

Student Id: W9621054

Chinnamsetti Sravan Sai

CHESS Winner’s Prediction Using Machine Learning Techniques

Chinnamsetti Sravan Sai

(W9621054)

*Abstract*— Chess is the most popular game that is played all over the world, it is played by two players, and it involves strategic decisions making while playing. There has been increased attention to applying machine learning (ML) techniques in cheese. There are several parameters that decide the winner in the chess game, the parameter like the time consumption to complete, the number of steps, defeating the opponent king, opening moves, and many more. The dataset consists of 20,058 data points and 17 features. The application of an ML algorithm makes the model prediction easier and faster, in this case predicting the winner is the target variable. ML classification techniques are used for predicting the winner. The outcomes display that the dataset can be used to gain visions into the planned actions of players in Cheese, and it can help as an appreciated resource for evolving novel machine-learning algorithms. In this scenario, the author used the ML classification techniques like K Neighbors, Gaussian Naive Bayes, Logistic Regression, Support vector, Random Forest, Decision Tree, AdaBoost Classifier, XGB, and Gradient Boosting, out of which the author had achieved the highest accuracy of 88 % by using XGB classifier.

Keywords— CHESS GAME, ML, XGBoost, classification report.

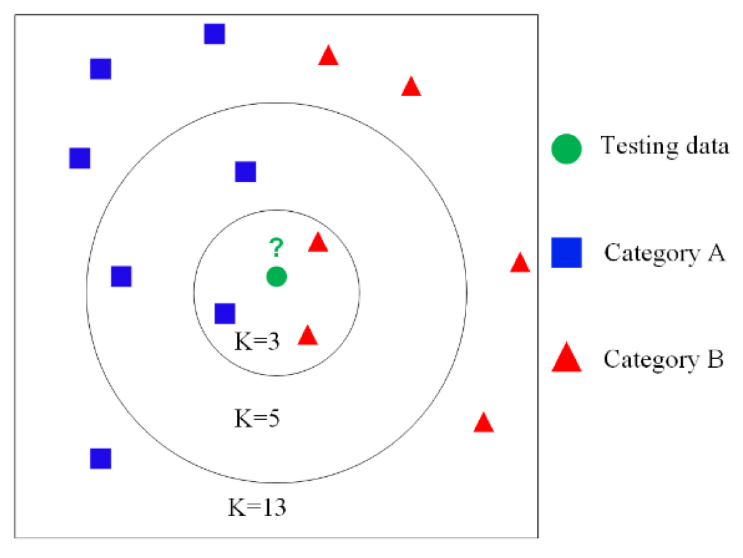
# Introduction

The chess game consists of 64 squares that are arranged in eight cross-eight grids. Each player has one king, one queen, two bishops, and eight pawns, two rooks, two knights. The main aim of the player is to checkmate the opponent’s king, it is like occupying the kingdom by defeating the other king (Baird and Baird 2020), The game can also end in a draw if one of several situations is met, like when a player has no permissible steps but is not in check when both players have incomplete pieces and cannot force a checkmate, or else a threefold repetition of the same position. In chess piece moves in an unlike way, and arresting an opponent's piece is typically done by relocating a piece onto the square engaged by the opponent's piece. Pawns are the only pieces that capture inversely from how they move, capturing crossways one square forward (Togelius 2019). Chess is a game of planning, strategies, and prudence. By keeping all this in mind the author developed a model that can able to predict the winner.

# Methodology

## KNN

KNN is known as the K Nearest Neighbors algorithm and it is a supervised nonparametric ML algorithm. KNN is the simplest algorithm out of all the ML algorithms present in the market. It is used for both classification and regression problems. Our current dataset is for the classification task (Sen et al., 2020). The drawback of KNN is its time complexity in predicting the outputs and KNN doesn’t learn any trends or patterns from the data, unlike the other algorithms. In the workings of the algorithm, K specifies the number of neighbors for a given data point. KNN computes the output based on simpler arithmetical techniques like mean, median, and mode, based on the units passed to K. The parameter is passed by the user and can be hyper-tuned. The neighbor selection happens through different kinds of distances like Euclidean distance, Manhattan distance, Lp norms, etc.

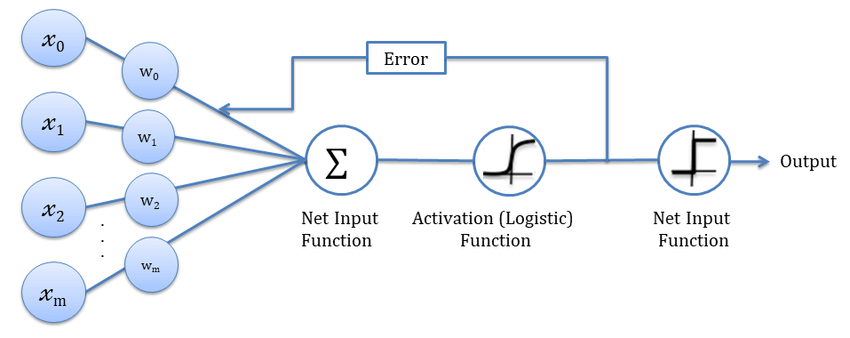


1. KNN Classifiers

Source: Qianwu Zhang (2019)

## Logistic Regression

LR is one of the supervised algorithms of ML. It is used in predicting classification problems. It helps to predict the relationship between the independent i.e., input variables, and dependent i.e., output variables of a given dataset. Here, the output variable can be of Binary class or multi-class classification, but highly accurate predictions for a binary class are easier in LR (Menard, 2022). It is easier as LR represents the best separator line on a graph between two classes of a dataset.

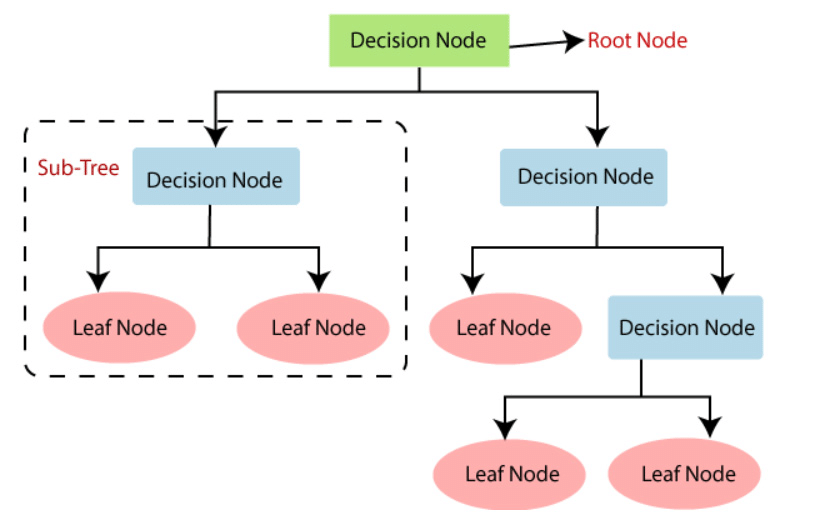


1. Logistic Regression.

Source: Md. Sabir Hossain (2019)

## Decision Tree Classifiers

A Decision Tree Classifier is a supervised ML algorithm, it is a tree-based approach. The whole dataset is represented in the form of a tree, where the output variable is known as a root node, rules formed according to the output variable are known as intermediary nodes, and the final one is known as a decision node (Safavian, 2018). The depth of the tree can be decided through max depth and by default, it is none. The rules of the tree are calculated based on the weighted entropies of the input variables. The weights can be calculated through entropy, Gini impurities, and log loss. Internally probability concept is also utilized in building the algorithm.

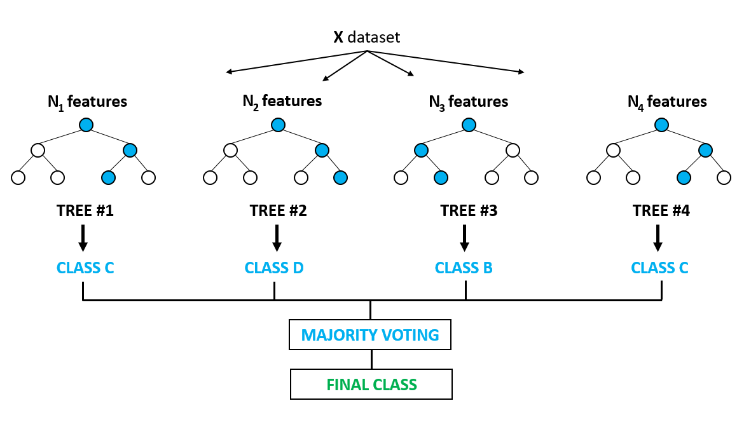


1. DT Classifier.

Source: Mohamed Hanafy (2021)

## Random Forest Classifier

Random forest helps solve both classification and regression problems, but as the given dataset is of classification task, the author uses a random first classifier to predict employee promotions (Parmar, 2019). Multiple decision trees are used in building this algorithm and each tree will be trained on a random subset of the dataset. Later, the aggregate of all the outcomes will be calculated from all the decision trees combined. Random solves the overfitting issue and so it is used by many authors for accurate predictions as it involves multiple decision trees.

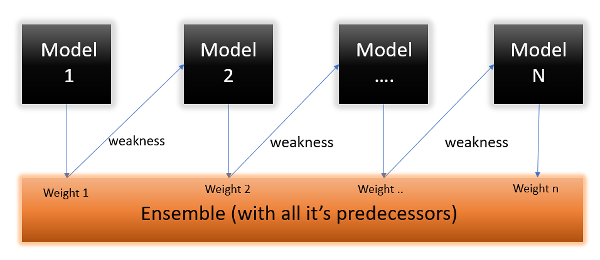


1. RF Classifier.

Source: Rachel Spiro (2021)

## Adaboost Classifier

Ada boost is known as Adaptive boosting. It identifies data points that are misclassified and adjust the weights to minimize their training error of them. It is used for both classifications as well as regression data. AdaBoost is executed by joining numerous weak learners into one strong learner. decision stump has one level of decision tree it makes decisions based on the input feature. Each and every observation is weighted similarly while drawing out the primary decision stump. if any observations are wrongfully classified after the primary decision stump, they are given higher weights. Again, a new decision stump is drawn by the higher-weight explanations as more important. Again, if any misclassified observations are given more weight, this process is continued till all the observations get the right class.

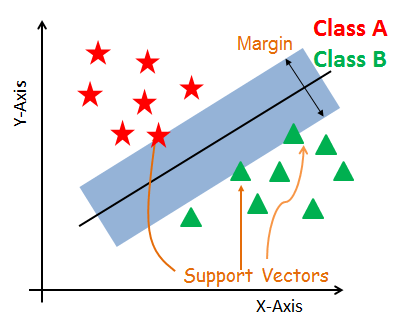


1. Adaboost classifier.

Source: Anshul Saini (2021).

## SVM Classifier

Support Vector Machine is also known as SVM, it is one of the supervised ML algorithms that is used for both classification as well as regression models, SVM algorithm is used to create decision boundaries that can separate n-dimensional space into different classes so that it can easily keep the unseen data point in the exact category in the future. The best decision boundary is known as a hyperplane. Its vectors help in creating the hyperplane. The data points that are nearer to the hyperplane influence the position and orientation of the hyperplane and are called as support vectors.

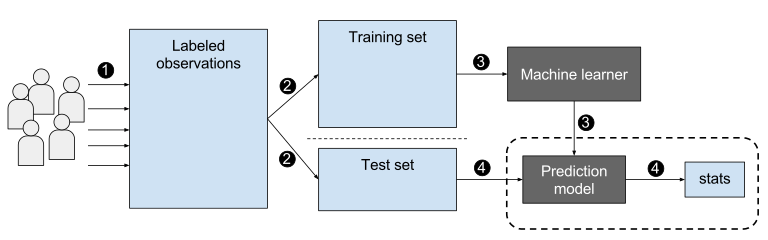


1. SVM classifier.

Source: DataCamp (2023).

## Gradient Boosting Classifiers

The key idea behind gradient boosting is to iteratively improve the model's predictions by adding new weak learners (i.e., decision trees) that focus on the examples that the model has the most trouble with. In other words, the model learns from its mistakes and gradually becomes better at predicting the gender label for new examples. In summary, gradient boosting is a powerful machine-learning technique that can be used to improve the accuracy of gender classification models by iteratively improving the model's predictions based on its mistakes.



1. GB Architecture.

Source: Dan Nelson (2023)

## XG Boost Classifiers

XGBoost Classifier is a kind of gradient-boosting technique that joins several ineffective predictive models to produce a powerful one. The approach updates the weights of the dataset instances depending on the errors of the previous trees and adds a new tree to the ensemble in each iteration. In addition, control overfitting is subject to regularization. In the context of gender classification, XGBoost can be used to build a high-performance model that can accurately predict the gender of a person based on their features (e.g., name, age, occupation).



1. XGB Architecture.

Source: Kinsuk Ghatak (2022)

# 3 Results and Analysis

## Basic Information

## 

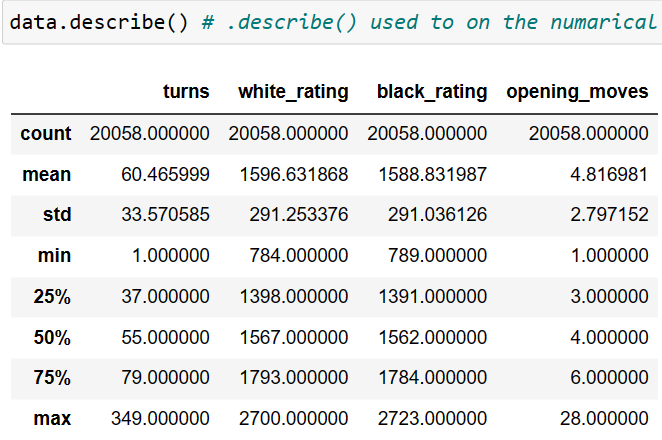
1. Basic information of the dataset.

Info is a function in pandas used to get a summary of a dataset. It provides information on the Datasets columns, data types, non-null values, and memory usage. The dataset consists of both the numerical and object datatypes and also the info gives the information on the count of non-null values that are present in each attribute.

## 

1. Null values in the dataset.

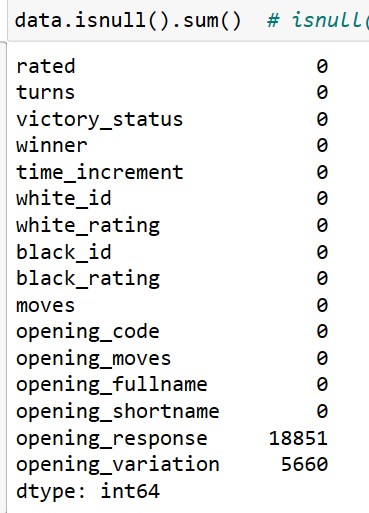
Is null will return the null values and the sum function will return the count of null values present in each column. In the given dataset there are 18851 missing values in the opening response and 5660 in the opening variation columns.



1. Statistical information of the dataset

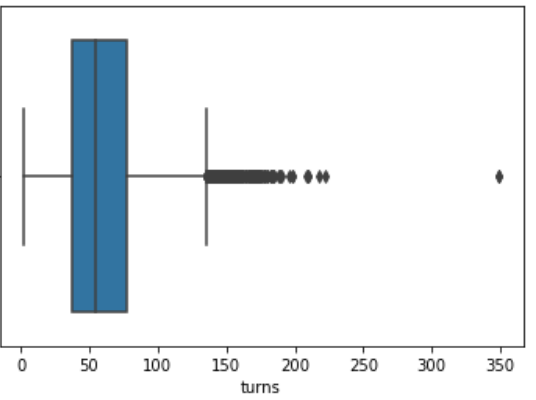
Describe function is used to generate the descriptive statistics of the dataset, it gives information on central tendency, count, std, percentiles minimum, and the maximum value in the attribute.

## Data Cleaning



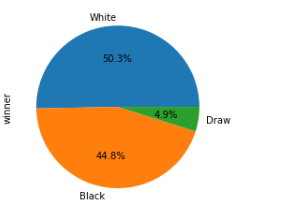
1. Missing values in each column.

Since there are only 18851 nonnull values in the opening response column the author dropped the entire column, and the author dropped all the other rows that contain missing values in the remaining column. The shape of the dataset changed to 14398 rows and 15 columns.



1. Outliers of the columns

The author had use IQR technique to treat the outliers and removed the outliers from all the other columns in the same way.

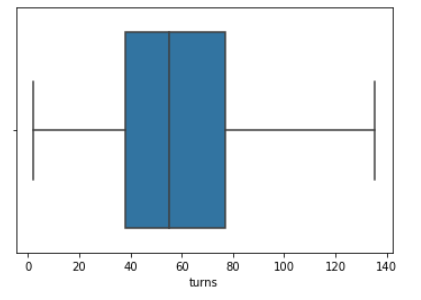


1. Pie chart of winners’ attribute

The author removed the rows that are drawn since the main intention is to find out the winner of the chess game.

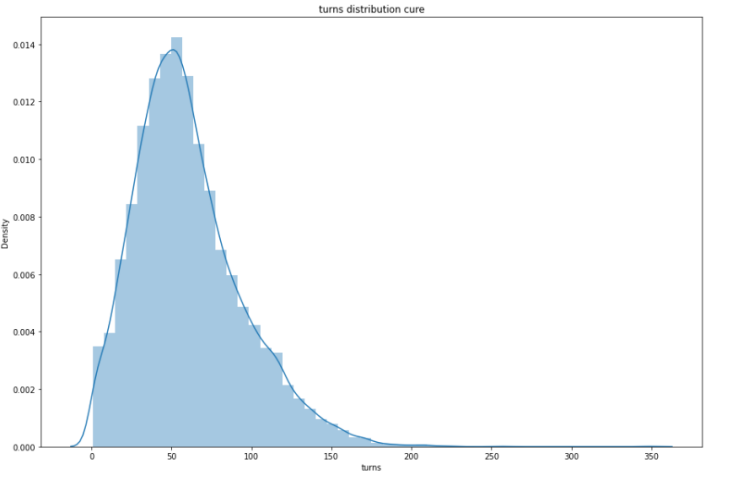
## Data visualization

Data visualization is an essential step in the data analysis process, including in predicting telecoms churn using machine learning.



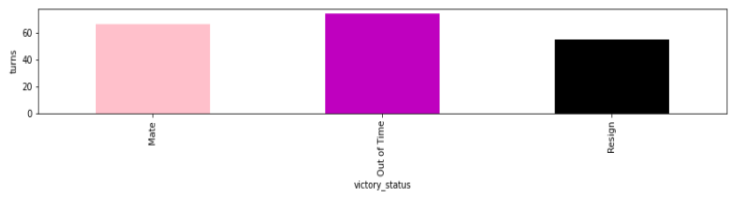
1. Boxplot on number of turns column

Figure 15 shows the box plot of the turns having no outliers for the given data. A box plot, also referred to as a box-and-whisker plot, is a style of chart used to show how a dataset is distributed over its quartiles. It displays the median, IQR, and any outliers or extremely high or low numbers.



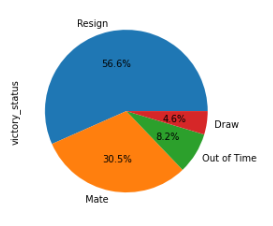
1. Distribution plot on a number of turns column.

Figure 16 describes the distribution, the distribution of the turns is in the range of 50 to 70. The probability distribution of a dataset is graphically depicted in a distribution plot. It aids in the visualization of the underlying distribution of data by showing the frequency of occurrence of various values within a dataset. Histograms, kernel density plots, and box plots are a few examples of distribution plots.



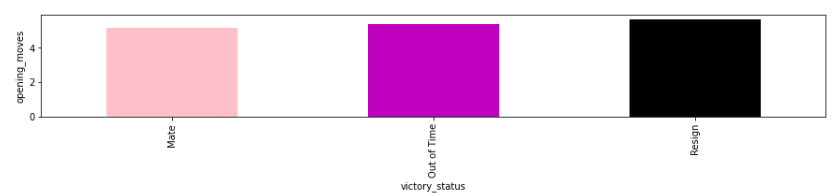
1. Bar plot on victory status and turns column.

From figure 17, the bar graph out of time is high when compared with the other’s victory status.



1. Pie plot on victory status column.

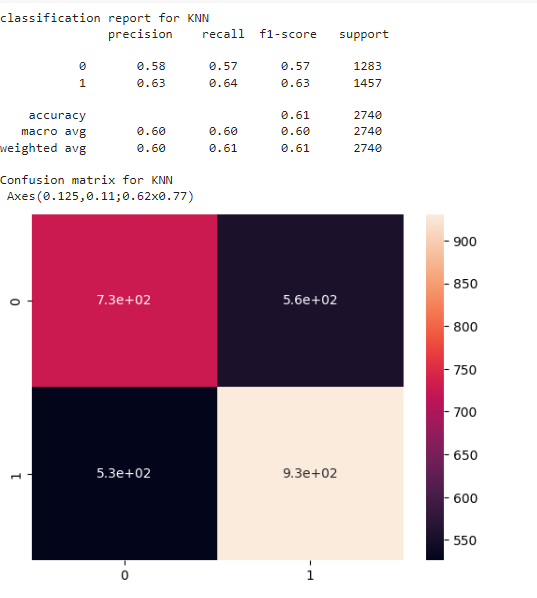
The data contains 56% of resign in the victory status column.



1. Bar plot on victory status and opening moves column

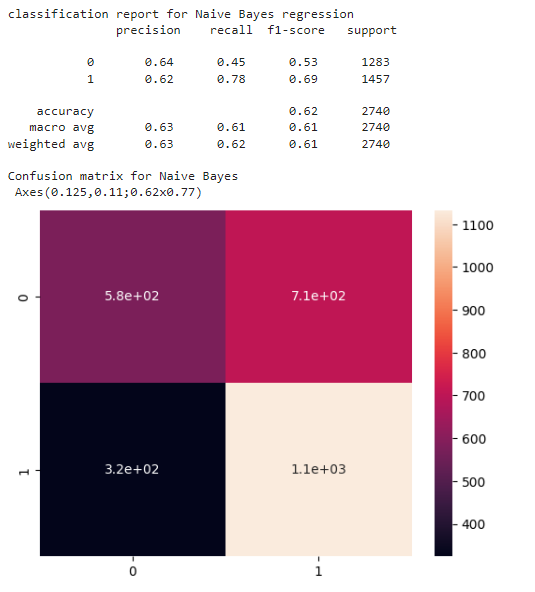
The resign is high when compared to mate and out of time, in victory status and opening moves bar plot.

## Model creation



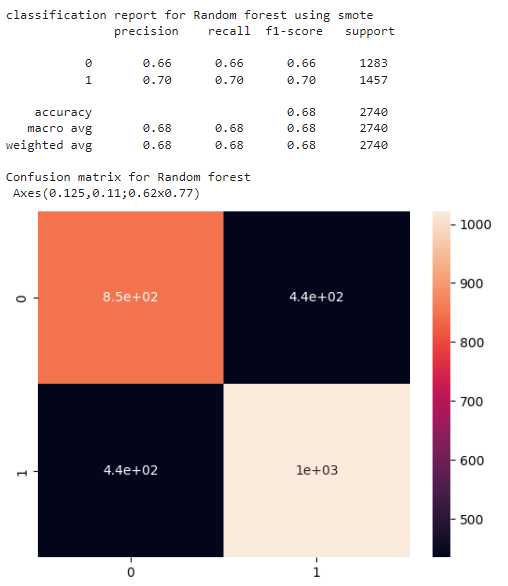
1. Classification report of KNN

The author achieved 61 percent accuracy by using known classifier.



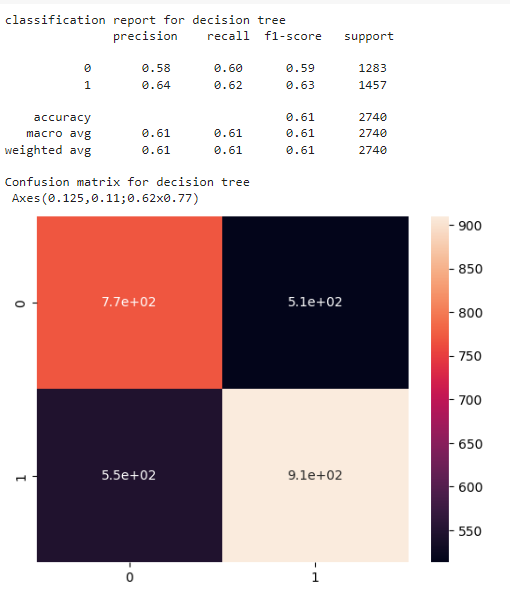
1. Classification report of Gaussian NB

The author achieved 62 percent accuracy by using Gaussian NB.



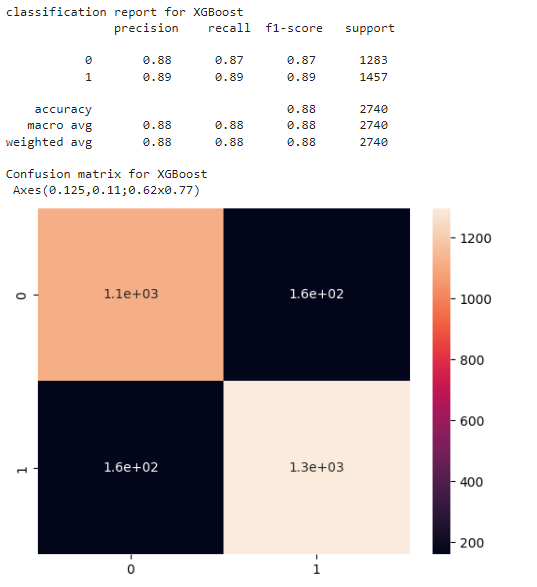
1. Classification report of RF model

The author achieved 68 percent accuracy by using a random forest regressor.

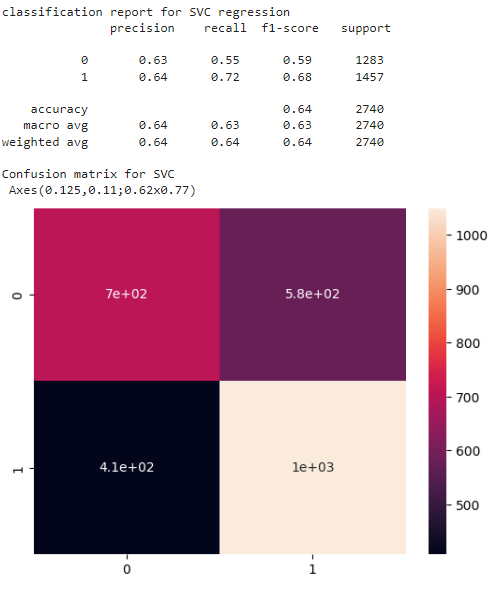


1. Classification Report of Decision Tree Classifiers

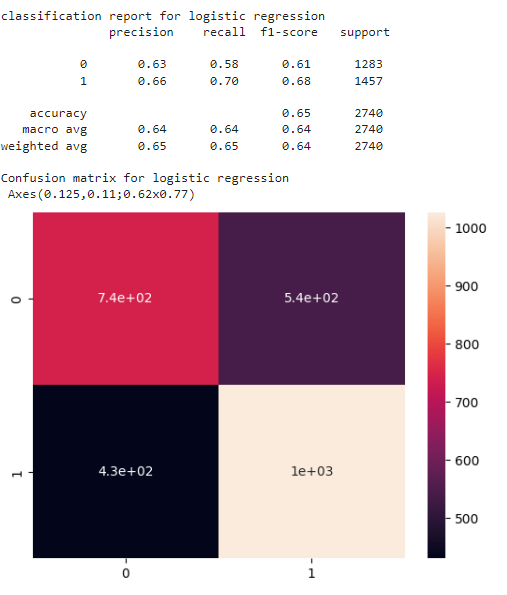
The author achieved 62 percent accuracy by using a Decision Tree classifier



1. Classification report of XGB regression code



1. Classification report of SVC code.

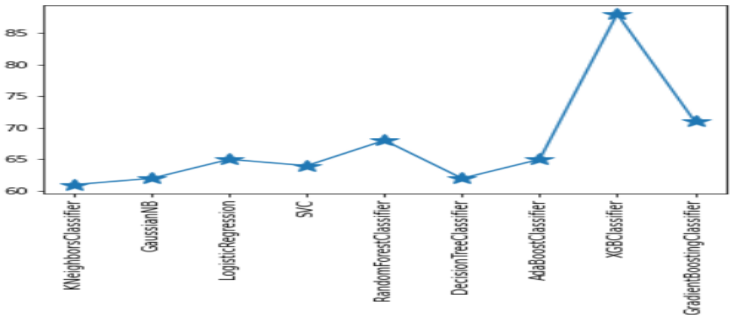


1. Classification Report of Logistic Regression

The author achieved 65 percent accuracy by using a Logistic Regression.

## Final model

The author had come to the conclusion that by using XGB regression the model had given the highest accuracy.



1. Acccuracy of all models

# Conclusion and Recommendations

Because of the abundance of attributions that must be inspected or surveyed for a rewarding prediction, chess game winner prediction will be a difficult function. The collecting and preparation of data is the major step in the prediction process. This Notebook evaluates used-chess game-winner prediction using the Kaggle dataset from kaggle.com which is transformed into CSV format and used to create ML algorithms. In this Study four algorithms used the three ML models’ accuracy was tested and compared against one another, this is an important comparison and balancing between one and multiple ML algorithms to check the accuracy, at last, the XGBoost regressor gives the best accuracy of 88%. Being an enlightened model, XGB Regressor gives the BEST accuracy in comparison to all prior works using these datasets.

# References

Anshul Saini (2021), Working of ADA Boost [Online], Available at:< https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/>, Accessed on: 01st Apr 2023.

Baird, C. and Baird, C., 2020. CHESS: War, Harmony, Sex and Politics. Games and Gaming in Early Modern Drama: Stakes and Hazards, pp.203-247.

Dan Nelson (2023), Working of Gradient Boosting [Online], Available at:< https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/ >, Accessed on: 01st Apr ,2023.

DataCamp (2023), Working of SVM [Online], Available at:< https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python>, Accessed on: 01st Apr ,2023.

Kinsuk Ghatak (2022), Working of XGB [Online], Available at:< https://www.shiksha.com/online-courses/articles/xgboost-algorithm-in-machine-learning/ >, Accessed on: 01st Apr ,2023.

Md. Sabir Hossain (2019), Development of an Intelligent Job Recommender System for Freelancers using Client’s Feedback Classification and Association

Mohamed Hanafy (2021), Machine Learning Approaches for Auto Insurance Big Data, [Online], Available at URL: < https://www.researchgate.net/figure/Structure-of-a-general-decision-tree\_fig1\_349499774>, Accessed on: 01st Apr ,2023.

Qianwu Zhang (2019), A Simple Joint Modulation Format Identification and OSNR Monitoring Scheme for IMDD OOFDM Transceivers Using K-Nearest Neighbor Algorithm, [Online], Available at URL: < https://www.mdpi.com/2076-3417/9/18/3892 >, Accessed on: 01st Apr ,2023.

Rachel Spiro (2021), Hyperparameter Tuning For Random Forest, [Online], Available at URL: < https://rspiro9.github.io/hyperparameter\_tuning\_for\_random\_forest>, Accessed on: 01st Apr ,2023.

Rule Mining Techniques, [Online], Available at URL: < https://www.researchgate.net/figure/Schematic-diagram-for-logistic-regression-classification\_fig2\_333982722 >, Accessed on: 01st Apr ,2023.

Sen, P.C., Hajra, M. and Ghosh, M., 2020. Supervised classification algorithms in machine learning: A survey and review. In Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018 (pp. 99-111). Springer Singapore.

Togelius, J., 2019. Playing smart: On games, intelligence, and artificial intelligence. MIT Press.