Eastern Mediterranean University

"For Your International Career"

Case Study: Type 2 Diabetes dataset

A Term Project of Pattern Recognition Course

Ali Moayedi Azarpour

ID: 16500560

December 2018

1. Introduction

Artificial Intelligence (AI) is the science of human's intelligence imitation by machines through programing them to act, learn and make decisions in the same way as humans do. Machines can be trained to develop cause-and-affect structures which let them to make decisions. However, AI is a very broad topic dealing with integration of intelligence systems. Machine Learning (ML), a subset of Al, is the science of learning from data. Algorithms and models develop over hidden patterns within datasets and automated detection of patterns is the subject of pattern recognition. In this filed, different algorithms are used to detect hidden relations and discovered ones can be used later by machines to make decisions. The process of feeding a set of data into a system is called learning which is done by using a set of labeled data known as train data. Whenever this set of labeled data is in hand and are given to a system, training is supervised. On the other hand, if labels are missing by any reason, algorithms use unsupervised learning techniques. These techniques develop models by training data and based on that, future unseen data or test data can be classified. Thus, regardless of learning technique, ultimate goal is to decide on the most probable class of test data.

There are different widely used algorithms to model data. They mostly have probabilistic nature and try to predicted classes using densities and probabilities. Naïve Bayes, Support Vector Machines (SVMs), logistic regression are examples of learning algorithms. However, it should be kept in mind, developed models should perform as well as possible and being able to generalize decision makings on new unseen data.

2. Case Study

With this brief introduction, as a practical study case, a type 2 diabetes dataset used for diabetic patients' identification considered. This dataset contains 1611 attributes, 4322 samples and two classes of positive and negative. Moreover, attributes are combination of different variable types

(nominal, continues and binary) along with missed values. Thus preprocessing is an inevitable step. In the following sections we first start with preprocessing and continue with fitting of different classification models. Through the study four classification algorithms, Naïve Bayes, logistic regression, Support Vector Machine (SVM) and k-Nearest-Neighbor (kNN) are used and results compared. Also in this study, Weka, a free machine learning software written in Java, and python for preprocessing and model construction are used respectively.

2.1. Preprocessing

The understudy dataset, as noted before, contains different types of values and different ranges. One of the preprocessing steps is to map data values into same range. However, before mapping data it has to be check for missing values. Missing values can make problems in two ways. First of all attributes with many missing values cannot be informative and helpful for training. And secondly, prediction of many missing values is not accurate. There are three techniques toward missing values. One technique is to remove samples with missing values. However, this approach results in data loss which is not of interest. The other way is imputation. And the third technique is attribute removal. In this study all of the techniques are used and they are discussed later at the time of usage. To evaluate our trained models performance, we divide dataset into two halves. In the first phase (phase 1), we build models using the first half and evaluate them on the second half. In the second phase (phase 2) train and test sets are replaced. However, by checking data it is observed that there are two classed of 'Negative' and 'Positive' and totally are separated. Thus we shuffle data to make sure classes are mixed properly. This is a vital step before splitting data to have balanced sets with almost equal ratio of classes.

According to discussed issues, dataset is initially checked for the number of missing values in each feature. If there were more than 30% of NA's in an attribute, we drop it. The rest of missing values based on the value type of attribute is imputed. If the feature was continues, average and if it was nominal, mode was used as the replacement. Besides, since data are from different ranges with different types, all of the continues attributes are passed through min-max scaling and mapped into range of 0 to 1. Nominal attributes in range of 0 to 4 by use of dummy variables presented by a set of binary attributes.

After splitting data, to have baseline for models evaluation and comparing them, a baseline prototype consists of first thirteen attributes including gender, age, race, education, family history, high blood sugar, body mass index (BMI), waist circumference, systolic blood pressure, diastolic blood pressure, hyper tension, physical activity and medicine is developed. In the next section we discuss how baseline is processed.

2.2. Baseline

Starting from baseline and examining features, it reveals that 49% of medicine attribute values are missing. Since we want to have a baseline of mentioned features, samples removal and imputation techniques regarding missing values are tested. By checking values of "medicine" attribute, there is no sample with value 0. Thus, NA's can be considered as 0. This means that patients with NA values probably did not receive any kind of treatment. Results of training after replacing NA's with 0 are given in Table 1.

Based on results, logistic regression and SVM have the highest accuracies and f-scores are almost identical. Naïve Bayes and kNN, respectively, come next. For SVM model linear kernel was set for training along with L2-norm penalties. In kNN classifier since value of k is a matter of experience and depends on the dataset, 15 models for different values of k in range of 1 to 14 developed and their performances are compared (Figure 1). It can be observed that changes of accuracy for k's greater than 9 are not meaningful. Hence, we set k to 9. Moreover, Euclidean distance used to measure distance between objects. In Logistic regression, third model, penalty function defined as L2-norm and Gaussian was the probability distribution considered in Naïve Bayes classifier. In Figure 2 ROC curves for four models in two phases are given.

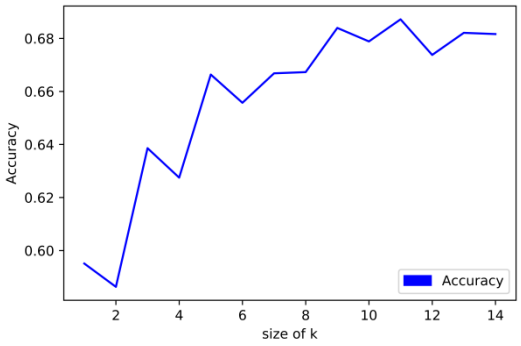

Figure 1 - This plot shows accuracy against different values of k in kNN algorithm

Table 1 - Developed models over baseline (Phase 1)

	Train by imputation				Train by removing			
	SVM	**KNN k=9**	**Logistic Regression**	**Naïve Bayes**	**SVM**	**KNN k=9**	**Logistic Regression**	**Naïve Bayes**
Accuracy	68.9496	62.5636	69.3660	65.5252	69.8751	63.5354	69.6900	65.4327
Precision	0.6910	0.6243	0.6949	0.6642	0.7001	0.6342	0.6981	0.6659
Recall	0.6912	0.6237	0.6952	0.6601	0.7004	0.6321	0.6984	0.6600
F score	0.6911	0.6240	0.6951	0.6622	0.7003	0.6332	0.6983	0.6629
AUC$_{(area\ under\ the\ curve)}$	0.6912	0.6237	0.6952	0.6601	0.7004	0.6321	0.6984	0.6600

Table 2 - Developed models over baseline (Phase 2)

	SVM	**KNN k=9**	**Logistic Regression**	**Naïve Bayes**
Accuracy	69.7825	64.6460	70.1527	67.0060
Precision	0.7041	0.6466	0.7037	0.6697
Recall	0.6934	0.6437	0.6984	0.6683
F score	0.6987	0.6451	0.7011	0.6690
AUC$_{(area\ under\ the\ curve)}$	0.6934	0.6437	0.6984	0.6683

In first phase first half of data is used for training and second half is used for testing. In second phase test and train data are replaced. ROC curves plots true positive rate against false positive rate. Thus a model is considered good if the curve skewed toward top left corner, which means higher True Positive Rate and lower False Positive Rate (FPR). In case of weak performance the curve gets close to the straight diagonal line. In table 1 on the right section, results in case of removing samples with missing values for attribute medicine are given. As we can see, changes in performance are not meaningful, and in fact we had some slight increase in values. However, since we prefer to keep our dataset larger, we continue with keeping missing values in our dataset and for medicine in as special case we impute NA's as 0.

In the preprocessing section, we mentioned dataset is divided into two halves. This is also true for our baseline. In table 1 first half of data used for training and second kept for testing. In table 2Error! Reference source not found. results of changing halves are given. Results show almost same levels of

performance for classifiers. Logistic regression and SVM have the highest performance. Naïve Bayes and kNN respectively have lower performances. Close results reveal data are uniformly split into two halves and they have almost same level of information for classifier to learn from. ROC plots are given in Figure 2, phase 2. It can be observed that curves are of approximately in same shape as in curves in phase 1.

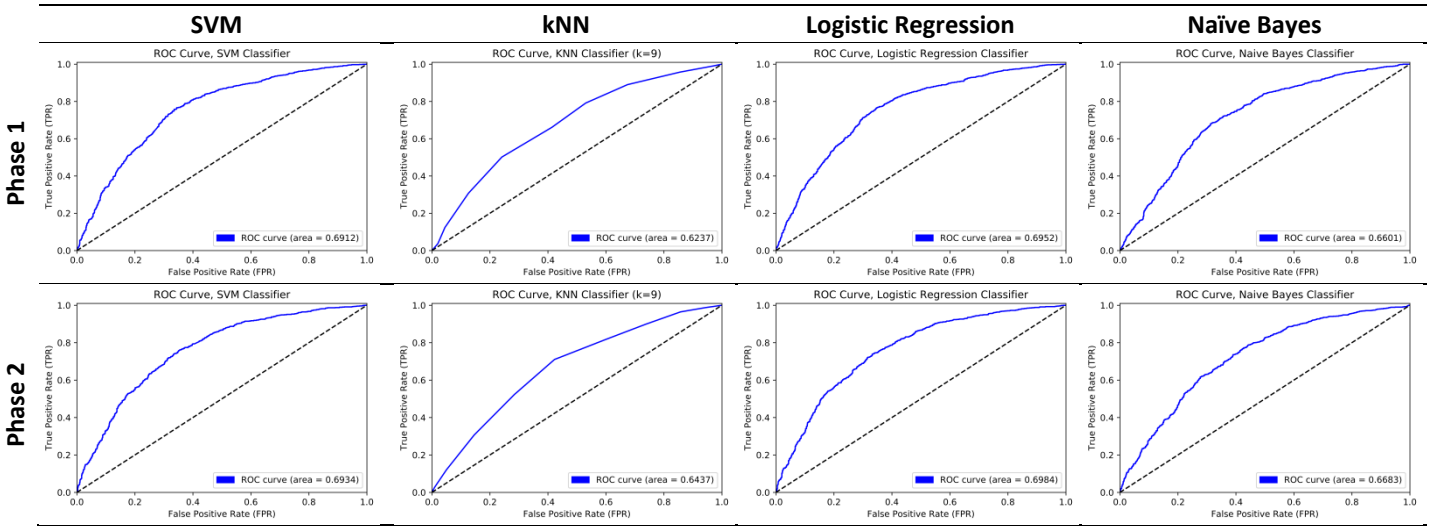

Figure 2 - ROC curves for two phases of training models on baseline features

2.3. Entire dataset

Now back to the whole set of attributes, we said that before splitting there were 1611 features with 4322 samples. In the first step before splitting dataset we checked for non-informative attributes to reduce dimensionality and have a set of informative features. Therefore, all attributes having more than 30% of missing ones removed. In addition to missing values, there exists some attributes having equal or very few distinct values for all samples. Such kinds of attributes cannot be informative to learn from. These attributes also eliminated from dataset. With these removals, dimensionality of dataset decreases to 561 attributes.

Although our dataset dimensionality reduced to almost one third, there might features that are not informative to learn from. Thus, further dimensionality reduction techniques are needs. These techniques are only performed on train set since test set is going to be unseen until time of testing. Thus we divide dataset and from now on, we work with 2161 samples and 561 features. To evaluate how dimensionality reduction techniques work, we train our models on the trainset after each technique implementation. For having a better view to compare outcomes, classification performance results on different dimensionality reduction techniques are provided in Table 4. Moreover, all of the remaining missing values based on the type attributes are imputed using mean or mode.

Before implementation of any reduction technique, performance of considered classifiers on 561 attributes checked and results are given in Table 3. As results show performance of almost all classifiers decreased in comparison with baseline. One of the reasons can be large number of attributes. There might be features that are misleading or they are in conflict with others in decision making for classifiers. Hence reducing dimensionality might be an effective step.

Table 3 - Classifiers performance on set of attributes after elimination of useless features

	SVM	KNN k=9	Logistic Regression	Naïve Bayes
Accuracy	65.1087	56.7330	67.9778	50.8561
Precision	0.6568	0.5873	0.6828	0.5748
Recall	0.6548	0.5768	0.6823	0.5294
f-score	0.6558	0.5820	0.6825	0.5511
AUC$_{(area\ under\ the\ curve)}$	0.6548	0.5768	0.6823	0.5294

2.4. Dimensionality reduction

There is a wide range of dimensionality reduction techniques. In this study 7 different methods including wrapper technique, inner correlation, person correlation, gain ratio, gain info and particle component analyses (PCA) are used. In general each method decides on a set of features that are most informative and have highest relation with the set of classes.

Wrapper Technique:

In this method, importance of each feature is check according to two evaluation measures. For discrete attributes such as nominal ones, accuracy is used to check how well a feature works in classification of object variable and root MSE considered for continues variables. For these measurement calculation, 10 fold cross validation on train dataset based on Naïve Bayes classifier is used. Also search technique to find the best attributes set to bidirectional. This technique reduces effective features to 28. Results of training are given in Table 4 .

Results show that although features decreased and performance increase in compare with consideration of all features, still accuracies and f-scores are below obtained results in baseline. Especially in Naïve Bayes classifier performance meaningfully decreased. By replacing train and test set and evaluating features' performance, number of selected features changes to 21 with 4 shared attributes with the set in phase 1. Testing trained models over selected features from second half reveals that performance of classifiers has not meaningfully changed. SVM and logistic regression have the highest performance in case of accuracy and f-score with around 68%. Therefore it can infer from results using wrapper technique for reduction along with Naïve Bayes as a feature importance measure cannot increase performance more than 68 percent. This performance is same as consideration of all features and less than baseline.

Inter Correlation

Another feature selection technique is to evaluate attributes according to their correlations with class feature. Highly correlated ones to the labels while having low inter-correlation with other features are preferred. By using this technique in phase 1, number of features reduces to 17. By comparing results, inner-correlation against wrapper method worked better, since performance of classifier increased and get to the same level as baseline. However, performance of Naïve Bayes classifier is still below. The point is that Naïve Bayes classifier work much better when it is train on the selected attributes from the second half of data(phase 2). Distribution of data in the second half caused more informative features being selected by this approach. This technique choose 20 features from the second half of data in which there were 8 identical attributes to the 17 selected ones from the first half.

Correlation

Pearson correlation between attributes and labels is another evaluation metric that can be used to select the most correlated ones. Thus the main difference of this technique with inner correlation is that correlation between attributes is not measured.

For nominal attributes a value by value basis checking is used to find correlation. In this study, top correlated features with correlation more than 0.15 are selected. Results of training on the top correlated features for phase 1 and phase 2 are given in Table 4. Correlation same as inner-correlation worked well and results are comparable with models developed over baseline. Moreover, Naïve Bayes model developed over features selected through correlation is absolutely better than the one developed over inner-correlation technique selected features.

Gain Ratio

Gain ratio is a measure indicates the amount of information we gain about an attribute (class) using another feature divided by the total information that attribute carries. Gain ratio is measured by following formula.

$$GainRatio(C, attribute) = \frac{H(C) - H(C|attribute)}{H(attribute)}$$

where C is a given class and H is the entropy of a class given an attribute (nominator) or entropy of an attribute regardless of a specified class (denominator).

In this study, features with more than 10% gain ratio selected for model training in phase 1 and phase 2. By testing models on the remaining half of data in each phase and comparing results, all classifiers have their lowest performance up to here. All classifiers have around 50% accuracy. These results indicate that selecting features based on their gain ratio are not informative for classifiers to learn. However, logistic regression and SVM same as other feature selection techniques did the best.

Gain information

In gain ratio decision on features was based on the proportion of condition information they carry to their total information. In gain information we only consider how informative each attribute is regarding class set and regardless of the total information it carries. Gain information is defined as

$$GainInfo(C, attribute) = H(C) - H(C|attribute)$$

By using this formula and selecting features with information gain more than 2%, 10 features selected. In phase 1 performance dramatically increased in compare with the results of grain ratio metric and reached to the same level as baseline models. Moreover, phase one and phase two had same performance although selected features were different.

Particle component analyses

The last dimensionality reduction technique in this study is Particle Component Analyses (PCA). PCA is a transformation technique to map data into a space with lower dimensions that are linearly separable. To map data it uses Eigen values and Eigen vector and these values determine how important a feature is. The feature with the larger Eigen value is more important than the feature with smaller value. Result of training on reduced sample space is given in last row of Table 4 . By

comparing results SVM and Logistic regression performed better than the other classifiers; however, their performance was much lower than models developed over selected attributes from other dimensionality reduction techniques. Naïve Bayes had the lowest performance with f-score less than 50%. Performance of PCA in phase 2 was not as well as other reduction methods too. The only point is kNN performed better than logistic regression with f-score of 0.60.

2.5. ROC Curve

Receiver Operating Curve is a graph that plots True Positive Rate (TPR) against False Positive Rate (FPR). TPR is also known as sensitivity since it is probability of correct detection, while FPR is defined as probability of false alarm. Therefore, a good classifier is the one with higher TPR and lower FPR. In ROC curve the best condition happens when the curve bended more toward left top corner. In ROC to measure how well a test works, the area under the curve (AUC) is considered as a comparison metric. The greater area means more useful a test is. In Figure 3 and 4 ROC curves from trained model in phase 1 and phase 2 are displayed. As we can see, SVM and logistic regression are the mostly bended curves toward left top corner and have the largest under curve area, while Naïve Bayes curve in many cases mapped on the diagonal line. This shows that Naïve Bayes was not successful in those cases. The area under the curve which is shown both in plots and tables with 'AUC' has larger value for classifiers that worked better. The largest AUC is for logistic regression on correlation reduction approach with 0.7095.

Along with ROC, correctly detection of classes specially positive cases is important. By comparing precision and recall of all classifiers in Table 4, we observe that the highest values are obtained by logistic regression classifier. Precision and recalls reported in this table are average two classes. For example in phase 1 of trained classifiers on reduced features using correlation technique, recall is 0.7. This value shows that 70% of all of the patients who had positive status were detected. Among classifiers, naïve Bayes had the lowest performance both in case of precision and recall in case of using PCA for dimensionality reduction. Precision in this case is 26% means that around one fourth of patients with diabetes where detected and recall of 50% means half of positive predicted samples were in fact negative.

3. Conclusion

This study started by discussing a dataset of diabetic patients consists of 4322 samples and 1611 features. Initially a baseline prototype using 13 features developed and performance of 4 classification algorithm evaluated. Then data passed through a series of preprocessing steps and finally to have a set of limited number of features different dimensionality reduction techniques used. Results revealed that although used classifiers are important in correctly classification of objects, however, reduction techniques play more significant role. Among all performed reduction methods, correlation and gain information had the best performance for all classifiers and performance of classifiers on selected features from these approaches were almost equal to the baseline and higher than training on all features. Moreover, results from phase 1 and phase 2 were almost equal which means that random split of data gives same amount of information to classifiers.

Table 4 - Accuracy, precision, recall, f-score and AUC of training classifiers on selected features from different reduction techniques. Results are two phases.

Techniques	Measures	Phase 1				Phase 2			
		SVM	kNN	Logistic Regression	Naïve Bayes	SVM	kNN	Logistic Regression	Naïve Bayes
Wrapper	Accuracy (%)	67.4688	65.4327	67.3299	51.5039	68.9033	66.9135	68.9496	53.9565
	Precision	0.6793	0.6568	0.6787	0.5826	0.7006	0.6775	0.7002	0.6116
	Recall	0.6779	0.6565	0.6769	0.5353	0.6832	0.6636	0.6839	0.5210
	F score	0.6786	0.6567	0.6778	0.5579	0.6918	0.6705	0.6919	0.5627
	$AUC_{(area\ under\ the\ curve)}$	0.6779	0.6565	0.6769	0.5353	0.6832	0.6636	0.6839	0.5210
Inner Correlation	Accuracy (%)	70.3378	66.3119	70.3841	52.8922	70.2453	64.5072	70.5229	61.7307
	Precision	0.7026	0.6621	0.7030	0.5498	0.7052	0.6444	0.7067	0.6562
	Recall	0.7028	0.6610	0.7030	0.5032	0.6992	0.6444	0.7025	0.6265
	F score	0.7027	0.6615	0.7030	0.5255	0.7022	0.6444	0.7046	0.6410
	$AUC_{(area\ under\ the\ curve)}$	0.7028	0.6610	0.7030	0.5032	0.6992	0.6444	0.7025	0.6265
Correlation	Accuracy (%)	70.6155	66.6821	71.0319	67.4225	70.4304	67.1911	70.0602	67.8852
	Precision	0.7053	0.6660	0.7095	0.6809	0.7064	0.6724	0.7024	0.6787
	Recall	0.7054	0.6642	0.7095	0.6782	0.7013	0.6693	0.6977	0.6790
	F score	0.7054	0.6651	0.7095	0.6796	0.7038	0.6709	0.7000	0.6788
	$AUC_{(area\ under\ the\ curve)}$	0.7054	0.6642	0.7095	0.6782	0.7013	0.6693	0.6977	0.6790
Gain Ratio	Accuracy (%)	53.0773	53.0773	53.0773	53.0773	52.4294	52.2906	52.4294	48.5886
	Precision	0.5723	0.5723	0.5723	0.5723	0.5883	0.5541	0.5883	0.6162
	Recall	0.5051	0.5051	0.5051	0.5051	0.5042	0.5029	0.5042	0.5060
	F score	0.5366	0.5366	0.5366	0.5366	0.5430	0.5273	0.5430	0.5557
	$AUC_{(area\ under\ the\ curve)}$	0.5051	0.5051	0.5051	0.5051	0.5042	0.5029	0.5042	0.5060
Gain Information	Accuracy (%)	70.6155	66.6821	71.0319	67.4225	70.1064	67.2374	70.6155	67.8390
	Precision	0.7053	0.6660	0.7095	0.6809	0.7040	0.6728	0.7083	0.6781
	Recall	0.7054	0.6642	0.7095	0.6782	0.6977	0.6698	0.7032	0.6783
	F score	0.7054	0.6651	0.7095	0.6796	0.7008	0.6713	0.7057	0.6782
	$AUC_{(area\ under\ the\ curve)}$	0.7054	0.6642	0.7095	0.6782	0.6977	0.6698	0.7032	0.6783
PCA	Accuracy (%)	55.7612	56.0389	56.5016	52.4757	58.5840	58.7691	56.2240	52.0592
	Precision	0.5578	0.5618	0.5624	0.4415	0.6716	0.6390	0.6580	0.2603
	Recall	0.5579	0.5478	0.5613	0.4985	0.5701	0.5737	0.5449	0.5000
	F score	0.5579	0.5547	0.5619	0.4683	0.6167	0.6046	0.5961	0.3424
	$AUC_{(area\ under\ the\ curve)}$	0.5579	0.5478	0.5613	0.4985	0.5701	0.5737	0.5449	0.5000

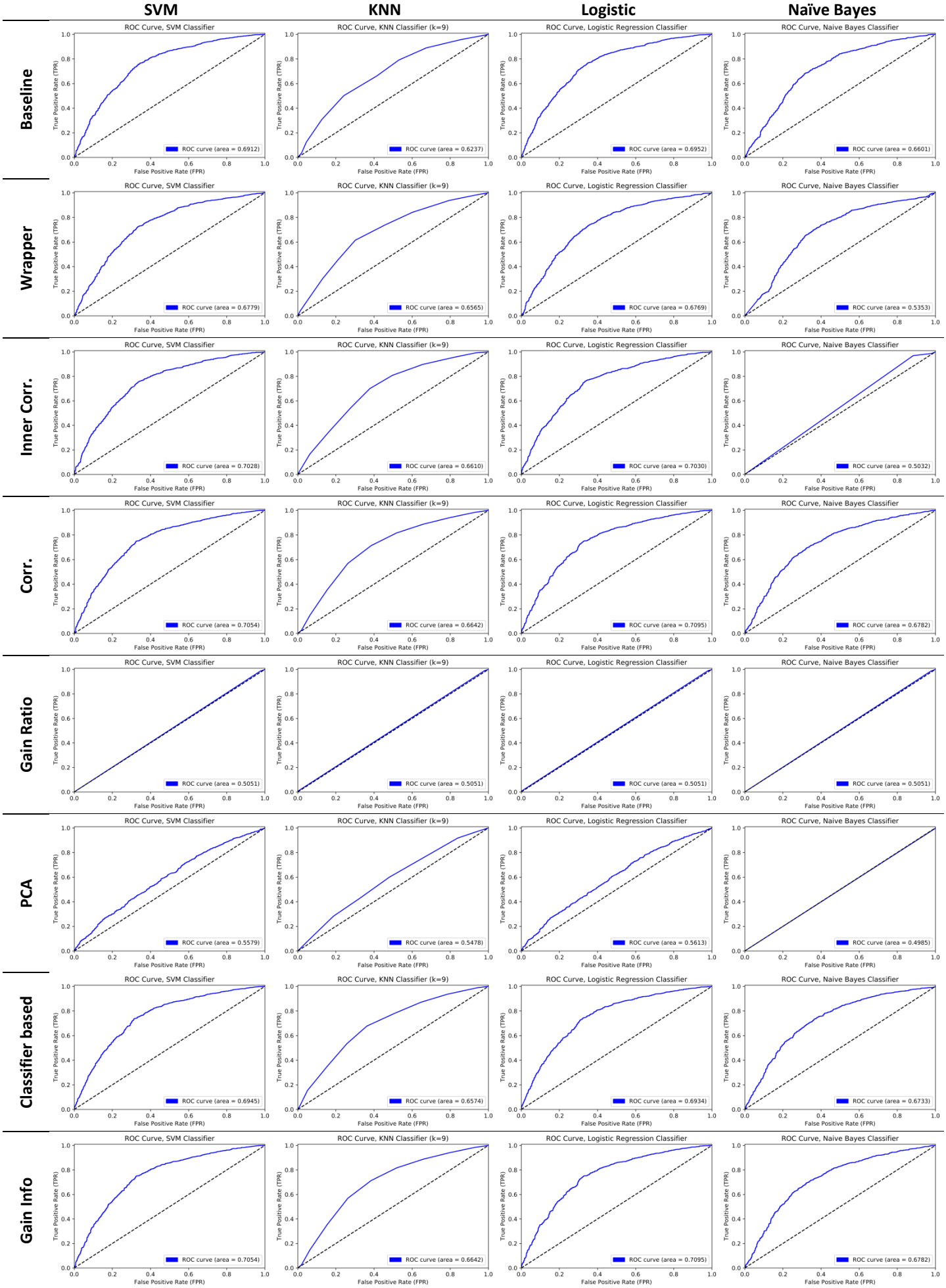

Figure 3 - ROC curves for all classifiers developed over different dimensionality reduction techniques in phase 1.

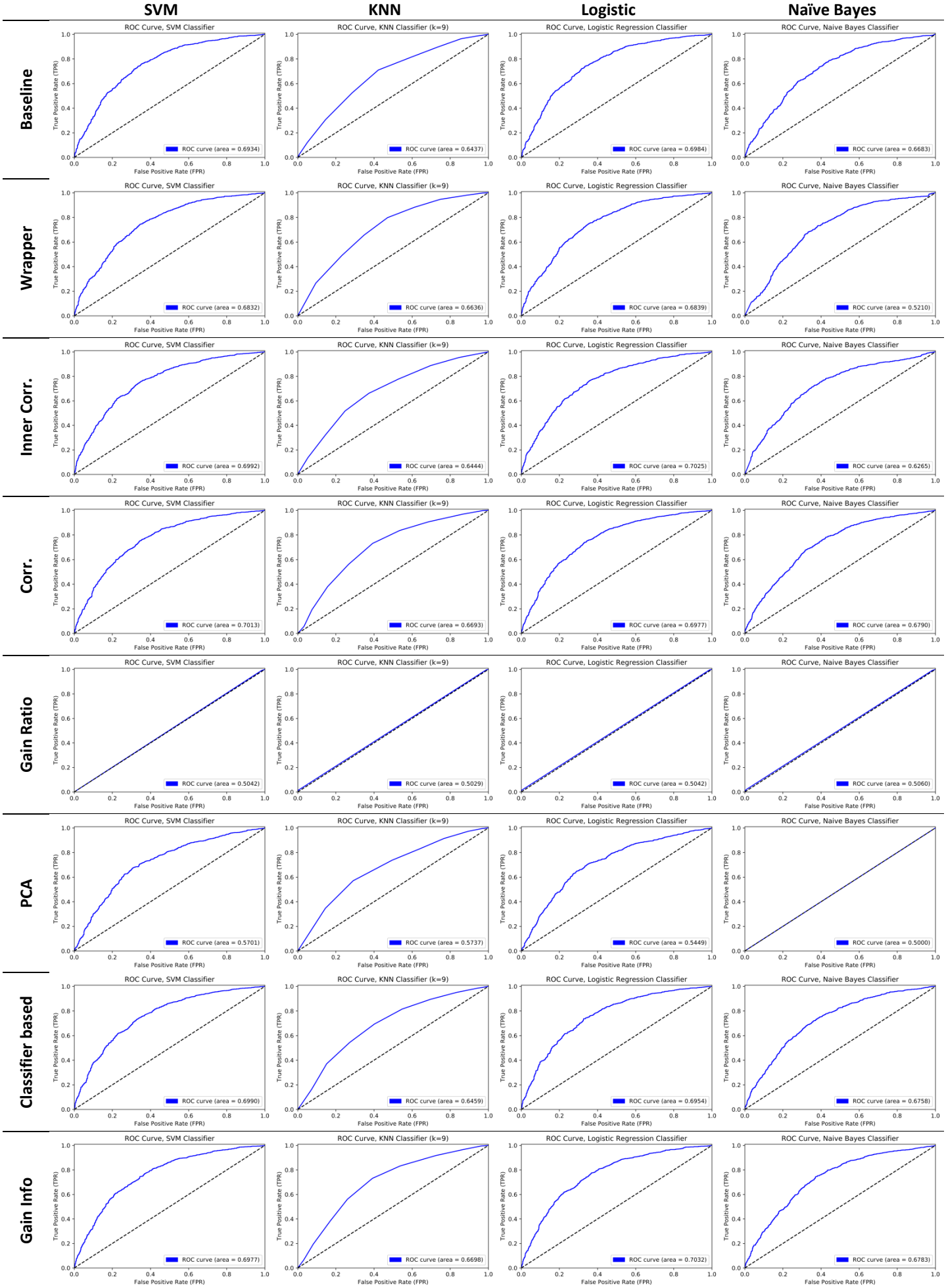

Figure 4 - ROC curves for all classifiers developed over different dimensionality reduction techniques in phase 2.