British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 4, Issue 2, 2021 (pp. 22-30)
www.abjournals.org

MOVIE SUCCESS PREDICTION USING DATA MINING

Olubukola D. Adekola[1], Stephen O. Maitanmi[2], Funmilayo A. Kasali[3], Ayokunle Omotunde[2], Oyebola Akande[2], Oduroye Ayorinde[4], Wumi Ajayi[2] and Yaw Mensah[2]

[1]Department of Software Engineering, Babcock University, Ilisan-Remo, Ogun State, Nigeria. E-mail: adekolao@babcock.edu.ng

[2]School of Computing and Engineering Sciences, Babcock University, Ilisan Remo, Ogun State, Nigeria

[3]Department of Computer Science and Mathematics, Mountain Top University, Ogun State, Nigeria

[4]Computer Science Department, Caleb University, Imota, Lagos, Nigeria

Cite this article:

Olubukola D.A., Stephen O.M., Funmilayo A.K., Ayokunle O., Oyebola A., Oduroye A., Wumi A., Yaw M. (2021), Movie Success Prediction Using Data Mining. British Journal of Computer, Networking and Information Technology 4(2), 22-30. DOI: 10.52589/BJCNIT-CQOCIREC.

Manuscript History

Received: 30 Aug 2021

Accepted: 17 Sept 2021

Published: 22 Sept 2021

ABSTRACT: *The movie industry is arguably one of the biggest entertainment sectors. Nollywood, the Nigerian movie industry produces tons of movies for public consumption, but only a few make it to box-office or end up becoming blockbusters. The introduction of movie success prediction can play an important role in the industry not only to predict movie success but to help directors and producers make better decisions for the purpose of profit. This study proposes a movie prediction model that applies data mining techniques and machine learning algorithms to predict the success or failure of an upcoming movie (based on predefined parameters). The parameters needed for predicting the success or failure of a movie include dataset needed for the process of data mining such as the historical data of actors, actresses, writers, directors, marketing and production budget, audience, location, release date, and competing movies on same release date. This model also helps movie consumers to determine a blockbuster, hit, success rating and quality of upcoming movies before deciding on a movie ticket. The data mining techniques was applied to Internet Movie Database MetaData which was initially passed through cleaning and integration process.*

KEYWORDS: Data Mining, Prediction, Machine Learning, Movies, Meta Data

Article DOI: 10.52589/BJCNIT-CQOCIREC
DOI URL: https://doi.org/10.52589/BJCNIT-CQOCIREC

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 4, Issue 2, 2021 (pp. 22-30)

www.abjournals.org

INTRODUCTION

In entertainment sector, movie industry is one of the biggest and most sought after. It is no surprise that technology-based startups like Netflix, Irokotv and Hulu are on the rise with the aim of expanding the reach of movies. However, only a small percentile achieves success and high ratings while a lot suffer low viewership, some never earn enough profit to cover their budget, while others never make it to box-office or attain considerable success. A movie budget covers various areas like marketing campaign, operational cost, videography, payment of team members and crew (Wales, 2017). If the movie ticket sales are lower than the cost of making the movie then the movie is a failure. Theatre or media that cannot make profit is bound to eventually fizzle out of existence (Ogunleye, 2004). A movie revenue is dependent on a number of factors namely cast of the movie (popular actors have propensity of increasing viewership), budget for the movie, budget for marketing and media campaign, film critics review, ratings and release date. While there might be no exact formula to calculate or provide analysis for predicting in value terms what a movie will generate even with all the parameters listed, data mining could be used to analyze dataset from previous movies, which can include revenue generated from pervious movies, release date and crews to predict the success of an upcoming movie. The use of data mining approach is very potent as it helps in identifying hidden patterns and relationship among various variables (Shu, 2020). The introduction of data mining techniques in predicting movie success would mean developing a model that could be used for future prediction. The technology that would help existing entertainment businesses thrive better by assisting better decision making before and during movie production is worth careful study. Whenever the proposed model is able to predict movie success rate with at least 97-98% accuracy, it means higher profits can be generated. The proposed study aims to develop a model based on data mining techniques that would be applied in predicting the success of upcoming movies. The use of historical data of movies to predict the success of upcoming movies such as director, crew and budget reduces certain level of uncertainty (*Nithin, Pranav,* Sarath Badu & *Lijiya*, 2014).

Problem Statement

Outside of the crew's popularity and movie popularity, the success of a movie is highly dependent on digital forces like online campaign, marketing, reviews, box office ratings and opening day ticket sales. However, before a movie is released, the success of the movie has to rely on media hype, previews and prelease marketing campaign, but these do not determine or translate to a movie's success when released. The problem here is most producers, directors and stakeholders end up expending millions on movie budget without knowing if the movie would be a success or a failure. This study proposes a model that analyzes the revenues generated from previous movies, the reviews, ticket sales, crew's popularity and marketing budget to predict the success of a movie. Such a prediction could be very useful for motion studios to make better intelligent decisions like improving content and creativity, artist compensations, advertising and marketing of the movie accordingly. This helps investors and stakeholders to predict an accurate return-on-investment (ROI) and other associated profit or loss.

Article DOI: 10.52589/BJCNIT-CQOCIREC
DOI URL: https://doi.org/10.52589/BJCNIT-CQOCIREC

LITERATURE REVIEW

The term Artificial intelligence (AI) was originally coined by John McCarthy in 1956 at a conference on the subject (McCarthy, 2000). AI is a wide aspect of computer science that deals with conception of intelligent machines that function like humans in workings and reactions (Grewal, 2014). Data mining is an aspect of AI that involves data processing using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large preexisting databases; a way to discover new meaning in data (Han, Kamber & Pei, 2011). It helps in discovering patterns, usually difficult to find, in order to decide upon the future trends in businesses. Data mined is traditionally viewed as a model of the semantic make-up of the dataset, where the model might be utilized on new information for forecast or classification. The main goal of data mining is to detect patterns that are previously unknown and further use them to make informed decisions for development of systems or models that improve businesses, decision making process and analysis (Ramageri, 2010). Movie data mining has employed methods as data processing, relationship mining, prediction and clustering. Clustering is an unsupervised method that breaks data into definite classes where the data in the same category are alike in properties but are different from other data in other clusters. Clustering plays a significant role especially in data mining applications, such as marketing, text mining, medical diagnostics, and so on (Mesakar, & Chaudhari, 2013). Clustering in movie data mining searches for resemblances and variations between movie and organizes them under similar categories for the purpose of movie success prediction. Discoveries with models may be applied well in detecting relationships among movies and movie success features or related variables. In movie data mining, predictions are used to study features of a model that significantly predicts crucial information on the essential construct. Distillation of data for human judgment focuses on making information comprehensible. Preparation process for building models uses Data distillation for classification while identification displays data in such a way that makes it simple to recognize with a well-known pattern (Baker, 2010). Relationship Mining (RM) operates with a focus of unearthing a huge number of variable connections among the variables in a dataset. In movie data mining, there are four frequently used types of RM; association rule mining, sequential pattern mining, correlation mining, and causal data mining. Relationship Mining is applied within the aspect of finding the link that depicts connections of the data of related movies based on certain features. This helps in enhancing the prediction model.

The three common learning types in Machine Learning (ML) approaches include supervised, unsupervised and reinforced learning. The learning process in supervised learning model are training and testing. The process of training involves taking samples in training data as input in which features are learned by the learning algorithm or learner and building the learning model (Dhage & Raina, 2016). The execution engine is used in the testing process by the learning model in order to predict the test or production data. The unsupervised learning type is of high importance within the process of multimedia content because partitioning of data when class labels are not involved is mostly a necessity (Greene, Cunningham & Mayer, 2008). Unsupervised ML techniques ease analysis of fresh datasets in a way that helps in creating analytic insights from unlabeled data (Usama et. al., 2017). There are various applications from unsupervised Machine Learning like feature learning, data clustering, dimensionality reduction, anomaly detection etc. In reinforced ML, there is an exposition of the machine to an environment where they get trained through trial and error. Reinforced learning helps to gain knowledge from the past experiences and uses it to get the most

Article DOI: 10.52589/BJCNIT-CQOCIREC
DOI URL: https://doi.org/10.52589/BJCNIT-CQOCIREC

appropriate knowledge in making appropriate choices; that is, based on the feedback received (Kumar, Ramakrishnan & Li, 2018).

MATERIALS AND METHODS

The methods capture different stages of model development consisting data collection, data preprocessing, generating training and testing dataset, model generation, prediction and outcomes. This helps to eliminate irrelevant data in order to enhance accurate prediction. Following this due process aids full cleaning of the dataset and discard of irrelevant data from the IMDB dataset as well as through detailed study of the dataset; only the attributes that could affect the prediction of success of a movie were used. The textual data was transformed into numeric data and converted into CSV (comma separated version) format. Most suited algorithm which yielded the highest accuracy and least error was employed. The model was developed using Python, IMDB dataset and python libraries. The model used historical data in order to successfully predict the possibility of success or failure of an upcoming movie. Data mining techniques were used for analyzing patterns in previous records of movie data and sales figures upon which classification was based.

The following (Figure1) represents the flow of activities in the system design:

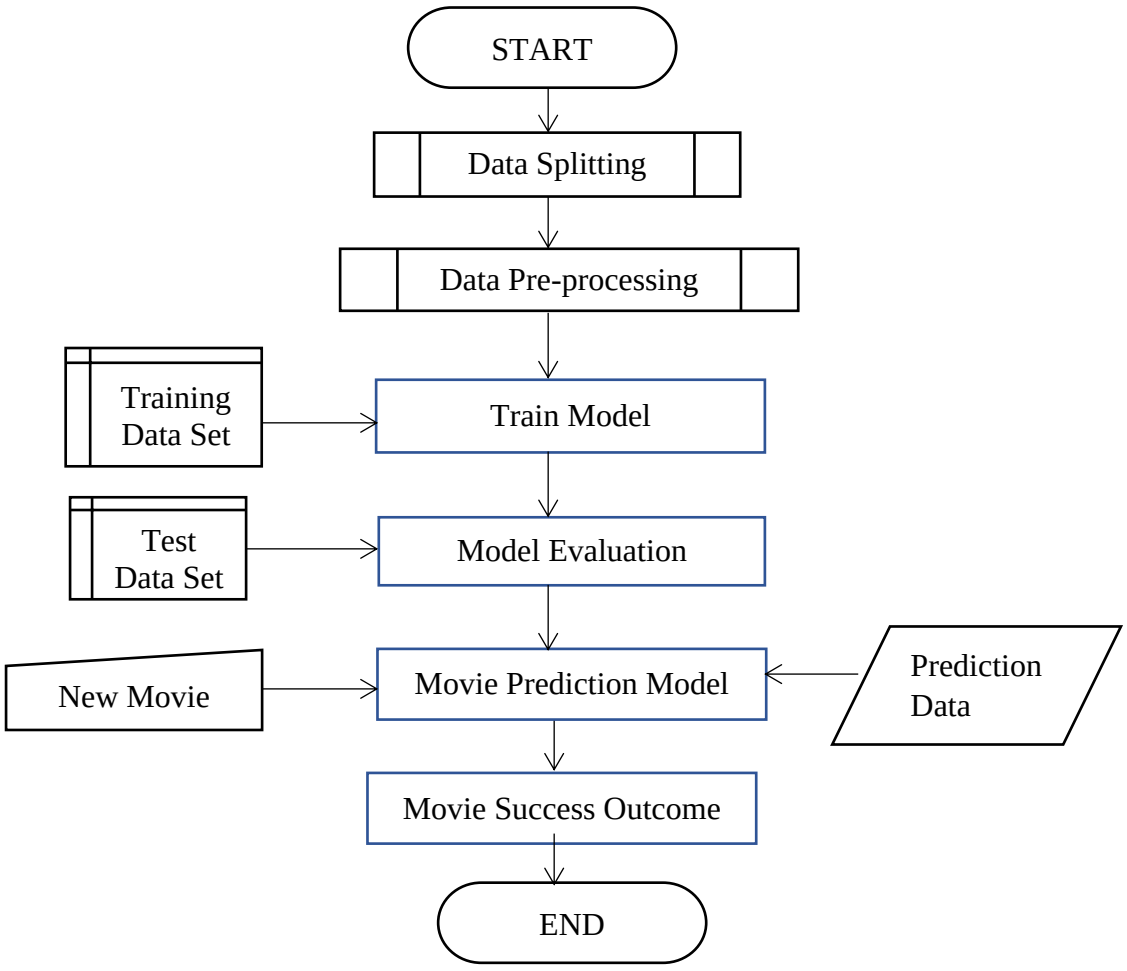

Figure 1: Diagram Illustrating the Flow of Activities in the System Design

Article DOI: 10.52589/BJCNIT-CQOCIREC
DOI URL: https://doi.org/10.52589/BJCNIT-CQOCIREC

Data Collection

Secondary data were pulled from online open source (IMDB) and used for training the predictive model. The structure of the available data is illustrated below:

Table 1: Structure of data fetched for the predictive model

S/N	VARIABLE NAME	VARIABLE FORMAT	VARIABLE TYPE
1	IMBD id	Characters	Continuous
2	Title	Characters	Continuous
3	Rating	Float	
4	Directors	Characters	Continuous
5	Writers	Characters	Continuous
6	Actors	Characters	Continuous
7	Budget	Int	Continuous
8	Gross	Int	Continuous
9	Financial Score	Float	Continuous
10	Genre	Characters	Categorical

Model Development

Data collection and cleaning are critical input for the successful development of any machine learning model. As data were extracted from online repository, irrelevant variables and features are eliminated. Cleaning of the dataset is aimed at removing inconsistences and other anomalies. Data analysis is done after the pre-processing stage. The data is partitioned into two sets. The main part of the dataset consists of 80% of the data used in the building of the model using appropriate machine learning algorithm. The remaining part of the dataset is used to train the model. Classification algorithms were deployed to make the built model predict the success of upcoming movies.

Selection of Classification Algorithm

Engaging the use of multiple models in Machine Learning to determine the one that works best for any set task (Cai, 2014) is the norm as no one algorithm can work best in any given situation according to the No Free Lunch theorem (NFLT) (Wolpert & Macready, 1997). Application of this rule is usually seen under Supervised Learning as validation and cross-validation are common techniques deployed in assessing the predictive accuracies of multiple models to determine which is best with regards to the task at hand. Training models that work well with multiple algorithms is also good practice. The classification algorithms chosen to train the models include Decision Trees, Gradient Boosting, Naïve Bayes and Multi-layer Perception (MLP).

Model Validation

This measures the accuracy of the data mining system against real data. This is applied to verify if the developed system built on training set actually has the capacity to handle general data (Alpaydin, 2010). Applying model validation ensures that the system avoids over-fitting. Some of the cross-validation techniques that can be used include hold-out method, leave-out validation technique and k-fold cross validation. K-fold validation would be deployed due to its accuracy level as well as its simplicity and wider adoption. It is a technique that involves

Article DOI: 10.52589/BJCNIT-CQOCIREC
DOI URL: https://doi.org/10.52589/BJCNIT-CQOCIREC

reserving a sample dataset that the model is not trained with which is then deployed on the model at the later stages of its development before it is finalized. The steps taken in cross validation are below:

i. Reservation of selected dataset.

ii. Training of the model with the unreserved dataset.

iii. Deployment of the validation set to ascertain the effectiveness of the model's performance.

The K-Fold cross validation method allows the data set to be partitioned into smaller k-subsets and the hold-out method is applied repeatedly 'k' times on the subsets. The other k-1 subsets are put together as a training set each time a k-subset is utilized at a test set. The average error from all set is later derived. The following set of steps is how a single run of k-fold cross validation goes for any set of 'n' training (Ray, 2018):

1. Break datasets into k-fold in a random manner
2. Build a model on k-1 folds of the dataset for each k-fold in the dataset then checking the model for its effectiveness for *kth* fold is done by testing.
3. Errors seen on each prediction is recorded.
4. Repeat 1 through 3 for all k-folds
5. The outcome is the average of the recorded errors and is called cross-validation error. This serves as the performance metric for the model.

Movie Success System Model

The success of a movie is a combination of ratings. Input data for each movie would be a vector of 3 dimensions containing the average score of its writers, actors and directors. The score for each actor, writer, director is a weighted sum of the scores of all the movies each of the dimensions have been involved in directly and indirectly. The predictor calculates the output score of a movie from the input vector using weighted sum of k-nearest neighbours (KNN). The weights is be optimized with cost function built by K-fold cross-validation.

The system model is designed to show an abstract representation of proposed system. That is, a model of major system capabilities. Unified modeling language (UML) tools such as the use case diagram and activity diagram were applied in modeling proposed system. The following Figure 2 illustrates major interactions between the Admin and the system model:

Article DOI: 10.52589/BJCNIT-CQOCIREC
DOI URL: https://doi.org/10.52589/BJCNIT-CQOCIREC

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 4, Issue 2, 2021 (pp. 22-30)
www.abjournals.org

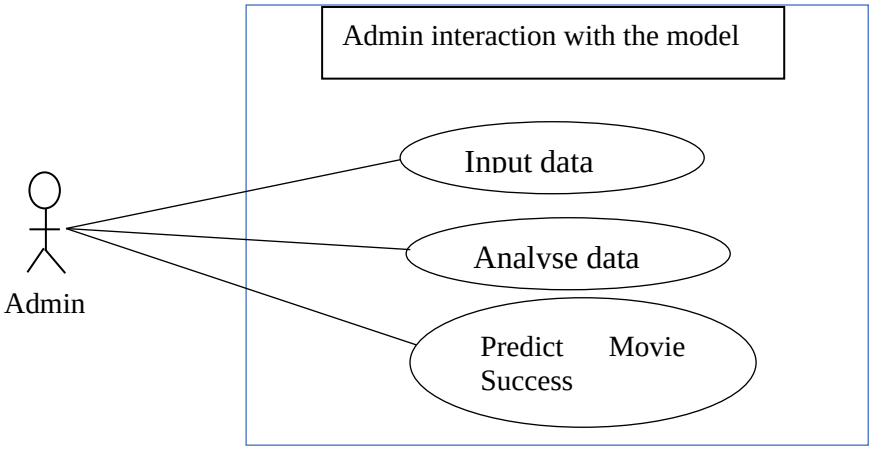

Figure 2: Use case diagram for Admin interaction with the model

The role of the developer is to build an interface that allows the admin to interact with the system. This system is designed for only the admin; there is no need to accommodate other user roles. The admin is capable of making changes and working with the entire system. The goal is to build a predictive model that can be used for movie success prediction and business analysis.

Figure 3 shows the activity diagram to capture the dynamic aspect of the system. This illustrates the dynamic behaviours of administrator's interaction with the Movie Success Prediction system.

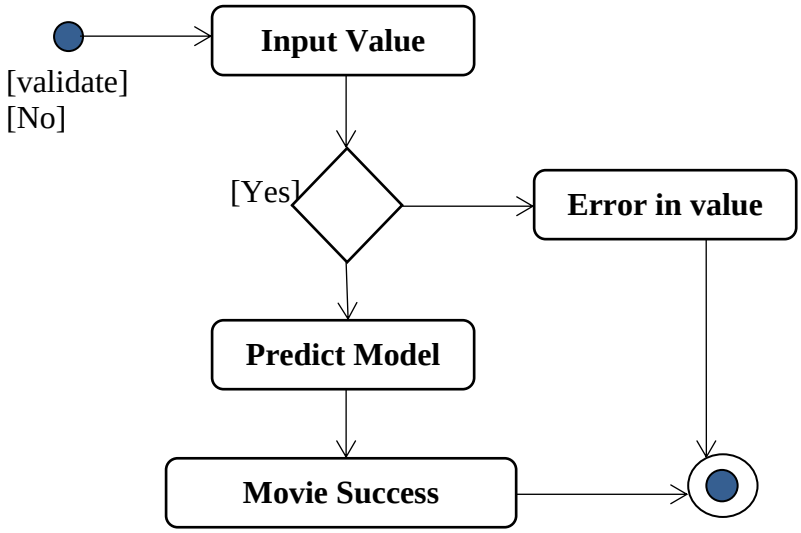

Figure 3 shows the activity diagram to capture the dynamic aspect of the system

Article DOI: 10.52589/BJCNIT-CQOCIREC
DOI URL: https://doi.org/10.52589/BJCNIT-CQOCIREC

RESULT AND DISCUSSION

In the movie success prediction model, a method for reading and learning the data contained in the IMBD dataset was implemented. Another key strategy is the choice of data mining technique used in extracting patterns which is crucial to the prediction process. This is important to data analysis and data evaluation. Existing dataset were analyzed to acquire trends and patterns that could help with the prediction. This included various techniques as data preprocessing and feature reduction which also promoted accuracy. Feature reduction is important in overcoming the cost of dimensionality. Having ascertained the prediction algorithm using a decision tree, the dataset were divided into train and test sets. After many attempts, precision, accuracy and impurity measure were satisfactory. This resulted in ability to predict the category which a movie would fall. This prediction is of great significance to movie stakeholders and producers and with this; they would be able to take the necessary steps to make a movie more profitable or decided against it. Also, the movie industry can use this design to modify the movie criteria for obtaining likelihood of blockbusters. Movie audience, in general, can also use the model to determine if a particular movie is worth it.

CONCLUSION AND RECOMMENDATION

The movie industry is one of the biggest sectors in the world. Audience acceptance of movies is important to both participants; the customers and the business stakeholders. A movie success prediction system capable of predicting the success of a movie is very vital especially in a competitive market like the movie industry. This study has been able address movie success prediction while it could be extended to tackle other related issues in the movie industry. More so, the study is geared towards helping the movie industry and stakeholders in reading the movie success prediction for better intelligent business decisions.

REFERENCES

Alpaydin, E. (2010). *Introduction to machine learning* (3rd ed.). New York: MIT Press.

Baker. R. (2010). Data Mining for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education (3rd edition), 7,* 112-118. Oxford, UK: Elsevier.

Cai, E. (2014). Machine learning of the day: The "no free lunch" theorem. [Online]. Retrieved from www.chemicalstatistician.wordpress.com/machine-learningof-the-day-The-no-free-launch-theorem.

Dhage, S. N. & Raina C. K. (2016). A review on Machine Learning Techniques. *International Journal on Recent and Innovation Trends in Computing and Communication, 4*(3), 395-399.

Greene D., Cunningham P. & Mayer R. (2008) *Unsupervised Learning and Clustering.* In: Cord M., Cunningham P. (eds) Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_3

Grewal, D. S. (2014). A Critical Conceptual Analysis of Definitions of Artificial Intelligence as Applicable to Computer Engineering. *IOSR Journal of Computer Engineering (IOSR-JCE), 16*(2), 9-13.

Article DOI: 10.52589/BJCNIT-CQOCIREC
DOI URL: https://doi.org/10.52589/BJCNIT-CQOCIREC

Han, J., Kamber, M. & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). New York: Morgan Kaufmann.

Kumar, V., Ramakrishnan, G., & Li, Y. (2018). A framework for automatic question generation from text using deep reinforcement learning. ArXiv, abs/1808.04961.

McCarthy, J. (2000). "Review of Artificial intelligence: A General Survey." [Online]. Available: http://wwwformal.stanford.edu/jmc/reviews/lighthill/lighthill.html.

Mesakar, S. & Chaudhari, M. (2013). A Review of Clustering Algorithms. *International Journal of Computer Science and Technology (IJCST), 4*(1), 255-257.

Nithin, V.R, Pranav, *M., Badu, P. B., &* Lijiya, A. (2014). Predicting movie success based on IMDB data. *International Journal of Business Intelligents, 3*(2), 34-36. 10.20894/IJBI.105.003.002.004.

Ogunleye, F. (2004). A Report from the Front: The Nigerian Videofilm. *Quarterly Review of Film and Video, 21*(2), 79-88. DOI: 10.1080/10509200490272991.

Ramageri, B. M. (2010). Data mining techniques and applications. *Indian Journal of Computer Science and Engineering. 1*(4), 301-305.

Ray, S. (2018) Improve Your Model Performance Using Cross Validation (in Python and R). https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r

Shu, X. (2020). *Knowledge Discovery in the Social Sciences: A Data Mining Approach.* Oakland, California: University of California Press.

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K., Elkhatib, …Al-Fuqaha, A. (2019). Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. *IEEE Access, 7,* 65579-65615. 10.1109/ACCESS.2019.2916648.

Wales, L.M. (2017). *The Complete Guide to Film and Digital Production: The People and the Process* (3rd ed.). New York: Routledge. https://doi.org/10.4324/9781315294896.

Wolpert, D. H. & Macready, W. G. (1997). "No free lunch theorems for optimization," in *IEEE Transactions on Evolutionary Computation, 1*(1), 67-82. doi: 10.1109/4235.585893.

Article DOI: 10.52589/BJCNIT-CQOCIREC
DOI URL: https://doi.org/10.52589/BJCNIT-CQOCIREC