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 ith 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 ith class minus the total number of training set instances for the ith 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%;

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

% 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.

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

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.

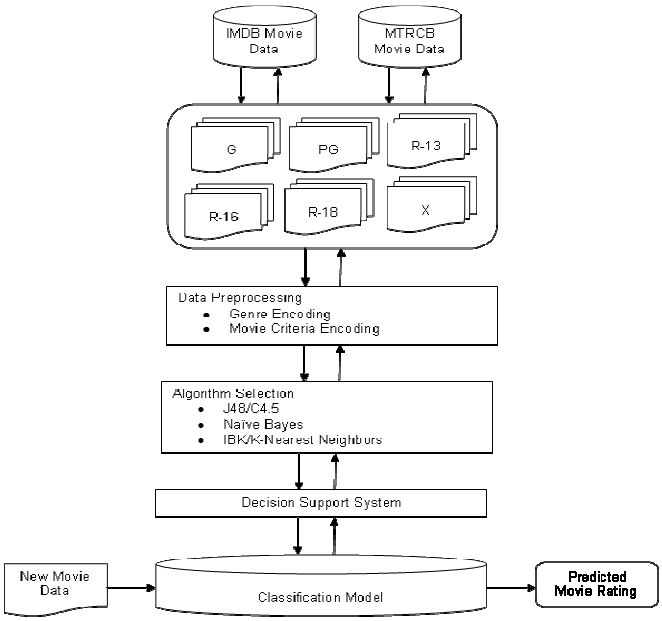

Fig. 6 System Architecture of the Developed Decision Support System

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;

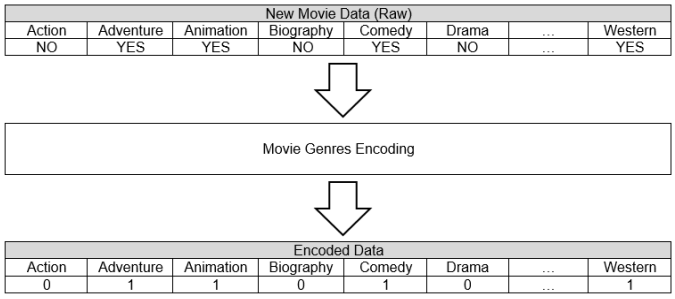

Fig. 7 Movie Genres Encoding

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.

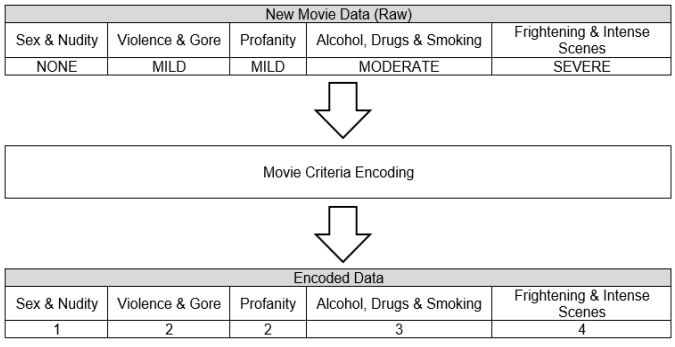

Fig. 8 Movie Criteria Encoding

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