Single and Ensemble Classification for Predicting User's Restaurant Preference

Esra'a Alshdaifat[1]
Department of Computer Information System
The Hashemite University
Zarqa, Jordan

Ala'a Al-shdaifat[2]
Department of Computer Science
The Hashemite University
Zarqa, Jordan

***Abstract*—Classification is one of the most attractive and powerful data mining functionalities. Classification algorithms are applied to real-world problems to produce intelligent prediction models. Two main categories of classification algorithms can be adopted for generating prediction models: Single and Ensemble classification algorithms. In this paper, both categories are utilized to generate a novel prediction model to predict restaurant category preferences. More specifically, the central idea espoused in this paper is to construct an effective prediction model, using Single and Ensemble classification algorithms, to assist people to determine the best relevant place to go based on their demographic data, income level and place preferences. Therefore, this paper introduces a new application of classification task. According to the reported experimental results, an effective Restaurant Category Preferences Prediction Model (RCPPM) could be generated using classification algorithms. In addition, Bagging Homogeneous Ensemble classification produced the most effective RCPPM.**

***Keywords*—Classification; data mining; ensemble algorithms; restaurant preferences**

I. Introduction

With the increasing accessibility of innumerable data collections, the extraction of interesting patterns from such data becomes a necessity. Data mining involves extracting interesting and helpful patterns from enormous amount of data [1]. Classification is a well-known data mining functionality that refers to the process of generating a prediction model and using it to predict categories for new unseen samples. More specifically, classification can be considered as a three-step process. The first step commences with generating the prediction model using the "training" dataset that comprises a set of samples, where each sample is associated with a categorical class label. The classification problems can be differentiated according to: (i) the number of class labels featured in the dataset and (ii) the number of the class labels associated with each sample in the dataset. With respect to the number of labels featured in the dataset two kinds of classification problems can be recognized: binary and multi-class classification problems. In binary classification problems, the considered dataset includes only two labels, while more than two labels featured in the multi-class classification problems. Regarding the number of labels associated with each sample in the dataset, also two types of classification problems can be distinguished: single-label and multi-label classification problems. When each sample in the dataset is associated with exactly one label then we have a single-label classification process. Whilst, if several labels can be associated with one sample then we have a multi-label classification process. Several classification algorithms can be utilized to produce the prediction model for each classification problem. After generating the prediction model, the next step is the evaluation in which the performance of the generated prediction model is assessed to determine its applicability to be used for predicting class labels for new samples. Several measures can be used to evaluate prediction models effectiveness; accuracy and Area Under the ROC Curve (AUC) are the most widely used measures [2], [3]. Based on the values obtained from evaluation measures, a decision can be drawn regarding whether or not to utilize the model for future prediction. The last step in the classification process is the model usage, where the prediction model is utilized to predict class labels for new unseen data. Classification has been employed in many application domains, examples of application domains include: text categorization [4], bioinformatics [5], manufacturing [6], e-learning evaluation system [7], medical diagnosis [8], data management [9], music categorization [10] and movie genre prediction [11]. Among these music categorization and movie genre predictions or genre preferences prediction [12], [13] could be considered as entertainment applications of classification. To the best of our knowledge, no previous work utilized classification algorithms for predicting restaurant category preferences.

In this paper, a novel application of classification is introduced. Classification algorithms are utilized to generate Restaurant Category Preferences Prediction Model (RCPPM). RCPPM could be considered as an entertainment application of data mining. Using RCPPM the category of the preferred restaurant could be predicted for the user relying on his/her demographic data, income level and place preferences. This would help people to know the most suitable restaurant category for them without wasting time trying several places or searching among a huge amount of the available options. To this end: (i) a novel dataset was collected, using a survey, in order to build the desired prediction model and (ii) several classification styles, i.e. single and ensemble classification algorithms were utilized. The RCPPM is a single-label multi-class classification. More specifically, each sample (user) is assigned with a single class label (preferred restaurant category) from several available categories. It is interesting to note here that RCPPM could be utilized as a "recommender system" that suggests a set of real places to the user. More specifically, RCPPM could be linked with a database comprising real places, in a specific country, that combined with categories (class labels). The recommendation process commences with acquiring features from the user, and then the RCPPM predicts the category of the preferred place relying on the given features. After that, all the real places stored in the database and categorized as the predicted category will be presented to the user.

The remainder of this paper is organized as follows. Section II supplies the reader with the essential background to the work presented in this study. Section III shows the methodology that has been followed to generate the RCPPM. Section IV presents an overview of the main characteristics of the dataset used to generate the RCPPM. Section V presents the obtained results followed by Section VI with the conclusion of the presented work and directions for future research.

II. Background

Classification is an interesting and challenging research area. Several researchers directed their research work on applying classification algorithms to real-word problems due to the potential benefits that can be summarized by producing prediction models that can predict a solution to each instance in the considered problem. As noted in the introduction to this paper, much research work has been conducted on various domains such as medical, biological, social and entertainment domains. In order to apply classification algorithms to real-world problems, the researcher should be knowledgeable about the available classification algorithms. In this section, the necessary background regarding classification algorithms is provided to the reader. Classification algorithms can be divided into two main categories: (i) “Single” classification algorithms and (ii) “Ensemble” classification algorithms. Commencing with Single classification algorithms, where only one classifier, that generated using one classification algorithm, is used for predicting output (class label). Several algorithms are available for this purpose, the most vastly used algorithms are:

- Naïve Bayes (NB) algorithms, which generate probabilistic classifiers relying on Bayes’ theorem.
- Decision Tree (DT) algorithms, which produce decision tree classifiers where none-leaf nodes represent features (input) and leaf-nodes represent class labels (output).
- Rule-Based (RB) algorithms, which generate classifiers comprised of a set of “If-Then” rules. Features (input) are presented at the If side, while class labels (output) at the Then side.
- k-Nearest Neighbor (kNN) algorithms, in which the generated classifiers are referred to as lazy classifiers, because no classification models are generated. Class labels (output) are predicted based on similarity.
- Artificial Neural Network (ANN) algorithms, which produce sophisticated mathematical classifiers that comprised of connected input/output units (neurodes) and communication channels (connections).
- Support Vector Machine (SVM) algorithms, which generate classifiers by finding a “hyperplane” that distinctly distinguishes the two classes featured in the dataset.

With respect to Ensemble classification, several classifiers cooperate together to output a more effective prediction than what can be acquired from using a single classifier. If the base classifiers within the Ensemble are generated using one classification algorithm, then the Ensemble is referred to as “Homogeneous”. While if the base classifiers are produced using more than one classification algorithm, then the ensemble is called “Heterogeneous” [14]. Any classification algorithm, such as DT, NB and SVM could be used to construct the base classifiers within the Ensemble. Three fundamental methods are usually used to combine the results produced by the individual classifiers: weighted averaging, majority voting and averaging [15]. Numerous researchers provided theoretical and practical evidences that Ensemble generally produces more effective prediction than their base classifiers when they are used alone (single classification) [14], [16], [17]. The most widely used Ensemble classification algorithms are:

- Bagging, in which several classifiers are constructed in parallel, using different variations of the considered dataset. To output prediction, voting is adopted to combine results from the trained classifiers [18], [19].
- Boosting, in which several classifiers are generated sequentially, the importance of the sequential connection is to use the information acquired by one classifier to enhance the training process of the next classifier [19], [20].

In this paper, several Single and Ensemble classification algorithms are utilized to generate the desired RCPPM.

III. The Adopted Experimental Methodology

This section presents the followed methodology to produce the desired RCPPM. The first and the main step in the adopted methodology is obtaining and preparing the dataset that will be used to train the classifier. The next section describes the main characteristics of the collected dataset and the considered preprocessing. Once the dataset is preprocessed, it will be fed to one of the classification algorithms to produce the prediction model. In this study, several Single and Ensemble classification algorithms have been utilized and this will be explained in the experiment section. The last step in the adopted methodology is to evaluate the effectiveness of the generated models, in order to decide the “best” model and its applicability to be used for future prediction. In this work, accuracy and Area Under the ROC Curve (AUC) metrics have been utilized for assessing the performance of the constructed prediction models. The accuracy is a simple metric that measures the percentage of the samples correctly predicted by the prediction model. While the AUC is a robust measure to evaluate the overall effectiveness of the prediction model by measuring the area under the ROC curve which plots true positive rate and false positive rate [1].

IV. Dataset Description

This section presents an overview of the main characteristics of the dataset that were used to generate the RCPPM. The considered dataset was collected using a survey that covers person demographic data, income level and place preferences. Table I presents the extracted features, with a brief description of each. The main goal is to build a prediction model to predict the user-preferred restaurant category.

TABLE I. THE EVALUATION DATASET DESCRIPTION

Feature	**Brief Description**	**Type**	**Values/Range**
Age	The age of the person	Nominal	{>18, 18-25, 26-35, >35}
Education Level	The educational level of the person	Nominal	{School, Collage, B.S, Master and PhD}
Work	Indicates whether the person works or not	Nominal/Binary	{yes, no}
Income level	The income level of the person per month in Jordanian Dinar	Nominal	{<50, 50-100, 100-300, 300-500, 500-1000, >1000}
Gender	The gender of the person	Nominal/Binary	{female, male}
Place Design	The design of the place preferred by the person	Nominal	{traditional, classic, modern}
Atmosphere	The preferred atmosphere for the person in terms of quiet or loud	Nominal/Binary	{ quiet, loud}
City	The city that the person prefers when he/she wants to go to a restaurant	Nominal	{Amman, Zarqa, Irbid, Jerash}
Average Spending	The average amount of money that the person spends, in Jordanian Dinar, when going to restaurants	Nominal	{<5, 5-10, 10-20, >20}
Hang-out reason	Indicates the usual reason(s) for going to restaurants with respect to the person	Nominal [*]	{Reading, Dating, Meeting, Parties, Studying}
Music Kind	Indicates the preferred person's music kind in the place he/she would like to go to	Nominal	{Background, DJ, No music, Live music}
Service	Indicates whether the person prefers table-service or self-service restaurants	Nominal/Binary	{Table-service, Self-service}
Go with	Indicates with whom the person prefers to go to restaurants	Nominal [*]	{family, friends, co-workers, nobody}
Food preferences	Refers to the person's preferred food kind(s)	Nominal [*]	{fast food, American, Italian, Middle East, Chinese}
Meal	Refers to the usual meal or food category the person prefers to eat at restaurants	Nominal [*]	{Breakfast, Lunch, Dinnar, Deserts, Drinks}
Sitting Preferences	Indicates whether the person prefers to sit in or out in the restaurant	Nominal	{Inside, Outside}
Seating Preferences	Refers to the kind of the furniture that available in the restaurant the person prefers to go to	Nominal	{Chairs, Couches, Both}
Parking	Indicates the availability of a parking service in the restaurant	Nominal/Binary	{yes, no}
Pay Method	Refers to the preferred payment method for the person	Nominal	{Cash, Card, Both}
Free Wi-Fi	Indicates if the person prefers free Wi-Fi to be available in the restaurant	Nominal/Binary	{yes, no}
Table Reservation	Indicates if the person can reserve a table before going to the restaurant	Nominal/Binary	{yes, no}
Open After Midnight	Indicates whether the person prefers restaurants that open after midnight	Nominal/Binary	{yes, no}
Speed	Indicates whether the speed of offering service is important to the person	Nominal/Binary	{yes, no}
Children seat	Indicates if the person prefers a children seat to be available in the restaurant	Nominal/Binary	{yes, no}
Wheelchair	Indicates if the person prefers a wheelchair seat to be available in the restaurant	Nominal/Binary	{yes, no}
Place Category	The category of the person's preferred restaurant (Class Label)	Nominal	{Fine Dining, National Dishes, Café Shop (Hookah), Café Shop (Study), Picnic, Jordan Folklore, Fast Food}

[*] the attribute is decomposed into a set of binary attributes during the preprocessing, because several options can be selected

Restaurants are categorized into seven categories (class labels): (i) Fine Dining, (ii) National Dishes, (iii) Café Shop (Hookah), (iv) Café Shop (Study), (v) Picnic, (vi) Jordan Folklore and (vii) Fast Food. Fig. 1 presents labels distribution in the considered dataset. As shown in the figure, the distribution of the labels is imbalanced, thus a preprocessing is required to resolve this issue and generate an effective prediction model. The well-known Minority Oversampling TEchnique (SMOTE) [21] was adopted. SMOTE is considered as an oversampling technique that produces artificial minority class samples.

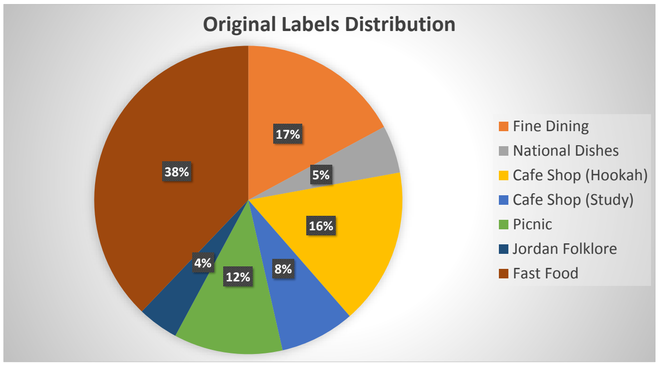

Fig. 1. Original Labels Distribution in the Considered Dataset

Fig. 2 represents labels distribution after applying SMOTE. In addition to SMOTE preprocessing, handling missing values, solving inconsistency and removing redundancy were also applied to the considered dataset. After preprocessing, the dataset features 25 dimensions and 344 data samples.

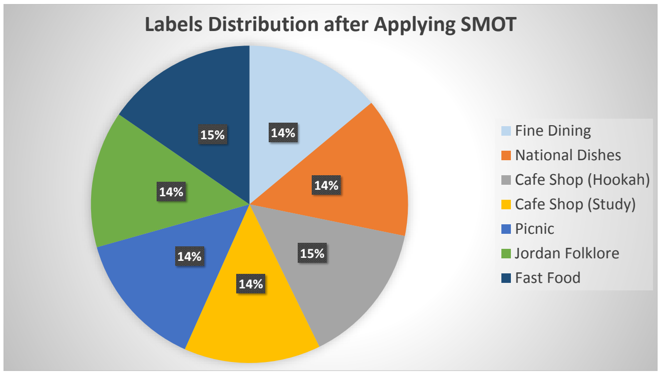

Fig. 2. Labels Distribution after Employing SMOTE

V. Experiments and Results

In this section, the obtained results from the undertaken experiments are presented. As noted earlier in the introduction to this paper, two categories of classification algorithms were utilized to generate the desired RCPPM: (i) Single classification and (ii) Ensemble classification. With respect to the first classification category; six well-known classification algorithms were used to produce the RCPPM: (i) Naïve Bayes (NB), (ii) Decision Tree (DT), (iii) Rule-Based (RB), (iv) k-Nearest Neighbor (kNN), (v) Artificial Neural Network (ANN) and (vi) Support Vector Machine (SVM). Regarding the second classification category, three algorithms were utilized to generate the RCPPM: (i) Bagging Ensemble Classification, (ii) Boosting Ensemble Classification, (iii) Heterogeneous Ensemble Classification. The well-known 10-fold cross validation technique was adopted to divide the dataset into training and testing sets and to obtain more accurate classification results. All classification experiments founded in this work were performed using the WEKA data mining tool [22].

Commencing with the results obtained from using single classification algorithms to construct the RCPPM. Table II presents the obtained results when using the six well-known classification algorithms. From the table it can be observed that DT and NB classifiers generated the same and the highest classification accuracy (Accuracy= 86.92 and AUC = 0.98).

TABLE II. Average Accuracy and AUC Results Obtained when using Single Classification Algorithms to Generate the RCPPM

Classification Algorithm	Accuracy	AUC
Simple Naïve Bayes (Naïve Bayes)	**86.9186**	**0.979**
Decision Tree (Hoeffding Tree)	**86.9186**	**0.979**
Rule-Based (Decision Table)	72.6744	0.917
k-nearest neighbor (IBK)	86.0465	0.976
Support Vector Machine (SMO)	84.0116	0.944
Artificial Neural Network (Multilayer Perceptron)	86.6279	0.961

Because the Ensemble model effectiveness is highly affected by the base classifiers [11], the Ensemble classification experiments were only conducted using DT and Naïve Bayes classifiers as base classifiers. Table III presents the obtained results from using ensemble classification to generate the RCPPM. Note here that Bagging (DT) refers to utilizing a set of DT classifiers as the base classifiers within the Bagging Ensemble to generate the RCPPM model. While Bagging (NB) refers to using Bagging Ensemble classification with NB classifiers as the base classifiers. Boosting (DT) refers to using Boosting Ensemble classification with DT classifiers as the base classifiers, while Boosting (NB) considers using NB classifiers as the base classifiers. Regarding Heterogeneous Ensemble classification, a combination of DT and NB classifiers were utilized to generate the model. Two Heterogeneous classification approaches were utilized, the first one adopts "Majority Voting" to combine results from the base classifiers, while the second one considers "Average Probability" to output the final prediction result. From the table, Bagging Ensemble classification outperforms Boosting and Heterogeneous Ensemble classification, in terms of average accuracy and AUC, for generating the RCPPM. The worst results obtained when using Boosting Ensemble classification to generate the RCPPM.

Fig. 3 presents a comparison between the performance of Single classification and Ensemble classification for generating RCPPM. From the figure, it is clearly observed that Bagging Ensemble classification outperforms Single classification algo-

TABLE III. Average Accuracy and AUC Results Obtained when using Ensemble Classification Algorithms to Generate the RCPPM

Classification Algorithm	Accuracy	AUC
Bagging (NB)	**87.2093**	**0.979**
Bagging (DT)	**87.2093**	**0.979**
Boosting (NB)	83.1395	0.952
Boosting (DT)	83.1395	0.943
Heterogeneous Ensemble (Average Probability)	86.9186	0.979
Heterogonous Ensemble (Majority Voting)	86.9186	0.923

rithms and other forms of Ensemble classification (Boosting and Heterogeneous). The reason behind the superiority of Bagging over Single classification and Boosting is the size of the considered dataset. More specifically, Bagging adopts the "Sampling with Replacement" technique to generate different variations of the dataset with the same size [1], and this technique works very well with small size datasets such as the dataset considered in this research. While the reason behind the superiority of Bagging over Heterogeneous Ensemble returns to the homogeneity of the base classifiers that can reduce prediction conflicts.

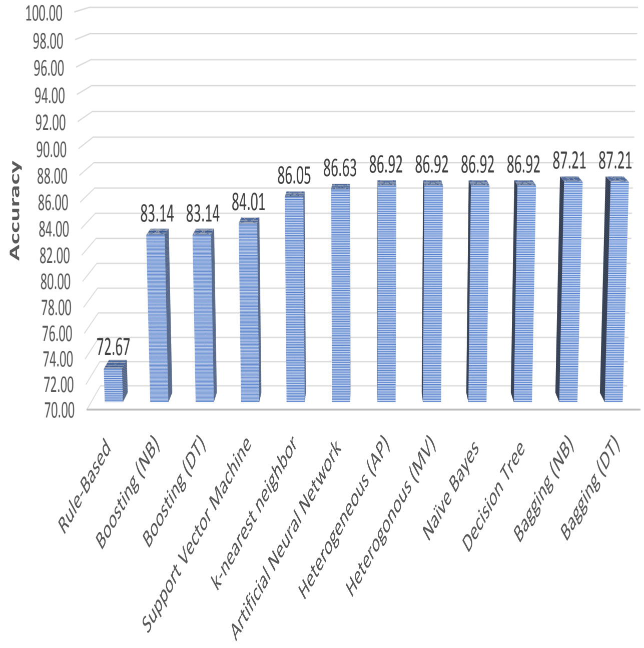

Fig. 3. Comparison between the Performance of Single Classification and Ensemble Classification for Generating RCPPM

VI. Conclusion

In this paper, Single and Ensemble classification algorithms have been utilized to generate a prediction model that aims to predict restaurant category preferences. The RCPPM is an intelligent prediction model that helps users to decide the best suitable place to go. The experiments have been accomplished using a novel dataset that covers person demographic data, income level and place preferences. From the reported experiments, supervised machine learning could be utilized to generate a high-performance RCPPM. Using ensemble of classifiers enhanced the classification effectiveness of the RCPPM. Moreover, Bagging Homogeneous Ensemble classification outperformed Single and Heterogeneous Ensemble classification. Although Heterogeneous Ensemble classification could be utilized to improve classification accuracy by using the power of completely different classifiers, it did not enhance the effectiveness of the RCPPM. The reason behind that could be the predictions conflict that generated by different kinds of classifiers. In the future, the authors plan to investigate the effect of using different features on predicting restaurant category preferences.

Acknowledgment

The authors would like to thank the Hashemite University for the continuous support. The authors would also like to thank Lina, Wasn, Alaa and Hala for distributing the survey and collecting the dataset.

References

[1] H. Jiawei, K. Micheline, and P. Jian, *Data Mining: Concepts and Techniques*. USA: Morgan Kaufmann, 2011.

[2] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[3] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Trans. on Knowl. and Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005. [Online]. Available: https://doi.org/10.1109/TKDE.2005.50

[4] L. Moreira-Matias, J. Moreira, J. Gama, and P. Brazdil, "Text categorization using an ensemble classifier based on a mean co-association matrix," vol. 7376, 07 2012.

[5] P. Yang, J. Yang, B. Zhou, and A. Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, 12 2010.

[6] O. Maimon and L. Rokach, "Ensemble of decision trees for mining manufacturing data sets," *Machine Engineering*, vol. 4, 01 2004.

[7] L. Kai and Z. Zhiping, "Using an ensemble classifier on learning evaluation for e-learning system," in *2012 International Conference on Computer Science and Service System*, 2012, pp. 538–541.

[8] P. Srimani and M. Koti, "Medical diagnosis using ensemble classifiers - a novel machine-learning approach," *J Adv Comput*, vol. 1, pp. 9–27, 01 2013.

[9] S. Kamatkar, A. Tayade, A. Viloria, and A. Hernández, *Application of Classification Technique of Data Mining for Employee Management System*, 06 2018, pp. 434–444.

[10] A. Elbir, H. Bilal Çam, M. Emre Iyican, B. Öztürk, and N. Aydin, "Music genre classification and recommendation by using machine learning techniques," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2018, pp. 1–5.

[11] H. Wang and H. Zhang, "Movie genre preference prediction using machine learning for customer-based information," *International Journal of Computer and Information Engineering*, vol. 11, no. 12, pp. 1329 – 1336, 2017. [Online]. Available: https://publications.waset.org/vol/132

[12] M. S. H. Mukta, E. M. Khan, M. E. Ali, and J. Mahmud, "Predicting movie genre preferences from personality and values of social media users," in *ICWSM*, 2017.

[13] H. Wang and H. Zhang, "Movie genre preference prediction using machine learning for customer-based information," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, 2018, pp. 110–116.

[14] Z.-H. Zhou, *Ensemble Learning*. Boston, MA: Springer US, 2009, pp. 270–273.

[15] I. Nti, A. Adekoya, and B. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," *Journal of Big Data*, vol. 7, pp. 1–40, 2020.

[16] T. G. Dietterich, "Ensemble methods in machine learning," in *MULTIPLE CLASSIFIER SYSTEMS, LBCS-1857*. Springer, 2000, pp. 1–15.

[17] N. Oza and K. Tumer, "Classifier ensembles: Select real-world applications," *Information Fusion*, vol. 9, pp. 4–20, 01 2008.

[18] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[19] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Amsterdam: Morgan Kaufmann, 2017. [Online]. Available: http://www.sciencedirect.com/science/book/9780128042915

[20] Y. Freund and R. E. Schapire, "A short introduction to boosting," in *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 1999, pp. 1401–1406.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278