A Comparison of the Effects of K-Anonymity on Machine Learning Algorithms

Hayden Wimmer
College of Business
Bloomsburg University
Bloomsburg, PA USA

Loreen Powell
College of Business
Bloomsburg University
Bloomsburg, PA USA

***Abstract*—While research has been conducted in machine learning algorithms and in privacy preserving in data mining (PPDM), a gap in the literature exists which combines the aforementioned areas to determine how PPDM affects common machine learning algorithms. The aim of this research is to narrow this literature gap by investigating how a common PPDM algorithm, K-Anonymity, affects common machine learning and data mining algorithms, namely neural networks, logistic regression, decision trees, and Bayesian classifiers. This applied research reveals practical implications for applying PPDM to data mining and machine learning and serves as a critical first step learning how to apply PPDM to machine learning algorithms and the effects of PPDM on machine learning. Results indicate that certain machine learning algorithms are more suited for use with PPDM techniques.**

Keywords*—Privacy Preserving; Data Mining; Machine Learning; Decision Tree; Neural Network; Logistic Regression; Bayesian Classifier

I. Introduction

Knowledge discovery in databases (KDD), or Data Mining (DM), seeks to uncover patterns and relationships contained in data. Privacy of information has come under increasing scrutiny with the advent of regulations such as HIPAA [1, 2]. Simply removing fields or obscuring the records would distort the knowledge contained within the data. This necessity led to the inception of privacy preserving in data mining, or PPDM. PPDM algorithms attempt to de-identify data while maintaining the knowledge contained within. The goal of PPDM research is minimal knowledge distortion; however, some knowledge may be lost when applying PPDM. Machine learning techniques are frequently employed in KDD, or data mining. This research aims to understand the effects of PPDM on common machine learning algorithms and serves as a first step toward mapping the effects of PPDM algorithms on machine learning algorithms. Specifically, this research compares artificial neural networks (ANN), Bayesian Classifier, Decision Stump, C4.5 Decision Tree Induction, Logistic Regression, and Classification and Regression Trees (CART). This work has practical implications for data science and analytics as applied by academics and practitioners alike. The remainder of this paper is structured as follows: section 2 provides a background of machine learning algorithms and privacy preserving in data mining, section 3 presents the methodology and results, and section 4 discusses conclusions and future directions.

II. Background

A. Neural Networks

Neural networks, artificial neural networks, ANN, or NN is a computational technique which is modeled after the human brain's neural pathways [3]. ANNs are frequently applied to pattern recognition and classification and have been applied to facial recognition [4]. An ANN has an input layer and an output layer with one or more (1...n) hidden layers. The hidden layers of the ANN apply a mathematical function to the input and are said to learn by employing techniques such as adjusting weights of the input in the hidden layer. A simple, single layer, ANN is shown as *Figure 1*. ANNs have been successfully applied to many scenarios that are of interest to data science. Examples of ANN applications include recognizing financial distress patterns [5], bankruptcy prediction [6-9], and decision support systems [10]. ANNs have been applied to classic problems such as stock price forecasting [11] and medical diagnoses [12].

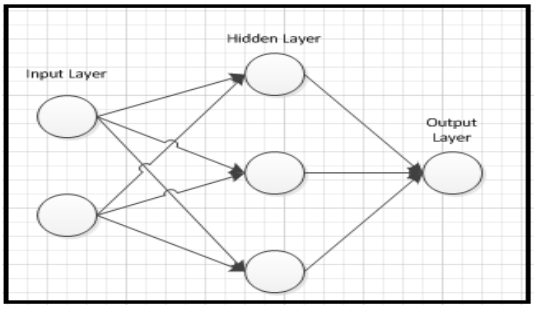

Fig. 1. A Simple Artificial Neural Network

B. Decision Trees

Decision tree algorithms are machine learning algorithms that accept data as an input and output a graph structure. Decision trees begin with a root node which branch into child nodes. A leaf node is a node with no children. Rules, as applied in expert systems, can be extracted from a decision tree. An example decision tree constructed from the classic weather data set as described by Livingston [13] is shown as *Figure 2*. The weather dataset has 14 instances and 5 features. The resulting decision tree is used to determine whether to perform a task, such as play a game, given weather conditions. One can extract rules such as *Table 1*.

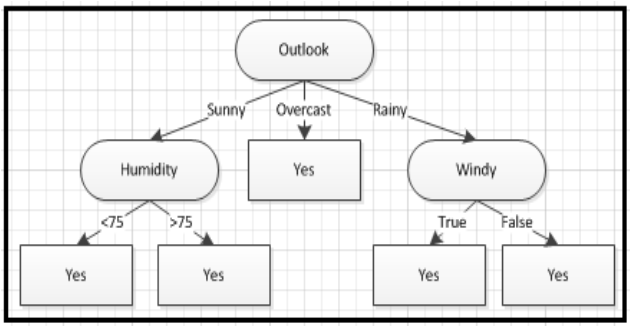

Fig. 2. A Decision Tree from Weather Data

TABLE I. RULE SET FROM DECISION TREE

If it is sunny and humidity is less than or equal to 75 then play
If it is sunny and humidity is greater than 75 the do not play
If it is overcase then play
If it is rainy and windy then do not play
If it is rainy and not windy then play

Decision tree learning algorithms include the classification and regression tree algorithm, ID3, C4.5, C5.0, CHAID, and decision stumps to name a few. Classification and regression trees, or CART, was conceived in 1984 by Breiman, et al. [14] and recursively works through data and using an index feature, the Gini index [15], and divides the data into a tree structure. ID3 [16], C4.5 [17], and C5.0 [18] are all related as C5.0 extends C4.5 and C4.5 extends ID3. ID3 and C4.5 are open source whereas C5.0 is proprietary. Like the CART algorithm, ID3 and C4.5 recursively employ a function to split the data into a tree structure; however, C4.5 and ID3 apply an entropy function which seeks to minimize the information loss occurring from each split of the data which is computed as the difference between the normalized information gains. C4.5 has been widely applied to domains such as network traffic classification[19], vehicle traffic pattern and driving behavior classification [20] patient classification [21], and organ classification [22]. CHAID is the Chi-Squared Automatic Interaction Detection algorithm and is similar to the aforementioned classification algorithms but is based in Bonferroni statistical testing [23, 24]. CHAID has a wide range of applications and has been used in financial distress classification [25].

C. Bayesian Classifier

Naïve Bayesian classifiers employ simple statistical assumptions to make classifications. These assumptions assist in increasing the performance of the classifier. The performance and assumptions make it an effective classifier for many applications such as junk mail filtering [26]. The Naïve Bayes classifier is considered one of the most efficient and effective classification algorithms [27]. The principle assumption made by a Naïve Bayes classifier is that all features, or independent variables, contribute equally to the target, or dependent variable. The effectiveness, regardless of the assumptions, is the optimality of classification is not necessarily related to the independence of the assumptions [28]. Despite its simplicity in its assumptions, it has been shown to outperform more powerful classifiers under many conditions which demonstrates Bayesian classifiers are a highly applicable classifier to many domains and classification problems [29]. Modern approaches include medical applications such as heart attack prediction [30], credit scoring [31], and social network analysis [32].

D. Logistic Regression

Logistic regression is a form of classifier that, given input independent variables, predict the target or dependent variable. Logistic regression is similar to linear regression with the exception that the target variable, or dependent variable, is categorical as opposed to continuous [33]. The dependent variable is a binary value {0, 1} frequently representing {yes, no}, {up, down}, or {good, bad}. Logistic regression has a wide range of applications such as making predictions in healthcare settings [34-36]. Logistic regression has seen modern applications in medical diagnoses [37] and in data science and analysis [38].

E. Privacy preserving data mining and K-Anonymity

Privacy and preserving in data mining, or PPDM, is a research stream that seeks to insert privacy into data mining while maintaining the integrity of the knowledge contained within the data [39]. The need for PPDM was emphasized in the late 1990s when the medical history of the governor of Massachusetts was uncovered by reassembling public census records with public medical data. This process, known as re-identification, detailed the need for anonymization when sharing medical data and hence the Datafly algorithm was introduced [40-42]. Some of the primary PPDM techniques include data perturbation, randomized response, condensation [43], data and rule hiding[44, 45], cryptography, noise adding, blocking, generative based, and sanitization based [46], and differential privacy [47, 48] to name a few.

Among the aforementioned PPDM techniques are algorithms that transform data to meet a standard, k-Anonymity. K-Anonymity states that each record must not be distinguishable from k respondents. In data records, there are attributes that uniquely identify individuals, such as a social security number. These attributes are considered identifier attributes. In addition to identifier attributes, there are attributes that, when combined with other attributes, uniquely identify individuals. These are referred to as quasi-identifiers. In 2000, it was found that 87% of individuals could be uniquely identified by the quasi-identifiers date of birth, zip code, and gender [49]. K-anonymity requires that, within a table, a set of quasi-identifying attributes must appear at least *k* times. For example, given the set or quasi-identifiers *S = {date of birth, zip code, gender}* and *t* is an instance in S such that *t= {2/20/1967, 98520, M}*, and *k=2*, then a minimum of 2 occurrences of tuple *t* is required for k-anonymity.

III. METHOD

A. Framework Experiment

The aim of this applied research is to begin mapping the effects of PPDM techniques on machine learning algorithms. First, a data file is read and classified with a machine learning algorithm. Second, the same data file is read, a privacy framework applied, and the same machine learning algorithm is applied. The general framework is detailed in *Figure 3*.

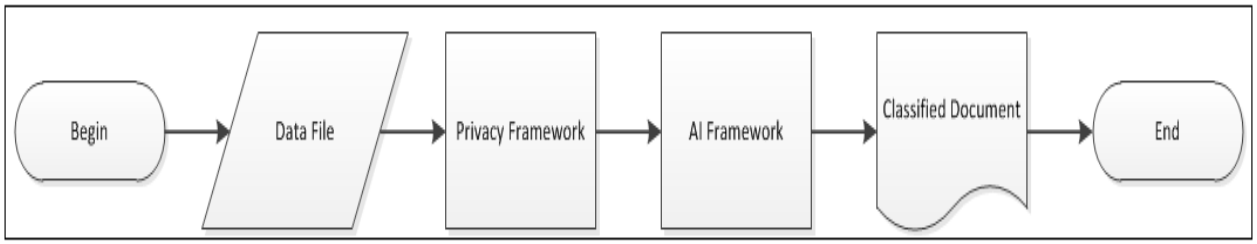

Fig. 3. General Framework for PPDM and ML

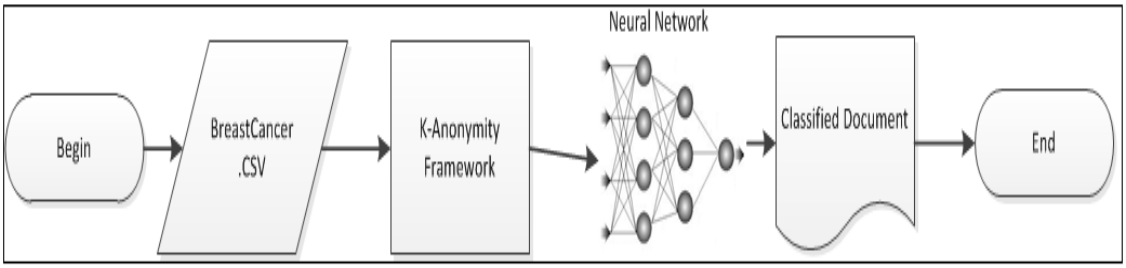

Fig. 4. Flow for ANN

Figure 4 shows the specific framework. In this work, k-anonymity is employed as the PPDM technique. Once k-anonymity with *k=2* is applied to the input dataset, the resulting anonymized dataset becomes input for 6 machine learning algorithms: 1)artificial neural network (ANN), 2) C4.5 decision tree, 3) decision stump algorithm, 4) classification and regression Tree, 5) Naïve Bayes classifier, and 6) a logistic regression. The flow for the ANN is shown as *Figure 4*.

B. Data Pre-Processing

The framework was applied to 3 separate datasets of differing size and attributes. The first dataset was extracted from [42]. The original Sweeney dataset had 5 features and 5 instances; however, an artificially generated first and last name and a target for classification were generated altering the dataset to 8 features and 5 instances. The second dataset was retrieved from the UCI machine learning repository [50] and was cited in [51] and in [52]. This dataset will be hereafter referred to as the cancer dataset. First, instances missing data were removed. Next, features for first, last, and middle name were added and randomly generated from a random name generator. The resulting dataset had 14 features and 699 instances. The final dataset was also extracted from the UCI machine learning repository and was originally extracted from census data and had 299999 instances and 6 features with 1 being an identifier and 1 being a target. Instances with missing data were removed. This dataset is named income.

C. Algorithm Parameters

All algorithms were run on a dedicated Windows 8 machine with an Intel i3 2.30GHZ processor and 8GB of physical memory. Machine learning algorithms examined include artificial neural networks, naïve Bayesian classifier, logistic regression, Decision Stump, Classification and Regression Trees (CART), and C4.5 decision tree induction. All algorithms were trained with 10 fold cross-validation. The artificial neural network was set to train through a maximum of 100 epochs and the number of hidden layers was set to a maximum of the average of the number of classes and the number of attributes. Back propagation was employed by the classifier and, if numeric, nodes are non-threshold linear units, otherwise, they are sigmoid. The naïve Bayes classifier employed is based on [53] where estimator values are chosen based on the training data. Logistic regression used multinomial regression paired with a ridge estimator as detailed in [54]. The Decision Stump algorithm is based on mean-squared error and information entropy. The minimum number of instances for a leaf was set as 1. The CART algorithm is based on [14] and implemented minimal cost-complexity pruning and 100% of the data was available to the 10 fold cross-validation process. The C4.5 decision tree algorithm [16, 17]was set to a minimum of 5 objects per leaf.

IV. Results

The results presented in Tables 2, 3, and 4 correspond to the datasets Sweeney, Cancer, and Income respectively. In reviewing the results it is necessary to consider the resulting confusion matrix from the classifier. A confusion matrix details how instances are classified. Specifically, the confusion matrix details true positives or instances correctly classified as positive, false positives or instances incorrectly classified as positive, false negatives or instances that were incorrectly classified as negative and true negatives or instances that were correctly classified as negative. In our example, the resulting confusion matrix is a 2x2 matrix and can be interpreted as *Figure 5*.

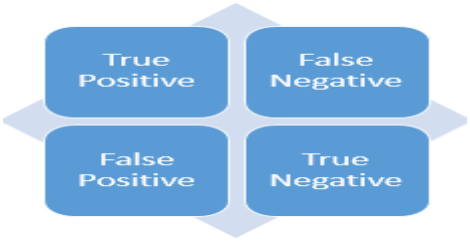

Fig. 5. Interpreting a Confusion Matrix

In each of the tables 2, 3, and 4 the resulting classification accuracy of the machine learning algorithm and confusion matrix is presented both before the PPDM technique (k-anonymity *k=2*) is applied. Any PPDM technique will make changes to the data and therefore the knowledge contained within. It is also theorized that the larger the dataset the less individual changes to individual instances a PPDM algorithm will have to introduce into the data. For example, with only 2 instances both instances will need to be changed for de-identification whereas a large dataset with a million instances will likely not have to change each individual instance to de-identify the data.

Minimizing the change in the hidden knowledge is the goal of any PPDM technique. The results show some machine learning algorithms performing better after the PPDM technique was applied which is suspect as it is natural for the knowledge to degrade after the PPDM technique.

The ANN was susceptible to this phenomenon with it improving for both the Sweeney and Income datasets and only slightly degrading for the Income dataset. In the cancer dataset prior to anonymization the resulting confusion matrix showed no true negatives or false negatives indicating that the classifier classified all data as positive with a high error rate (34.48%) which also happened in the Sweeney dataset after the PPDM technique. The confusion matrices of the income dataset reveal a decrease in true positives and false positives but an increase in false negatives and true negatives. Based on the aforementioned figures artificial neural networks did not perform well on the datasets when combined with PPDM.

The performance of the C4.5 decision tree algorithm was unchanged post PPDM for the Sweeney dataset and had a decrease in classification accuracy for the cancer and income datasets. The confusion matrix was unchanged for the Sweeney dataset but only the number of false positives and true positives changed for the cancer dataset indicating a shift in records being incorrectly classified as positive. This was 23 instances out of 699 or 3.1%. The results were different for the income dataset with a shift with an increase in false negatives and true negatives. There were 246432 correctly classified instances before and 244312 correctly classified instances after PPDM for a change in less than 1%. This indicates that C4.5 performed well with the datasets and k-anonymity.

The decision stump algorithm remained unchanged post-PPDM for the Sweeney dataset and decreased for the cancer dataset. There was a decrease in true positives and true negatives and corresponding increases in false positives and false negatives. In the case of the income dataset the decision stump simply classified everything as negative. Based on this, there are concerns with the performance of the decision stump algorithm with the datasets and PPDM technique.

The Classification and Regression Tree, or CART, algorithm was unable to make any classifications in the Sweeney dataset due to the small size. CART demonstrated a large improvement in classification accuracy in the cancer dataset indicating a potential concern. Applied to the income dataset there was a decrease in classification accuracy (82.16 to 81.43) and there was a reduction in true positives but an increase in true negatives. This indicates CART has potential on only 1 of the 3 datasets.

Naïve Bayes showed decreases in all classification accuracies for the 3 datasets which, as previously stated, is to be expected. The confusion matrices showed a decrease in true positives and true negatives for the Sweeney and cancer datasets with a decrease in true positives and increase in true negatives for the income dataset. The changes in the cancer dataset were small with a true positive reduction of less than 1% and a 14% reduction in true negatives. In the income dataset there was a 76% reduction in true positives with only a 4% increase in true negatives.

Logistic regression was unchanged for the Sweeney dataset and classification accuracy decreased for both the cancer and income datasets. Logistic regression demonstrated a decrease in true positives and true negatives in both the cancer and income datasets. True positive reduction was 2.4% and 95% and true negative reduction was 12% and 2%.

The aforementioned results are open to interpretation and can be interpreted differently based on the objectives of the PPDM and machine learning project. The results indicate that C4.5 performs best with K-anonymity, with Naïve Bayes second, and logistic regression third. The remaining 3 approaches (ANN, Decision Stump, and CART) seemed to be problematic among the 3 datasets and, while pairing any machine learning algorithm with PPDM care is critical, pairing k-anonymity with ANN, decision stump, and CART should be performed with additional cautionary measures.

TABLE II. RESULTS FROM SWEENEY DATASET

Sweeny Dataset				
	Before K anonymity		After K anonymity *k=2*	
	Classification Accuracy	*Confusion Matrix*	*Classification Accuracy*	*Confusion Matrix*
ANN	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$	60	$\begin{bmatrix}0 & 2\\0 & 3\end{bmatrix}$
C4.5	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$	40	$\begin{bmatrix}0 & 2\\2 & 1\end{bmatrix}$
Decision Stump	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$
CART	NA	NA	NA	NA
Naïve Bayes	80	$\begin{bmatrix}1 & 1\\0 & 3\end{bmatrix}$	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$
Logistic Regression	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$	40	$\begin{bmatrix}0 & 2\\1 & 2\end{bmatrix}$

TABLE III. RESULTS FROM CANCER DATASET

Cancer Dataset				
	Before K anonymity		After K anonymity *k=2*	
	Classification Accuracy	*Confusion Matrix*	*Classification Accuracy*	*Confusion Matrix*
ANN	65.52	$\begin{bmatrix}458 & 0\\241 & 0\end{bmatrix}$	90.96	$\begin{bmatrix}433 & 25\\38 & 201\end{bmatrix}$
C4.5	94.42	$\begin{bmatrix}436 & 22\\17 & 224\end{bmatrix}$	91.1	$\begin{bmatrix}436 & 22\\40 & 199\end{bmatrix}$
Decision Stump	92.41	$\begin{bmatrix}417 & 41\\12 & 229\end{bmatrix}$	87.52	$\begin{bmatrix}446 & 12\\75 & 164\end{bmatrix}$
CART	64.95	$\begin{bmatrix}454 & 4\\221 & 0\end{bmatrix}$	90.96	$\begin{bmatrix}435 & 23\\40 & 199\end{bmatrix}$
Naïve Bayes	95.99	$\begin{bmatrix}436 & 22\\6 & 235\end{bmatrix}$	91.25	$\begin{bmatrix}433 & 25\\36 & 203\end{bmatrix}$
Logistic Regression	96.14	$\begin{bmatrix}444 & 16\\11 & 230\end{bmatrix}$	91.23	$\begin{bmatrix}433 & 25\\36 & 203\end{bmatrix}$

TABLE IV. RESULTS FROM INCOME DATASET

Income Dataset				
	Before K anonymity		After K anonymity *k=2*	
	Classification Accuracy	*Confusion Matrix*	*Classification Accuracy*	*Confusion Matrix*
ANN	81.91	[10494 45618; 8661 235226]	81.25	[1591 54521; 1722 242165]
C4.5	82.14	[13034 43078; 10489 233398]	81.44	[3870 52242; 3445 240442]
Decision Stump	81.3	[0 56112; 0 243887]	81.3	[0 56112; 0 243887]
CART	82.16	[12353 43759; 9774 234113]	81.43	[4263 51849; 3867 240020]
Naïve Bayes	81.94	[14214 41898; 12292 231595]	81.36	[3413 52699; 3220 240667]
Logistic Regression	81.49	[6992 49120; 6396 237491]	81.19	[344 55768; 650 243237]

V. DISCUSSIONS AND FUTURE DIRECTIONS

The aim of the research presented in this work is to developing an understanding of the effects of PPDM techniques on machine learning algorithms. Specifically, the effects of K-Anonymity were tested against artificial neural networks (ANN), Bayesian classifiers, Decision Stump algorithm. C4.5 Decision tree induction, logistic regression, and classification and regression tree (CART) algorithm. The machine learning algorithms were tested on datasets of differing sizes and features before and after a privacy preserving data mining algorithm was applied. Results indicate that certain machine learning algorithms are more suited to use with PPDM techniques than others. This research opens the possibility for other researchers to continue and contribute by applying different PPDM techniques with machine learning algorithms. Limitations include a lack of additional datasets with a higher number of features, theoretical justification on performance, and a lack of other PPDM and machine learning algorithms. Future work will include more extensive datasets, a deeper theoretical justification, and comparing additional PPDM techniques and machine learning algorithms, specifically frequent itemset hiding and the A-priori algorithm.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of [1] G. J. Annas, "HIPAA regulations-a new era of medical-record privacy?," New England Journal of Medicine, vol. 348, pp. 1486-1490, 2003.

[2] C. F. D. Control and Prevention, "HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services," MMWR: Morbidity and Mortality Weekly Report, vol. 52, pp. 1-17, 19, 2003.

[3] M. T. Hagan, H. B. Demuth, and M. H. Beale, Neural network design: Pws Pub. Boston, 1996.

[4] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, pp. 23-38, 1998.

[5] P. K. Coats and L. F. Fant, "Recognizing financial distress patterns using a neural network tool," Financial Management, pp. 142-155, 1993.

[6] K. C. Lee, I. Han, and Y. Kwon, "Hybrid Neural Network Models for Bankruptcy Prediction," Decision Support Systems, vol. 18, pp. 63-72, 1996.

[7] K. Tam, M, "Predicting Bank Failures: A Neural Network Approach," Applied Artificial Intelligence: An International Journal, vol. 4, pp. 265-282, 1990.

[8] K. Tam and M. Kiang, "Managerial Applications of Neural Networks: The Case of Bank Failure Predictions," Management Science, vol. 38, pp. 926-948, 1992.

[9] R. Wilson and R. Sharda, "Bankruptcy Prediction Using Neural Networks," Decision Support Systems, vol. 11, pp. 545-557, 1994.

[10] N. Kumar, R. Krovi, and B. Rajagopalan, "Financial decision support with hybrid genetic and neural based modeling tools," European Journal of Operational Research, vol. 103, pp. 339-349, 1997.

[11] E. Hadavandi, H. Shavandi, and A. Ghanbari, "Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting," Knowledge-Based Systems, vol. 23, pp. 800-808, 2010.

[12] Q. K. Al-Shayea, "Artificial neural networks in medical diagnosis," International Journal of Computer Science Issues, vol. 8, pp. 150-154, 2011.

[13] F. Livingston, "Implementation of Breiman’s random forest machine learning algorithm," ECE591Q Machine Learning Journal Paper, 2005.

[14] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and regression trees: CRC press, 1984.

[15] R. I. Lerman and S. Yitzhaki, "A Note on the Calculation and Interpretation of the Gini Index," Economics Letters, vol. 15, pp. 363-368, 1984.

[16] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, pp. 81-106, 1986.

[17] J. R. Quinlan, C4. 5: programs for machine learning vol. 1: Morgan kaufmann, 1993.

[18] J. R. Quinlan. (2012). C5.0: An Informal Tutorial.

[19] Y. Zhang, H. Wang, and S. Cheng, "A method for real-time peer-to-peer traffic classification based on C4. 5," in Communication Technology (ICCT), 2010 12th IEEE International Conference on, 2010, pp. 1192-1195.

[20] Z.-W. Yuan and Y. Dong, "Research the association of dangerous driving behavior and traffic congestion based on C4. 5 algorithm," Computer, Intelligent Computing and Education Technology, p. 403, 2014.

[21] A. G. Karegowda, V. Punya, M. Jayaram, and A. Manjunath, "Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4. 5," International Journal of Computer Applications, vol. 45, 2012.

[22] M. K. Ross, K.-W. Lin, K. Truong, A. Kumar, and M. Conway, "Text categorization of Heart, Lung, and Blood studies in the Database of Genotypes and phenotypes (dbGap) Utilizing n-grams and Metadata Features," Biomedical informatics insights, vol. 6, p. 35, 2013.

[23] E. Antipov and E. Pokryshevskaya, "Applying CHAID for logistic regression diagnostics and classification accuracy improvement," Journal of Targeting, Measurement and Analysis for Marketing, vol. 18, pp. 109-117, 2010.

[24] J. Magidson, "The chaid approach to segmentation modeling: Chi-squared automatic interaction detection," Advanced methods of marketing research, pp. 118-159, 1994.

[25] N. Ozgulbas and A. S. Koyuncugil, "Developing Road Maps for Financial Decision Making by CHAID Decision Tree: CHAID Decision Tree Application," in Information Management and Engineering, 2009. ICIME '09. International Conference on, 2009, pp. 723-727.

[26] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in Learning for Text Categorization: Papers from the 1998 workshop, 1998, pp. 98-105.

[27] H. Zhang, "The optimality of naive Bayes," A A, vol. 1, p. 3, 2004.

[28] I. Rish, "An empirical study of the naive Bayes classifier," presented at the IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001.

[29] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Machine learning, vol. 29, pp. 103-130, 1997.

[30] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," International Journal on Computer Science and Engineering (IJCSE), vol. 2, pp. 250-255, 2010.

[31] N.-C. Hsieh and L.-P. Hung, "A data driven ensemble classifier for credit scoring analysis," Expert Systems with Applications, vol. 37, pp. 534-545, 2010.

[32] M. Al Hasan and M. J. Zaki, "A survey of link prediction in social networks," in Social network data analytics, ed: Springer, 2011, pp. 243-275.

[33] D. W. Hosmer Jr and S. Lemeshow, Applied logistic regression: John Wiley & Sons, 2004.

[34] J. J. Tomaszewski, R. G. Uzzo, N. Kocher, T. Li, B. Manley, R. Mehrazin, et al., "Patients with anatomically “simple” renal masses are more likely to be placed on active surveillance than those with anatomically “complex” lesions," in Urologic Oncology: Seminars and Original Investigations, 2014.

[35] A. C. Davis, G. Watson, N. Pourat, G. F. Kominski, and D. H. Roby, "Disparities in CD4 Monitoring among HIV-Positive Medicaid Beneficiaries: Evidence of Differential Treatment at the Point of Care," in Open Forum Infectious Diseases, 2014, p. ofu042.

[36] E. Dahlén, C. Almqvist, A. Bergström, B. Wettermark, and I. Kull, "Factors associated with concordance between parental‐reported use and dispensed asthma drugs in adolescents: findings from the BAMSE birth cohort," Pharmacoepidemiology and Drug Safety, 2014.

[37] D. Timmerman, B. Van Calster, A. C. Testa, S. Guerriero, D. Fischerova, A. Lissoni, et al., "Ovarian cancer prediction in adnexal masses using ultrasound‐based logistic regression models: a temporal and external validation study by the IOTA group," Ultrasound in obstetrics & gynecology, vol. 36, pp. 226-234, 2010.

[38] J. Sall, A. Lehman, M. L. Stephens, and L. Creighton, JMP start statistics: a guide to statistics and data analysis using JMP: SAS Institute, 2012.

[39] R. Agrawal and R. Srikant, "Privacy-preserving data mining," ACM Sigmod Record, vol. 29, pp. 439-450, 2000.

[40] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the Datafly System," in Proceedings of the AMIA Annual Fall Symposium, 1997, p. 51.

[41] L. Sweeney, "Datafly: a system for providing anonymity in medical data," Database Security, XI: Status and Prospects, 1998.

[42] L. Sweeney, "Computational disclosure control for medical microdata: The Datafly system," in Record Linkage Techniques 1997: Proceedings of an International Workshop and Exposition, 1997, pp. 442-453.

[43] G. Nayak and S. Devi, "a survey on privacy preserving data mining: approaches and Techniques," International Journal of Engineering Science and Technology (IJEST), vol. 3, pp. 2117-2133, 2011.

[44] V. S. Verykios, "Association rule hiding methods," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, pp. 28-36, 2013.

[45] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," ACM Sigmod Record, vol. 33, pp. 50-57, 2004.

[46] R. Natarajan, R. Sugumar, M. Mahendran, and K. Anbazhagan, "A survey on Privacy Preserving Data Mining," International Journal on Advanced Research in Computer and Communications Engineering, vol. 1, 2012.

[47] C. Dwork, "Differential privacy: A survey of results," in Theory and Applications of Models of Computation, ed: Springer, 2008, pp. 1-19.

[48] C. Dwork, "Differential privacy," in Automata, languages and programming, ed: Springer, 2006, pp. 1-12.

[49] L. Sweeney, "Simple demographics often identify people uniquely," Health (San Francisco), pp. 1-34, 2000.

[50] (2013). UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/ml/

[51] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," Proceedings of the national academy of sciences, vol. 87, pp. 9193-9196, 1990.

[52] J. Zhang, "Selecting typical instances in instance-based learning," 1992.

[53] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995, pp. 338-345.

[54] S. Le Cessie and J. Van Houwelingen, "Ridge estimators in logistic regression," Applied statistics, pp. 191-201, 1992.