

**CLASSIFICATION OF OUTCOMES OF ENGLISH PREMIER  
LEAGUE MATCHES USING MACHINE LEARNING MODELS**

**BY**

**IYIOLA, TOMILAYO PROMISE  
(20PCD02190)**

**AUGUST, 2022**

**CLASSIFICATION OF OUTCOMES OF ENGLISH PREMIER  
LEAGUE MATCHES USING MACHINE LEARNING MODELS**

**BY**

**IYIOLA TOMILAYO PROMISE**

**(20PCD02190)**

**B.Sc Mathematics, Bowen University, Iwo.**

**A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADU-  
ATE STUDIES IN PARTIAL FULFILMENT OF THE REQUIREMENTS  
FOR THE AWARD OF THE MASTER OF SCIENCE (M.Sc.) DEGREE  
IN INDUSTRIAL MATHEMATICS IN THE DEPARTMENT OF MATH-  
EMATICS, COLLEGE OF SCIENCE AND TECHNOLOGY, COVENANT  
UNIVERSITY.**

**AUGUST, 2022**

## ACCEPTANCE

This is to attest that this dissertation is accepted in partial fulfilment of the requirements for the award of the degree of Master of Science in Industrial Mathematics in the Department of Mathematics, College of Science and Technology, Covenant University, Ota, Nigeria.

**Mr. Taiwo B. Erewunmi**

(Secretary, School of Postgraduate Studies)

Signature and Date

**Prof. Akan B. Williams**

(Dean, School of Postgraduate Studies)

Signature and Date

## DECLARATION

**I, IYIOLA, TOMILAYO PROMISE (20PCD02190)** declare that this research was carried out by me under the supervision of Dr. Hilary I. Okagbue of the Department of Mathematics, College of Science and Technology, Covenant University, Ota, Nigeria. I attest that the dissertation has not been presented either wholly or partially for the award of any degree elsewhere. All sources of data and scholarly information used in this dissertation are duly acknowledged.

**IYIOLA, TOMILAYO PROMISE**

**(Student)**

**Signature and Date**

## CERTIFICATION

We certify that this dissertation titled ”**CLASSIFICATION OF OUTCOMES OF ENGLISH PREMIER LEAGUE MATCHES USING MACHINE LEARNING MODELS**” is an original research work carried out by **IYIOLA, TOMILAYO PROMISE (20PCD02190)** in the Department of Mathematics, College of Science and Technology, Covenant University, Ota, Ogun State, Nigeria under the supervision of Dr. H. I. Okagbue. We have examined and found this work acceptable as part of the requirements for the award of Master of Science in Industrial Mathematics.

**Dr. Hilary I. Okagbue**  
(Supervisor)

Signature and Date

**Prof. Samuel A. Iyase**  
(Head of Department)

Signature and Date

**Prof. Kayode Ayinde**  
(External Examiner)

Signature and Date

**Prof. Akan B. Williams**  
(Dean, School of Postgraduate Studies)

Signature and Date

## DEDICATION

To the Almighty God for grace, opportunity, and divine wisdom, To my lovely parents and siblings.

## ACKNOWLEDGEMENTS

I want to give glory and praise to the Almighty God for helping me through the course of this programme.

I appreciate the Chancellor, Dr. D. O. Oyedepo, for establishing this great institute of learning and providing an enabling environment for research.

I sincerely appreciate Prof. S. A. Iyase, Head of the Department of Mathematics, Covenant University, for his words of encouragement and wise counsel all through the programme.

My appreciation also goes to the post-graduate coordinator of the Department of Mathematics, Covenant University, Dr. S. O. Edeki for his contributions and guidance throughout this research work.

I acknowledge the intellectual tutelage of my supervisor, Dr. H. I. Okagbue and his immense support and guidance towards the commencement and completion of this dissertation.

I extend my appreciation to the post-graduate committee of the Department of Mathematics, Covenant University, Prof. S. A. Iyase, Prof. T. A. Anake, Prof. A. M. Oke-doye, Dr. M. C. Agarana, Dr. S. O. Edeki, Dr. A. A. Opanuga, Dr. J. G. Oghonyon, Dr. O. O. Agboola, Dr. O. A. Odetunmibi, Dr. G. O. Alao, Dr. H. I. Okagbue, Dr. A. F. Adedotun, and Dr. O. F. Imaga, for their contributions to the success of this work in one way or the other.

I am also thankful to the administrative officer of the Department of Mathematics, Covenant University, Mrs. T. A. Ajayi for her assistance during this research work.

I will also like to appreciate the entire faculty and staff of the Department of Mathematics. My appreciation also goes to my wonderful course-mate, Lekan Igbekele, for his contribution to this academic journey.

The appreciation will not end without acknowledging my lovely parents, Dr. O. A. Iyiola and Mrs. M. A. Iyiola for their ever-present love, care, prayers, support and assistance, may God continue to bless and keep them in Jesus' name. I also extend my appreciation to my lovely siblings, Olumayowa Aina (Nee Iyiola), Ayobamiji Iyiola, and Ayanfeoluwa Iyiola, for their support, assistance and words of encouragement

during the course of this research work.

## TABLE OF CONTENTS

	PAGES
COVER PAGE	i
TITLE PAGE	ii
ACCEPTANCE	iii
DECLARATION	iv
CERTIFICATION	v
DEDICATION	vi
ACKNOWLEDGEMENTS	vii
LIST OF FIGURES	xiii
LIST OF TABLES	xvi
LIST OF SYMBOLS	xviii
ABSTRACT	xix
<b>CHAPTER ONE : INTRODUCTION</b>	<b>1</b>
1.1 Background to the Study	1
1.2 Statement of Research Problem	2
1.3 Aim and Objectives of the Study	3
1.4 Justification for the Study	4
1.5 Scope of the Study	4
1.6 Definition of Terms	4
<b>CHAPTER TWO : LITERATURE REVIEW</b>	<b>6</b>
2.1 English Premier League	6
2.1.1 Competition Format	7
2.1.2 Promotion and Relegation	7
2.1.3 Simplified Football Rules	8
2.1.4 Domestic Leagues Competition Format	8

2.1.5	Main Match Events	8
2.2	Research on English premier league	9
2.3	Football Classification and Prediction using Machine Learning Models	10
2.4	Gaps Identified in Literature	12
<b>CHAPTER THREE : METHODOLOGY</b>		<b>13</b>
3.1	Variables in the Data	13
3.2	Data Sources	14
3.2.1	Software used for the Analysis	15
3.3	Descriptive Statistics	15
3.3.1	Arithmetic Mean	16
3.3.2	Bar Chart	16
3.3.3	Frequency Analysis	17
3.4	Kruskal-Wallis Test	17
3.5	Feature Selection	19
3.5.1	Relief and ReliefF Algorithm	20
3.5.1.1	Relief Algorithm	20
3.5.1.2	ReliefF Algorithm Explained	21
3.5.1.3	Reliable probability estimation	21
3.5.1.4	Incomplete data	21
3.5.1.5	Multi-class problems	21
3.5.2	Information Gain Ratio	22
3.5.2.1	Information Gain	22
3.5.2.2	Information Gain Calculation	22
3.5.2.3	Split Information	22
3.5.2.4	Split Information Calculation	23
3.5.2.5	Information Gain Ratio Calculation	23
3.6	Classification Models	23
3.6.1	Logistic Regression	24
3.6.1.1	Ordinal Logistic Regression	24
3.6.1.2	Multinomial Logistic Regression	24
3.6.1.3	Binary Logistic Regression	24
3.6.1.4	Model	24

3.6.2	Adaptive Boosting	25
3.6.2.1	Training	26
3.6.3	Random Forest	26
3.6.4	Gradient Boosting	27
3.7	Level of Significance	28
3.8	Classification(Performance) Metrics	28
3.8.1	Accuracy	29
3.8.2	Confusion Matrix	29
3.8.3	F1 Score	30
3.8.4	Precision	31
3.8.5	Recall	31
3.8.6	ROC AUC (Reciever Operating Characteristics Curve)	31
<b>CHAPTER FOUR : RESULTS AND DISCUSSION</b>		<b>33</b>
4.1	Frequency Analysis & Chart Representation	33
4.1.1	Frequency analysis & chart representation for 2016/2017 season	33
4.1.2	Frequency analysis & chart representation for 2017/2018 season	35
4.1.3	Frequency analysis & chart representation for 2018/2019 season	37
4.1.4	Frequency analysis & chart representation for 2019/2020 season	39
4.1.5	Frequency analysis & chart representation for 2020/2021 season	41
4.2	Test of Equality of Median on the Outcomes for the Five Seasons	43
4.3	Feature Selection of the Outcomes	44
4.3.1	Criteria for selection	44
4.4	Classification Of Outcomes Using The Variables Recommended Via Feature Selection	49
4.4.1	Cross Validation	49
4.4.2	Parameter settings of the machine learning (ML) models	49
4.4.3	Performance metrics and classification analysis	50
4.4.3.1	2016/2017 season performance metrics	50
4.4.3.2	2017/2018 season performance metrics	58
4.4.3.3	2018/2019 season performance metrics	61
4.4.3.4	2019/2020 season performance metrics	64
4.4.3.5	2020/2021 season performance metrics	67

4.5	Determination of Outcome using the First and Second Half Results	71
4.5.1	Relationship between the first and second halves	72
4.5.2	Classification	73
4.5.3	Cross Validation	73
4.5.4	Classification analysis and evaluation metrics	74
4.5.5	2016/2017 season performance metrics	74
4.5.6	2017/2018 season performance metrics	78
4.5.7	2018/2019 season performance metrics	83
4.5.8	2019/2020 season performance metrics	87
4.5.9	2020/2021 season performance metrics	92
<b>CHAPTER FIVE : CONCLUSION AND RECOMMENDATIONS</b>		<b>97</b>
5.1	Summary of Findings	97
5.2	Conclusion	99
5.3	Contributions to Knowledge	99
5.4	Recommendations	100
5.5	Limitations of the Study	100
5.6	Areas for Further Research	100
<b>REFERENCES</b>		<b>102</b>

## LIST OF FIGURES

FIGURES	TITLE OF FIGURES	PAGES
3.1	Bar chart representation	17
3.2	Confusion matrix of binary classification	30
3.3	The Receiver Operating Characteristics (ROC) curve	32
4.1	The bar-chart representation of the outcome for the 2016/2017 season	34
4.2	The bar-chart representation of the outcome for the 2017/2018 season	36
4.3	The bar-chart representation of the outcome for the 2018/2019 season	38
4.4	The bar-chart representation of the outcome for the 2019/2020 season	40
4.5	The bar-chart representation of the outcome for the 2020/2021 season	42
4.6	Diagrammatic view of the 16 variables and its contribution to the outcome	48
4.7	The top six variables recommended via feature selection	49
4.8	The classification of instances of the outcome for the LR, GB, AB models	51
4.9	The classification of instances of the outcome for the RF model	51
4.10	2016/2017 feature importance for Logistic Regression (LR)	52
4.11	2016/2017 feature importance for Adaptive Boosting (AB)	53
4.12	2016/2017 feature importance for Gradient Boosting (GB)	53
4.13	2016/2017 feature importance for Random Forest (RF)	54
4.14	ROC curve for LR, GB and AB machine learning models	55
4.15	Lift curve for LR, GB and AB machine learning models	56
4.16	Calibration plot for LR, GB and AB machine learning models	57
4.17	The classification of instances of the outcome for the four models	58
4.18	2017/2018 feature importance for Logistic Regression (LR)	59
4.19	2017/2018 feature importance for Adaptive Boosting (AB)	59
4.20	2017/2018 feature importance for Gradient Boosting (GB)	60
4.21	2017/2018 feature importance for Random Forest (RF)	60

4.22	The classification of instances of the outcome for the LR, GB, AB models	61
4.23	The classification of instances of the outcome for the RF model	62
4.24	2018/2019 feature importance for Logistic Regression (LR)	62
4.25	2018/2019 feature importance for Adaptive Boosting (AB)	63
4.26	2018/2019 feature importance for Gradient Boosting (GB)	63
4.27	2018/2019 feature importance for Random Forest (RF)	64
4.28	The classification of instances of the outcome for the four ML models	65
4.29	2019/2020 feature importance for Logistic Regression (LR)	65
4.30	2019/2020 feature importance for Adaptive Boosting (AB)	66
4.31	2019/2020 feature importance for Gradient Boosting (GB)	66
4.32	2019/2020 feature importance for Random Forest (RF)	67
4.33	The classification of instances of the outcome for the LR, GB, AB models	68
4.34	The classification of instances of the outcome for the RF model	68
4.35	2020/2021 feature importance for Logistic Regression (LR)	69
4.36	2020/2021 feature importance for Adaptive Boosting (AB)	69
4.37	2020/2021 feature importance for Gradient Boosting (GB)	70
4.38	2020/2021 feature importance for Random Forest (RF)	70
4.39	The classification of instances of the outcome for the RF Model	74
4.40	The classification of instances of the outcome for the LR model	75
4.41	The classification of instances of the outcome for the GB model	75
4.42	The classification of instances of the outcome for the AB model	76
4.43	2016/2017 feature importance for Logistic Regression (LR)	76
4.44	2016/2017 feature importance for Adaptive Boosting (AB)	77
4.45	2016/2017 feature importance for Gradient Boosting (GB)	77
4.46	2016/2017 feature importance for Random Forest (RF)	78
4.47	The classification of instances of the outcome for the RF model	79
4.48	The classification of instances of the outcome for the LR model	79
4.49	The classification of instances of the outcome for the GB model	80
4.50	The classification of instances of the outcome for the AB model	80
4.51	2017/2018 feature importance for Logistic Regression (LR)	81

4.52	2017/2018 feature importance for Adaptive Boosting (AB)	81
4.53	2017/2018 feature importance for Gradient Boosting (GB)	82
4.54	2017/2018 feature importance for Random Forest (RF)	82
4.55	The classification of instances of the outcome for the RF model	83
4.56	The classification of instances of the outcome for the LR model	84
4.57	The classification of instances of the outcome for the GB model	84
4.58	The classification of instances of the outcome for the AB model	85
4.59	2018/2019 feature importance for Logistic Regression (LR)	85
4.60	2018/2019 feature importance for Adaptive Boosting (AB)	86
4.61	2018/2019 feature importance for Gradient Boosting (GB)	86
4.62	2018/2019 feature importance for Random Forest (RF)	87
4.63	The classification of instances of the outcome for the RF model	88
4.64	The classification of instances of the outcome for the LR model	88
4.65	The classification of instances of the outcome for the GB model	89
4.66	The classification of instances of the outcome for the AB model	89
4.67	2019/2020 feature importance for Logistic Regression (LR)	90
4.68	2019/2020 feature importance for Adaptive Boosting (AB)	90
4.69	2019/2020 feature importance for Gradient Boosting (GB)	91
4.70	2019/2020 feature importance for Random Forest (RF)	91
4.71	The classification of instances of the outcome for the RF model	92
4.72	The classification of instances of the outcome for the LR model	93
4.73	The classification of instances of the outcome for the GB model	93
4.74	The classification of instances of the outcome for the AB model	94
4.75	2020/2021 feature importance for Logistic Regression (LR)	94
4.76	2020/2021 feature importance for Adaptive Boosting (AB)	95
4.77	2020/2021 feature importance for Gradient Boosting (GB)	95
4.78	2020/2021 feature importance for Random Forest (RF)	96

## LIST OF TABLES

TABLES	TITLE OF TABLES	PAGES
3.1	Table of variables	13
3.2	Table of variables cont'd	14
4.1	Frequency table for the independent variables of the 2016/2017 season	33
4.2	Frequency table for the independent variables of the 2017/2018 season	35
4.3	Frequency table for the independent variables of the 2018/2019 season	37
4.4	Frequency table for the independent variables of the 2019/2020 season	39
4.5	Frequency table for the independent variables of the 2020/2021 season	41
4.6	Normality test for the outcomes of the five seasons	43
4.7	The Post-hoc result for the outcomes of the five seasons	44
4.8	Feature selection output of the variables for 2016/2017 season	45
4.9	Feature selection output of the variables for 2017/2018 season	45
4.10	Feature selection output of the variables for 2018/2019 season	46
4.11	Feature selection output of the variables for 2019/2020 season	46
4.12	Feature selection output of the variables for 2020/2021 season	47
4.13	The SRC and Cod summary table	48
4.14	Parameter settings of the four models	50
4.15	Performance metrics on the ML models for 2016/2017 season	50
4.16	Performance metrics on the ML models for 2017/2018 season	58
4.17	Performance metrics on the ML models for 2018/2019 season	61
4.18	Performance metrics on the ML models for 2019/2020 season	64
4.19	Performance metrics on the ML models for 2020/2021 season	67
4.20	Table of outcomes for the five seasons	71
4.21	Chi-square values of the pairs of the first and second halves and outcome	72
4.22	Classification table for first half, second half and outcome	73
4.23	Performance metrics on the ML models For 2016/2017 season	74
4.24	Performance metrics on the ML models For 2017/2018 season	78

4.25	Performance metrics on the ML models for 2018/2019 season	83
4.26	Performance metrics on the ML models for 2019/2020 season	87
4.27	Performance metrics on the ML models for 2020/2021 season	92

## LIST OF SYMBOLS

$\bar{x}$  : Arithmetic mean.

$\mu$  or  $\mu_x$  : Population mean.

$\sum$  : Summation.

$R_i$  : Sum rank of the  $i_{th}$  group.

$\tau$  : Threshold.

$X$  : Discrete random variable

$N(t_i)$  : No of times that  $t(i)$  occurs.

$N(t)$  : Total no of counts.

$t$  : Set of events.

$F_T(x)$  : Boosted classifier.

$F_{t-1}(x)$  : Previous stage trained boosted classifier.

$IG$  : Information gain.

$N$  : No of data points in a data set.

$y_i$  : Real value of y

$\emptyset$  : Empty set.

$\Omega$  : Nonempty/sample space.

LR :Logistic regression.

ML : Machine learning.

AB : Adaptive boosting.

GB : Gradient boosting.

CA : Classification accuracy.

AUC : Area under curve.

ANN : Artificial neural network.

SVM : Support vector machine

## ABSTRACT

Football remains an important sport in the world and it has a lot of followers. Researchers are often interested in the analysis of the results of football matches, which helps in the prediction or classification of outcomes (results) of football matches based on some variables. Most of the available models of prediction and classification of outcomes are based on a selected variable or a large number of variables. The use of a few variables can not predict accurately and the use of large variables leads to the problem of interpretation (Parsimony). This work used feature selection methods to reduce sixteen selected independent variables (football related) to six variables in the classification of the outcome variable (home win, away win, and draw) of five seasons of English premier league matches. As expected, a home win is a modal observation in all five seasons. The Kruskal Wallis test showed that the median outcome was not the same for the five seasons, while four machine learning models classified the outcome using the six best variables recommended via the feature selection. Furthermore, the result of the first half and second half was used to classify the final outcome. Five performance metrics attest that the ML models are good in the classification. Cross-Validation ensured that the issues of over-fitting were adequately addressed. Bookmakers may find this research interesting as some variables were identified as key to the classification of outcomes of football matches.

***Keywords: Algorithm, Classification, Cross-Validation, English Premier League, Feature Selection, Frequency, Machine Learning, Kruskal Wallis.***