Indonesian Journal of Data and Science

Volume 4 Issue 3
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i2.166

Research Article

Predicting Online Gaming Behaviour Using Machine Learning Techniques

Nurul Rismayanti [1,*]

[1] Universitas Negeri Malang, Malang, Indonesia, nurulrismayanti.labfik@umi.ac.id

Correspondence should be addressed to Nurul Rismayanti; nurulrismayanti.labfik@umi.ac.id

Received 11 May 2024; Accepted 28 June 2024; Published 31 July 2024

Abstract:

Understanding player behaviour in online gaming is essential for enhancing user engagement and retention. This study utilizes a dataset from Kaggle, capturing a wide range of player demographics and in-game metrics to predict player engagement levels categorized as 'High,' 'Medium,' or 'Low.' The dataset includes features such as age, gender, location, game genre, playtime, in-game purchases, game difficulty, session frequency, session duration, player level, and achievements. The research employs a Gaussian Naive Bayes model, with data pre-processing steps including feature selection, categorical data encoding, and scaling of numerical features. The dataset is split into training (80%) and testing (20%) sets, and a 5-fold cross-validation is used to ensure model robustness. The model's performance is evaluated using accuracy, precision, recall, and F1-score. The results show consistent performance across different folds, with an average accuracy of 84.27%, precision of 85.59%, recall of 84.27%, and F1-score of 83.98%. These findings indicate that the Gaussian Naive Bayes model can reliably predict player engagement levels, identifying significant predictors such as session frequency and in-game purchases. The study contributes to game analytics by providing a predictive model that can help game developers and marketers design more engaging gaming experiences. Future research should incorporate a broader range of features, including psychological and social factors, and explore other machine learning algorithms to enhance predictive accuracy. This study's insights are valuable for developing strategies to improve player retention and satisfaction in the gaming industry.

Keywords: Online Gaming, Player Engagement, Machine Learning, Gaussian Naive Bayes, Game Analytics.
Dataset link: https://www.kaggle.com/datasets/rabieelkharoua/predict-online-gaming-behavior-dataset

1. Introduction

In recent years, the online gaming industry has experienced exponential growth, attracting millions of players worldwide. As this digital landscape expands, understanding player behaviour has become increasingly crucial for game developers and marketers. Comprehensive metrics and demographics related to player behaviour can provide invaluable insights into how games are played and what drives player engagement. The dataset used in this study, sourced from Kaggle, captures a broad spectrum of player demographics and in-game metrics. This includes variables such as age, gender, location, game genre, playtime, in-game purchases, game difficulty, session frequency, session duration, player level, achievements, and engagement levels. These variables are essential for developing predictive models that can enhance user experience and retention strategies. One of the primary challenges in the gaming industry is predicting player engagement levels. Engagement is a critical factor as it directly influences a game's popularity and revenue generation. However, due to the complexity and variability of player behaviour, accurately predicting engagement levels is a challenging task. This study aims to address this problem by leveraging machine learning techniques to predict player engagement levels based on a variety of demographic and in-game metrics. The goal is to develop a robust predictive model that can provide actionable insights for game developers and marketers.

The primary objective of this research is to utilize the Gaussian Naive Bayes algorithm to predict player engagement levels, categorized as 'High', 'Medium', or 'Low'. This involves pre-processing the dataset, including feature selection, encoding categorical data, and scaling the features. By employing a 5-fold cross-validation method, the study aims to ensure the robustness and reliability of the predictive model. The research also aims to evaluate the model's performance using key metrics such as accuracy, precision, recall, and F1-score [1]–[3]. Through this approach, the study seeks to identify which features are most influential in predicting player engagement. To guide this research, several key questions and hypotheses have been formulated. Firstly, the study seeks to determine how accurately the Gaussian Naive Bayes model can predict player engagement levels [4]–[6]. Secondly, it aims to identify the most significant features that contribute to these predictions. Hypothetically, it is expected that certain demographics, such as age and location, along with specific in-game behaviours like session frequency and in-game purchases, will play a crucial role in determining engagement levels. By addressing these questions, the research hopes to provide deeper insights into player behaviour and engagement.

The scope of this research is confined to the dataset provided, which includes a specific set of features and variables. While the dataset offers a comprehensive view of player behaviour, it may not encompass all possible factors that influence engagement. Additionally, the study acknowledges the limitations inherent in using a single dataset, such as potential biases and the challenge of generalizing findings across different gaming platforms or genres. Nonetheless, the research aims to provide a foundational understanding that can be built upon in future studies with more diverse datasets and advanced machine learning techniques [7]–[9]. Overall, this research contributes to the field of game analytics by developing a predictive model for player engagement and identifying key factors that influence engagement. The findings can assist game developers in designing more engaging and immersive experiences, ultimately enhancing player retention and satisfaction. Additionally, the study provides a methodological framework for using machine learning in gaming analytics, which can be adapted and extended in future research. By addressing the complex challenge of predicting player engagement, this research offers valuable insights that can drive innovation and growth in the online gaming industry.

2. Method:

This study employs a quantitative research design using a machine learning approach to predict player engagement levels in online gaming. The primary algorithm used for prediction is the Gaussian Naive Bayes model, which is suitable for handling the various categorical and continuous variables present in the dataset [10]–[12]. The research design includes data pre-processing, feature selection, model training, and evaluation using cross-validation [13]–[15]. The overall goal is to create a robust model that accurately predicts player engagement levels and identifies significant predictors of engagement. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

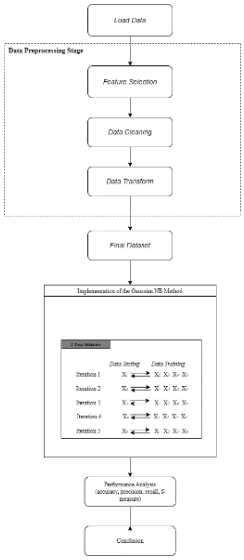

Figure 1. General Research Design Stages

Data Collection Process

The dataset used in this study is sourced from Kaggle, titled "Predict Online Gaming Behaviour Dataset." It includes a comprehensive set of features related to player demographics and in-game behaviour. The dataset was downloaded from Kaggle, ensuring it contains a broad and representative sample of online gamers. The data collection process involved gathering detailed metrics on player behaviour and demographics, which are crucial for building an accurate predictive model. Given the nature of the dataset, various pre-processing steps are required to prepare the data for analysis. This includes handling missing values, encoding categorical variables, and scaling numerical features. The key features are:

Table 1. Feature Descriptions

Column Name	Type	Description
PlayerID	Categorical	Unique identifier for each player
Age	Numerical	Age of the player
Gender	Categorical	Gender of the player (Male, Female)
Location	Categorical	Geographic location (USA, Europe, Asia, Other)
GameGenre	Categorical	Genre of the game Sports, Action, Strategy, Simulation, RPG)
PlayTimeHours	Numerical	Average hours spent playing per session
InGamePurchases	Categorical	Indicates if the player makes in-game purchases (No, Yes)
GameDifficulty	Categorical	Difficulty level of the game (Easy, Medium, Hard)
SessionsPerWeek	Numerical	Number of gaming sessions per week
AvgSessionDurationMinutes	Numerical	Average duration of each gaming session in minutes
PlayerLevel	Numerical	Current level of the player in the game
AchievementsUnlocked	Numerical	Number of achievements unlocked by the player
EngagementLevel	Categorical	Categorized engagement level (Low, Medium, High)

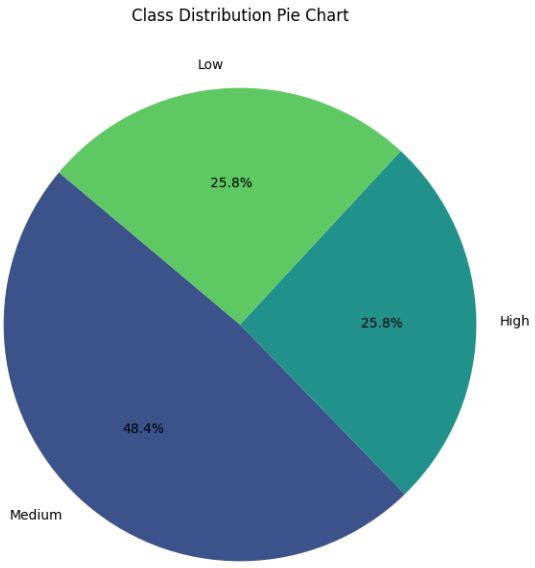

Figure 2. Class Distribution Pie Chart by Engagement Level

Data Analysis Methods

Feature Selection and Pre-processing [16]–[18]:

- **Removing Irrelevant Features:** The `PlayerID` column is removed as it does not contribute to predictive modelling.
- **Encoding Categorical Data:** Categorical variables such as `Gender`, `Location`, `GameGenre`, `GameDifficulty`, and `EngagementLevel` are encoded using numerical values. For example:

$$Gender = \begin{matrix} 0 & if\ Male \\ 1 & if\ Female \end{matrix}$$

$$Location = \begin{matrix} 0 & if\ USA \\ 1 & if\ Europe \\ 2 & if\ Asia \\ 3 & if\ Other \end{matrix} \quad (1)$$

- **Scaling Features:** All numerical features are scaled to have a mean of 0 and variance of 1 to standardize the data:

$$X_{scaled} = \frac{X - \mu}{\sigma} \quad (2)$$

Where μ is the mean and σ is the standard deviation of the feature.

Table 2. Feature Descriptions after Pre-processing

Column Name	Type	Description
`Age`	Numerical	Age of the player
`Gender`	Numerical	Gender of the player (0 = Male, 1 = Female)
`Location`	Numerical	Geographic location (0 = USA, 1 = Europe, 2 = Asia, 3 = Other)
`GameGenre`	Numerical	Genre of the game (0 = Sports, 1 = Action, 2 = Strategy, 3 = Simulation, 4 = RPG)
`PlayTimeHours`	Numerical	Average hours spent playing per session
`InGamePurchases`	Numerical	Indicates if the player makes in-game purchases (0 = No, 1 = Yes)
`GameDifficulty`	Numerical	Difficulty level of the game (0 = Easy, 1 = Medium, 2 = Hard)
`SessionsPerWeek`	Numerical	Number of gaming sessions per week
`AvgSessionDurationMinutes`	Numerical	Average duration of each gaming session in minutes
`PlayerLevel`	Numerical	Current level of the player in the game
`AchievementsUnlocked`	Numerical	Number of achievements unlocked by the player
`EngagementLevel`	Numerical	Categorized engagement level (0 = Low, 1 = Medium, 2 = High)

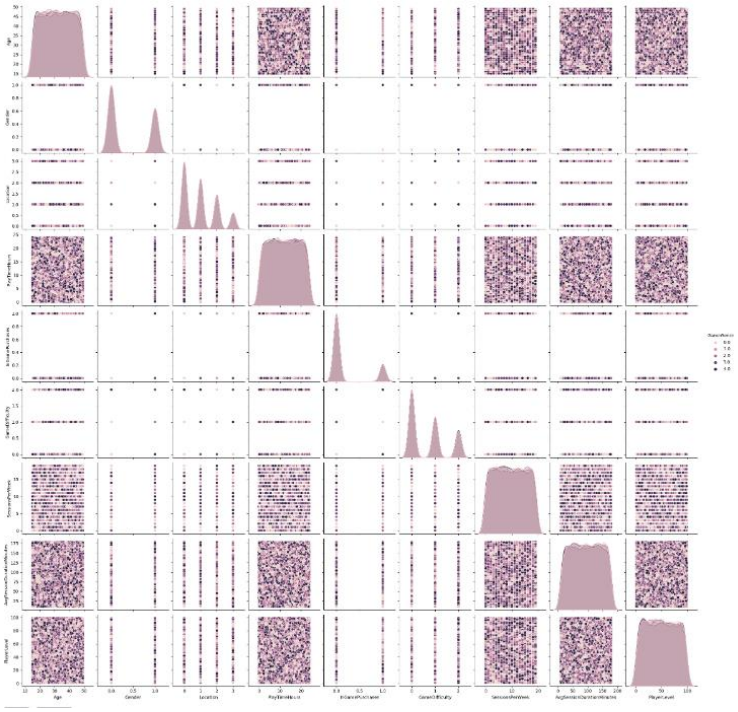

Figure 3. Scatter plots

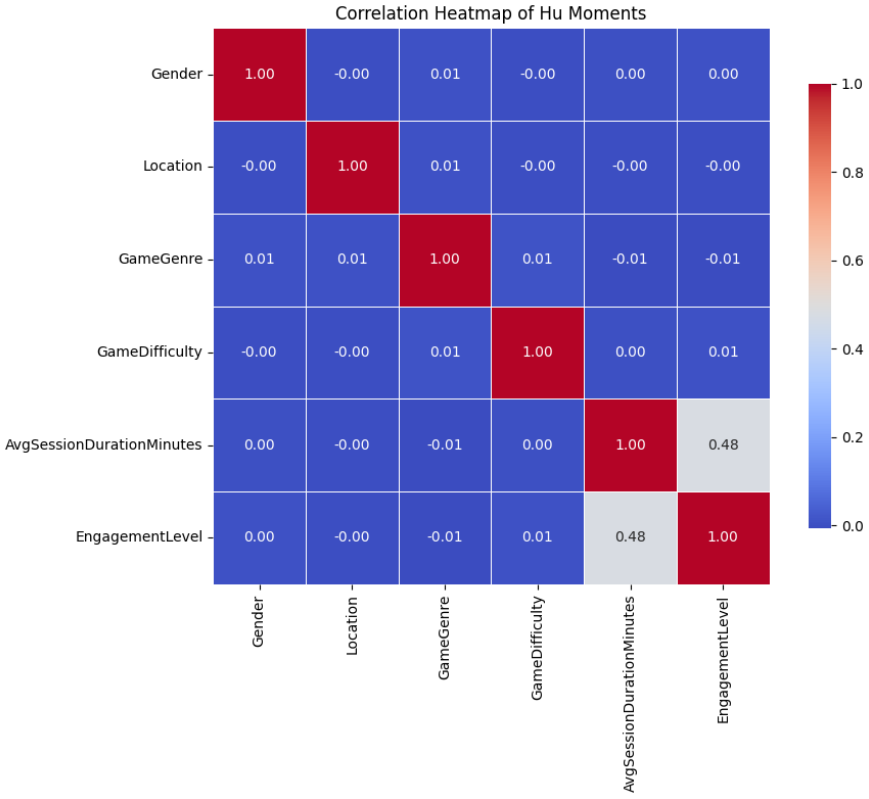

Figure 4. Correlation Heatmap of Hu Moments

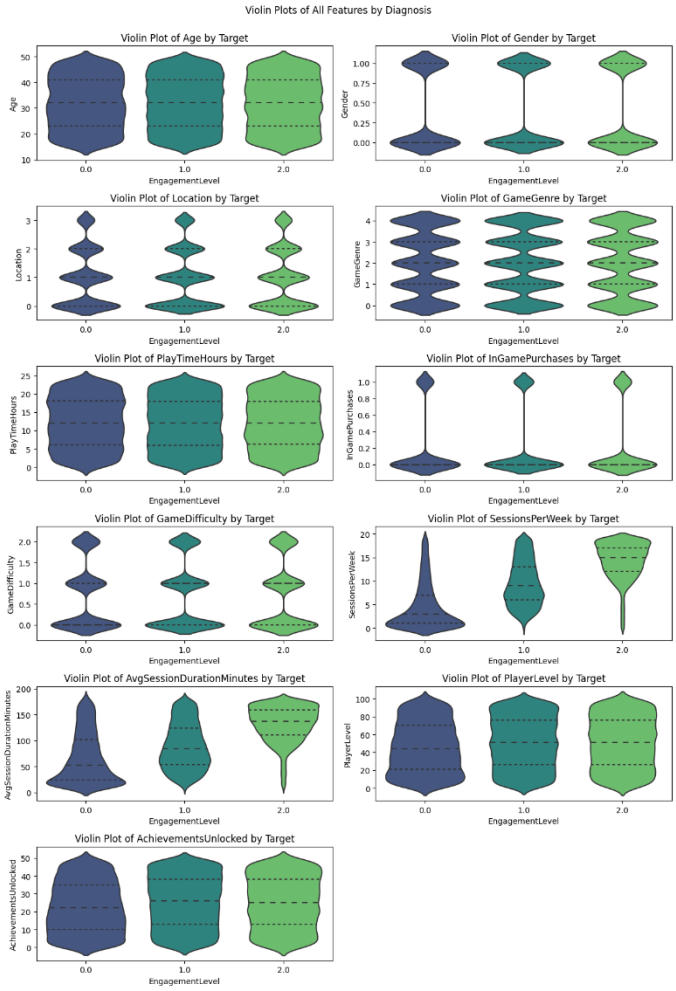

Figure 5. Violin Plots of All Features

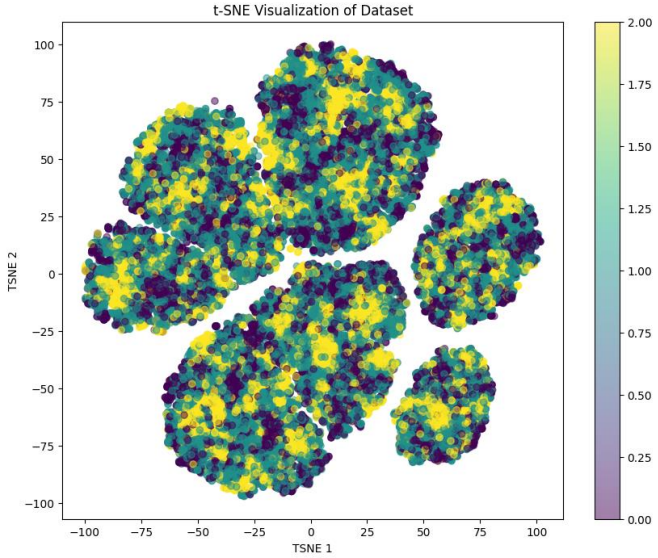

Figure 6. 3D t-SNE Visualization of Dataset

The visualizations provide a comprehensive view of the dataset's structure and feature relationships. **Figure 4** reveals the interdependencies between features. **Figure 6** visualization highlights the clustering of data points based on high-dimensional features. Lastly, the **Figure 5** depict the distribution of each feature, offering a clear view of data spread and outliers.

Model Implementation:

- Data Splitting: The dataset is split into training (80%) and testing (20%) sets to facilitate model validation and testing.

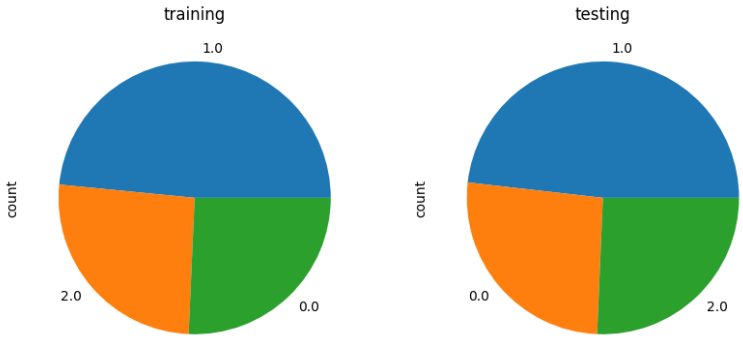

Figure 7. Splitting Dataset 20 % testing, 80% training

- **Gaussian Naive Bayes Model:** The Gaussian Naive Bayes algorithm is used due to its efficiency in handling both categorical and continuous data [12], [19]. The algorithm calculates the probability of each class label based on the input features using the formula:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{3}$$

- **Cross-validation:** A 5-fold cross-validation technique is employed to ensure the robustness of the model. The dataset is split into 5 subsets, and the model is trained and validated 5 times, each time using a different subset as the validation set and the remaining subsets as the training set [20].

$$\mathrm{CV}_{(K)} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{Error}_i \tag{4}$$

Performance Evaluation

- **Accuracy:** The ratio of correctly predicted instances to the total instances:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{5}$$

- **Precision:** The ratio of true positive predictions to the total positive predictions:

$$Precision = \frac{TP}{(TP + FP)} \tag{6}$$

- **Recall:** The ratio of true positive predictions to the total actual positives:

$$Recall = \frac{TP}{(TP + FN)} \tag{7}$$

- **F1-Score:** The harmonic mean of precision and recall:

$$F - measure = \frac{2(presisi \times recall)}{(presisi + recall)} \quad (5)$$

The above formulas explain:
True Positive (TP): The number of cases correctly predicted as positive by the model.
True Negative (TN): The number of cases correctly predicted as negative by the model.
False Positive (FP): The number of cases incorrectly predicted as positive by the model.
False Negative (FN): The number of cases incorrectly predicted as negative by the model.

These metrics provided a comprehensive understanding of the model's performance, highlighting its strengths and areas of improvement [21]–[23].

3. Results and Discussion

Results

The data processing involved several crucial steps, including feature selection, encoding categorical variables, and scaling numerical features to prepare the dataset for modelling. Initially, the `PlayerID` column was removed due to its irrelevance in predictive modelling. Categorical variables such as `Gender`, `Location`, `GameGenre`, `GameDifficulty`, and `EngagementLevel` were encoded numerically to facilitate the application of the Gaussian Naive Bayes algorithm. Subsequently, all numerical features were scaled to ensure a mean of 0 and a variance of 1, standardizing the data for effective model training and evaluation.

The Gaussian Naive Bayes model was trained and evaluated using a 5-fold cross-validation technique to ensure robustness and reliability. The performance metrics, including accuracy, precision, recall, and F1-score, were calculated for each fold. The results are summarized in the **Table 3**, with performance metrics expressed as percentages:

Table 3. Performance Metrics Across 5-Fold Cross-Validation for the GNB

K-n	Metrics			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	84.05%	85.32%	84.05%	83.76%
K-2	84.51%	85.98%	84.51%	84.18%
K-3	84.25%	85.57%	84.25%	83.95%
K-4	84.31%	85.69%	84.31%	84.02%
K-5	84.22%	85.40%	84.22%	83.97%
$\sum$ ***Avg***	84.27%	85.59%	84.27%	83.98%

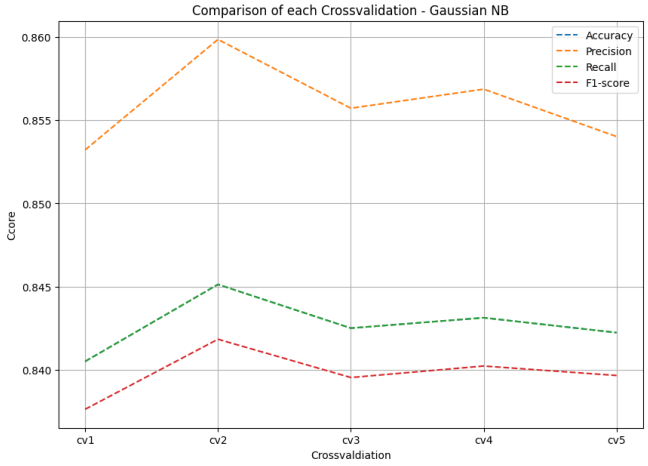

Figure 8. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the GNB

To provide a visual representation of the model's performance, the accuracy, precision, recall, and F1-score across the 5 folds are depicted in the following **Figure 8**. Additionally, a confusion matrix illustrating the model's performance on the test set is presented to show the distribution of true positives, true negatives, false positives, and false negatives in **Figure 9**.

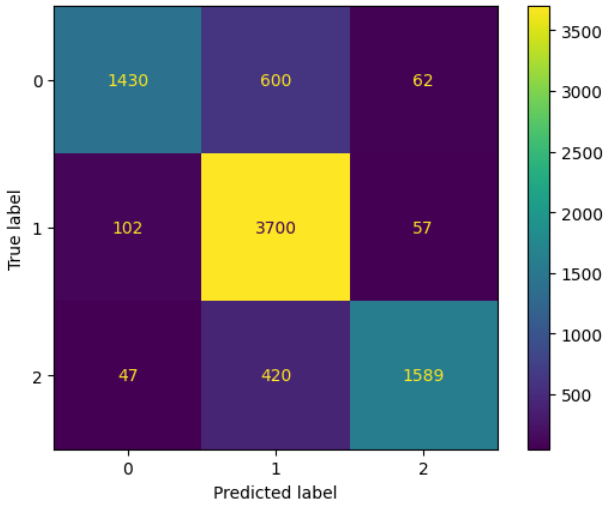

Figure 9. Visualisation of Confusion Matrix

The interpretation of these results indicates that the Gaussian Naive Bayes model performs consistently across the different folds, with accuracy and recall averaging approximately 84.27%. Precision and F1-score also show consistency, with averages around 85.59% and 83.98%, respectively. These metrics demonstrate the model's ability to predict player engagement levels with a reasonable degree of accuracy and reliability.

Discussion

The results from the Gaussian Naive Bayes model reveal several significant findings regarding player engagement prediction. Firstly, the model's consistent performance across all folds of cross-validation suggests that it generalizes well to different subsets of the data. This consistency is crucial for practical applications where the model would need to perform reliably across various gaming environments and player demographics. The precision and recall metrics, both averaging above 84%, indicate that the model is effective at identifying true positives while minimizing false positives and false negatives. This balance is particularly important in scenarios where accurately identifying highly engaged players can lead to targeted marketing strategies and personalized gaming experiences, thereby enhancing overall player satisfaction and retention.

These findings are consistent with previous research in the field of game analytics, which highlights the effectiveness of machine learning models in predicting player behaviour based on demographic and in-game metrics. The practical implications of this research are significant for game developers and marketers. By leveraging such predictive models, they can design more engaging and immersive gaming experiences tailored to the preferences and behaviours of different player segments. However, there are several limitations to this study that must be acknowledged. The dataset used, while comprehensive, may not encompass all factors influencing player engagement. External factors such as social influences, gaming trends, and individual player motivations were not considered in this model. Additionally, the use of a single dataset limits the generalizability of the findings across different gaming genres and platforms.

Future research should focus on incorporating a broader range of features, including psychological and social factors, to enhance the predictive power of the models. Moreover, exploring other machine learning algorithms and ensemble methods could provide further improvements in accuracy and reliability. Expanding the dataset to include a more diverse array of games and player demographics would also help in developing more generalizable models. Overall, this research contributes valuable insights into the predictive modelling of player engagement in online gaming, offering a foundation for future studies and practical applications in game development and marketing.

4. Conclusion

This study investigated the prediction of player engagement levels in online gaming using a Gaussian Naive Bayes model, applied to a comprehensive dataset sourced from Kaggle. The results demonstrated that the model could reliably predict engagement levels, with consistent performance metrics across 5-fold cross-validation, including an average accuracy of 84.27%, precision of 85.59%, recall of 84.27%, and F1-score of 83.98%. These findings validate the hypothesis that demographic and in-game metrics can effectively predict player engagement, with features such as session frequency, in-game purchases, and player level being particularly influential.

The research contributes to the field of game analytics by providing a predictive model that can assist game developers and marketers in enhancing user retention and satisfaction. By identifying key factors that influence engagement, the study offers practical insights for designing more personalized and engaging gaming experiences. However, the study's limitations, such as the use of a single dataset and the exclusion of social and psychological factors, suggest avenues for further research. Future studies should incorporate a broader range of features and explore additional machine learning algorithms to improve predictive accuracy and generalizability across different gaming contexts. Expanding the dataset to include diverse gaming genres and player demographics would also provide more comprehensive insights, ultimately contributing to more effective engagement strategies in the gaming industry.

References:

[1] K. V Swamy, "Skin Disease Classification using Image Preprocessing and Machine Learning," *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, IATMSI 2024*. 2024, doi: 10.1109/IATMSI60426.2024.10502445.

[2] A. A. Kolchev, "Classification of benign and malignant solid breast lesions on the ultrasound images based on the textural features: the importance of the perifocal lesion area," *Comput. Opt.*, vol. 48, no. 1, pp. 157–165, 2024, doi: 10.18287/2412-6179-CO-1244.

[3] A. D. Purwanto, "Decision Tree and Random Forest Classification Algorithms for Mangrove Forest Mapping in Sembilang National Park, Indonesia," *Remote Sens.*, vol. 15, no. 1, 2023, doi: 10.3390/rs15010016.

[4] Y. Shi, "The Iris Cassification Based on Gaussian Naive Bayes Agorithm," *ACM International Conference Proceeding Series*. pp. 732–736, 2022, doi: 10.1145/3573428.3573559.

[5] M. V Anand, "Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer," *Mob. Inf. Syst.*, vol. 2022, 2022, doi: 10.1155/2022/2436946.

[6] A. J. Meerja, "Gaussian naïve bayes based intrusion detection system," *Adv. Intell. Syst. Comput.*, vol. 1182, pp. 150–156, 2021, doi: 10.1007/978-3-030-49345-5_16.

[7] R. Setiawan and H. Oumarou, "Classification of Rice Grain Varieties Using Ensemble Learning and Image Analysis Techniques," *Indones. J. Data ...*, 2024.

[8] I. P. A. Pratama, E. S. J. Atmadji, and ..., "Evaluating the Performance of Voting Classifier in Multiclass Classification of Dry Bean Varieties," *Indones. J. ...*, 2024.

[9] M. D. Genemo, "Federated Learning for Bronchus Cancer Detection Using Tiny Machine Learning Edge Devices," *Indones. J. Data Sci.*, 2024.

[10] S. Naiem, "Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in the Detection of DDOS in Cloud Computing," *IEEE Access*, vol. 11, pp. 124597–124608, 2023, doi: 10.1109/ACCESS.2023.3328951.

[11] K. Sen, "Heart Disease Prediction Using a Soft Voting Ensemble of Gradient Boosting Models, RandomForest, and Gaussian Naive Bayes," *2023 4th Int. Conf. Emerg. Technol. INCET 2023*, 2023, doi: 10.1109/INCET57972.2023.10170399.

[12] I. Sulistiani, "Breast Cancer Prediction Using Random Forest and Gaussian Naïve Bayes Algorithms," *2022 1st Int. Conf. Inf. Syst. Inf. Technol. ICISIT 2022*, pp. 170–175, 2022, doi: 10.1109/ICISIT54091.2022.9872808.

[13] R. A. Azdy, R. F. Syam, E. Faizal, and ..., "Performance Evaluation of Bagging Meta-Estimator in Lung Disease Detection: A Case Study on Imbalanced Dataset," *Int. J. ...*, 2023.

[14] T. E. Tarigan, E. Susanti, M. I. Siami, I. Arfiani, and ..., "Performance Metrics of AdaBoost and Random Forest in Multi-Class Eye Disease Identification: An Imbalanced Dataset Approach," *... Artif. Intell. ...*, 2023.

[15] A. Naswin and A. P. Wibowo, "Performance Analysis of the Decision Tree Classification Algorithm on the Pneumonia Dataset," *... Artif. Intell. Med. ...*, 2023.

[16] I. Alwiah, U. Zaky, and A. W. Murdiyanto, "Assessing the Predictive Power of Logistic Regression on Liver Disease Prevalence in the Indian Context," *... J. Data Sci.*, 2024.

[17] F. T. Admojo and N. Rismayanti, "Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions," *Indones. J. Data ...*, 2024.

[18] I. G. I. Sudipa, R. A. Azdy, I. Arfiani, and ..., "Leveraging K-Nearest Neighbors for Enhanced Fruit Classification and Quality Assessment," *Indones. J. ...*, 2024.

[19] A. Krysovatyy, "Classification Method of Fictitious Enterprises Based on Gaussian Naive Bayes," *Int. Sci. Tech. Conf. Comput. Sci. Inf. Technol.*, vol. 2, pp. 224–227, 2021, doi: 10.1109/CSIT52700.2021.9648584.

[20] S. Ortiz-Toquero, "Classification of Keratoconus Based on Anterior Corneal High-order Aberrations: A Cross-validation Study," *Optom. Vis. Sci.*, vol. 97, no. 3, pp. 169–177, 2020, doi: 10.1097/OPX.0000000000001489.

[21] H. Azis, P. Purnawansyah, N. Nirwana, and ..., "The Support Vector Regression Method Performance Analysis in Predicting National Staple Commodity Prices," *Ilk. J. ...*, 2023.

[22] H. Azis, L. Syafie, F. Fattah, and ..., "Unveiling Algorithm Classification Excellence: Exploring Calendula and Coreopsis Flower Datasets with Varied Segmentation Techniques," *2024 18th Int. ...*, 2024.

[23] H. Azis, F. Fattah, and P. Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020.