COMP 551
Mini Project 1: Machine Learning 101

Marilia Karla, Nithilasaravanan Kuppan, Jerry Dong *McGill University*

***Abstract*—In this project, the performance of two Machine Learning (ML) algorithms, namely Logistic Regression (LR) and Gaussian Naive Bayes(GNB) was investigated on four benchmark datasets. The data was preprocessed where missing or irrelevant variables were removed and certain datasets were normalized accordingly. Furthermore, all the datasets were split into training, testing and validation sets. The models were implemented and K-fold cross validation was used. Moreover, some additional features were proposed for means of improving the performance, which was chosen to be the log, quadratic and interactive terms of the most discriminant features, these were added only to GNB.**

I. INTRODUCTION

Fundamentally, the difference between generative and discriminative classification is that generative models learn the joint distribution whereas discriminative models map the conditional distribution[1]. Hence, generative models consider the functional form of P($x|y$)P(y) picking the most likely label y. On the other hand, discriminative models consider the functional form of P($x|y$) estimating directly from the dataset[2]. The generative classifier, Naive-Bayes (NB), and its discriminative counterpart, Logistic Regression (LR), are popular classification techniques. The decision on which classification technique to use is not always clear, however, many exhibit a preference for discriminative models. Ng and Jordan show that NB reaches its asymptomatic error quickly in relation to the number of training sets[3]. However, as one increases the number of training sets, LR will achieve a lower asymptomatic error rate.

A. The Task

This project investigates NB and LR exploring the relationship between them as we compare the classification methods on 4 different datasets. These included data captured from a radar of the ionosphere, the income on census data, different aspects of breast cancer and information regarding wine samples.

B. Related Work

On the ionosphere dataset, Sigillito et al. test several different feedforward neural networks to discriminate the 'good' from 'bad' radar returns of the ionosphere[4]. On the adult dataset, Kohavi demonstrates that the accuracy of Naive Bayes does not scale up as well as decision trees in larger data bases [5]. In the breast cancer dataset, Ratanamahatana compared C4.5 decision trees against Selective Bayesian Classifiers which demonstrated that Selective Bayesian Classifiers typically learn faster than both NB and C4.5 decision trees[6]. Finally, for the wine dataset, various datamining approaches have been used to predict taste preferences[7].

II. DATASETS

The data was first cleaned, processed and organized to suit the learning. The datasets was then split into a training, testing and validation set. The distribution of the data was 75%, 15% and 10%, respectively for all the sets for LR. Finally, 'fit' and 'predict' functions were implemented on the dataset and the classification was evaluated.

A. Ionosphere Dataset

The Space Physics Group of the John Hopkins University Applied Physics Laboratory collected this radar data[8]. The radar system which is located in Goose Bay, Labrador, employs a phased array of 16 high frequency antennas. In ionosphere research, a human must classify radar returns of the ionosphere as either suitable for further processing. This task can be extremely time consuming[x]. From the radar system, this dataset contains 34 attributes with the final attribute describing information on whether the radar returns 'Good' or 'Bad' evidence of some type of structure in the ionosphere. The first 34 attributes were labelled '1,2,3...34' with the final attribute being assigned 'target'. '0' was mapped with 'b' (bad) evidence (36%) and '1' was mapped with 'g' (good) evidence (64%).

B. Adult Dataset

The extraction of this data was done by Barry Becker from the 1994 Census database[9]. This dataset contains 14 attributes with the final attribute describing information on whether income was greater than or less than 50K. The attributes were assigned accordingly to the information they pertain to while the final attribute was assigned 'target'. '0' was mapped with target income less than '50K' (75%) and '1' was mapped with those with targets more than 50K(25%).

C. Breast Cancer Wisconcin Dataset

Breast Cancer Wisconcin is a dataset that contains 699 instances derived from clinical cases of breast cancer from Dr. Wolberg[10]. 8 groups of samples were collected chronologically. This dataset contains 10 attributes with the final attribute describing information on whether the cancer is benign (assigned the value of 2) (68%) or malignant (assigned the value of 4) (34%).

D. Wine Quality Dataset

Portugal is a top ten wine exporting country, with 3.17% of the market share in 2005. Exports of its 'Vinho Verde' wine (from the northwest region) have increased by 36% from 1997 to 2007[11]. Wine Qualtiy is a dataset that contains information of white variants of the Portuguese 'Vinho Verde' wine. Hence, wine certification and assessment through physicochemical and sensory tests are important for the growth of the wine industry. This dataset contains 12 attributes with the final attribute describing information on the quality of the wine assigning a score from 1-10 where 1-5 is judged as bad (47%) and 6-10 is judged as good (53%).

E. Data Preprocessing

An important step before the training machine learning classifier is to pre-process the data. In this subsection, it will be explained how the data cleaning was performed.

1) Unknown Data: Some data sets showed unknown data that were replaced with "?". These data were ignored and removed from the classification analysis.

2) Remove Data: On adult-dataset, the attribute 'education' was dropped as 'education_num' contained the same relevant information. In adult.test, the first row ('1x3 Cross validator') was dropped and then adult.test and adult.data were combined together. The additional spaces found in the attribute names were also removed. The presence of missing or null values were checked for and resolved by removing them from the data. Attributes 'fnlwgt','education_num','capital_gain','capital_loss','hours _per_week' were normalized.

On Ionosphere-dataset, the second feature was remove because all values was '0'.

3) Categorical to Continuous type: On adult-dataset, it was used the concept of one-hot encoding to convert categorical type in continuous. Thus, was possible to be an analysis of the correlation between different type of features.

4) Binarized Output Labels: On wine-red dataset, the output was composed by the scalar number up to 10, which corresponded to the quality of the wine. To binarize the output, all labels greater than 6 were considered as '1' and the rest as '0'.

For the cancer data set, the output was formed for '4' indicating malignant breast cancer and '2', indicating benign breast cancer, so the same process was used, where it was replaced by '1' if they were equal to '4' and '0' if they were equal to '2'.

The Adult data set and the Ionosphere data set have output labels such as less than or equal to '50K' or more than '50K' and 'b and g' respectively. Both cases are binary, but the output is a string.

F. Data Statistical Analysis

Prior to implementing classification analysis, the interaction between each of the features was analyzed using a covariance matrix, as illustrated in Figure 1, 3, 4 and 5.

As observed in Figure 1, the alcohol is positively correlated with wine quality, while being negatively correlated with volatile acidity.

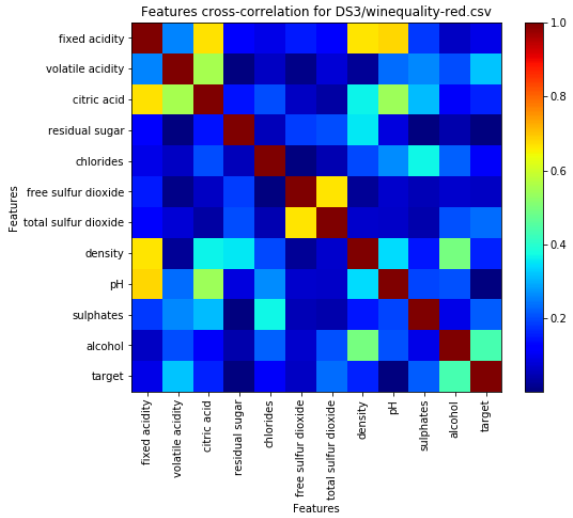

Figure 1: Covariance matrix for wine dataset. The target vector correspond to quality

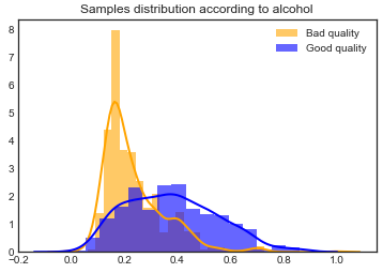

Figure 2: Distribution of samples for good quality and bad quality wine, for the alcohol features

In addition, to explore the discriminative power it is possible to observe the distribution of one of the features according to the correlation matrix, to exemplify this, a plot was made for the wine dataset with respect to the amount of alcohol present, Figure 2.

On the Cancer-dataset the features Bare Nuclei, Uniformity of Cell Shape and Uniformity of Cell Size are highly correlated with malignant breast cancer.Therefore, these features are considered to have high discriminative power.

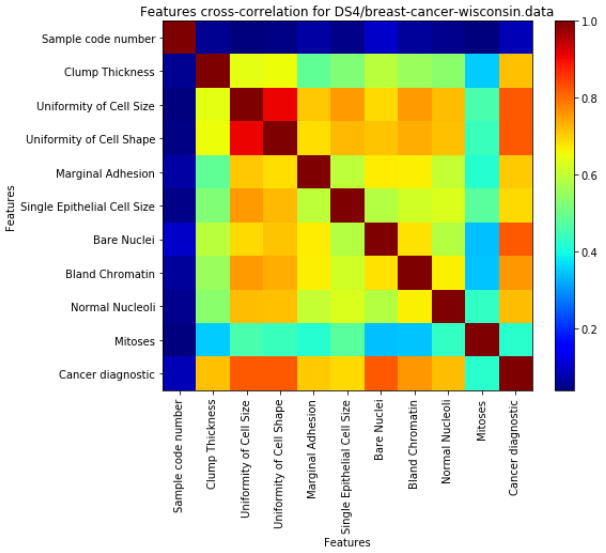

Figure 3: Covariance matrix for Cancer dataset.

For the Ionosphere dataset, it is possible to note that Feat2 and Feat4 were the ones that had the high discriminative power.

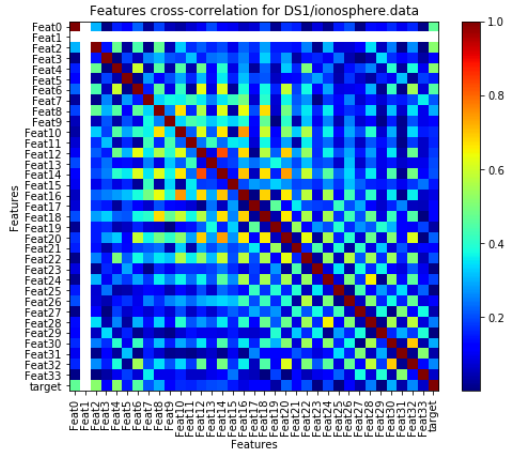

Figure 4: Covariance matrix for Ionosphere dataset.

For the Adult dataset, the age and sex features were those with the greatest discriminative power.

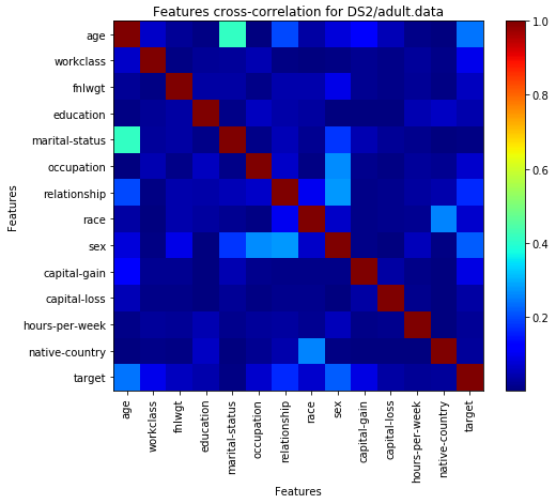

Figure 5: Covariance matrix for Adult dataset.

III. Results

In this section we shall briefly discuss about the logic of the algorithms implemented along with the model metrics associated with each dataset.

A. Logistic Regression

The Logistic Regression logic class implemented in this report has 5 distinct functions - *sigmoid_func* to calculate the squashing function, *LossFunc* to calculate the cost at a given set of inputs, *Fit* to perform full batch gradient descent and as a result, figure out the optimal model weights for the given inputs, *predict* to apply the best weights and predict the class and *cv* to perform cross validation.

After cleaning, normalizing and pre-processing the data, the first step is to find the best model weights with the appropriate learning rate (alpha) and number of iterations. The *cv* function does exactly this - we provided it with the Input features, target (*all from training*) and iteratively provided different alphas and calculated the accuracies. Once we found the best alpha, we used the same to train our model with the training data, find the correct predictions with the test data and then calculated the model evaluation metrics.

The following table summarizes the model metrics for all four datasets:

Table I: Model evaluation metrics

Dataset	Accuracy	Precision	Recall
Ionosphere	92.45%	90.24%	100%
Adult	85.46%	72.99%	62.29%
Breast Cancer	93.33%	90.32%	93.33%
Wine Quality	74.58%	74.6%	76.42%

We can see that Logistic Regression does perform well, in fact, exceptionally well in the first and third datasets to correctly classify the data. It is also interesting to see how the cost (*calculated on the training set*) varies with respect to the number of iterations of gradient descent.

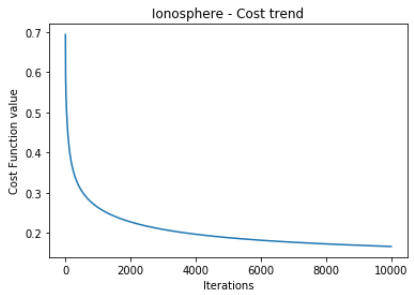

Figure 6

We can see that the cost value decreases with increasing iterations and it settles after reaches an optimum value. The table below shows the learning rates and the #iterations of gradient descent that gave the above mentioned accuracies for all the datasets:

Table II: Model parameters

Dataset	**Learning rate**	**#iterations**
Ionosphere	0.1	10,000
Adult	1	12,000
Breast Cancer	0.05	5,000
Wine Quality	0.6	6,000

B. Naive Bayes

Contrary to others generative learning models LDA and QDA, Naive Bayes assume that each features are independent from the others which simplify the computation of the conditional probability $P(x \mid y)$ in the Bayes rule Equation 1 in the Equation 2.

$$P(y = k|x) = \frac{P(x|y = k) \times P(y = k)}{P(x)} \tag{1}$$

$$P(x|y) = P(x_1|y) \times ... \times P(x_m|y) \tag{2}$$

1) Gaussian Naive Bayes: In the case of Gaussian Naive Bayes each conditional probability presented in Equation 2 is assumed to be a Gaussian distribution. The parameters computed during fitting were the mean and variance of the samples across each features for each class. Equation 3 shows the expression of the conditional probability during prediction.

$$P(x_j|y) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{0.5*(\frac{x-\mu}{\sigma})^2} \tag{3}$$

The result presented in Table III are obtained with a 5 Stratified fold cross-validation for GNB. In addition, through the analysis of the correlation matrix, explained above, a feature selection (FS) was carried out with the most interesting features of each dataset, in order to compare the accuracy of the original data set and the dataset with the FS Table IV.

In addition, new features were added to the dataset, Quadratic, Interactive and Log, aiming to compare the results of the dataset with the original features and the dataset with new features. Where the FS was also held. The result is shown in the Table V.

Table III: Original (ori) dataset - GNB

Dataset	avg_acc	std_acc	avg_pre	std_pre	avg_rcl	std_rcl
Cancer_ori	94.7%	2.4%	87.3%	5.1%	99.6%	0.9%
Wine_ori	72%	2.5%	76.1%	6.1%	71.8%	9.2%
Adult_ori	79.2%	0.6%	62.6%	2.1%	41%	0.8%
Ion_orig	59.7%	5.6%	47.2%	3.6%	99.2%	1.6%

Table IV: Result of FS with 5 and 3 features in the original dataset- GNB

Dataset	avg_acc	std_acc	avg_pre	std_pre	avg_rcl	std_rcl
Cancer_ori_fs5	95.4%	2.6%	89%	5.8%	99.6%	0.9%
Wine_ori_fs5	72.2%	3.5%	76.6%	6.4%	70.9%	7.9%
Adult_ori_fs5	78.3%	7.5%	64.8%	13.5%	78.4%	7.4%
Cancer_ori_fs3	96.6%	2.2%	92.3%	4.9%	98.7%	1.7%
Wine_ori_fs3	73.5%	2.3%	77.5%	6%	73.2%	10.4%
Adult_ori_fs3	77.8%	0.4%	66.7%	3.1%	21.4%	0.3%
Ion_ori_fs3	80.3%	6.4%	73.9%	12.6%	72.8%	9.9%

Table V: Result with the augmentation of Features) and FS - GNB

Dataset	avg_acc	std_acc	avg_pre	std_pre	avg_rcl	std_rcl
Cancer_aug	96%	1.5%	90.1%	3.5%	99.6%	0.9%
Wine_aug	70.3%	3.3%	77%	5.6%	64.8%	8.5%
Adult_aug	78.3%	0.4%	55.7%	0.8%	63.9%	0.6%
Ion_aug	64%	5.8%	50.2%	4.5%	88.8%	13.9%
Cancer_aug_fs5	96.7%	1.6%	93.6%	3%	97.4%	3.1%
Wine_aug_fs5	69.8%	7%	75%	5.4%	66.1%	16.5%
Adult_aug_fs5	74.2%	0.6%	48.1%	1.2%	43.2%	0.9%
Ion_aug_fs5	84.3%	5%	73%	7.2%	90.4%	4.8%
Cancer_aug_fs3	96.7%	1.6%	93.7%	4.3%	97.4%	2.5%
Wine_aug_fs3	69.5%	6.6%	74.3%	4.8%	66.5%	16.8%
Adult_aug_fs3	72.2%	0.9%	43.3%	1.9%	37.3%	1.3%
Ion_aug_fs3	86%	4.6%	75.8%	5.7%	89.6%	7.4%

Finally, for a better view, the Table VI shows the best results for each dataset.

Table VI: Best results - GNB

Dataset	avg_acc	std_acc	avg_pre	std_pre	avg_rcl	std_rcl
Cancer_aug_fs5	96.7%	1.60%	93.6%	3%	97.4%	3.1%
Wine_ori_fs3	73.5%	2.30%	77.5%	6%	73.2%	10.4%
Adult_ori	79.2%	0.6%	62.6%	2.1%	41%	0.8%
Ion_aug_fs3	86%	4.6%	75.8%	5.7%	89.6%	7.4%

IV. Experiments

A. Comparing Accuracies

Both the models were first put through K Fold cross validation before calculating the final metrics using the test data. Table VII compares the average accuracies computed from both Logistic Regression and Naive Bayes (*K was set as 5*)

This shows that for the four datasets that we analyzed, Logistic Regression outperformed Naive Bayes, except the case of Breast cancer data where Naive bayes did better because of the lack of data. This proves the fact that when the

Table VII

Dataset	Logistic Regression	Naive Bayes
Ionosphere	83.91%	59.7%
Adult	84.91%	79.2%
Breast Cancer	89.9%	94.7%
Wine Quality	74.53%	72%

training size reaches infinity the discriminative model: logistic regression performs better than the generative model Naive Bayes [1].

B. Varying learning rates

This experiment investigates the effect of learning rates on accuracies and then investigating how accuracies change with changing #iterations of gradient descent.

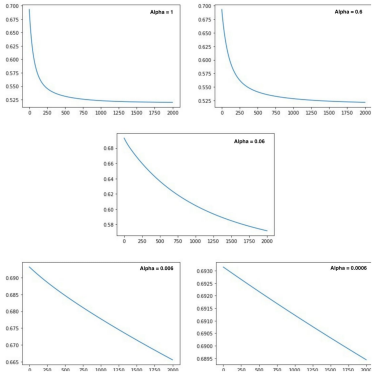

Figure 7

For this purpose, we took the Ionosphere dataset and performed K Fold analysis for various alphas.

Table VIII: Ionosphere - Varying learning rates

Alpha	K = 1	K = 2	K = 3	K = 4	K = 5	Avg
1	80%	83.33%	88.33%	81.67%	86.21%	83.9%
0.1	80%	86.67%	90%	78.33%	89.66%	**84.93**%
0.01	73.33%	81.67%	86.67%	73.33%	82.76%	79.55%
0.001	66.67%	68.33%	78.33%	66.67%	72.41%	70.48%
0.0001	65%	68.33%	75%	65%	70.69%	68.8%

We can clearly see that Alpha = 0.1 gives us the best accuracy therefore, we used the same to do a K Fold CV study of its trends with respect to #iterations of gradient descent. The figure below elucidate the fact that at a constant learning rate, the increase in number of iterations increases the accuracy till it reaches the optimum value. We also implemented a stopping criteria based on the training error - if the difference between two consecutive cost values is less than 0.00001 then it terminates the loop and outputs the optimum weights. This method of early stopping is essential because it saves crucial run time and resources getting occupied while the gradient descent iteration runs without fruition.

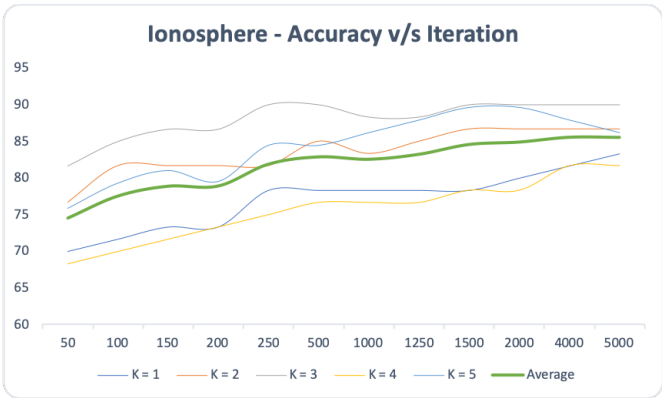

Figure 8

C. Train size variation

For logistic regression, 90% of the total data was allocated as the training data and the rest was kept aside for testing. Out of that 90%, we allocated 100%, 90%, 75%, 60%, 50% and 30% iteratively for training and calculated the average accuracy on the test set. As expected, we see a dip in the accuracy as we decrease the training set size.

Table IX: Accuracy trends with # iterations
Alpha = 0.05

% of Training data	Test Accuracy
100%	73.12%
90%	71.88%
75%	67.5%
60%	66.25%
50%	66.25%
30%	58.75%

This experiment strengthens the fact that a model must be fed as much of training data as possible, but not at the expense of scarce testing data. We must find a balance between the ratios of the different sets to ensure that a) the model has enough data to read the patterns and compute the weights optimally and b) the model has enough testing data to reliably compute the test metrics.

V. Discussion and Conclusion

The primary motive of this project is to understand the crux of two major classification algorithms - Logistic Regression and Naive Bayes. These two algorithms were employed on four different datasets.

- It is important to pre-process data in order to get accurate results
- Small alpha will not assist in attaining the optimum but large alpha will overshoot it
- K Fold CV is useful in figuring out the hyperparameters, thus avoiding errors
- Logistic Regression outperforms Naive Bayes except when the model ran on breasy cancer data - this is true because of the scarcity of training data

One of the things that should be given more importance is the Data pre-processing part. This report has hinted how powerful feature selection and feature creation can be in terms of making the model accurate as well as more efficient. Another aspect that would improve our models would be regularization. This work has helped us incisively understand these algorithms and the associated steps involved in a Machine Learning project.

We believe we've covered all the tasks enlisted in the exercise which engendered in us delving deeper into these algorithms and try out feature engineering on the Naive Bayes models.

In the original dataset, LR achieved better results than in GNB. In addition, the two models were able to learn, obtaining values greater than the flow chance. which were 75.1% (Adult-dataset), 64.1% (Ionesphere-dataset), 53.47% (Wine-dataset) and 65% (Cancer-dataset).

It was also observed that adding new features to the datasets led to to better results.This can be explained by the fact that the classifier has increased complexity, which in some cases allows it to learn better.

VI. Statement of Contributions

The collaborative work of this team involved everyone contributing to data processing. Jerry contributed to the literature survey, class splits and writing of the report.

Nithilasaravanan contributed to implementing Logistic Regression, kfold cross validation, experimentation and assisted on writing the report.

Marilia worked on implementing Naive Bayes, kfold cross validation, experimentation and assisted on writing the report.

References

[1] L.M. Gladence, M.Karthi and V. Maria, A statistical comparison of logistic regression and different bayes classification methods for machine learning, *ARPN Journal of Engineering and Applied Sciences*, vol.10, pp.5947-5953, Jan. 2015.

[2] J.W. Halloran, Classification: Naive Bayes vs Logistic Regression,*Technical Report*, 2009.

[3] A.Y. Ng and M.I. Jordan,On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes, *Neural Information Processing Systems*, 2002.

[4] V. Sigillito et al. Classification of radar returns from the ionosphere using neural networks, *Computer Science*, 1989.

[5] . R.Kohavi. Scaling up the accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid, 1996.

[6] . C.A. Ratanamahatana and D. Gunopulos Scaling up the Naive Bayesian Classifier: using Decision Trees for feature selection, 2002.

[7] P.Cortez et al. Modeling wine preferences by data mining from physicochemical properties, *Decision Support Systems*, vol.47, pp.547-553, Nov. 2009.

[8] https://archive.ics.uci.edu/ml/datasets/ionosphere

[9] https://archive.ics.uci.edu/ml/datasets/Adult

[10] https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[11] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29