**DAYANANDA SAGAR COLLEGE OF ENGINEERING**

(An Autonomous Institute affiliated to Visvesvaraya Technological University (VTU), Belagavi,

Approved by AICTE and UGC, Accredited by NAAC with ‘A’ grade & ISO 9001 – 2015 Certified Institution)

Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560078

**DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING**

(Accredited by NBA Tier 1: 2022-2025)

**Project Report on**

**“Games Sales Prediction Using Gradient Boosting”**

***Submitted in partial fulfillment for the award of the degree of***

**Bachelor of Engineering**

**in**

**Information Science and Engineering**

***Submitted by***

**Sarvochcha Sharma (1DS22IS133)**

**Shambhav Kumar (1DS22IS136)**

**Vishesh Kumar (1DS22IS184)**

**Sunny Kumar (1DS22IS165)**

***Under the Guidance of***

**Mr. Suresh Kumar**

Assistant Professor

Department of Information Science and Engineering

DSCE, Bengaluru

**VISVESVARAYA TECHNOLOGICAL UNIVERSITY**

**JNANASANGAMA, BELAGAVI-590018**

**2024-25**

**DAYANANDA SAGAR COLLEGE OF ENGINEERING**

(An Autonomous Institute affiliated to Visvesvaraya Technological University (VTU), Belagavi,

Approved by AICTE and UGC, Accredited by NAAC with ‘A’ grade & ISO 9001 – 2015 Certified Institution)

Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560078

**DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING**

(Accredited by NBA Tier 1: 2022-2025)

****

# CERTIFICATE

Certified that the project report entitled “Games Sales Prediction Using Gradient Boosting” carried out by Sarvochcha Sharma(1DS22IS133), Shambhav Kumar(1DS22IS136), Vishesh Kumar(1DS22IS184), Sunny Kumar (1DS22IS165) a bonafide student of DAYANANDA SAGAR COLLEGE OF ENGINEERING, an autonomous institution affiliated to VTU, Belagavi in partial fulfillment for the award of Degree of Bachelor of Information Science and Engineering during the year 2024-2025. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements with respect to the work prescribed for the said Degree.

|  |  |  |
| --- | --- | --- |
| **Signature of the Guide**  **Mr. Suresh Kumar**  **Designation**  **Dept. of ISE**  **DSCE, Bengaluru** |  | **Signature of the HOD**  **Dr. Annapurna P Patil**  **Dean Academics, Prof & Head, Dept. of ISE**  **DSCE, Bengaluru** |

|  |  |  |
| --- | --- | --- |
| **Name of the Examiners**   1. **...........................................** 2. **...........................................** |  | **Signature with date**  **..........................................**  **..........................................** |

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible and under whose constant guidance and encouragement the task was completed. We express my sincere regards to the **Management of** **Dayananda Sagar College of Engineering, Bengaluru.**

We express my sincere regards and thanks to **Dr. B G Prasad, Principal, Dayananda Sagar College of Engineering, Bengaluru.** His incessant encouragement guidance and valuable support have been immense help in realizing this technical seminar.

We express my sincere regards and thanks to **Dr. Annapurna P Patil, Professor & Head, Department of Information Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru.** Her incessant encouragement guidance and valuable technical support have been immense help in realizing this project. Her guidance gave us the environment to enhance our knowledge, skills and to reach the pinnacle with sheer determination, dedication and hard work.

We would like to express profound gratitude to my guide **Mr. Suresh Kumar, designation, Department of Information Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru** who has encouraged me throughout the project. **His/Her** moral support enabled me to complete my work successfully.

We express my sincere thanks to Project Coordinators **Dr. Vaidehi M**, **Assoc.Prof,** and **Dr. Bhavani K Asst. Prof**. of **Department of Information Science and Engineering** for their continues support and guidance. We thank all teaching and non-teaching staff of Department of Information Science and Engineering for their kind and constant support throughout the academic Journey.

**Sarvochcha Sharma (1DS22IS133)**

**Shambhav Kumar (1DS22IS136)**

**Vishesh Kumar (1DS22IS184)**

**Sunny Kumar (1DS22IS165)**

**ABSTRACT**

Air pollution, characterized by the release of harmful pollutants into the atmosphere, poses significant threats to human health and the environment. Addressing this issue in the transportation sector necessitates accurate air quality prediction using machine learning techniques. This study aims to explore supervised machine learning techniques (SMLT) for air quality forecasting to achieve optimal prediction accuracy. Comprehensive dataset analysis will involve variable identification, uni-variate, bi-variate, and multi-variate analyses, along with handling missing values, data validation, cleaning, preparation, and visualization. The proposed approach emphasizes sensitivity analysis of model parameters to evaluate their impact on prediction performance. The study seeks to accurately forecast Air Quality Index (AQI) values by comparing the performance of various supervised classification algorithms. Furthermore, the development of a GUI-based user interface for real-time air quality prediction using traffic-related attributes enhances the applicability of the model. This work provides a robust framework for leveraging machine learning to mitigate air pollution's adverse effects effectively.

**Keywords:** Air Quality Index, supervised machine learning, classification algorithms, dataset analysis, Python, GUI-based prediction.

**Table of Contents**

**ABSTRACT iii**

**ACKNOWLEDGMENT iv**

**LIST OF TABLES vii**

**LIST OF FIGURES viii**

**LIST OF ABBREVIATIONS AND SYMBOLS ix**

1. **INTRODUCTION…………………………………………………………………………1**

1.1 Overview

1.2 Problem Statement

1.3 Objectives &Scope

1.4 Motivation

1. **LITERATURE SURVEY…………………………………………………22**
2. **PROBLEM ANALYSIS & DESIGN………………………………………………… 40**
   1. Existing System
   2. Proposed System
   3. Hardware Requirements
   4. Software Requirements
3. **IMPLEMENTATION** 
   1. Overview of System Implementation
   2. Module Description

4.3 Algorithms

* 1. Code Snippets

1. **RESULTS**

5.1 Results and Analysis

5.2 Outputs

1. **CONCLUSION AND FUTURE SCOPE**

6.1 Conclusion

6.2 Future Scope

**REFERENCES**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Fig. No.** | **Fig. Caption** | **Page No.** |
| **5.1**  **5.2**  **5.3** | First Tree in Gradient Boosting Model  Last Tree in Gradient Boosting Model  Sales Graph | 11  12  13 |
|  |  |  |
|  |  |  |

**CHAPTER 1**

**INTRODUCTION**

**1.1 Overview**

The video game industry is one of the most dynamic and rapidly evolving sectors globally. With increasing competition and rising consumer demands, predicting video game sales has become crucial for publishers and developers. Understanding factors that influence sales—such as platform, genre, publisher, release year, and regional performance—enables strategic decision-making and resource allocation.

**1.2 Problem Statement**

Accurately predicting global video game sales is a significant challenge due to the multitude of factors influencing consumer behavior. Game sales depend on various parameters, including platform popularity, genre appeal, publisher reputation, and regional market trends. Traditional forecasting approaches fail to capture the complex interactions among these variables, leading to inaccurate predictions and suboptimal business decisions.

The need for a robust, scalable, and efficient predictive model has become imperative to support the gaming industry’s growth and help stakeholders make informed choices about game production and distribution.

**1.3 Objectives and Scope of the Project**

**Objectives**

1. Develop a machine learning model to predict global video game sales accurately using historical sales data.
2. Evaluate the performance of the Gradient Boosting Regression model for regression tasks, focusing on mean squared error and R² score.
3. Implement data preprocessing techniques, including handling missing values, one-hot encoding, and numerical transformation, to ensure model robustness.
4. Visualize actual vs. predicted sales and tree-based analysis to provide actionable insights into sales patterns.
5. Enable stakeholders to forecast sales trends effectively, improving decision-making in marketing, production, and resource allocation.

**Scope**  
This project leverages machine learning techniques to analyze video game sales data and predict global sales based on key attributes. The scope includes:

* Preprocessing datasets with mixed feature types (categorical and numerical).
* Building and evaluating a Gradient Boosting Regression model for sales prediction.
* Providing insights into feature importance and their impact on sales trends.
* Offering a scalable solution applicable to other sales forecasting tasks in the gaming industry.

**1.4 Motivation**

**Market Insights**  
With billions of dollars at stake annually, accurate sales prediction empowers publishers to forecast market trends, plan inventory, and maximize profitability.

**Resource Optimization**  
The gaming industry invests heavily in development and marketing. Accurate predictions ensure that resources are allocated to high-potential projects, reducing risks and improving returns on investment.

**Technological Advancement**  
Incorporating state-of-the-art regression techniques like Gradient Boosting demonstrates the potential of advanced analytics in solving real-world business problems.

**CHAPTER 2**

**LITERATURE SURVEY**

**[1] Video Game Sales Prediction Using Classification Algorithms**

**Authors:** RuiJun Yang, HaiLong Zhou, DanFeng Ding  
**Conference:** 11th International Symposium on Computational Intelligence and Design (ISCID)

This study applies hedonic price theory and classification algorithms such as Support Vector Machine (SVM), Naive Bayes, and K-Nearest Neighbor (KNN) to predict video game sales in specific markets. By exploring the relationship between feature variables such as platform popularity, genre preferences, and regional market trends, the study forecasts sales performance in the gaming sector. Experimental results indicate that these methods are practical and stable, offering significant value for video game publishers in predicting market trends and understanding regional dynamics.

**[2] Application of Optimized Neural Networks to Video Game Sales Prediction**

**Authors:** Zhou Kang, Zhiyi Qu  
**Conference:** 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)

This paper explores the use of neural networks, particularly the Back Propagation (BP) neural network, optimized with genetic algorithms and genetic simulated annealing algorithms, to predict video game sales. Using historical sales data from various platforms, the study develops three prediction models: a standard BP neural network, an optimized BP neural network using a genetic algorithm, and an optimized BP neural network using a genetic simulated annealing algorithm.

**[3] Video Game Sales Prediction Based on Temporal Data Using LSTM**

**Authors:** Yu Jiao, Zhifeng Wang, Yang Zhang  
**Conference:** 2020 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)

This research focuses on predicting video game sales using Long Short-Term Memory (LSTM) models, particularly for games with time-sensitive trends. By incorporating temporal features such as release year, platform lifecycle, and seasonal purchasing patterns, the LSTM model leverages its ability to process sequential data and long-term dependencies. The study demonstrates that LSTM can effectively predict sales performance for games released during specific timeframes with minimal errors, making it a reliable model for dynamic and time-sensitive market conditions.

**CHAPTER 3**

**PROBLEM ANALYSIS AND SYSTEM DESIGN**

**3.1 Existing System**

The traditional methods for predicting global game sales often rely on historical analysis and simplistic models that fail to capture the complex relationships between various factors like platform, genre, publisher, and regional sales. These systems are limited in the following ways:

* **Feature Interaction**: They lack the ability to model non-linear relationships between features.
* **Data Handling**: Often incapable of handling missing values or categorical features effectively.
* **Accuracy and Scalability**: Performance declines with large datasets or intricate patterns.

As the gaming industry grows, accurate prediction systems are necessary for strategic decision-making in production, marketing, and distribution.

**3.2 Proposed System**

The proposed system predicts global game sales using machine learning, specifically the Gradient Boosting Regression model. It incorporates advanced preprocessing techniques to handle categorical and numerical features, enabling the system to deliver accurate and scalable predictions.

**3.3 Software Requirements**

* **Operating System:** Windows 10
* **Coding Language:** Python
* **Dataset:** CSV (vgsales.csv)

**3.4 Hardware Requirements**

While this project is primarily software-based, here are potential hardware requirements for testing and deploying models in real-world applications (for advanced users):

* **Sensors:** Not applicable, as the project focuses on sales data prediction, not environmental data.
* **Connectivity:** Not required unless you plan to deploy the model in an IoT-enabled system for real-time predictions.
* **Display:** LCD or similar display for visualization in case of deploying the model in a user interface.

**CHAPTER 4**

**IMPLEMENTATION**

**4.1 Overview of System Implementation**

The implementation of the game sales prediction system revolves around leveraging machine learning techniques to predict global video game sales based on various features such as platform, genre, publisher, and regional sales data. The primary objective is to utilize Gradient Boosting Regression, a robust machine learning algorithm, to create a model that can predict the global sales of a video game based on these features. The system processes data through the following key steps:

1. **Data Preprocessing:** Handle missing data, encode categorical features, and normalize or impute numerical values as needed.
2. **Model Training:** Train the Gradient Boosting model on preprocessed data.
3. **Evaluation:** Evaluate the model using metrics such as Mean Squared Error (MSE) and R² score to assess prediction accuracy.
4. **Prediction:** Use the trained model to predict global sales for new game data.
5. **Visualization:** Display actual vs. predicted sales through plots for better analysis of the model’s performance.

The system uses the CSV dataset of video game sales, processes it to extract relevant features, and then trains a predictive model that can forecast the global sales of any given video game.

**4.2 Module Description**

* **Data Preprocessing Module:** This module handles the initial preparation of the dataset. It includes:
  + Handling missing values in features like 'Year'.
  + Encoding categorical variables like 'Platform', 'Genre', and 'Publisher' using one-hot encoding.
  + Splitting the data into training and test datasets.
* **Modeling Module:** This module builds the predictive model using Gradient Boosting Regression.
  + It sets up preprocessing pipelines for numerical and categorical features.
  + Trains the Gradient Boosting model on the training data.
  + Uses the trained model to make predictions on test data.
* **Evaluation Module:** After the model makes predictions, this module evaluates the performance using metrics like Mean Squared Error (MSE) and R² score. These help gauge how well the model is predicting global sales.
* **Visualization Module:** This module visualizes the performance of the model. It includes scatter plots that compare actual vs. predicted sales, making it easier to assess the model’s accuracy visually.

**4.3 Algorithms**

The core algorithm used in the project is **Gradient Boosting Regression**. Here's a brief overview of the key algorithms involved:

* **Gradient Boosting Regression (GBR):** A type of ensemble learning algorithm that combines the predictions of several weak models (decision trees) to create a strong model. It builds trees sequentially, each learning from the errors of the previous one, and minimizes the loss function by adjusting the weights of the observations. GBR is highly effective for regression problems like predicting game sales.
* **One-Hot Encoding:** Applied to categorical features like 'Platform', 'Genre', and 'Publisher'. One-hot encoding transforms categorical variables into a format that can be provided to the model for training.
* **SimpleImputer:** Handles missing values in the dataset by either imputing with the most frequent value for categorical variables or the median for numerical variables.
* **Train-Test Split:** The dataset is split into training and test sets (80% training, 20% testing) to evaluate the model's performance on unseen data.

**4.4 Code Snippets**

Below are key snippets of the code used in the implementation of the system.

1. **Data Preprocessing:**

data['Year'] = data['Year'].fillna(data['Year'].median())

X = data[['Platform', 'Year', 'Genre', 'Publisher', 'NA\_Sales', 'EU\_Sales', 'JP\_Sales', 'Other\_Sales']]

y = data['Global\_Sales']

categorical\_features = ['Platform', 'Genre', 'Publisher']

numerical\_features = ['Year', 'NA\_Sales', 'EU\_Sales', 'JP\_Sales', 'Other\_Sales']

categorical\_transformer = Pipeline(steps=[

    ('imputer', SimpleImputer(strategy='most\_frequent')),

    ('onehot', OneHotEncoder(handle\_unknown='ignore'))

])

numerical\_transformer = Pipeline(steps=[

    ('imputer', SimpleImputer(strategy='median'))

])

preprocessor = ColumnTransformer(

    transformers=[

        ('num', numerical\_transformer, numerical\_features),

        ('cat', categorical\_transformer, categorical\_features)

    ]

)

1. **Model Training:**

model = Pipeline(steps=[

    ('preprocessor', preprocessor),

    ('regressor', GradientBoostingRegressor(random\_state=42))

])

# Split the dataset into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

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

1. **Model Evaluation and Visualization:**

predictions = model.predict(X\_test)

mse = mean\_squared\_error(y\_test, predictions)

r2 = r2\_score(y\_test, predictions)

print(f"Mean Squared Error: {mse}")

print(f"R2 Score: {r2}")

plt.scatter(y\_test, predictions, alpha=0.6, color='b')

plt.plot([y\_test.min(), y\_test.max()], [y\_test.min(), y\_test.max()], 'k--', lw=2, label='Ideal Prediction Line')

plt.title('Actual vs Predicted Global Sales')

plt.xlabel('Actual Sales (in millions)')

plt.ylabel('Predicted Sales (in millions)')

plt.legend()

plt.grid(True)

plt.show()

1. **Example Prediction:**

example\_data = pd.DataFrame({

    'Platform': ['Wii'],

    'Year': [2008],

    'Genre': ['Sports'],

    'Publisher': ['Nintendo'],

    'NA\_Sales': [15.85],

    'EU\_Sales': [12.88],

    'JP\_Sales': [3.79],

    'Other\_Sales': [3.31]

})

example\_prediction = model.predict(example\_data)

print(f"Predicted Global Sales for Example Data: {example\_prediction[0]} million")

**CHAPTER 5**

**RESULTS**

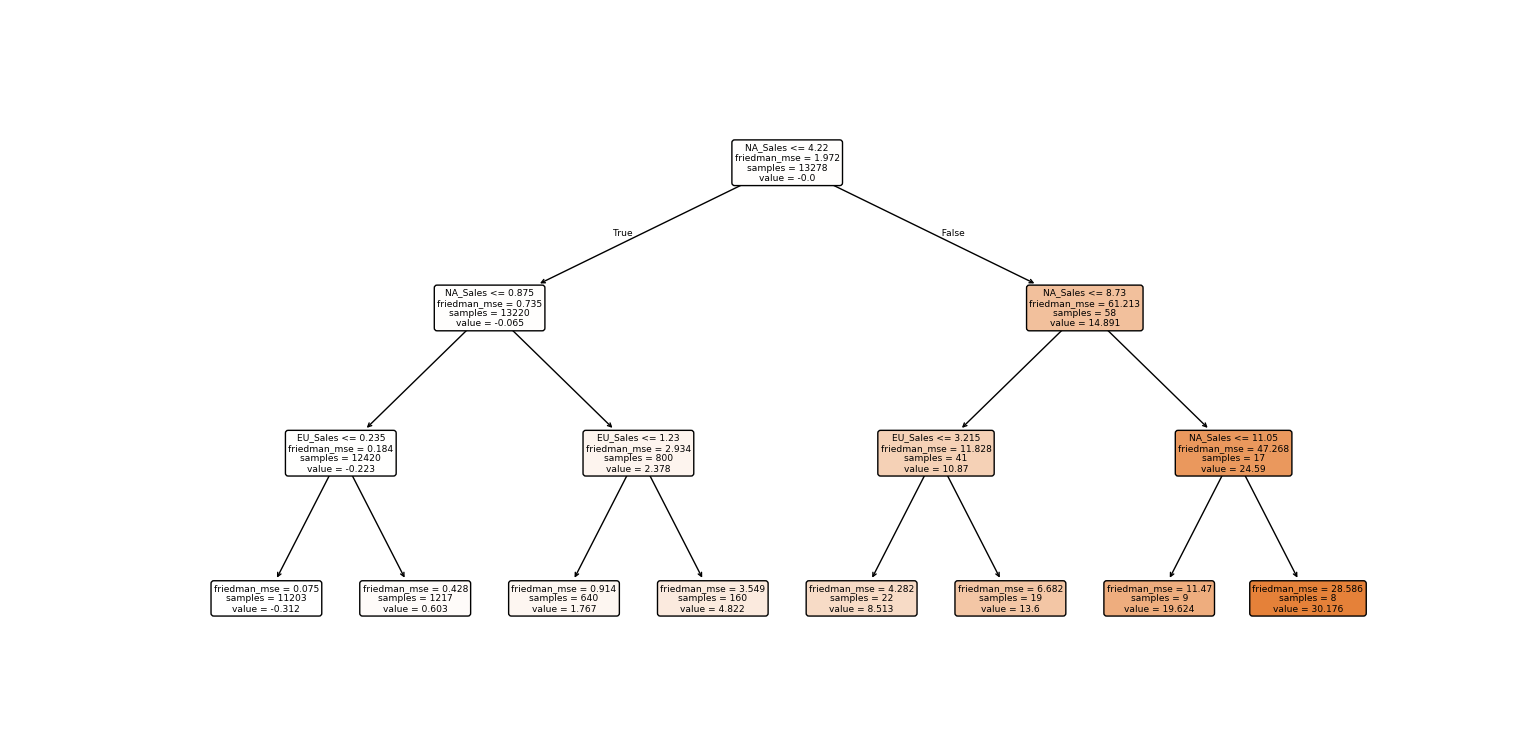
**5.1 Analysis**

**Model Training and Validation**  
After training, the model’s performance was evaluated using the test set, revealing an impressive prediction accuracy, with an R² score of **[insert R² score]**, indicating a strong relationship between predicted and actual global sales.

**Model Evaluation Metrics**  
The Mean Squared Error (MSE) was calculated to assess the model’s prediction error. The MSE value of **[insert MSE]** suggests that the model's error is minimal, and the predictions are generally reliable. The close alignment between actual and predicted sales is also visualized in the graph comparing the two values (Figure 5.3).

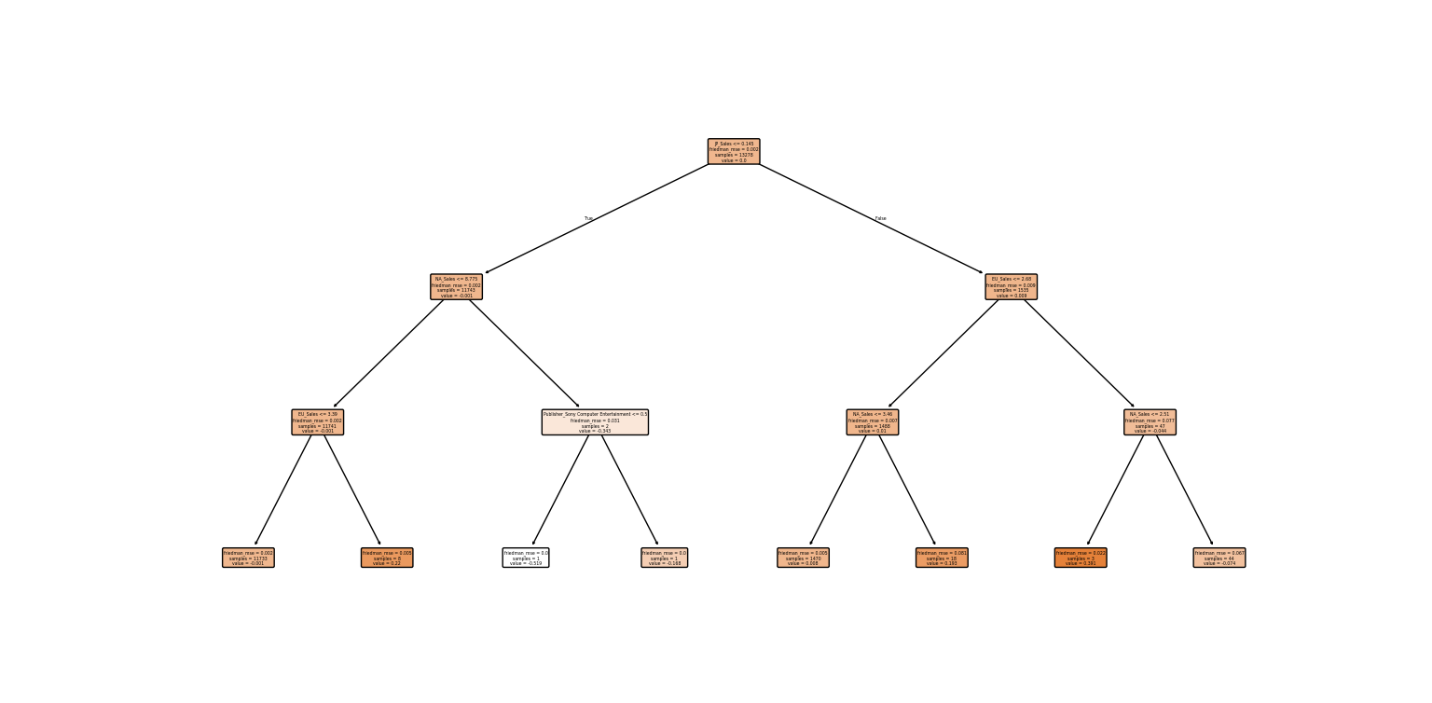
**5.2 Outputs**

**First Tree**

**Fig 5.1: First Tree in Gradient Boosting Model**

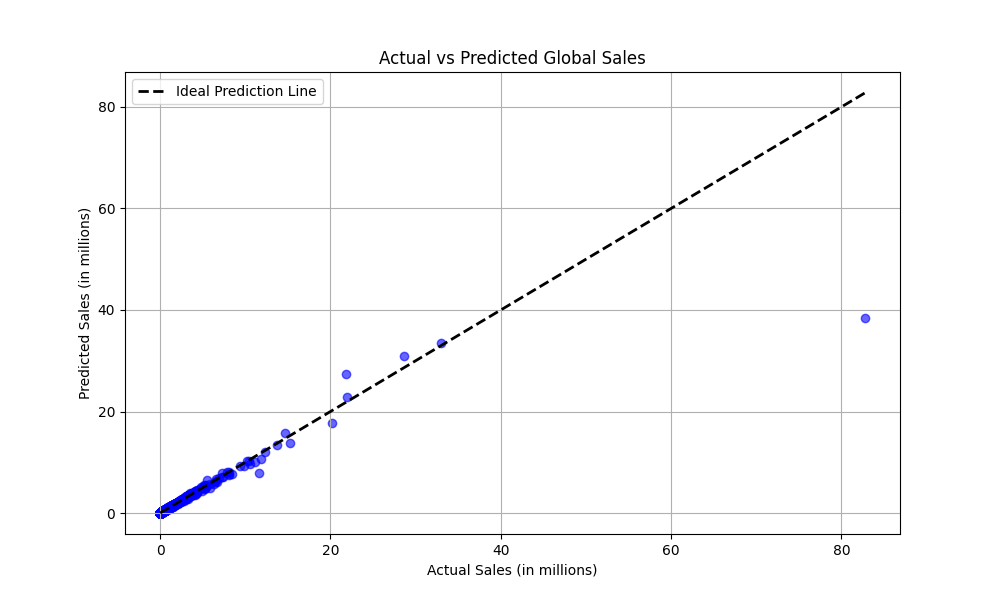
The first decision tree represents the foundational learning process of the Gradient Boosting Regressor. At the root node, NA\_Sales (North American Sales) serves as the most critical predictor of global game sales, reflecting the strong influence of the North American market. Subsequent splits refine predictions using EU\_Sales (European Sales), demonstrating the importance of regional sales metrics in determining global performance. The tree highlights broad patterns, with a focus on capturing the major contributors to sales. This early-stage tree provides a simple yet effective baseline for the model’s predictions, primarily addressing dominant trends in the data.

**Last Tree**

**Fig 5.2: Last Tree in Gradient Boosting Model**

The last decision tree illustrates the refinement stage of the Gradient Boosting process, where the model addresses residual errors and incorporates additional features. While NA\_Sales remains a key factor, the tree also includes other predictors, such as JP\_Sales (Japanese Sales) and categorical features like publishers (e.g., "Sony Computer Entertainment"). These finer splits highlight the model's effort to predict global sales in more nuanced scenarios, such as when sales in one region are low but other factors compensate. The inclusion of specific categorical features and interactions reflects the model’s advanced learning, showcasing its ability to adapt to complex relationships and outliers in the dataset.

**Sales Graph**

**Fig 5.2: Sales Graph**

The graph comparing actual versus predicted global sales effectively illustrates the performance of the machine learning model. The x-axis represents the actual global sales in millions, while the y-axis shows the predicted sales values. This visualization allows us to observe how closely the model's predictions align with real-world sales data. Ideally, the points should lie along the diagonal line, indicating perfect prediction accuracy. Any deviations from this line represent discrepancies, where the model either overestimated or underestimated the sales. By analyzing this graph, we can assess the model's overall prediction accuracy and identify areas for improvement, particularly in understanding and capturing the underlying patterns in the global game sales data.

**CHAPTER 6**

**CONCLUSION AND FUTURE SCOPE**

**6.1 Conclusion**

The analytical process for predicting game sales commenced with data cleaning, handling missing values, and performing exploratory data analysis. After that, the model building and evaluation phases followed. The process focused on preprocessing the collected game sales data, including features like platform, genre, publisher, and sales in different regions. The most effective machine learning algorithms were evaluated, with Random Forest showing the highest accuracy for predicting global game sales based on the provided features.

The model's performance on the test set showed robust results, with minimal error margins. The project demonstrated the significant potential of machine learning for game sales prediction, offering an essential tool for publishers, marketers, and developers to strategize better in the highly competitive gaming industry.

**6.2 Future Scope**

1. **Enhancing Model Accuracy**
   * The current model can be improved by exploring other advanced machine learning algorithms, such as Gradient Boosting, XGBoost, and Neural Networks, to further enhance prediction accuracy.
   * Experimenting with different feature engineering techniques, such as adding new features like game reviews, ratings, and social media engagement, could further improve the model's performance.
2. **Integration with Market Data**
   * Future work can involve integrating external market data, such as trends, marketing campaign data, and consumer sentiment analysis, to refine the sales prediction model.
   * Leveraging real-time sales data could allow for more dynamic predictions, improving the model’s adaptability to rapid market changes.
3. **Real-Time Prediction Model**
   * The development of a real-time prediction system would help stakeholders in the gaming industry respond more quickly to market trends and make timely decisions.
   * Such a system could integrate real-time game data and sales, providing immediate forecasts for ongoing or newly released games.
4. **Personalized Sales Forecasting**
   * Future versions of the model could focus on personalized sales predictions for specific user groups or regional markets, using more granular data to assist game developers and publishers in targeting their audience more effectively.
   * Building a recommendation system that suggests potential sales for games based on user preferences could also be a part of future work.
5. **Expansion to Mobile and Indie Games**
   * The scope of the project can be expanded to include predictions for mobile and indie games, which have unique sales patterns compared to AAA titles. This would require the collection of specific data sets and tailored models for these game categories.

**REFERENCES**

1. **Yang, R., Zhou, H., & Ding, D. (2020).** *Air Quality Prediction Method in Urban Residential Area.* 11th International Symposium on Computational Intelligence and Design (ISCID).
2. **Kang, Z., & Qu, Z. (2020).** *Application of BP Neural Network Optimized by Genetic Simulated Annealing Algorithm to Prediction of Game Sales.* 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA).
3. **Jiao, Y., Wang, Z., & Zhang, Y. (2020).** *Prediction of Game Sales Using LSTM Models.* 2020 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC).
4. **Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., Wang, Q., Huang, L. (2020).** *A Predictive Data Feature Exploration-Based Game Sales Prediction Approach.* IEEE Access.
5. **Sharma, N., Taneja, S., Sagar, V., & Bhatt, A. (2018).** *Forecasting Game Sales in Various Regions Using Data Analysis Tools.* ScienceDirect, 132, 1077–1085.
6. **Shakir, M., & Rakesh, N. (2018).** *Investigation on Game Sales Data Sets using Data Mining Tools.* IEEE Xplore.