



ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

ACCA6011, Ryan Michael Hopkins

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.]

When you're ready to add your content, just click here and start typing.]

Ryan Hopkins (2221956)

[Email address]

Table of Contents

1. Introduction	2
2. Problem Statement & Literature Review	2
3. Methodology	2
3.1 Dataset Description.....	2
3.2 Data Preprocessing	2
3.3 Decision Tree	2
3.4 Support Vector Machine	2
3.5 Hyperparameter Tuning	3
3.6 Clustering with K-Means	3
3.7 Evaluation Metrics	3
4. Experiments & Results	3
4.1 Exploratory Data Analysis	3
4.2 Decision Tree Results	5
4.3 SVM Results	6
4.4 Comparison of Supervised Models	5
4.5 Clustering Results	7
5. Discussion	8
5.1 Interpretation of Supervised Learning Results.....	5
5.2 Interpretation of Clustering Results	5
7. Conclusion.....	6

1. Problem Statement & Literature Review

The rapid growth of the online gaming industry has led to an increasing need for data-driven approaches to understand player behaviour and improve engagement, this growth has roughly doubled in the last 10 years, with the world economic forum stating that revenue has doubled from an estimated \$100 billion to \$298-\$500 billion [1] [2]. Player engagement is a key factor influencing retention, in-game spending, and overall game success.

This report aims to address this challenge by developing predictive models to classify player engagement levels High, Medium, Low using behavioural and demographic data from the Predict Online Gaming Behaviour dataset available on Kaggle. Specifically, Decision Tree and Support Vector Machine (SVM) classifiers are applied to evaluate their effectiveness in predicting engagement levels.

In addition to supervised learning, this report also employs clustering techniques to uncover hidden patterns and group players based on similarities in their behaviour. The combination of classification and clustering approaches enables both prediction and deeper understanding of player segments.

The objective is not only to assess model performance but also to generate actionable insights that can support game developers in improving player retention, personalisation, and overall user experience.

2. Methodology

2.1 Dataset Description

The dataset being used has been created by Rabie El Kharoua and hosted on Kaggle, it hosts data which provides demographics related to player behaviour in online gaming environments. This dataset includes features such as age, gender, and location, alongside behavioural variables including playtime, session frequency, and in-game purchases. The target variable represents player engagement level, which is categorised into three classes High, Medium, and Low.

2.2 Data Preprocessing

Before development, the dataset was preprocessed to improve data quality and ensure compatibility with selected models. Missing values were examined and handled to reduce the risk of biased model performance. Numerical missing values were replaced using the mean of the respective feature, while categorical missing values were replaced using the most frequent category.

2.3 Decision Tree

A Decision Tree classifier was implemented as one of the supervised learning models for predicting player engagement levels. the algorithm selects the feature that produces the greatest reduction in impurity, commonly measured using the Gini Index or entropy. This process continues until a stopping condition is reached [3], resulting in a tree-like structure of decision rules. The decision tree was trained by using a random state which was equal to 42, all other settings were left default.

2.4 Support Vector Machine

A support vector machine is a supervised learning algorithm, it works by finding an optimal hyperplane that will maximise the distance between classes. The SVM used a linear kernel, this kernel was chosen as it lowers processing power and loads the processed data faster. If the dataset was larger and more complex, RBF would have been chosen as the RBF sorts complex data better, however with the trade-off with efficiency.

2.5 Hyperparameter Tuning

To improve performance, hyperparameter tuning was conducted using GridSearchCV, this is from the Scikit-learn library. GridSearchCV tests multiple combinations of predefined parameter values and identifies the combination that produces the best performance. This process helps optimise model settings while reducing the likelihood of overfitting as described by [geeks4geeks \[4\]](#).

2.6 Clustering with K-Means

K-Means clustering was applied as an unsupervised learning technique. K-Means partitions the dataset into a predefined number of clusters by assigning each observation to the nearest cluster. The optimal number of clusters was determined using the elbow method and silhouette score analysis. [5] Behavioural variables such as playtime, session frequency, and in-game purchases were selected for clustering because these features directly reflect player activity patterns. This approach allowed meaningful player groups to be identified based on behavioural similarity. See figure ___ for the elbow method and silhouette score

2.7 Evaluation Metrics

To evaluate how well the models performed, a range of metrics was used for both classification and clustering. For the classification models, accuracy, precision, recall, F1-score, and AUC-ROC were used. Accuracy shows how often the model made correct predictions overall. Precision looks at how many of the predicted results were correct, while recall focuses on how well the model identified all relevant cases. The F1-score provides a balance between precision and recall, making it useful when the data is not evenly distributed across classes, AUC-ROC was also used to assess how well the model can distinguish between different engagement levels.

For the clustering analysis, silhouette score and inertia were used to measure the quality of the clusters. The silhouette score indicates how well each data point fits within its assigned group compared to other groups, while inertia measures how closely grouped the data points are within each cluster. Together, these metrics give a clearer picture of how effective the clustering approach was in identifying meaningful player segments.

3. Experiments & Results

3.1 Exploratory Data Analysis

Initial data exploration was conducted to better understand the structure and characteristics of the dataset. The distribution of the target variable showed that player engagement levels was slightly imbalanced, with 25.7847% of players classified as High, 48.3973% as Medium, and 25.8180% as Low. This distribution is important as it could influence the model performance, particularly for metrics such as precision and recall.

Analysis of feature distributions revealed that behavioural variables such as playtime and session frequency were right-skewed, indicating that a small proportion of players contributed disproportionately high activity levels. In contrast, demographic features such as player level were more evenly distributed across the dataset. See figures 1 and 2 for the box plot for age

Correlation analysis showed that there were no strong linear relationships between the features and player engagement level. Both demographic variables and behavioural features such as playtime and in-game purchases displayed weak correlations with the target variable. see figure 3 for the correlation heatmap

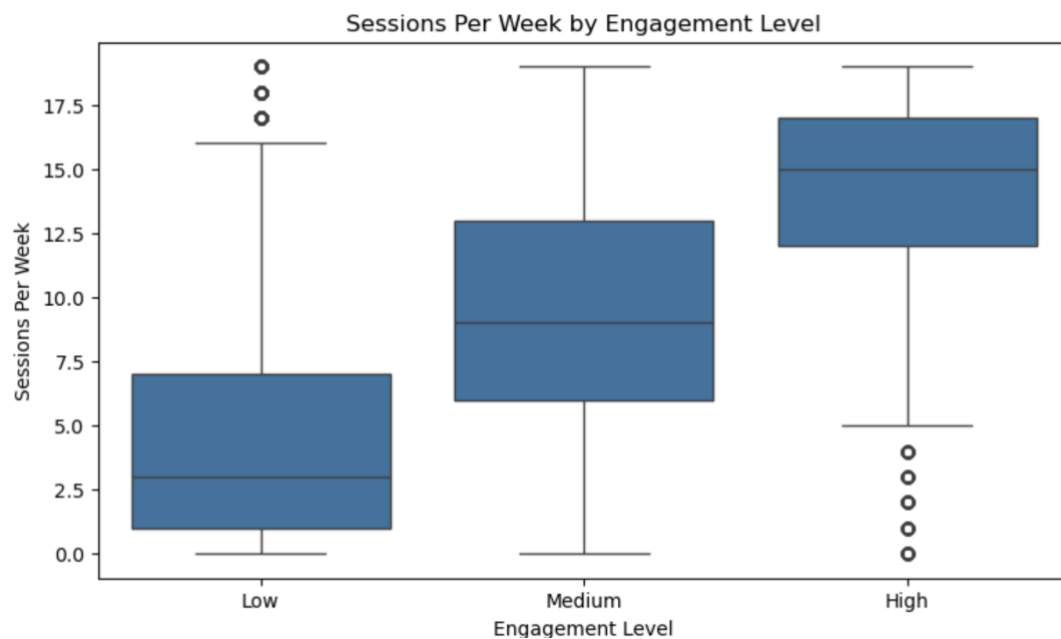


Figure 1 box plot Sessions per week by engagement level

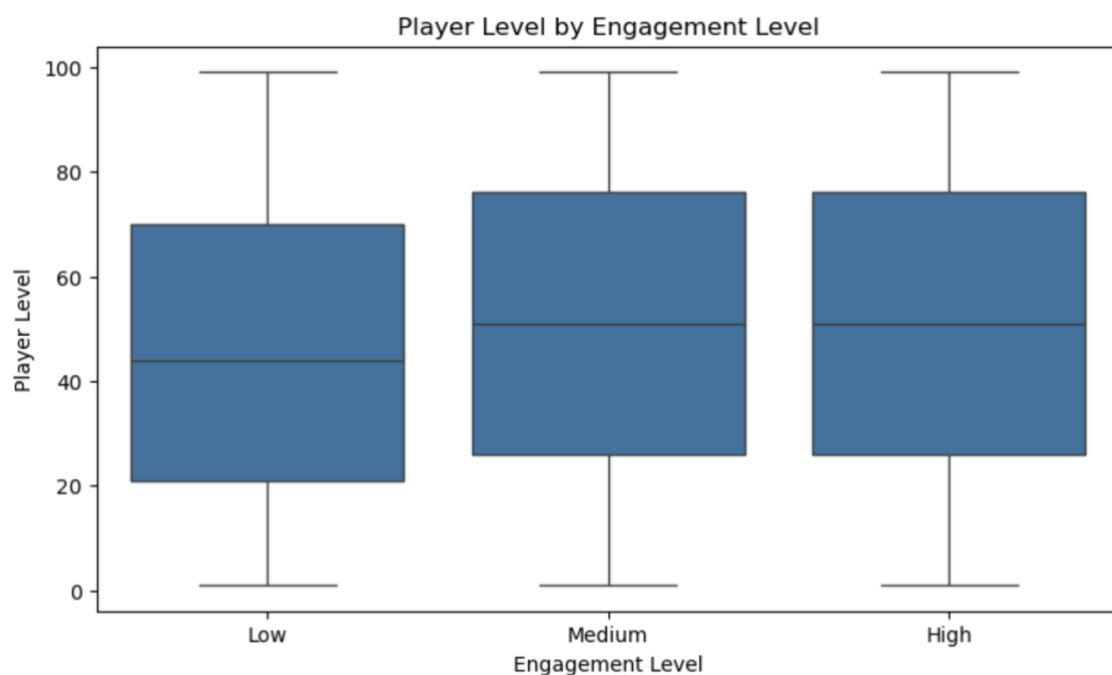


Figure 2 box plot player level by engagement level

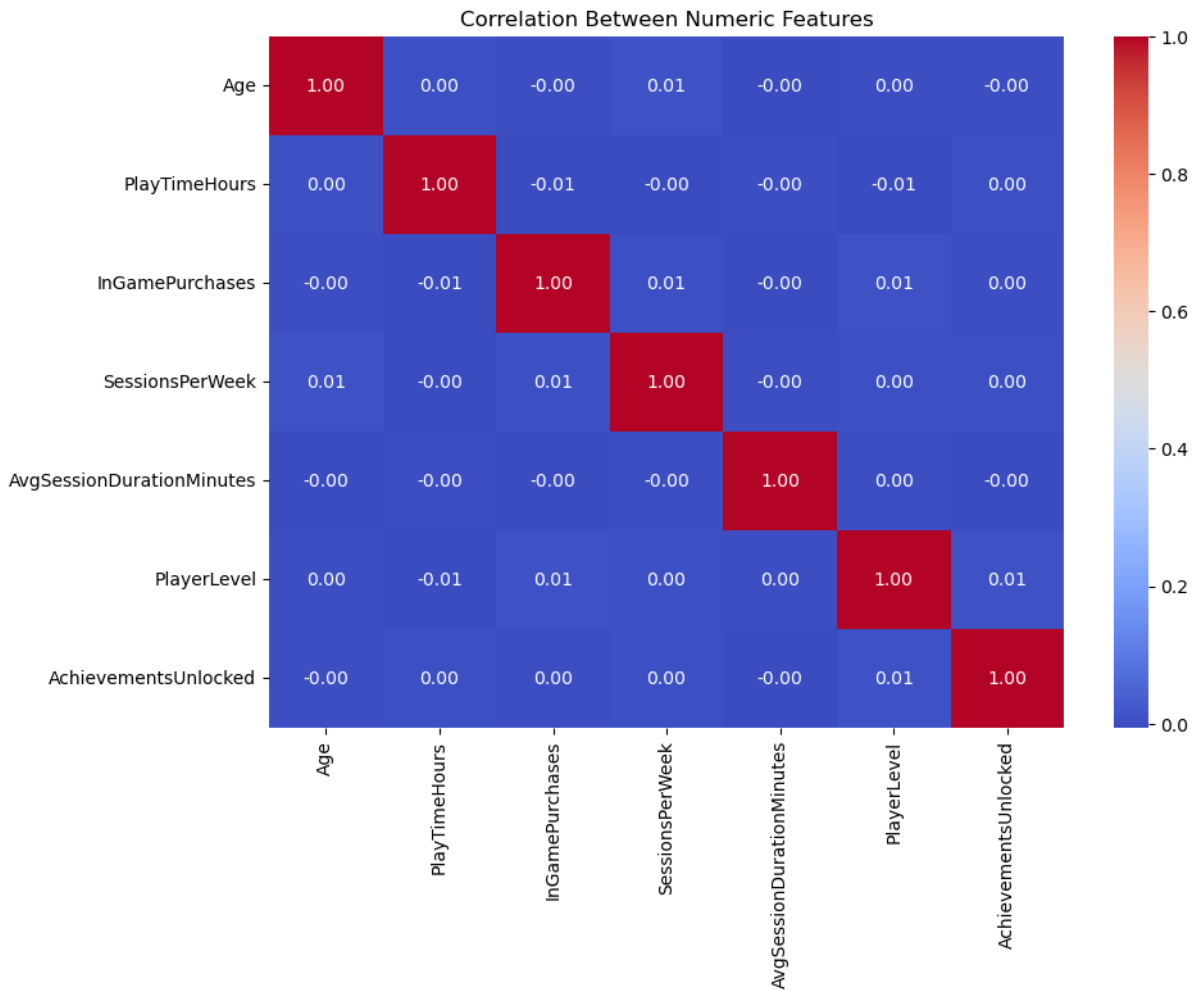


Figure 3 correlation heat map by numeric figures.

3.2 Decision Tree Results

The Decision Tree classifier was evaluated before and after hyperparameter tuning to assess performance improvements. The baseline model achieved an accuracy of 82.62%, which improved to 90.57% after tuning. See figure 4 for the results of decision tree, and figure 5 for the depth of the tree.

The optimal hyperparameters identified through GridSearchCV included a maximum depth of 10, minimum samples split of 2, and minimum samples per leaf of 5. These adjustments helped reduce overfitting and improved the model's ability to generalise to unseen data.

Evaluation metrics showed improvements across precision, recall, and F1-score after tuning. The confusion matrix indicated that the model performed well in identifying medium engagement players. However misidentified and struggled when trying to identify high engagement and low engagement users.

```

=== Decision Tree Results ===
Accuracy: 0.8262
Precision: 0.8268
Recall: 0.8262
F1 Score: 0.8264
AUC-ROC: 0.8637

```

Figure 3 Decision tree results before hyperparameters in place.

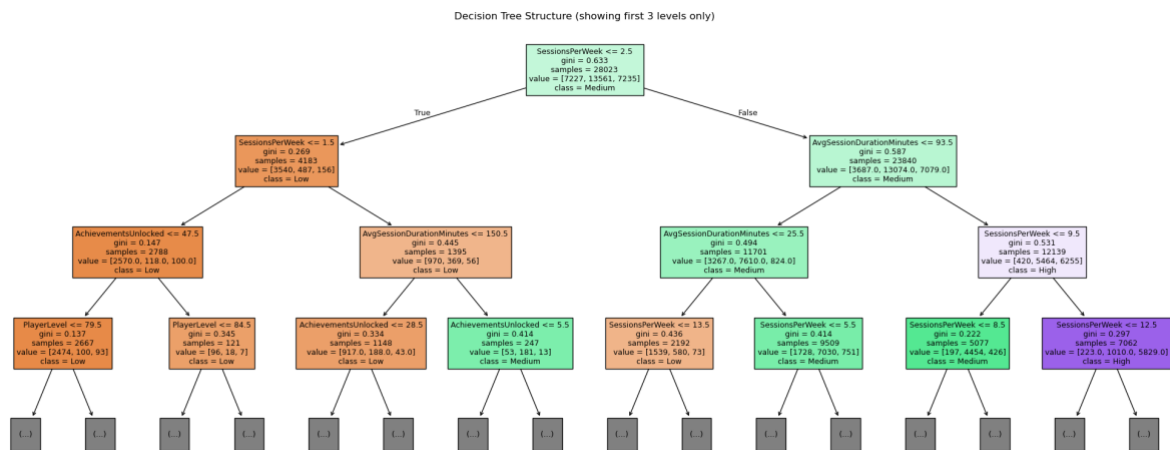


Figure 5: Visualised decision tree only top 3 rows are visible

3.3 SVM Results

The Support Vector Machine model was also evaluated using both baseline and tuned configurations. The baseline model achieved an accuracy of 82.49%, which increased to 89.62% after hyperparameter tuning.

Compared to the Decision Tree, the SVM model demonstrated lower performance across precision, recall, and F1-score. Overall, the SVM model provided less consistent predictions across all engagement levels. However, the reason for this could be due to the linear Kernel choice over the RBF kernel, the linear sacrificed performance for efficiency which in hindsight reduced the overall precision and accuracy scores. See figure 6 for SVM Results and figure 7 for the SVM confusion matrix a way to visualise SVM data

```

=== SVM Results ===
Accuracy: 0.8249
Precision: 0.8254
Recall: 0.8249
F1 Score: 0.8239
AUC-ROC: 0.9113
  
```

Figure 6: SVM Results before hyperparameters

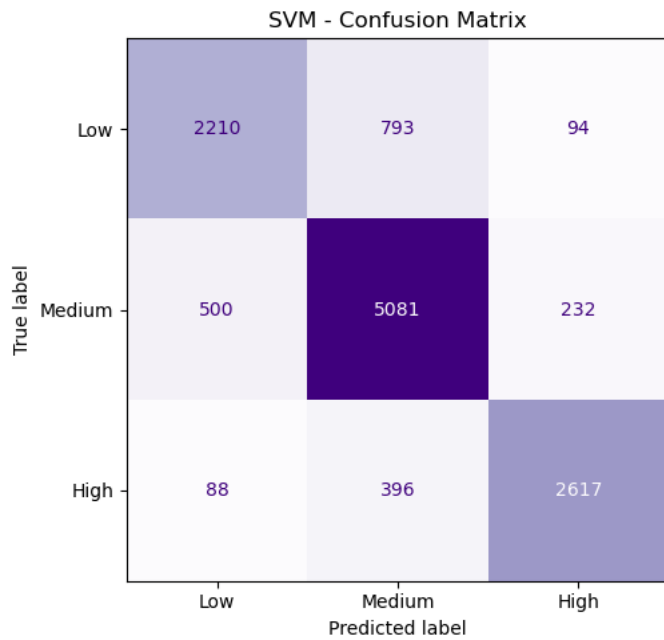


Figure 7: SVM Confusion Matrix

3.4 Comparison of figures

Model	Accuracy	F1 Score	AUC-ROC
Decision Tree (Default)	0.8262	0.8264	0.8637
Decision Tree (Tuned)	0.9059	0.9057	0.9333
SVM (Default)	0.8249	0.8239	0.9113
SVM (Tuned)	0.8962	0.8956	0.9441

Key observations

- Tuning improved both models significantly, Decision Tree by ~8%, SVM by ~7%
- Tuned Decision Tree has the highest accuracy and F1 (90.59% / 90.57%)
- Tuned SVM has the highest AUC-ROC (94.41%) meaning it's best at distinguishing classes
- Default models performed similarly to each other before tuning
- This shows hyperparameter tuning was highly effective for both classifier

Principal Component Analysis (PCA) was used to reduce the dimensionality of the dataset and visualise the player clusters in two dimensions. This allowed for a clearer understanding of how the clusters are distributed. The visualisation displays the overlap between clusters and furthers the point of a low silhouette score. See figure 8 for the PCA graph

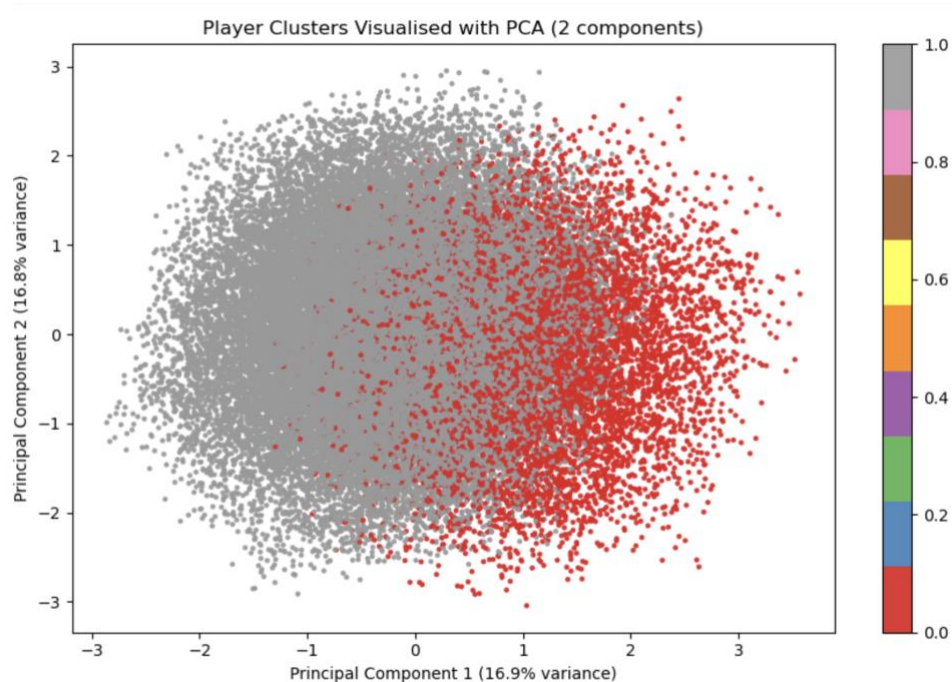


Figure 8: Player clusters visualised with PCA

4. Discussion

These results demonstrate the unpredictable of these models where the expectations can be subverted, the expectation is that SVM will be more accurate using the evaluation metrics however the tuned decision tree had the highest accuracy at 90.57% after tuning, both models performed very similarly and were within 1% accuracy. SVM had a higher score with the AUC-ROC at 94.41% over 93.33% having a higher score means it can deal with distinguishing between classes without problem. The lower than expected performance of SVM can be attributed to the choice of kernel.

The clustering results identified two main groups of players based on their patterns. Although, the low silhouette score of 23.66% indicates that these clusters are not strongly separated. This suggests that the clustering algorithm did not successfully separate players based on engagement level. Instead, the clusters appear to group players in a way that does not strongly reflect the target variable. This finding is consistent with the relatively low silhouette score and indicating weak cluster separation. The low separation between clusters could be due to the nature of the dataset where players behaviours overlap a significant proportion.

The supervised learning models used have offered advantages and disadvantages, The Decision Tree model was easy to interpret and provided clear decision rules, making it useful for understanding how different features influenced engagement predictions. Although Decision Trees

are generally prone to overfitting, this was not a significant issue in this case, as the performance on the test data remained consistent with the training results after hyperparameter tuning

In contrast, the Support Vector Machine model was more effective at finding complex patterns within the data, resulting in stronger performance across evaluation metrics such as 94.41% for AUC-ROC. the SVM model required careful tuning of parameters such as the kernel, which unfortunately reduced the overall effectiveness of the SVM model.

The results of the implemented models provide a few key insights that can support decision making for the development studio. The supervised learning models demonstrate player engagement can be predicted with a decent level of accuracy allowing the studio strategies on how to keep those players on for longer, perhaps by increased rewards, promotions or content updates, with the main goal of increased player retention.

However, the clustering results indicate that player segments are not clearly defined, as shown by the low silhouette score and overlapping cluster characteristics. This suggests that player behaviour exists on a spectrum rather than in distinct groups meaning that broad segmentation strategies may be less effective. Instead, the studio may benefit from more personalised approach to player engagement.

Overall, these findings support a data-driven approach to game design and player retention. By combining predictive modelling with behavioural insights, the studio can better understand its player base improve user experience, and develop targeted strategies to maximise long-term engagement and retention

Bibliography

- [1] O. Wallach, "The history of the gaming industry in one chart," 27 November 2020. [Online]. Available: <https://www.weforum.org/stories/2020/11/gaming-games-consels-xbox-play-station-fun/>. [Accessed 15 April 2026].
- [2] S. Read, "Gaming boomed in lockdown and market value will reach \$320bn," World Economic Forum, 28 July 2022. [Online]. Available: <https://www.weforum.org/stories/2022/07/gaming-pandemic-lockdowns-pwc-growth/>. [Accessed 15 April 2026].
- [3] IBM, "What Is a Decision Tree?," 2 November 2021. [Online]. Available: <https://www.ibm.com/think/topics/decision-trees>. [Accessed 22 April 2026].
- [4] GeeksforGeeks, "Performing Feature Selection with gridsearchcv in Sklearn," 19 July 2024. [Online]. Available: <https://www.geeksforgeeks.org/machine-learning/performing-feature-selection-with-gridsearchcv-in-sklearn/>. [Accessed 22 April 2026].

- [5] A. singh, "K-Means Clustering: A Deep Dive into Unsupervised Learning," 25 July 2025. [Online]. Available: <https://medium.com/@abhaysingh71711/k-means-clustering-a-deep-dive-into-unsupervised-learning-81213f56cfc9>. [Accessed 22 April 2026].