# Machine Learning Approaches to Forecasting Surf Competitions.

This project investigates the use of machine learning to predict the winners of surf competition heats. Three machine learning models, Artificial Neural Networks (ANNs), Random Forests, and XGBoost, are employed to analyse historical competition data, surfer statistics, and wave conditions. The project aims to determine the feasibility and effectiveness of machine learning in forecasting surf competition outcomes.

The project uses a dataset consisting of more than 19,000 Championship Tour heats conducted by the World Surfing League from 2012 to 2024. The dataset contains three primary sources of information:

- **Historical Competition Data:** This segment encompasses details such as outcomes of heats, locations of competitions, among other relevant metrics.
- **Surfer Statistics:** This category provides detailed information regarding the competitors. It includes metrics such as the surfers' rankings, stances, average heat totals, and historical performance at various competition venues.
- **Wave Conditions:** The dataset also includes quantifiable metrics related to the surfing environment, covering aspects like wave height, wind direction, and the type of waves encountered.

Before training the models, the data underwent a preprocessing stage involving feature scaling and encoding categorical features. For each model a training-validation split was employed and hyperparameter tuning was performed to optimise the performance on the validation set.

Each model (ANN, Random Forest, XGBoost) was trained on the training set and evaluated on the validation set using the following classification metrics:

- **Accuracy:** Measures the overall proportion of correctly predicted heat winners.

- **Precision:** Indicates the ratio of true positives (correctly predicted winners) to all positive predictions by the model.

- **Recall:** Represents the ratio of true positives (correctly predicted winners) to all actual winners in the validation set.

- **F1-Score:** A harmonic mean of precision and recall, providing a balanced view of model performance.

| Metric | | Precision | Recall | F1-Score | OA Accuracy | Support |
|---|---|---|---|---|---|---|
| ANN | 0 | 0.72 | 0.74 | 0.73 | 0.70 | 3135 |
| | 1 | 0.66 | 0.65 | 0.66 | 0.70 | 2396 |
| Random Forest | 0 | 0.71 | 0.74 | 0.73 | 0.70 | 3135 |
| | 1 | 0.69 | 0.64 | 0.67 | 0.70 | 2396 |
| XGBoost | 0 | 0.73 | 0.75 | 0.74 | 0.71 | 3135 |
| | 1 | 0.68 | 0.66 | 0.68 | 0.71 | 2396 |

Fig. 1 compares the accuracy of three machine learning models in predicting surf competition outcomes. The ANN model achieved 0.70 overall accuracy, with an F1-score of 0.73 for class 0 (surfer losing the heat) and 0.66 for class 1 (surfer winning the heat). The Random Forest model also reached 0.70 accuracy, with F1-scores of 0.73 for class 0 and 0.67 for class 1. XGBoost performed slightly better, with 0.71 accuracy and F1-scores of 0.74 for class 0 and 0.68 for class 1.
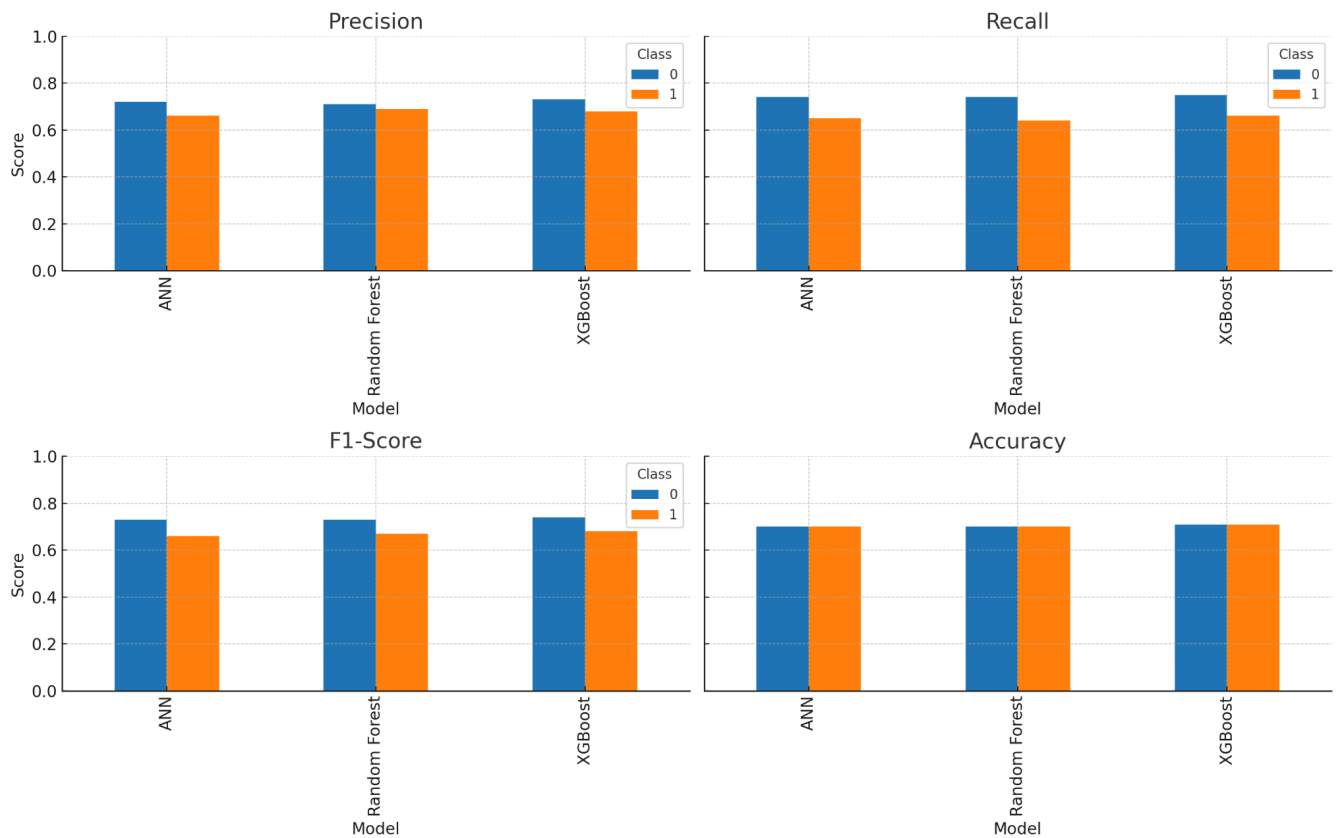
The analysis demonstrates that among the models tested, XGBoost yielded the highest accuracy in addressing the complex and variable nature of competitive surfing. Although ANNs and Random Forest models displayed comparable levels of accuracy, XGBoost distinguished itself with marginally superior F1-scores. This indicates that machine learning, particularly XGBoost, has the capacity to navigate the unpredictable dynamics of surf competitions. To further refine the precision of these predictions, future work should focus on enlarging the dataset and integrating a broader spectrum of variables in order to capture the high variability characteristic of the sport.