International Journal on
Advanced Science
Engineering
Information Technology

Vol.9 (2019) No. 1
ISSN: 2088-5334

Rate Movie App: Implementation of K-Nearest Neighbors Algorithm in the Development of Decision Support System for Philippine Movie Rating and Classification

Ian Dexter M. Siñel[1], Benilda Eleonor V. Comendador[2]

[#] Polytechnic University of the Philippines Graduate School, Sta. Mesa, Manila, 1016, Philippines
E-mail: [1]iandextersinel@outlook.com; [2]bennycomendador@yahoo.com

***Abstract* — Movies that are publicly exhibited in the Philippine Cinema, regardless if produced locally (local films) and/or outside the country (foreign films) undergo a thorough evaluation before public exhibition to properly identify suited audiences. There are many factors that contribute to the classification and rating of a specific movie. Movies play a vital role for Filipino culture as for some people; these serve as their leisure activity, for other people, these are not just a leisure activity instead a form of visual art that may send important messages to the audiences and/or may re-enact human personal experiences. It is very important that movie(s) will be classified accordingly without any form of biases. This paper promotes a Decision Support System that can be used in predicting movie classification and rating using historically evaluated movies from 2010 to 2017. The study considers the user ratings on the following attributes: Sex & Nudity, Violence & Gore, Profanity, Alcohol, Drugs & Smoking and Frightening and Intense Scenes scrapped from a public movie database. Along with these considerations are the genre(s) associated with a movie. The study conducted revealed that K-Nearest Neighbors Algorithm outperforms Naive Bayes and J48/C4.5 Algorithm in classifying Philippine Movie rating with 92.80% accuracy as compared to 68.70% and 56.79% for Naive Bayes and J48/C4.5 algorithm respectively. The developed decision support system implements the K-Nearest Neighbors algorithm to satisfy the objectives mentioned. With this, Review Committees who evaluate movies may have guides in making critical decisions in the domain of movie evaluation.**

***Keywords*— movie application; K-Nearest Neighbors algorithm; C4.5 algorithm; naïve bayes algorithm; decision support system.**

I. INTRODUCTION

Philippine local movies play a significant role in the country’s economic status. The Creative Economy is defined as industries which have their origin in individual creativity, skill, and talent and which have the potential for wealth and job creation through the generation and exploitation of intellectual property. The rising interest in the Philippine Creative Industry contributes to the country’s economic growth [1]. Based on the study by Intellectual Property of the Philippines (IPOPHL) and World Intellectual Property Organization (WIPO), the Creative Industry contributes an estimation of 7.34% in the country’s Gross Domestic Product (GDP) [2]. Also, it helps the employment sector accounted for 14.14% or 560,665 workers in the labor force. Filipinos enjoy watching television programs and local and foreign movies. Films are social mirrors as they provide a media landscape where values and culture may be used for critical thinking. It recreates the experiences of others and even one’s struggles as an individual [3]. Thus, most Filipinos are engaged in such activities.

In the Philippines, many production houses exist which produce quality movies that Filipinos watch (i.e., Star Cinema, GMA Films, Regal Entertainment, Viva Films, Sampaguita Pictures, and OctoArts Films). As such, the Philippine government established an organization which will be responsible for the classification and rating of the Philippine movies subject for showing in the local media and regulate all the products and movies that will be released in public. The Movie and Television Review and Classification Board or locally known as *Lupon sa Pagrerepaso ng Sine at Telebisyon* [4] is a government agency that is responsible for screening, reviewing and examine all motion pictures as herein defined television programs. It includes publicity materials such as advertisements, trailers, and stills, whether such motion pictures and publicity materials be for theatrical or non-theatrical distribution, for television broadcast or general viewing, imported or produced in the Philippines, and the latter case, whether that be for local viewing or export. It is essential to have a proper classification of movies and television programs to make sure that the digital material suits the audiences. The said agency is also mandated to classify motion pictures, television programs and similar shows into categories such as “G” or “For General Patronage” (all ages admitted), “P” or “Parental Guidance Suggested”, “R” or “Restricted” (for adults only), “X” or “Not for Public Viewing”. Based on seven (7) years of data requested from MTRCB [4], it has been found out that it was 2016 when the agency had classified the most number of approved materials with 241 total counts while it was 2013 when a total of 13 movies were rejected or not allowed for public viewing. While ensuring the quality of movies to be shown publicly, MTRCB [4] and the Review Committee still, the agency encountered appeals for reclassification of the movies. For instance, in seven (7) years, a total of 34 movies were reclassified which entail 2% from the 1729 total of classified movies, which required extra effort in performing a second review of the movies being evaluated.

A logical process of searching patterns and useful data through large databases is called Data Mining [5]. The framework that was used in the study is presented in Figure 1 which includes the collection of the data. From the data collected, three important activities were performed.

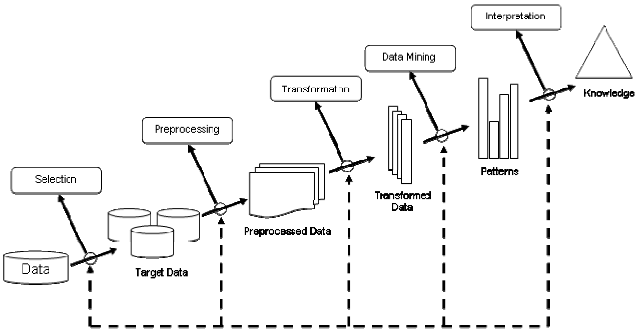


Fig. 1 Data Mining Framework [5]

Exploration, which is the first step that will clean and transform the raw data; next step is Pattern Identification which will identify the patterns existing and choose and select which pattern will possibly make the best prediction; lastly, is Deployment, selected patterns and models will be deployed to get the desired outcome. To execute the framework the following tools are used: The developed Decision Support System used the dataset provided by the Movie and Television Review and Classification Board [4] via the electronic portal under the Philippine Freedom of Information Act (2016). The data obtained from MTRCB [4] was cross-matched against the data scrapped from the publicly available movie database that contains the user ratings and genres from the Internet Movie Database (IMDB) [6]. Waikato Environment for Knowledge Analysis (WEKA) [7] was used to experiment with selecting the best-suited algorithm using the performance rate as its metrics [8]. The selected algorithms included the three (3) data mining algorithms such as C4.5/J48 Decision Tree Classifier, Naïve Bayes Algorithm and IBK (K-Nearest Neighbors). Furthermore, Visual Studio [9] using C#.Net was used to develop the Graphical User Interface (GUI) in which the classifier or the model created was deployed.

Films are believed to be an instrument to express one’s emotion or experience with the help of the digital media and later on will be released to the public. Filipino films come in different forms and types. The principal type of films are categorized into Action Pictures, Social, Historical, Psychological, Comedy, Religious, Theatrical, Documentary, Dramatic, Factual Cartoons, Non-artistic, Pictorial Reports and Travelogues [10].

Presidential Decree No. 1986 is legally known as “Creating the Movie and Television Review and Classification Board.” The Movie and Television Review and Classification Board (MTRCB) [4] is the agency responsible for the rating, reviewing and classifying Movies/Films, other Television programs, and digital materials. According to the Presidential Decree No. 1986, Chapter VI., Section 1, Review Committees will be determined by the Chairman of the Board to which a total of three (3) Board Members will compose the Committee on First Review as indicated in Section 2b of the same article. Furthermore, in Section 2c, it has been clearly stated that the decision will come from the majority vote otherwise the members need to continue deliberating until a majority decision is reached. All decisions will come in writing called Committee Report containing a detailed explanation of the committees’ classification to which shall be made available to the applicant whenever requested. Motion Pictures can be classified with the standard ratings established by the MTRCB [4] including the following: (a) General Audience (“G”); (b) Parental Guidance – 13 (“PG”); (c) Restricted – 13 (“R-13”); (d) Restricted – 16 (“R-16”); (e) Restricted – 18 (“R-18”); and lastly, (f) Not for Public Exhibition (“X”). Moreover, the agency implements the use of the following icons as per movie classification ratings which are presented in Figure 2.

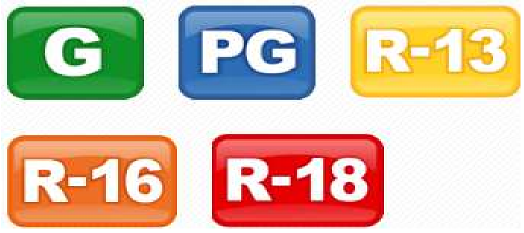


Fig. 2 MTRCB Movie Classification Rating Icons

Data Mining is no longer new in the field of Information Technology. Data Mining is believed to be a powerful tool that automatically summarizes data, extracts useful information and discover any existing interesting patterns from raw data [11]. It is the process of digging an added value to the data collected in the past. Data Mining plays around databases that contain the complex and large amount of data. The purpose of data mining is to find connections or patterns that may provide useful indications [12]. Knowledge Discovery in Databases (KDD) is another term that is often referred to as Data Mining that seeks to uncover patterns and predictive information. Data mining is a very useful tool as it can be used in a wide range of dataset depending on its purpose thus which includes the following: (1) Exploratory Data Analysis that examines data distribution, outliers and anomalies [13]; (2) Descriptive Modeling which summarizes a given data [14]; (3) Predictive Modeling which involves finding mathematical relationships that exists with the goal of identifying future values of the target variable [15]; (4) Discovering Patterns and Rules that finds unusual behavior where the data points are significantly different from the rest [11] and lastly, (5) Retrieval by Content. From the time being, an increasing interest in the study of using data mining technologies in different domains due to its flexibility and powerful approaches that helps support businesses. Flexibility of data mining technologies help serve different functions including (1) Classification which aims to generate models that can be used for classifying a data item; (2) Regression which aims to map a specific data into a prediction variable; (3) Clustering that helps create different categories that may describe a data item; (4) Dependency Modeling or Association Rule Learning, that analyzes significant relationship from one variable to another; (5) Deviation Detection which uncovers suspicious changes from the data or also known as Anomaly Detection; and lastly, (6) Summarization that aims to create a statement that will describe the data [16].

In general, to successfully perform data mining on large databases, a set of defined steps can be followed i: e Exploration, Pattern Identification, and Deployment. Exploration is the step that aims to preprocess and transform raw data into another form while Pattern Identification tries to form a pattern and identifies the pattern that creates the best result and Deployment which is the actual deployment of the selected pattern for the desired outcome [17]. Many data are now publicly available in different forms, structures, and formats. Data may come as supervised or unsupervised by its nature. Supervised data means that an identified class is already associated with the data otherwise unsupervised data.

Given that we have any data stored, we can use different data mining techniques or algorithms to perform a specific task depending on the data available which may include the prediction task. There are two (2) types of data mining methods: (1) Eager Learner in which, from the given training set, the algorithm constructs a classification model before receiving a new data to classify. Few to mention are the Decision Trees, Support Vector Machine (SVM), Neural Networks (NN) and Naïve Bayes Algorithm; and Lazy Learners, whereas it simply stores data and waits until it is given a new data to classify. The most common and most popular algorithm under Lazy Learners is the K-Nearest Neighbors algorithm [18]. Another data mining algorithm is K-Nearest Neighbors or k-NN. k-NN is a lazy classifier that works by measuring the distance between the query instance and every instance in the training data set; it finds the k^{th} training instance which has the least distance from the query instance [19]. This algorithm aims to create a classifier for the new data based on the attributes associated with the training set of samples [20].

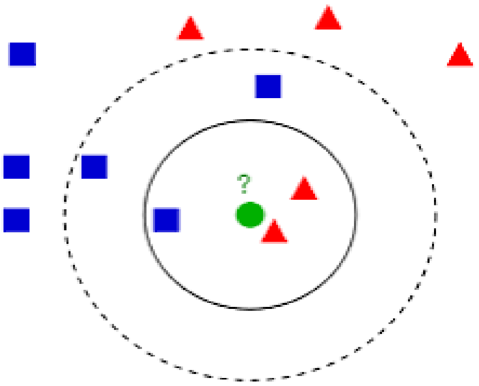
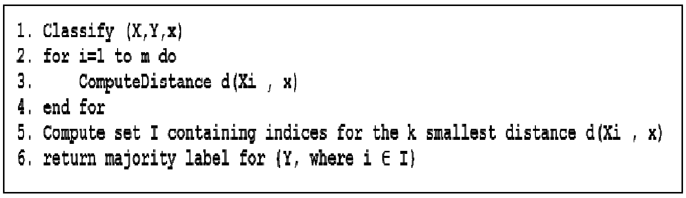


Fig. 3 K-Nearest Neighbors Algorithm Graphical Representation

The algorithm uses the following parameters, k, which represents the number of the nearest neighbors to be considered; d_{ij}, distance measure, which measures the similarity of the query instance to the training instances; query instance, refers to the object or data item that needs to be classified. From Figure 3, the "?" in this case, the circular shape, refers to the query instance. The remaining geometric shapes (square and triangle) are the available classes in the training set. The first inner circular dashed line represents the k-nearest neighbors using k value set to 1. The second circular dashed line signifies the k-nearest neighbors using k value set to 2. The k value can be set to any value less than or equal to 1. The pseudo code of the K-NN algorithm [21] is reflected in Figure 4 below.



```
Classify (X,Y,x)
for i=1 to m do
    ComputeDistance d(Xi , x)
end for
Compute set I containing indices for the k smallest distance d(Xi , x)
return majority label for {Y, where i ∈ I}
```

Fig. 4 K-Nearest Neighbors Algorithm Pseudo code

Whereas:

X refers to the training data;
Y refers to the class labels of *X*; and
x refers to the unknown sample

The prediction algorithm identifies and classifies a new vector input *x* by examining the taking the most frequently occurring class that is nearest on the k-value set from the training sample data points [22]. Since that many data about movie analysis are publicly available on the internet, many, studies have been recently conducted to assess what can be done with these data. A study developed methodologies for automatic movie rating prediction [23] using the IMDB [6] database. It used baseline methods, content-based methods (e.g., Regression Trees, Neural Networks), collaborative methods (e.g., kNN, Latent Semantic Analysis) and hybrid methods (SVD-kNN). Among the mentioned methods, the probabilistic latent semantic analysis outperforms other methodologies presented [23]. Another study was conducted in movie analysis which used social networks in improving movie rating predictions by implementing the most common collaborative filtering technique, kNN in the development of a recommender system [24].

Further researches and advances have been completed with interest in movie data mining. As a proof of this, another study was conducted that built a recommender system and gave recommendations taking the user's preferences into consideration by training models using combined Latent Factor, K-Nearest Neighbors, and Binary Decision Tree Regression. It is concluded that the hybrid model showed better performance than the solely content-based model taking into consideration users and movie features [25].

II. MATERIAL AND METHOD

The research explicitly used the data collected from primary and secondary sources. Primary data sources included the movie data provided by MTRCB [4]. Additionally, the IMDB [6] database was also used which provided the user ratings for each movie. Figure 5 displays the sources of primary data that were used in the study which depicted that the dataset was built by combining the data obtained from MTRCB [4] and included the list of movies with corresponding ratings (G, PG, R-13, R-16, R-18 and X) and the data obtained from IMDB (Internet Movie Database) [6] which included the Movie Criteria used by the users when rating a movie and the associated genre(s).

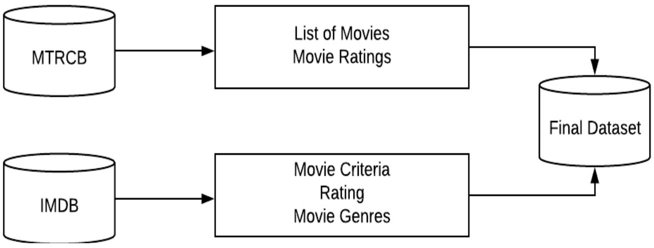


Fig. 5 Dataset Sources

Technical and conference documents, articles, discussions, and reports regarding decision support systems, data mining and algorithms were used as the secondary sources of data. Using these sources, the data set for the developed decision support system was built. Furthermore, all related literature, studies, and articles were used as the guideline throughout the development of the research. Interview with the experts was initially conducted to have an overview of the existing practices and workflow of assessing Philippine movies. To further have an idea on the technology that was used in the study, reading of related literature and studies was conducted. This helped the researchers to gain more insights into what is needed and what needs to be done to complete the study. Technical requirements were derived from this and were very helpful in the development of the proposed Decision Support System. Additionally, to conduct the experiment in the study, the dataset obtained was divided which was used as the Training and Testing dataset. A total of 1799 instances of movie data were divided into training, and testing data set displayed in Table 1.

TABLE I
TRAIN AND TEST DATASET PER CLASS

Class	Total Available Data	Train (80%)	Test (20%)
G	274	219	55
PG	738	590	148
R-13	353	282	71
R-16	132	106	26
R-18	299	239	60
X	3	2	1
Total	1799	1439	361

From Table 1, it can be seen that the total Training Dataset is 1439 instances from the original dataset and the remaining were used as the Test Dataset with a total of 361instances. The instances for training and testing dataset were assigned with unique numerical and incrementing ID's and were used to randomly select data from the original dataset. Each of the Training and Testing datasets shared the same attributes which are enumerated in Table 2 describing the data types and the possible values. Attribute ID numbers 1 to 32 were all sourced from IMDB [6] while Attribute ID number 33 was sourced from MTRCB [4].

TABLE II
DATASET ATTRIBUTES DESCRIPTION

ID	Attribute Name	Data Type	Possible Values
1	A01_SexNudity	Nominal	1 – None 2 – Mild 3 – Moderate 4 – Severe
2	A02_ViolenceGore	Nominal	
3	A03_Profanity	Nominal	
4	A04_Alcohol, Drugs Smoking	Nominal	
5	A05_FrighteningIntenseScenes	Nominal	
6	A06_Action	Binary	0 – False 1 – True
7	A07_Adventure	Binary	
8	A08_Animation	Binary	
9	A09_Biography	Binary	
10	A10_Comedy	Binary	
11	A11_Crime	Binary	
12	A12_Documentary	Binary	
13	A13_Drama	Binary	
14	A14_Family	Binary	0 – False 1 – True
15	A15_Fantasy	Binary	
16	A16_FilmNoir	Binary	
17	A17_GameShow	Binary	
18	A18_History	Binary	
19	A19_Horror	Binary	
20	A20_Musical	Binary	
21	A21_Mystery	Binary	
22	A22_News	Binary	
23	A23_RealityTV	Binary	
24	A24_Romance	Binary	
25	A25_SciFi	Binary	
26	A26_Short	Binary	
27	A27_Sport	Binary	
28	A28_Superhero	Binary	
29	A29_TalkShow	Binary	
30	A30_Thriller	Binary	
31	A31_War	Binary	
32	A32_Western	Binary	

33	A33_Rating	Nominal	1 – G 2 – PG 3 – R13 4 – R16 5 – R18 6 – X

The study used a historical dataset of the classified movies. The dataset was split using the 80-20 rule. To further perform the experiment needed in the study, the following formula was used to derive the total number of instances for training and testing set.

$$Yi = ABS(Xi * 0.80) \quad (1)$$

$$Zi = ABS(Xi - Yi) \quad (2)$$

Where X_i refers to the total number of available instances for i^{th} class label; Y_i refers to the total number of instances for the Training Set which can be computed by getting the absolute value of 80% (0.80) from the total number of instances available for the i^{th} class (X_i); and Z_i refers to the total number of instances for the Training set which can be computed by getting the absolute value of total number of instances available for the i^{th} class minus the total number of training set instances for the i^{th} class (Y_i). The analysis of data was done using MS Excel [26] and the Waikato Environment for Knowledge Analysis (WEKA) [7]. These tools were used for statistical analysis and model/classifier evaluation. It helped the researchers explore the data available and transform nominal or categorical data into numerically encoded data so that a model can be easily built from the available data. To further experiment, the following formulas were used to evaluate the Accuracy Rate (% Accuracy), Error Rate (% Inaccuracy), TP Rate, FP Rate, Precision, and Recall.

%Accuracy, which refers to the prediction accuracy of the model which can be computed as the total number of all correct classification and prediction, divided by the size of the Testing Class. The best % Accuracy Rate is 100%;

$$\%\ \text{Accuracy} = \frac{\#\ of\ Correct\ Classification}{Size\ of\ Testing\ Class} \times 100\% \quad (3)$$

% Error, which refers to the probability of error of the model or classifier, can be calculated as the total number of incorrect classification divided by the total number of records or instances from the testing class. A value of 0% describes the best % Error Rate.

$$\%\ \text{Error} = \frac{\#\ of\ Incorrect\ Classification}{Size\ of\ Testing\ Class} \times 100\% \quad (4)$$

To further perform a deep dive analysis on the results, **TP Rate (True Positive Rate) or Recall** can be calculated as the number of correct positive classification divided by the total number of positives (True Positives & False Negatives combined). It is also called as Sensitivity whereas the best value is 1.0;

$$\text{TP Rate or Recall} = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \quad (5)$$

FP Rate or False Positive Rate can be computed as the total number of incorrect positive classification divided by the total number of negatives or the sum of True Negatives (TN)and False Positive (FN). It is often referred to as Specificity to which a value of 0.0 means best FP Rate.

$$\text{FP Rate} = \frac{False\ Positive\ (FP)}{True\ Negative\ (TN) + False\ Positive\ (FP)} \quad (6)$$

Precision or Positive Predictive Value (PPV) can be calculated as the total number of correct positive classifications or predictions divided by total number of positive predictions (True Positives and False Positives combined) whereas the best value is 1.0;

$$\text{Precision} = \frac{True\ Positive\ (FP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \quad (7)$$

Additionally WEKA [7] was used to evaluate the classification performances of the mentioned algorithms (J48/C4.5, Naïve Bayes and K-Nearest Neighbors Algorithm). Whichever algorithm gives the best results; the model was saved and loaded to the decision support system. Additionally, the System Architecture which was used in the study is portrayed in Figure 6.

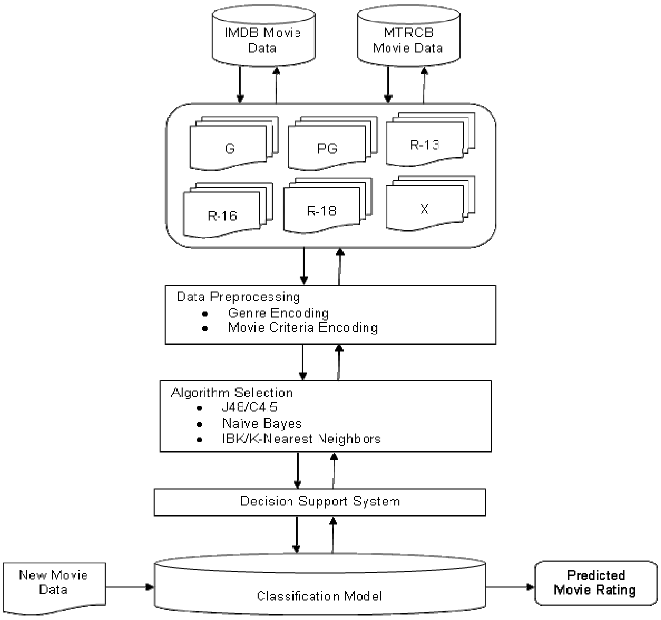
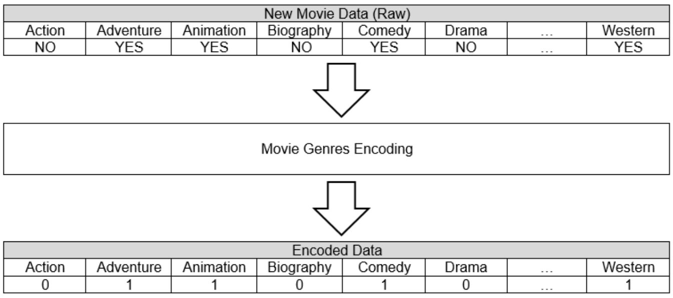


Fig. 6 System Architecture of the Developed Decision Support System

The study made use of the historical dataset of Classified Philippine Movies. The collected data passed through different stages in order to build the classifier which was ingested in the developed Decision Support System to predict and classify class labels of new unseen movie data. The System Architecture is broken down into the following components:

Data Collection. Raw data about the movie data was collected. The first step was to collect a list of movies classified by MTRCB [4] with its corresponding rating or classification. The list obtained from MTRCB [4] was cross-matched to the IMDB [6] database to get other movie data especially the movie genre(s) and the movie criteria rating;

Text Processing. This stage was used to pre-process and transform raw data. There are two (2) activities included in this stage including (a) Genre Encoding, in which raw genre text is encoded as binary values. In this study, the following Genres were used: Action, Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, Film-Noir, Game-Show, History, Horror, Musical, Mystery, News, Reality-TV, Romance, Science Fiction (Sci-Fi), Short, Sport, Superhero, Talk-Show, Thriller, War and Western. Each new movie data can be assigned or associated with multiple genres. The Genre Encoding activity is displayed in Figure 7;



New Movie Data (Raw)

Action	Adventure	Animation	Biography	Comedy	Drama	...	Western
NO	YES	YES	NO	YES	NO	...	YES

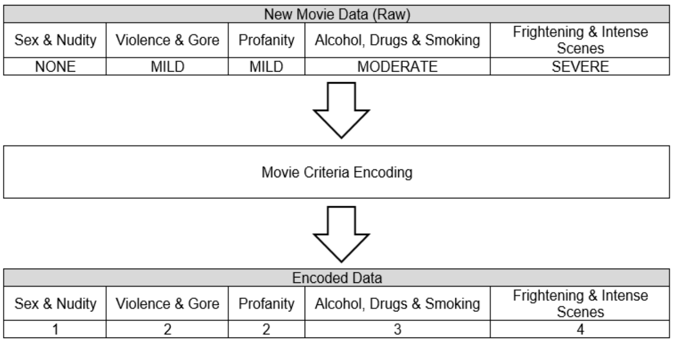
Movie Genres Encoding

Encoded Data

Action	Adventure	Animation	Biography	Comedy	Drama	...	Western
0	1	1	0	1	0	...	1

Fig. 7 Movie Genres Encoding

From Figure 7, it can be seen that the Movie Genres Encoding is responsible for encoding texts of new movie data into binary values. Genres with zero (0) values indicate that a specific genre is not depicted in the movie, otherwise one (1). Another important preprocessing activity is the (b) Movie Criteria Encoding which translates the raw data into a numerical value. The study used five (5) criteria in assessing the movie as reflected in Figure 8 including Sex & Nudity, Violence & Gore, Profanity, Alcohol, Drugs & Smoking, and Frightening & Intense Scenes. Each new movie data was encoded and converted to a numerical value indicating 1 as "None" which means there is no depiction of the criteria in the movie; 2 as "Mild", which means that little by little, the criteria is depicted in the movie; 3 as "Moderate" which means a moderate and frequent depiction of criteria in the movie is exhibited; and 4 as "Severe" which means that there is a strong depiction of the criteria in the movie.



New Movie Data (Raw)

Sex & Nudity	Violence & Gore	Profanity	Alcohol, Drugs & Smoking	Frightening & Intense Scenes
NONE	MILD	MILD	MODERATE	SEVERE

Movie Criteria Encoding

Encoded Data

Sex & Nudity	Violence & Gore	Profanity	Alcohol, Drugs & Smoking	Frightening & Intense Scenes
1	2	2	3	4

Fig. 8 Movie Criteria Encoding

After the raw data were encoded, the full dataset was split into 80% training dataset and 20% training data set which were fed to WEKA [7] in modeling the classifier and testing the model respectively.

Algorithm Selection. In this stage, three (3) data mining techniques' classification performance were compared with each other to identify which one best works with the given dataset. The algorithm that obtained the highest classification accuracy was used as the intelligent component in building the classifier and model for the Decision Support System; and

Decision Support System. The Decision Support System was developed using C#.Net. It used the classification model generated built using the most accurate algorithm identified from the previous stage. The decision support system accepted new movie data. From the new unseen movie data, the decision support system classified the movie giving a probability that the given input data belonged to a specific class.

III. Results and Discussion

A. The Evaluated Performance of three (3) Data Mining Algorithms in Classifying Movies

An experiment was conducted to identify which among the three data mining algorithms worked best in classifying new movie data. From this point, a model was built per data mining algorithm and evaluated each classification performances. Each model used the same data set and was run and simulated using WEKA [7]. The experiment used the dataset split for Training and Testing dataset which was previously presented in Table 1. Table 3 presents a summary of the comparison of the three data mining algorithms in the classifying Philippine movie.

TABLE III
SUMMARY OF COMPARISON OF THE THREE DATA MINING ALGORITHM IN CLASSIFYING PHILIPPINE MOVIE

Algorithm	% Accuracy	% Error	Rank
Naive Bayes	68.70	31.30	2
J48/C4.5	56.79	43.21	3
K-NN	92.80	7.20	1

Using Table 3, it can be observed that among the three (3) data mining algorithms, K-Nearest Neighbors Algorithm outperforms J48/C4.5 Algorithm and Naïve Bayes Algorithm. K-Nearest Neighbors Algorithm obtained an accuracy rate of 92.80% followed by 68.70% of Naïve and 56.79% from J48/C4.5. To further check the performance, a breakdown of classification performance of the algorithms per class with their corresponding confusion matrices are presented in Table 4 – 9.

TABLE IV
NAIVE BAYES ALGORITHM CLASSIFICATION PERFORMANCE PER CLASS

Class	TP Rate	FP Rate	Precision	Recall
G	0.745	0.052	0.719	0.745
PG	0.709	0.235	0.677	0.709
R-13	0.324	0.121	0.397	0.324
R-16	0.038	0.018	0.143	0.038
R-18	0.583	0.156	0.427	0.583
X	0.000	0.006	0.000	0.000

Table 4 presents that using Naïve Bayes Algorithm, the class that obtained the highest TP Rate is class G equivalent to 0.745 followed by PG, R-18, R-13, R-16 and X with a TP Rate of 0.709, 0.583, 0.324, 0.038 and 0 respectively. This can be interpreted that Naïve Bayes can have good predictions to G and PG-rated movies only. Using the values for FP Rate, it can be seen that the order of which the classes attained the highest FP Rate starts with class PG with a 0.235 followed by R-18, R-13, G, R-16 and X with FP Rate equivalent to 0.583, 0.324, 0.745, 0.038 and 0 respectively which means that class PG obtained the highest rate for False Positives or incorrect positive predictions among others. Moreover, to further understand the distribution of classification per each class by Naïve Bayes, a confusion matrix is presented in Table 5.

TABLE V
NAIVE BAYES ALGORITHM CONFUSION MATRIX

Actual Rating	Predicted Rating					
	G	PG	R13	R16	R18	X
G	**41**	13	0	0	1	0
PG	13	**105**	16	0	13	1
R-13	3	24	**23**	3	18	0
R-16	0	3	8	**1**	14	0
R-18	0	10	11	3	**35**	1
X	0	0	0	0	1	**0**

Table 5 depicts that Naïve Bayes correctly classified 41 G-rated instances, 105 PG-rated instances, 23 R-13 rated instances, only 1 R-16 rated instances, 35 R-18 rated instances and no correct predictions for X-rated instances.

TABLE VI
J48/C4.5 ALGORITHM CLASSIFICATION PERFORMANCE PER CLASS

Class	TP Rate	FP Rate	Precision	Recall
G	0.764	0.033	0.808	0.764
PG	0.838	0.277	0.678	0.838
R-13	0.535	0.090	0.594	0.535
R-16	0.346	0.000	1.000	0.346
R-18	0.583	0.060	0.660	0.583
X	0.000	0.000	0.000	0.000

Using Table 6 which presents the J48/C4.5 Algorithm Performance, it can be seen that the class that obtained the highest TP Rate is class PG equivalent to 0.838 followed by G, R-18, R-13, R-16 and X with a TP Rate of 0.764, 0.583, 0.535, 0.346 and 0 respectively. This can also be interpreted that J48/C4.5 can have good predictions to PG and G rated movies only as the remaining classes did not have good performance prediction. Using the values for FP Rate, it can be seen that the order of which the classes obtained the highest FP Rate starts with class PG with a 0.277 followed by R-13, R-18, G, R-16 and X with FP Rate equivalent to 0.535, 0.583, 0.764, 0.346 and 0 respectively which means that class PG has obtained the highest rate for False Positives or incorrect positive predictions among others. Moreover, to further understand the distribution of classification per each class by the J48/C4.5 Algorithm, a confusion matrix is presented in Table 7 below.

TABLE VII
J48/C4.5 ALGORITHM CONFUSION MATRIX

Actual Rating	Predicted Rating					
	G	PG	R13	R16	R18	X
G	**42**	12	0	0	1	0
PG	9	**124**	9	0	6	0
R-13	1	28	**38**	0	4	0
R-16	0	4	6	**9**	7	0
R-18	0	14	11	0	**35**	0
X	0	1	0	0	0	**0**

Table 7 shows that upon using J48/C4.5, class PG had the highest correct prediction with a total of 124 instances followed by G with 42 correct predictions, R-13 with 38 correct predictions, R-18 with 35 and class R-16 and class X with 9 and 0 correct predictions respectively.

TABLE VIII
K-NEAREST NEIGHBORS ALGORITHM CLASSIFICATION PERFORMANCE PER CLASS

Class	TP Rate	FP Rate	Precision	Recall
G	0.764	0.033	0.808	0.764
PG	0.838	0.277	0.678	0.838
R-13	0.535	0.090	0.594	0.535
R-16	0.346	0.000	1.000	0.346
R-18	0.583	0.060	0.660	0.583
X	0.000	0.000	0.000	0.000

K-Nearest Neighbors Algorithm classification performance per class is depicted in Table 8. Using K-Nearest Neighbors algorithm, the computed TP Rate or True Positive Rate for each class are as follows, class G with 0.927, PG with 0.953, R-13 with 0.930, R-16 with 0.808, R-18 with 0.917 and class X with a TP rate of 1. From the results of computation for TP Rate, using it as an evaluation metrics, it can be seen that all classes had a good, correct prediction. Additionally, the computed False Positive Rate or FP Rate for K-Nearest Neighbors can also be seen in Table 14, of which, class X obtained an FP Rate of 0 which means that K-Nearest Neighbors had no error in predicting in X-rated movies. The remaining FP rate for each class is presented with R-16 with 0.006, 0.010 for R-18, R-13 with 0.014 and G with 0.026 and PG with 0.042 FP Rate. The FP Rate dictates the rate to which the built model or classifier has errors in prediction. As per interpretation, the closer the computed value for the FP rate to 0 means good predictions. Moreover, to further understand the distribution of classification per each class by the KNN Algorithm, a confusion matrix is presented in Table 9.

TABLE IX
J48/C4.5 ALGORITHM CONFUSION MATRIX

Actual Rating	Predicted Rating					
	G	PG	R13	R16	R18	X
G	**51**	4	0	0	0	0
PG	7	**141**	0	0	0	0
R-13	0	3	**66**	1	1	0
R-16	1	1	1	**21**	2	0
R-18	0	1	3	1	**55**	0
X	0	0	0	0	0	**1**

Table 9 depicts the prediction distribution for each class using the K-Nearest Neighbors algorithm from the Testing set with a total of 361 instances. It can be observed that class G had a total of 51 correct predictions out of 55 instances from the training set, 141 correct predictions out of 148 instances for PG-rated movies, 66 correct predictions out of 71 instances for R-13 rated movies, 21 correct predictions out of 26 instances for R-16 rated movies, R-18 rated movies with 55 correct instances out of 60 instances and lastly for X rated movies with 1 correct prediction out of 1 instance. Aggregating the results, the total number of correct prediction is 335 instances out of 361 total sizes of the testing set giving an accuracy rate of 92.80%

IV. Conclusions

Based on the experiment, K-Nearest Neighbors worked well with the Philippine movie dataset. K-Nearest Neighbors outperformed J48/C4.5 and Naïve Bayes algorithm. This showed as well that the K-Nearest Neighbors algorithm worked well with datasets with full binary data and nominal data. Also based on the results obtained from the Pre and Post Survey conducted, it can be concluded that the respondents from the time that the study was conducted, maximized the full potential of hardware technology available but had limited and few utilization into software technology. From this point forward, it can be said that the software technology might mean that there is no available software or decision support system that helped the respondents automatically classify and rate Philippine movies.

Moreover, the developed Decision Support System has been evaluated with features that made it acceptable for the respondents. Based on the results, it revealed that the respondents perceived that the developed decision support system was usable, functional, efficient, portable and reliable. The researchers recommend performing exploratory analysis of different genre included and removing unnecessary genre that may or may not help the classification of a movie. Experiment with trying to increase the accuracy rate of the three data mining algorithms by getting the relevance of a specific attribute by using Information Gain, Gain Ratio and Chi-Square. As the study focused on predicting class labels of Philippine movies only, include predicting rating of television programs and series, Television commercials, advertisements, theatrical plays, and other related public materials. Have movie posters evaluated as well during the assessment of the movie which might have hidden additional useful information that can be used in predicting the class label of a movie? Explore the possibility of correlating the cast, crew, production house and directors of the movie upon assessing rating of a movie.

References

[1] B. Dalisay. (2014) The Wealth Within Us. [Online]. Available: http://www.philstar.com/arts-and-culture

[2] L. Desiderio. (2015) DTI Seeks Investments in Creative Industries. [Online]. Available: http://www.philstar.com:8080/business/2012/07/15/827808/dti-seeks-investments-creative-industries

[3] I. E. Valera. (2015) “Perceived Status of the Filipino Film Industry: Implications for Media Education” International Conference in Language Learning and Teaching at HCT Men’s College UAE, vol. 8(1), pp. 85-89.

[4] (2018) the MTRCB Official Website. [Online]. Available: http://www.mtrcb.gov.ph

[5] S. Rasheedudin. (2013) “The Theoretical Framework of Data Mining and Its Techniques” International Journal of Social Science, vol. 2(1), pp. 81-85. [Online]. Available: http://www.indianresearchjournals.com/pdf/IJSSIR/2013/January/9.pdf

[6] (2018) Internet Movie Database Official Website. [Online]. Available: http://www.imdb.com

[7] (2018) Waikato Environment for Knowledge Analysis Official Website. [Online]. Available: http://www.cs.waikato.ac.nz/weka/

[8] F. Eibe. (2016) Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kau). [Online]. Available: ftp://ftp.ingv.it/pub/manuela.sbarra/Data Mining Practical Machine Learning Tools and Techniques - WEKA.pdf

[9] (2018) Visual Studio Express Official Website. [Online]. Available: http://www.visualstudio.microsoft.com

[10] L. Garcia and C. Masigan. (2001). An In-depth Study on the Film Industry in the Philippines. [Online]. Available: https://www.dirp4.pids.gov.ph/ris/taps/tapspp0103.pdf

[11] S. Deshpande and V. Thakare. (2010) “Data Mining System and Applications: A Review” International Journal of Distributed and Parallel Systems, vol. 1(1), pp. 32-44. [Online]. Available: https://doi.org/10.5121/ijdps.2010.1103

[12] L. Marlina, M. Lim and A. P. Utama Siahaan. (2016) “Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms) International Journal of Engineering Trends and Technology, vol. 38(7), pp. 380-383. [Online]. Available: https:doi.org/10.14445/22315381/IJETT-V38P268

[13] M. Komorowski, D. Marshall, J. Salciccioli and Y. Crutain. (2016) “Exploratory Data Analysis” Research Gate DOI: 10.1007/978-3-319-43742-2_15. pp. 185-203

[14] D. Madigan. (2010) Descriptive Modeling. [Online]. Available: http://www.stat.columbia.edu/~madigan/DM08/descriptive.ppt.pdf

[15] F. Halili and A. Rustemi. (2016) “Predictive Modeling: Data Mining Regression Technique Applied in Prototype” International Journal of Computer Science and Mobile Computing, vol. 5(8), pp. 207-215

[16] T. Silwattananusarn and K. Tuamsuk. (2012) “Data Mining and Its Application for Knowledge Management: A Literature Review from 2007 to 2012” International Journal of Data Mining and Knowledge Management Process (IJDKP), vol. 2(5), pp. 13-24. [Online]. Available: https://doi.org/10.5121/ijdkp.2012.2502

[17] M. Ramageri. (2010) “Data Mining Techniques and Its Application” Indian Journal of Computer Science and Engineering, vol. 1(4), pp. 301-305

[18] F. Solomon, D. Abebe, and S. Bhabani. (2014) “A Comparative Study on Performance Evaluation of Eager versus Lazy Learning Methods” International Journal of Computer Science and Mobile Computing, vol. 3(3), pp. 562-568

[19] A. Kulkarni and R. Maclin. (n.d.) The Nearest Neighbor Approach using Clustering in Netflix Prize Data. [Online]. Available: https://dirp4.pids.gov.ph/ris/taps/tapspp0103.pdf

[20] K. Teknomo. (2010) K-Nearest Neighbors Tutorial. [Online]. Available: https://people.revoluedu.com/kardi/tutorial/KNN

[21] B. Tay, J. K. Hyun, and S. Oh. (2014) “A Machine Learning Approach for Specification of Spinal Cord Injuries using Fractional Anisotropy Values Obtained from Diffusion Tensor Images” Computational and Mathematica Methods in Medicine, Hindawi Publishing Corporation, vol. 2014. [Online]. Available: https://doi.org/10.1155/2014/276589

[22] D. Sontag. (n.d.) Nearest Neighbor Methods. [Online]. Available: http://people.csail.mit.edu/dsontag/courses/ml13/slides/lecture11.pdf

[23] M. Marovic, M. Mihokovic, M. Miksa, S. Pribil, and A. Tus. (2011) “Automatic Movie Rating Prediction using Machine Learning” in 2011 Proceedings of the 34th International Convention MIPRO, pp. 1640-1645

[24] S. P. Sahu. (2017) “Machine Learning Algorithms for Recommender System – A Comparative Analysis” International Journal of Computer Applications Technology and Research, vol. 6(2), pp. 97-100

[25] Y. Peng, Y. Duan and Z. Zou. (n.d.) Movielens: Several Approaches to Rating Prediction. [Online]. Available: https:// https://dirp4.pids.gov.ph/ris/taps/tapspp0103.pdf

[26] (2018) MS Excel Official Website. [Online]. Available: https://www.office.live.com/start/Excel.aspx