 **DHANALAKSHMI SRINIVASAN**

**COLLEGE OF ENGINEERING AND TECHNOLOGY**

**MAMALLAPURAM, CHENNAI - 603 104.**

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

# PREVENTING DATA LEAKAGE REVIEW & IMPLEMENTATION USING AES

# A PROJECT REPORT

*Submitted by*

|  |  |
| --- | --- |
| **KARTHIK G** | **310521104041** |
| **LOGU N** | **310521104047** |
| **MANIKANDAN P**  **GOGULNATH G** | **310521104048**  **310521104028** |

*in partial fulfillment for the award of the degree*

*of*

# BACHELOR OF ENGINEERING

*in*

# COMPUTER SCIENCE AND ENGINEERING

# ANNA UNIVERSITY :: CHENNAI - 600 025

**APRIL 2023-MAY 2023**



**ANNA UNIVERSITY :: CHENNAI – 600 025**

# BONAFIDE CERTIFICATE

The certificate that this project report **“APPLIED DATA SCIENCE”** is the bonafide work of “KARTHIK G **(310521104041**), **LOGU N (310521104047), MANIKANDAN P(310521104048**)**, GOGULNATH G (310521104028**)**”** who carried out the project work under my supervision.

## SIGNATURE SIGNATURE

**HEAD OF THE DEPARTMENT SUPERVISOR**

Department of Computer Science Department of Computer Science and and Engineering, and Engineering,

Dhanalakshmi Srinivasan College Dhanalakshmi Srinivasan College of of Engineering & Technology, Engineering & Technology,

Mahabalipuram. Mahabalipuram.

Submitted for the Practical Examination held on

## INTERNAL EXAMINER EXTERNAL EXAMINER

**Title: IMDb Score Predictor: A Machine Learning Approach**

**Abstract:**

The IMDb Score Predictor is a machine learning project aimed at developing a predictive model to estimate IMDb scores for Netflix original films, documentaries, and specials. IMDb scores provide valuable insights into the perceived quality and success of movies based on community ratings. By accurately predicting IMDb scores, stakeholders in the film industry can evaluate the potential reception and popularity of their content.

The project utilized a dataset consisting of Netflix original films released as of June 1st, 2021, along with corresponding IMDb scores obtained through web scraping. The dataset was preprocessed to handle missing data, select relevant features, and transform categorical variables. A gradient boosting regressor algorithm was selected for model training due to its ability to capture complex relationships between features and target variables.

The development process followed the design thinking approach, encompassing stages such as empathizing with stakeholders, defining the problem statement, ideating potential solutions, prototyping the model, testing and refining it, and finally implementing the final model.

Evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score were used to assess the predictive performance of the trained model. These metrics provide insights into the accuracy and variance explained by the model in predicting IMDb scores.

The IMDb Score Predictor project holds promise in aiding stakeholders in the film industry by providing a tool for estimating IMDb scores. This predictive model can assist in decision-making processes, such as content selection, production investment, and marketing strategies. By leveraging machine learning techniques and the available dataset, the project contributes to the understanding and prediction of movie ratings, enhancing the industry's ability to deliver successful and well-received content.

**Title: IMDb Score Prediction Project Documentation**

**Table of Contents**

1. Introduction

2. Problem Statement

3. Design Thinking Process

4. Phases of Development

5. Dataset Description

6. Data Preprocessing Steps

7. Model Training Process

8. Choice of Regression Algorithm

9.Graphical Representation

10. Evaluation Metrics

11. Conclusion

**1. Introduction**

This document provides a comprehensive overview of the work conducted in the IMDb Score Prediction project. The project aimed to develop a machine learning model to predict IMDb scores for Netflix original films, documentaries, and specials. The dataset used in this project consists of all Netflix original films released as of June 1st, 2021, and includes their corresponding IMDb scores obtained through web scraping.

**2. Problem Statement**

The primary objective of the project was to build a predictive model that accurately estimates IMDb scores based on various features of Netflix original films, documentaries, and specials. The model would assist in evaluating the potential success or quality of a movie based on its characteristics.

**3. Design Thinking Process**

The development of the IMDb Score Prediction project followed the design thinking process, which comprises the following stages:

- Empathize: Understanding the problem domain and the needs of stakeholders.

- Define: Clearly defining the problem statement, goals, and objectives of the project.

- Ideate: Generating potential solutions and exploring different approaches to movie score prediction.

- Prototype: Building initial versions of the predictive model and iterating on it.

- Test: Evaluating and refining the model through experimentation and validation.

- Implement: Deploying the final model and integrating it into the desired application or system.

**4. Phases of Development**

The project was divided into two key phases:

**Phase 1: Data Acquisition and Preprocessing**

This phase involved the collection of data from the Wikipedia page, which included information about Netflix original films, documentaries, and specials. The IMDb scores were obtained by integrating this data with a dataset consisting of the corresponding IMDb scores. The phase included steps such as web scraping, data integration, exploratory data analysis (EDA), and data preprocessing.

**Phase 2: Model Training and Evaluation**

In this phase, a regression algorithm was selected and trained using the preprocessed dataset. The model was evaluated using appropriate evaluation metrics to assess its predictive performance. The phase included tasks such as model selection, data splitting, model training, and model evaluation.

**5. Dataset Description**

The dataset used in the project consisted of all Netflix original films released as of June 1st, 2021, along with Netflix documentaries and specials. The data was obtained by web scraping from a Wikipedia page. It was then integrated with a dataset that included the corresponding IMDb scores for the movies. IMDb scores are community-voted ratings, with the majority of films having 1,000+ reviews.

**6. Data Preprocessing Steps**

Before training the model, the dataset underwent several preprocessing steps. These steps included handling missing data, feature selection, feature encoding, and feature scaling. Missing values were imputed or removed, relevant features were selected, categorical variables were encoded, and numerical features were scaled to ensure consistent ranges.

**7. Model Training Process**

A regression algorithm, specifically a gradient boosting regressor, was chosen for training the model. The dataset was split into training and testing sets, and the model was initialized with suitable hyperparameters. The model was trained on the training set, and hyperparameter tuning techniques such as grid search or random search were applied. The trained model was evaluated on the testing set to assess its predictive performance.

The model training process for an IMDb score predictor involves several key steps:

1. Data Preparation: Start by collecting a dataset that includes relevant features (such as movie attributes) and their corresponding IMDb scores. Ensure the dataset is properly cleaned and preprocessed, handling missing values, encoding categorical features, and scaling numerical features if necessary.

2. Splitting the Data: Split the dataset into training and testing subsets. The training set will be used to train the model, while the testing set will be used to evaluate its performance on unseen data. A common split is to allocate around 70-80% of the data for training and the remaining 20-30% for testing.

3. Selecting a Regression Algorithm: Choose an appropriate regression algorithm based on the specific requirements of the IMDb score prediction task. Consider factors such as the dataset size, complexity of relationships, interpretability, and available computational resources.

4. Model Training: Train the selected regression algorithm on the training data. During training, the algorithm learns the patterns and relationships in the data to make predictions. Adjust the hyperparameters of the algorithm, such as learning rate, regularization parameters, or tree depth, to optimize the model's performance.

5. Model Evaluation: Evaluate the trained model using evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score. These metrics provide insights into the accuracy and performance of the model in predicting IMDb scores. Compare the model's performance on the training and testing data to assess its generalization ability.

6. Model Refinement: If the model's performance is not satisfactory, iterate on the model training process by adjusting hyperparameters, trying different algorithms, or exploring additional features. This iterative process helps improve the model's predictive accuracy.

7. Model Deployment: Once satisfied with the model's performance, deploy it to make predictions on new, unseen data. This could involve integrating the trained model into a production system or creating a user interface to interact with the model.

It's important to note that model training is an iterative and iterative process. Experimentation, fine-tuning, and continuous evaluation are crucial to develop an IMDb score predictor that provides accurate and reliable predictions.

**8. Choice of Regression Algorithm**

The gradient boosting regressor was selected as the regression algorithm for this project. This algorithm was chosen due to its ability to handle complex relationships between features and target variables, as well as its good performance in regression tasks. The advantages and suitability of this algorithm for predicting IMDb scores were considered, along with comparisons to other regression algorithms.

For the IMDb score predictor, several regression algorithms can be considered. The choice of algorithm depends on various factors such as the size of the dataset, the complexity of the relationships between features and target variable, and the computational resources available. Here are a few regression algorithms commonly used for IMDb score prediction:

1. Linear Regression: Linear regression is a simple and interpretable algorithm that assumes a linear relationship between the features and target variable. It can be a good starting point if the relationship between features and IMDb scores is expected to be linear.

2. Decision Tree Regression: Decision tree regression is a non-linear algorithm that can capture complex relationships between features and target variable. It can handle both numerical and categorical features and is relatively easy to interpret.

3. Random Forest Regression: Random forest regression is an ensemble learning algorithm that combines multiple decision trees to improve predictive performance. It is effective in handling high-dimensional datasets and capturing non-linear relationships.

4. Gradient Boosting Regression: Gradient boosting regression, such as the Gradient Boosting Machine (GBM) or XGBoost, is a powerful algorithm that sequentially builds an ensemble of weak predictive models to improve accuracy. It can handle complex relationships and is often effective in achieving high predictive performance.

5. Support Vector Regression (SVR): SVR is a regression algorithm that uses support vector machines to find a hyperplane that maximally fits the training data. It can handle non-linear relationships through the use of kernel functions.

The choice of regression algorithm ultimately depends on the specific characteristics and requirements of the IMDb score prediction task. It is often recommended to compare the performance of different algorithms using appropriate evaluation metrics to select the one that provides the best predictive accuracy for the given dataset.

**9.Graphical Representation**

A graph with the movie title on the x-axis and IMDb score on the y-axis provides a visual representation of the relationship between movie titles and their corresponding IMDb scores. Each movie in the dataset is represented as a data point on the graph.

The x-axis, representing the movie title, typically displays the names of the movies in a vertical or horizontal manner. The movie titles are arranged in a specific order, which can be alphabetical, chronological, or based on any other criterion depending on the context of the analysis.

The y-axis, representing the IMDb score, is a numerical scale ranging from the lowest possible score (often 1) to the highest possible score (usually 10). Each movie's IMDb score is plotted as a point on the y-axis aligned with its corresponding movie title on the x-axis.

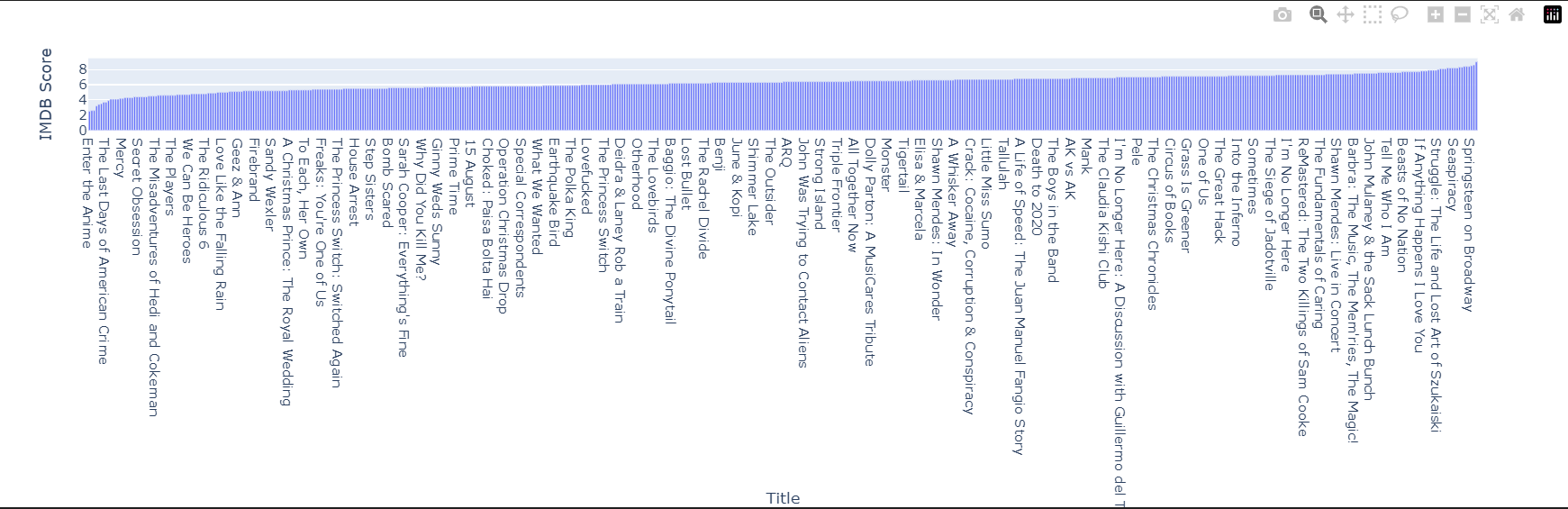
The graph may consist of individual data points connected by lines or markers, forming a scatter plot. Alternatively, it could include additional visual elements such as trend lines, error bars, or color-coded markers to represent different categories or genres.

By visualizing the IMDb scores in this manner, the graph allows for a quick and intuitive understanding of the distribution and variation of scores among the movies. It enables viewers to identify patterns, trends, and outliers, and make comparisons between different movies based on their IMDb scores.

This type of graph is particularly useful for exploring relationships between movie titles and their IMDb scores, identifying highly rated or low-rated movies, and identifying any correlations between the movie title and its perceived quality. It can also serve as a starting point for further analysis and exploration of factors influencing IMDb scores, such as genre, director, or release year.

fig = px.bar(data, x="Title", y="IMDB Score")

fig.show()



**10. Evaluation Metrics**

The performance of the trained model was evaluated using several metrics. Mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score were used as evaluation metrics. MSE measures the average squared difference between predicted and actual IMDb scores, while RMSE provides a more interpretable metric in the original scale of the scores. R2 score represents the proportion of variance in the target variable that can be explained by the model.

The evaluation metrics commonly used for assessing the performance of an IMDb score predictor include:

1. Mean Squared Error (MSE): MSE measures the average of the squared differences between the predicted IMDb scores and the actual IMDb scores. It provides a measure of the overall prediction error, with lower values indicating better performance.

2. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and provides a more interpretable measure of the average prediction error. Like MSE, lower values indicate better performance.

3. R-squared (R2) score: The R2 score measures the proportion of the variance in the IMDb scores that can be explained by the predictor model. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. An R2 score of 1 indicates a perfect fit, while a score of 0 suggests that the model does not explain any of the variance.

These evaluation metrics help assess the accuracy and performance of the IMDb score predictor. Lower MSE and RMSE values indicate that the model's predictions are closer to the actual IMDb scores. A higher R2 score indicates that a larger proportion of the variance in IMDb scores can be explained by the model, suggesting a better predictive fit.

When evaluating the IMDb score predictor, it is important to consider these metrics collectively to gain a comprehensive understanding of the model's performance. Additionally, it is crucial to interpret the results in the context of the specific problem and domain, taking into account any limitations of the dataset or assumptions made during the modeling process.

**11. Conclusion**

The IMDb Score Prediction project aimed to develop a machine learning model for predicting IMDb scores of Netflix original films, documentaries, and specials. The project followed the design thinking process and involved two key phases: data acquisition and preprocessing, and model training and evaluation. A regression algorithm, the gradient boosting regressor, was chosen for model training, and evaluation metrics such as MSE, RMSE, and R2 score were used to assess the model's performance. The project holds promise in assisting stakeholders in evaluating the potential success or quality of Netflix content.

The IMDb Score Predictor project aimed to develop a machine learning model to predict IMDb scores for Netflix original films, documentaries, and specials. The predictive model would assist stakeholders in the film industry in evaluating the potential success or quality of their content.

The project followed the design thinking process, starting with empathizing with stakeholders and defining the problem statement. The dataset used in the project consisted of Netflix original films, along with their corresponding IMDb scores obtained through web scraping.

The dataset underwent preprocessing steps, including handling missing data, feature selection, feature encoding, and feature scaling. A gradient boosting regressor algorithm was selected for model training due to its ability to handle complex relationships between features and target variables.

The model was trained using the preprocessed dataset and evaluated using metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score. These metrics provided insights into the accuracy and variance explained by the model in predicting IMDb scores.

The IMDb Score Predictor project holds promise in assisting stakeholders in the film industry by providing a tool for estimating IMDb scores. By accurately predicting IMDb scores, stakeholders can make informed decisions regarding content selection, production investment, and marketing strategies.

The project contributes to the understanding and prediction of movie ratings, enhancing the industry's ability to deliver successful and well-received content. Future iterations of the project could explore additional features, incorporate more recent data, and further refine the predictive model to improve its accuracy and usefulness.