained we create 5 different machine learning algorithms and they are **KNN, SVM, LOGISTIC REGRESSION, DECISION TREE, RANDOM FOREST.** We examined the outcomes of accuracy and found that which models performs the best**.**

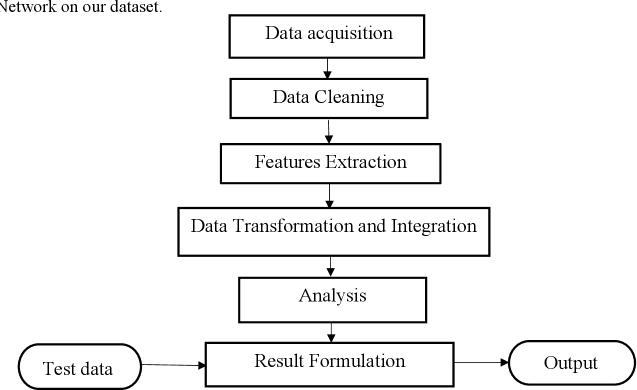
**1.4-OBJECTIVES:**

The purpose of the box office prediction is to find a machine learning model to identify the box office success of a movie using the given dataset with accuracy. The model should be able to determine the correctly as the movie’s box office success. The dataset is divided into two training and testing datasets.

**1.5-ARCHITECTURE:**

This is a Supervised learning approach. As the data have categorical values, we used regression-based machine learning algorithms to predict the condition of movie whether it reached its target or not.

We collected the data set firstly and after collecting the data pre-processing is done.Then the data set is get divided into 2 sets(i.e training and testing). Using classification based machine learning models we trained the model after finding the accuracy on the training data set we found the accuracy on the testing model. Based on those conclusions we evaluated the model. And this is the architecture followed by us.



**CHAPTER-2**

**LITERATURE SURVEY**

Makers and film producers need to predict the profit they are gonna get so that they can invest their money and time accordingly.

Previous studies that have been conducted show that it is possible to predict the success of movie using attributes such as budget, rating, actors and director. Thus the goal of this thesis is to further examine the possibility of using a greater data set with features previously not used with machine learning.

Movies generate income from many revenue streams, movie rentals and purchases, theatre ticket/box office sales, merchandising and home video are a few. To be able to compare a movies, one revenue stream was selected, which is the income generated by theatre ticket sales, or box office sales. The figures used in this study are used from box-office mojo, the leading online box-office reporting service, owned and operated by IMDb.

**2.1-Analysis of the Survey:**

The study can be used as a proof of concept for applications in other areas, and should highlight some of the challenges one needs to overcome to successfully create a prediction model. This idea could in theory be extended to predict credit ratings, the stock market or housing market. The only requirement being a vast and reliable data source. When combining the questions mentioned above to form a problem statement, formulating good as a measurement of a movies rating and sales, the following problem statement was produced.

**6**

**CHAPTER-3**

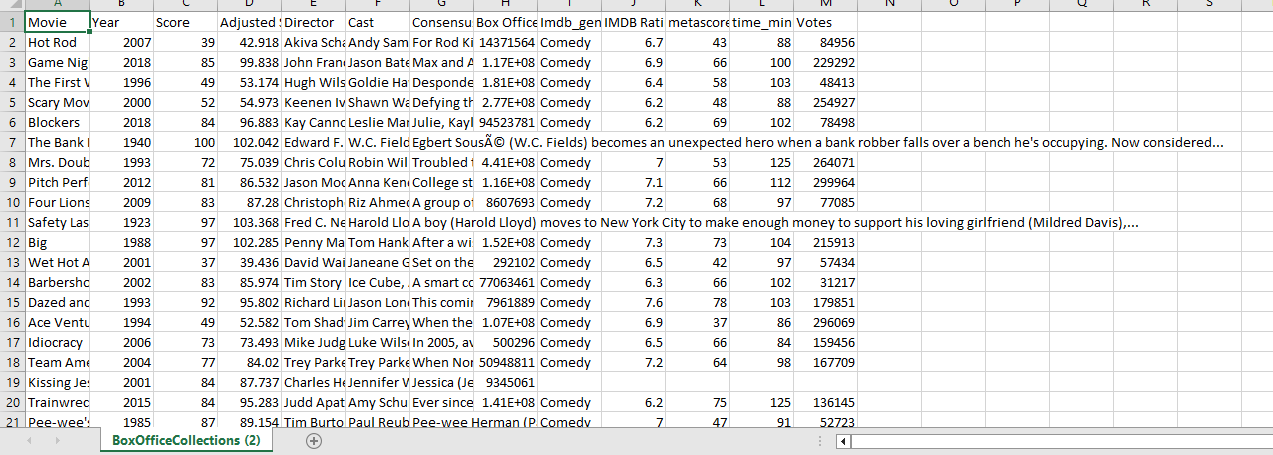
**DATA PRE-PROCESSING**

Box office prediction dataset consists of nearly 2000 rows and 13 columns.

The names of columns are listed below:

MOVIE, YEAR,SCORE,ADJUSTED YEAR,DIRECTOR,CAST,CONSENSUS, BOX OFFICE COLLECTION, IMDB\_GENRE, RATING,METASCORE,TIME\_MINUTE,VOTES .

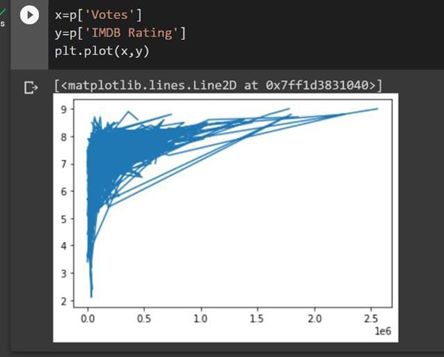
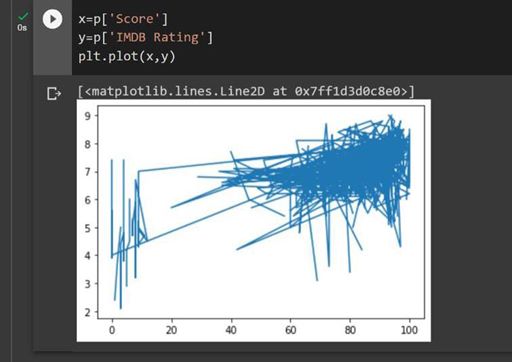
**3.1 DATA DESCRIPTION**

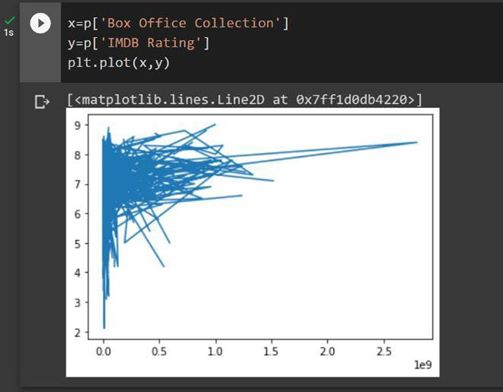
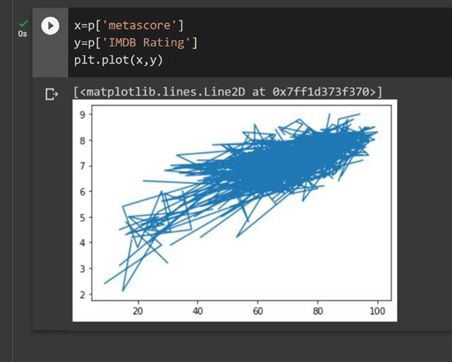
****

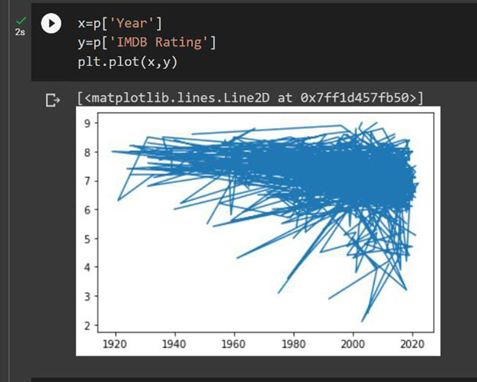
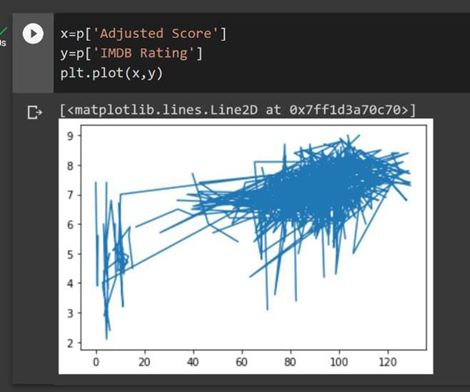
The goal of data preprocessing is to ensure that the data is accurate, complete, and consistent before any analysis is performed.

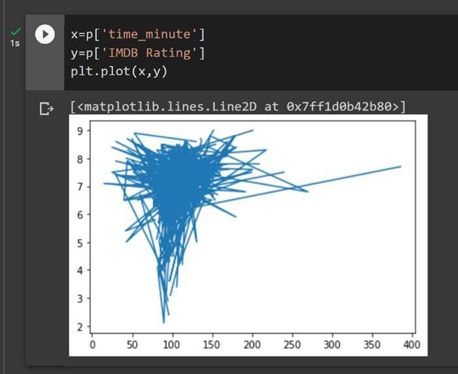
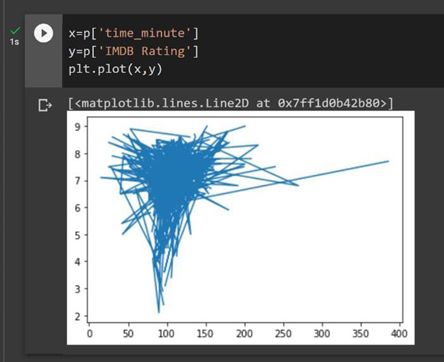
In our dataset we removed a null value and we added predicted values to it. This is the preprocessing we used in our data set.

3.2 DATA VISUALIZATION

**CHAPTER 4**

**METHODOLOGY**

**4.1. PROCEDURE TO SLOVE THE GIVEN PROBLEM:**

**4.1.1 K-Nearest Neighbors:**

The K-NN working can be explained on the basis of the below algorithm:

**Step-1:** Select the number K of the neighbors

**Step-2:** Calculate the Euclidean distance of K number of neighbors

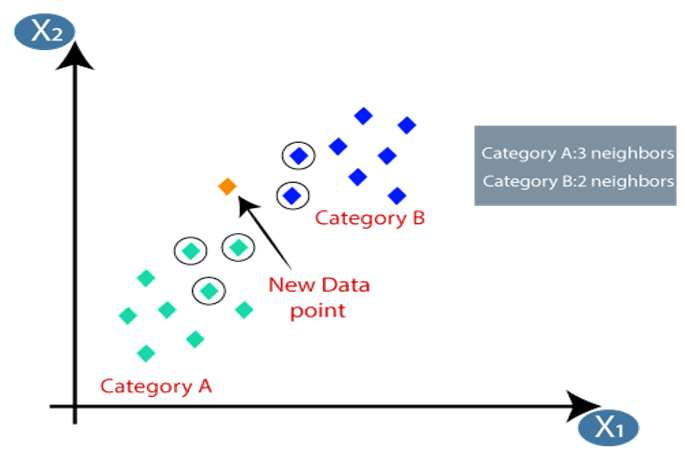
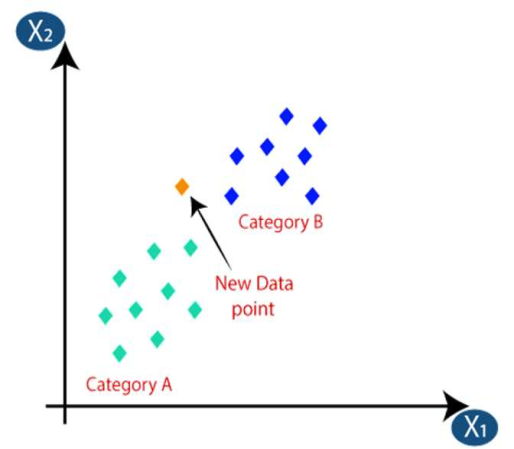
**Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.

**Step-4:** Among these k neighbors, count the number of the data points in each category.

**Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.

**Step-6:** Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below image:



Firstly, we will choose the number of neighbors ,so we will choose the k=5.Next, we will calculate the **Euclidean distance**between the data points. The Euclidean distance is the distance between points, which we have already studied in geometry. By calculating the Euclidean distance, we get the nearest.Neighbors,as three nearest neighbors in category A and two Nearest neighbors in category B.

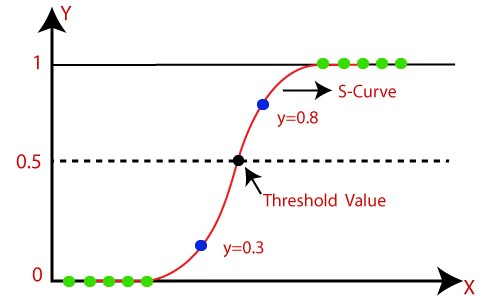
As we can see the 3 nearest neighbors are from category A,hence this new data point must belong to category A.

**4.1.2. Logistic Regression:**

Logistic regression uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.

Logistic regression is used when the dependent variable is binary such as click on a given advertisement link or not, spam detection, Diabetes prediction, the customer will purchase or not, an employee will leave the company or not.

Logistic regression uses Maximum Likelihood Estimation (MLE) approach i.e., it determines the parameters (mean and variance) that are maximizing the likelihood to produce the desired output.



Logistic Regression uses a sigmoid or logit function which will squash the best fit straight line that will map any values including the exceeding values from 0 to 1 range. So, it forms an “S” shaped curve.

Sigmoid function removes the effect of outlier and makes the output between 0 and 1.

**The logistic function is of the form:**

**P(X)=1/1+e^-z**

where μ is a location parameter

s is a scale parameter.

**4.1.3 SVM(Support Vector Machine):**

* Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. In the SVM method,we plot each data item as a point in n-dimensional space(where n is no.of features you have).
* We perform classification by finding the hyperplane that differentiates the two classes very well.
* We have three hyper planes(A,B and C). Now,identify the right hyper plane to classify star and circle.
* We want our data points to be as far away from the hyperplane as possible,while still being on the correct side of it.
* The distance between the hyperplane and the nearest data point from either set is known as margin.
* The goal is to choose a hyperplane with the greatest possible margin.
* There will never be any data point inside the margin.



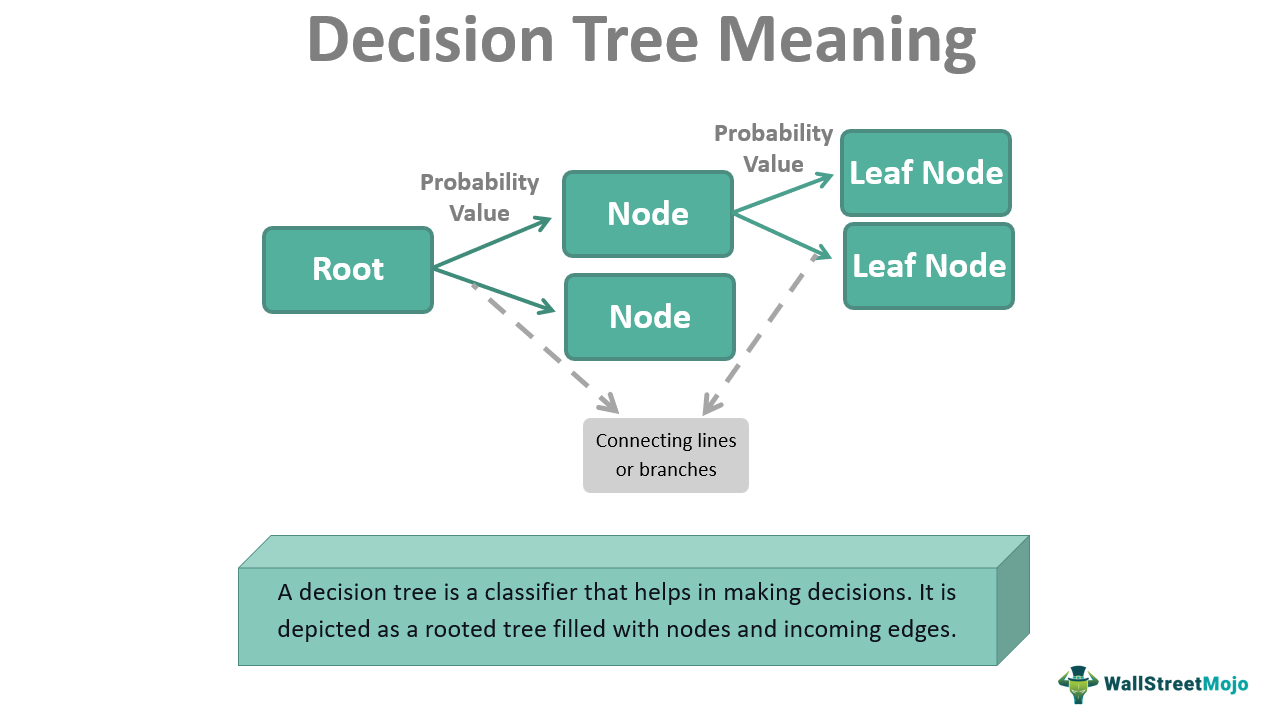
**4.1.4Decision Tree:**

* It is a non parametric method used for supervised learning method used for both classification and regression.
* It uses tree representation to solve the problem.As deeper the tree goes, the more complex the decision rules and the fitter the model.
* Entropy and Information Gini are used to calculate root node among all the nodes.
* Hence,an optimal tree can be formed.
* From given data a tree can be formed and using entropy & information gini we can calculate accuracy.
* Tree is a hierarchical representation(pictorial representation).
* ENTROPY:Entropy is the measure of uncertainity of a random variable.The higher entropy results in more information.
* **Entropy=**summation(-p\*log(p ))

i

i

* INFORMATION GINI:Information gini is the measure of changes in the entropy.
* **Information Gini=**E(T)-E(T,X)

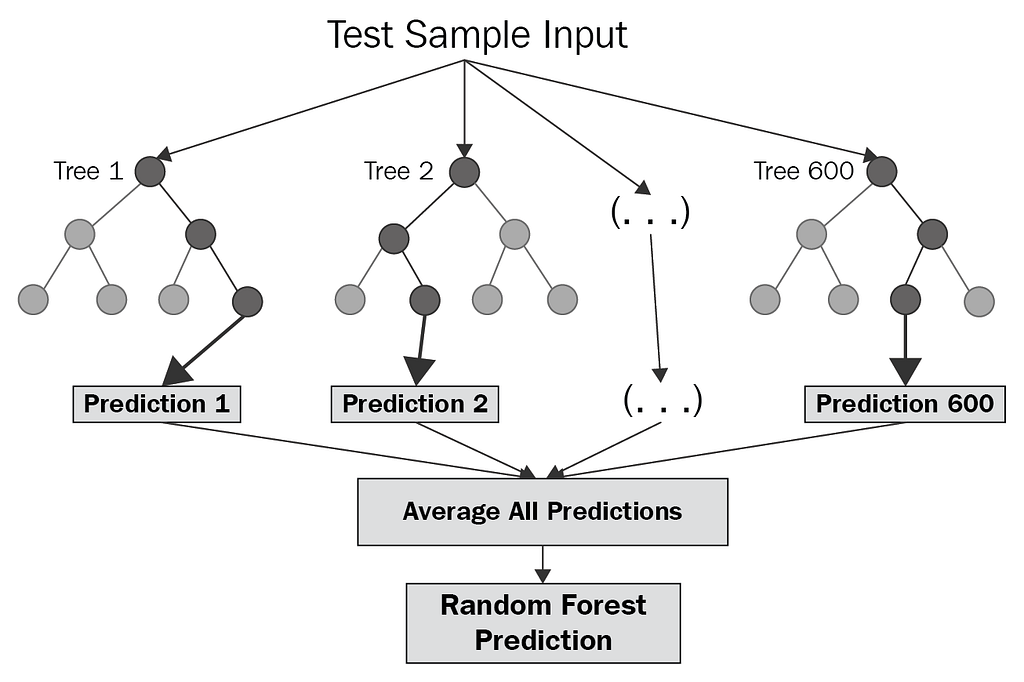


**4.1.4Random Forest:**

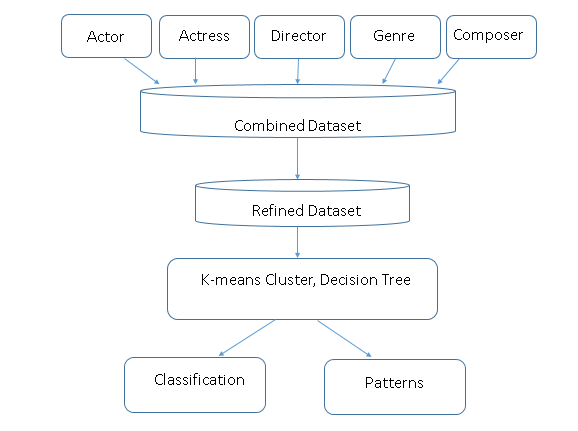
Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which we’ll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged.



**MODEL ARCHITECTURE:**



**4.2-SOFTWARE DESCRIPTION:**

**Software requirements:**

**Operating systems: windows**

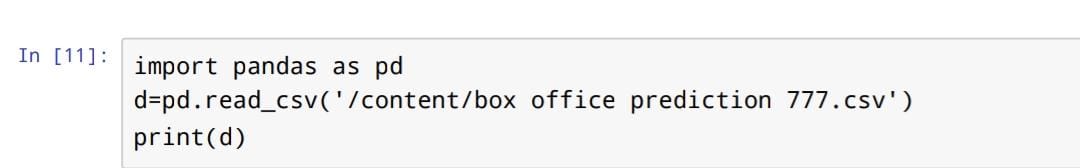
**Platform: google collab**

**Programming languages: python.**

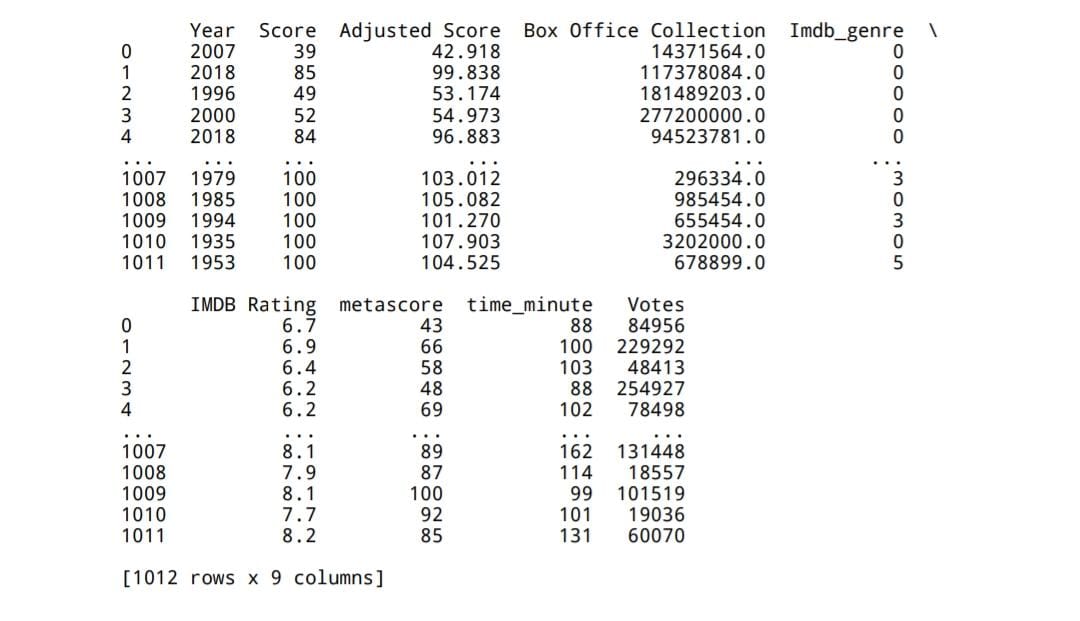
**5.RESULTS**

**CODE**

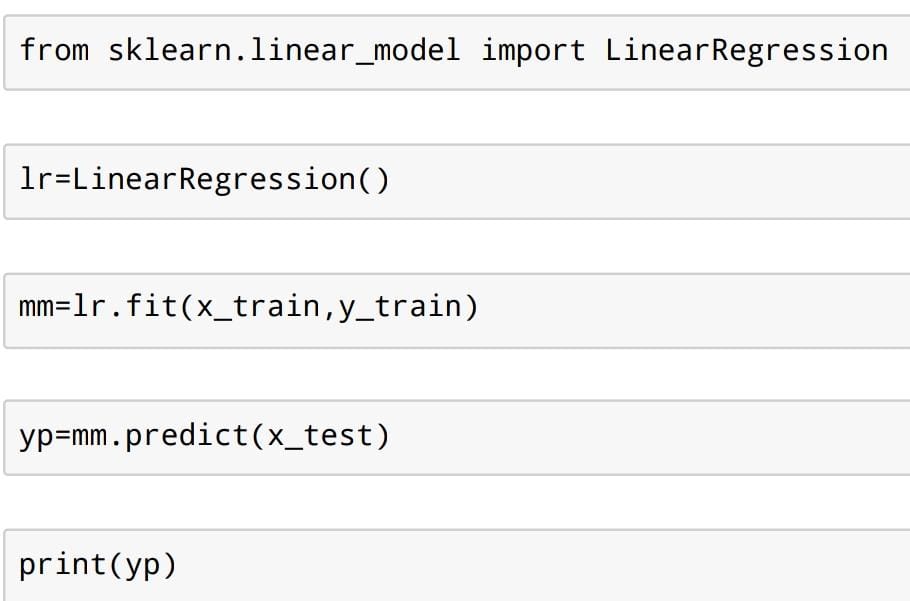
**Dataset:**



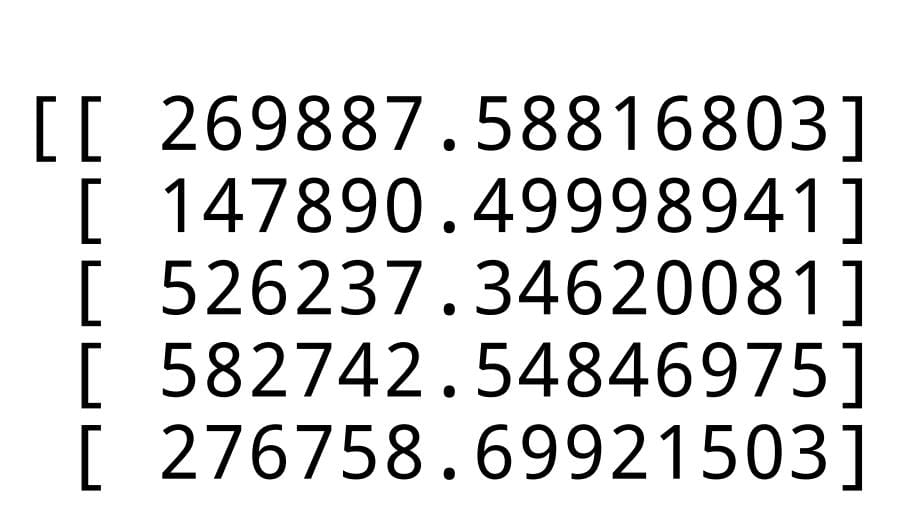
**Output:**

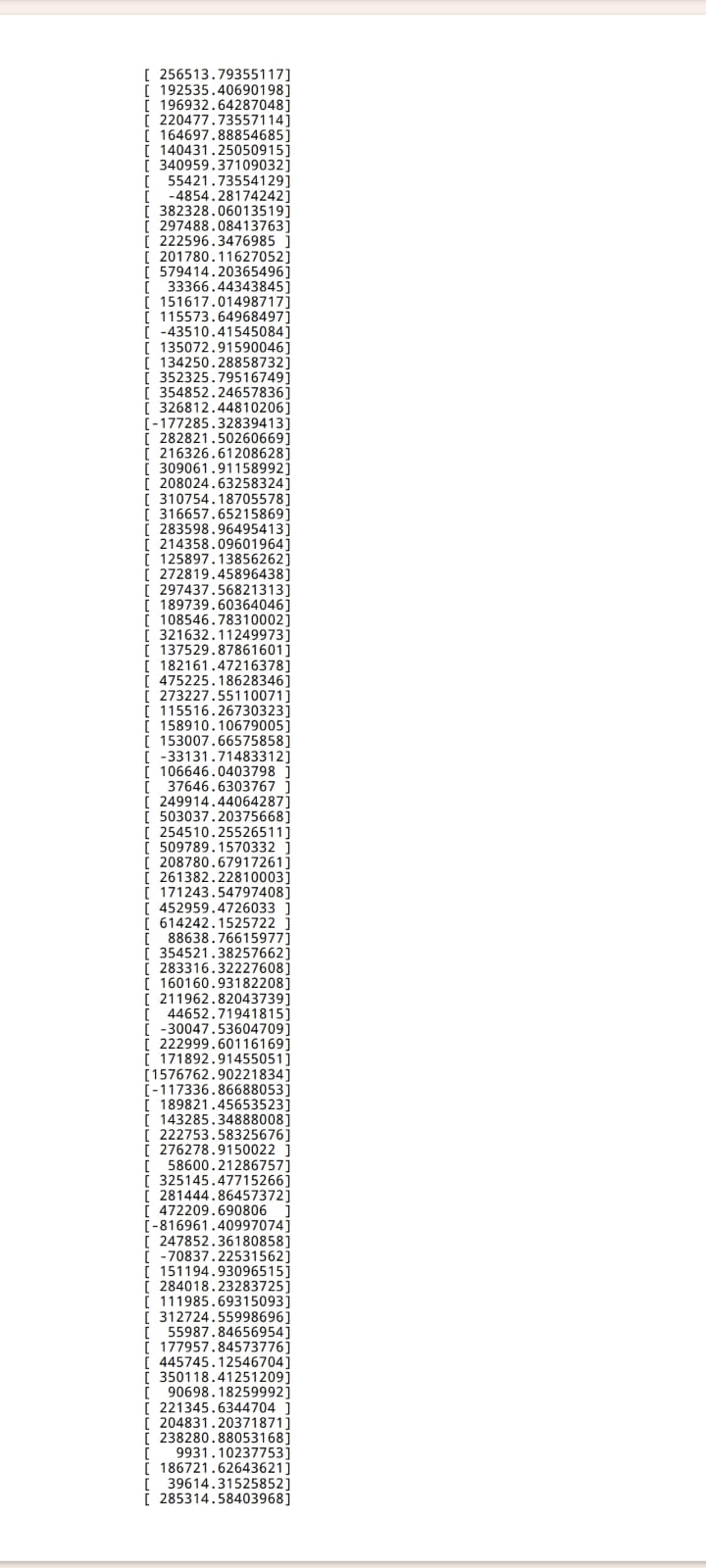


**Linear Regression:**



**Output:**





# K nearest neighbors:

from sklearn.neighbors import KNeighborsRegressor

model=KNeighborsRegressor()

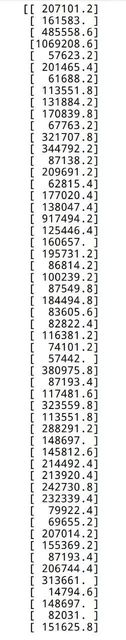
model.fit(x\_train,y\_train)

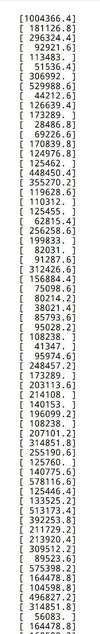
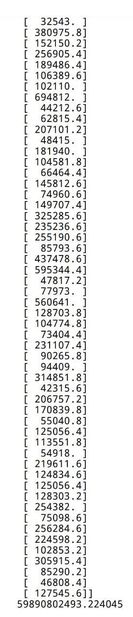
y\_pred=model.predict(x\_test)

print(y\_pred)

from sklearn.metrics import mean\_squared\_error

print(mean\_squared\_error(y\_test,y\_pred))

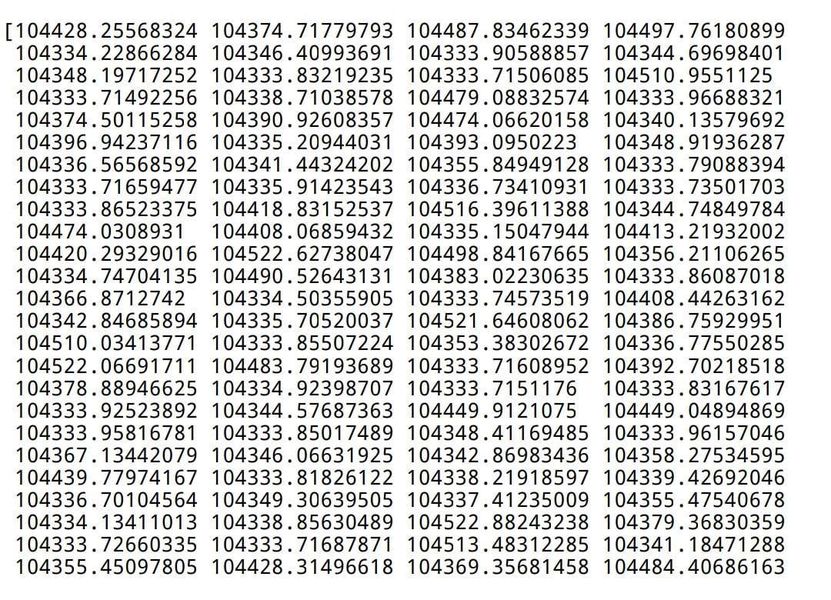


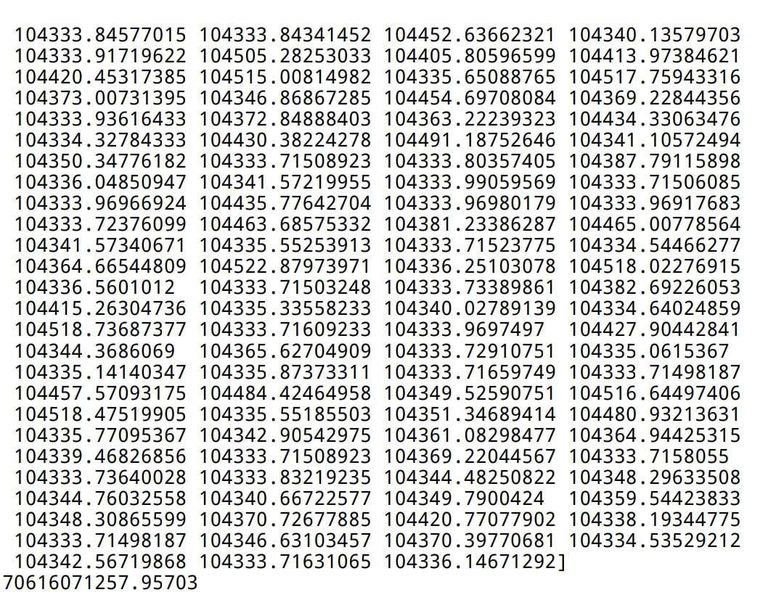


**Support vector machine:**

from sklearn.svm import SVR  
model=SVR()  
[model.fit](http://model.fit/)(x\_train,y\_train)  
y\_pred=model.predict(x\_test)  
print(y\_pred)  
from sklearn.metrics import mean\_squared\_error  
print(mean\_squared\_error(y\_test,y\_pred))

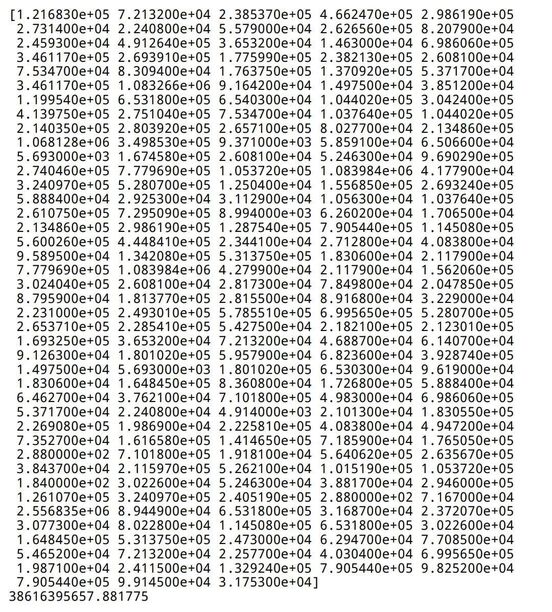
output:

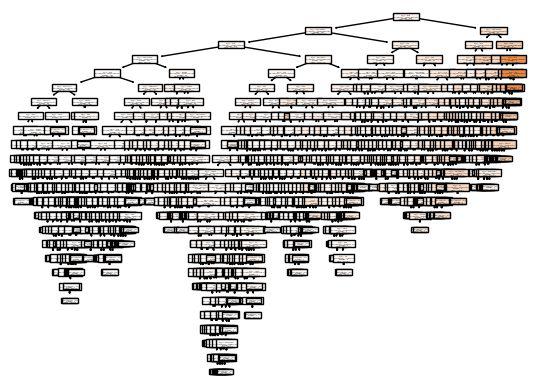
****

****

**Decision tree:**

from sklearn.tree import DecisionTreeRegressor  
model=DecisionTreeRegressor()  
[model.fit](http://model.fit/)(x\_train,y\_train)  
y\_pred=model.predict(x\_test)  
print(y\_pred)  
from sklearn.metrics import mean\_squared\_error  
print(mean\_squared\_error(y\_test,y\_pred))  
from sklearn import tree  
tree.plot\_tree(model,filled=True)

****

****

**6. CONCLUSION**

In conclusion, predicting box office collections using machine learning (ML) is a complex and challenging task that involves analyzing a wide range of variables such as historical box office data, movie genre, release date, cast and crew, marketing efforts, and social media sentiment analysis. However, by leveraging the power of ML algorithms such as regression analysis, decision trees, and neural networks, it is possible to accurately predict box office collections for a given movie.

One of the main advantages of using ML for box office prediction is the ability to process vast amounts of data quickly and efficiently, allowing studios and investors to make informed decisions about which movies to invest in and how to market them. ML also enables the identification of patterns and trends that might not be immediately apparent to human analysts, thereby providing insights into the key factors that drive box office success.

Despite the advantages, it's important to note that predicting box office collections using ML is not a foolproof method. There are still many variables that can impact the success of a movie, and unexpected events such as the COVID-19 pandemic can disrupt traditional movie release patterns and make predictions difficult. Nonetheless, the use of ML for box office prediction represents an important tool for the film industry, and advances in technology and data analysis will continue to improve the accuracy of these predictions in the future.

**7.FUTURE SCOPE**

The future scope of box office prediction in machine learning (ML) is quite promising. With the advancement of ML algorithms and techniques, it has become easier to analyze large amounts of data and make accurate predictions. Here are some possible areas where ML could be applied to predict box office performance:

1. Predicting movie revenue:

ML algorithms can analyze various factors such as genre, cast, release date, marketing budget, and social media buzz to predict the box office revenue of a movie. This can help studios make informed decisions about movie production, marketing, and distribution.

2. Optimizing marketing campaigns:

ML algorithms can analyze audience behavior and preferences to help studios tailor their marketing campaigns to specific demographics. This can help studios make more efficient use of their marketing budget and increase the chances of a movie's success.

3. Identifying potential blockbusters:

ML algorithms can analyze historical box office data to identify patterns and trends that can predict the success of a movie. This can help studios identify potential blockbusters early on and invest in them accordingly.

4. Forecasting the impact of external factors: ML algorithms can analyze external factors such as weather, economic conditions, and competition to predict how they might affect box office performance. This can help studios make informed decisions about when to release a movie and how to position it in the market.

Overall, the future of box office prediction in ML is bright, with potential applications in movie production, marketing, and distribution.

Top of Form

**REFERENCES**

[1] https://www.kaggle.com/code/akkibansal/box-office-prediction

[2] http://www.ijraset.com/research-paper/imdb-box-office-prediction-using-ml-algorithms

[3] https://github.com/BrianRuizy/box-office-prediction

[4] http://www.researchgate.net/publication/348120181\_Prediction\_of\_Box- office\_Success\_A\_Review\_of\_Trends\_and\_Machine\_Learning\_Computational\_Models

[5] <https://link.springer.com/article/10.1007/s10796-016-9689-z>

[6] https://www.ijraset.com/research-paper/imdb-box-office-prediction-using-ml-algorithms