An Automated Technique using Gaussian Naïve Bayes Classifier to Classify Breast Cancer

B. M. Gayathri
Research Scholar,
M.S.University
S.D.N.B Vasihnav College for Women,Chromepet,Ch-44

C. P. Sumathi, PhD
Associate Professor and Head
S.D.N.B Vaishnav College for Women
Chromepet,Chennai-44

ABSTRACT
Objectives: The proposed work is to classify breast cancer with few attributes. Reducing the attributes reduces the time, so that the patient need not wait for result for a long time. For classification, the user friendly environment is created. The user can enter the details of the patient such as Clumpthickness, Uniformity in cell size etc., and the result is classified as benign or malignant. **Statistical analysis:** Variable selection is done by one of the variable reduction algorithm called Linear Discriminant Analysis (LDA). LDA is one of the statistical method. The dataset is passed to LDA function repeatedly and the combination of variables which gave the good accuracy is selected. The variables that are selected by using LDA are used in classifying breast cancer. **Findings:** This application is created to find whether the given record is benign or malignant tumor. In this proposed work, the dataset from UCI repository for breast cancer detection is used. There are many other works done for finding breast cancer risk, diagnosing breast cancer etc., and there may be at least ten variables used for classification which may be time consuming. But in this proposed work, only four are used and it gave the accuracy of up to 96%. Hence this may be the first step or idea for detecting breast cancer with lesser variables, so that this may be helpful for the doctors. **Improvements:** The proposed work is done based on the UCI machine learning repository dataset, which was uploaded by Wisconsin Hospitals, Madrid. Some changes can be made in the coding and this methodology can also be implemented in other dataset also by reducing the attributes.

General Terms
Naive bayes classifier, Linear Discriminant Analysis, Wisconsin, Machine learning.

Keywords
Classification, Mahalanobis, Normalization, Fisher, data-preprocessing

1. INTRODUCTION
Breast cancer is caused due to several reasons, with possibility due to different life styles and eating habits. Other than this, it may occur because of genetic disorder, getting older, alcohol consumption etc. There are different kinds of breast cancer such as DCIS (Ductal Carcinoma in Situ), IDC (Invasive Ductal Carcinoma), ILC (Invasive Lobular Carcinoma), LCIS (Lobular Carcinoma in Situ) etc.; Classifying cancer has many challenges in the field of data mining. Researchers have developed many techniques for diagnosing breast cancer. Machine Learning is used mostly in cancer research. This paper deals with Bayesian statistics- The Naive bayes classifier which is used to classify the given data as benign or malignant. There are many approaches used for breast cancer diagnosis using machine learning techniques and data mining. Mostly these techniques are used by using mammography images in which they use image processing method in combination with machine learning technique to classify tumor. Breast cancer risk evaluator tool for evaluating multiple breast cancer factors was developed[1]. The author used BIRAD (Breast Imaging Reporting and Data System) score for evaluating breast cancer risk and the tool was developed using Java programming language. The digital mammography technique for detecting breast cancer in its benign stage was designed[2]. The K-Nearest Neighborhood classifier was used for calculating the breast tissue density. The features were extracted by using Region of Interest (ROI) and the new approach using wavelet transforms for detecting breast cancer[3]. The authors have proposed this tool for diagnosing the breast cancer in earlier stage. Various segmentation techniques were used and both microcalcification and masses are detected from the image. Next Histopathological images[4] was used for diagnosing breast cancer using feed forward back propagation neural network technique for finding breast cancer.

The algorithm was developed to predict the recurrent events[5] in breast cancer using the Wisconsin prognostic dataset. The author used naive bayes classifier for classifying the recurrent events. The new algorithm using kernel based naïve bayes classifier for breast cancer prediction[6] has been developed. The author used mammography data for tumor prediction. GUI was designed for detecting probability of having breast cancer in women[7] in future. They have used totally nine attributes for prediction. A new method for variable selection by using clustering and classification[8] was developed. In this work the author focuses on model based learning and the authors have compared with other variable selection techniques on the real time dataset.

A comparative study on different classification techniques on breast cancer using FNA data[9] was done. The authors have analyzed the dataset using Support vector machines and Bayesian network approach and SVM gave better performance than Bayesian approach. The design for computer detection system for breast cancer using naïve bayes and KNN (K-Nearest Neighborhood) [10] approaches has been developed. In this system mammographic mass dataset is used which is based on BI-RADS (Breast Imaging Reporting and Data System). The missing values in this dataset are replaced by missing value imputation techniques. In the conclusion author shows the difference between the accuracy before and after using imputation technique. In this work, the accuracy was up to 83%.

The paper was presented by comparing three different methods for classification[11] of breast cancer data. These classifiers can be used in machine learning. They are Naïve Bayes, C4.5 decision tree algorithm and Multilayer perceptron

function. The dataset used for this work was collected from Nottingham Tenovus Primary Breast Carcinoma Series from 1076 patients. The system was developed using Weka software. All the three classifiers were compared and Naïve bayes gave the accuracy of up to 86.9%. The feature selection and classification for diagnosing breast cancer[12] was done using Wisconsin diagnostic data. The author have reduced the attributes to 3 out of 30 candidates and used multilayer perceptron as a classifier. The proposed method is different from the previous works. In this technique Wisconsin original data set is used and the system is designed as a user friendly environment .User can enter the details of cancer, which calculates and predicts whether the entered data is benign or malignant.

2. NAÏVE BAYES CLASSIFIER

Naïve bayes classifier is a simple probabilistic classifier based on applying bayes theorem. Naïve bayes considers each and every feature variable as independent variable. This classifier can be trained very efficiently in supervised learning and also can be used in complex real world situations. The main advantage of Naïve bayes is that it requires small amount of training data which are necessary for classification. The classification is done by bayes rule to calculate the probability of class label C, given that the particular instance $X_1 ... X_n$, by the formula $P(C=c|X_1=x_1,\ldots\ldots, X_n=x_n)$. Classifier can be defined as Classify $(F_1 \ldots F_n)$ =argmax $P(C=c)\prod_{i=1}^{n} P(F_i = fi|C=c)$ where $F_1 \ldots F_n$ =Feature Variables and C = Class label.

2.1 Gaussian Naïve Bayes

When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to Gaussian distribution. The training data are segmented by class and the mean and variance of each class is calculated. Therefore to estimate the probabilities of continuous dataset the following formula can be used.

$P(v_j|c_i) = \frac{1}{\sqrt{2\pi\sigma_{ji}}} e\left(-\frac{(v_j-\mu_{ij})^2}{2\sigma^2_{ji}}\right)$ Where v = variable, c = class.

3. DATA PREPROCESSING

Data preprocessing is an important step in data mining. The dataset that are used in research must be pre-processed before applying it for any kind of classification. Those dataset may contain impossible data combination, missing values, redundant information etc., .This may lead to misleading results. Therefore the quality of data must be improved before running an analysis. The following section shows the data preprocessing done for this work[2].

3.1 Replacing missing values

The dataset used in this algorithm consist of 699 sample records and 10 attributes excluding Sample id number, in which one of the attribute have missing data, which are replaced by one of the data preprocessing method called data cleaning, by which the missing data are replaced by finding median of that attribute.

3.2 Variable selection

The variables that are used in this analysis are selected by one of the feature extraction method called Linear Discriminant Analysis (LDA). LDA is used in statistics, pattern recognition and in machine learning, to find the combination of features which separates two classes and its results can be used as a linear classifier. In this dataset, different combinations of variables are selected and its results are analyzed. The combinations of variables that explains the data best is used in this analysis. The dataset is normalized by finding the Log_{10} value for each and every attributes, so that the dataset becomes consistent[15]

3.2.1 Feature selection using LDA

LDA is based upon the concept of searching for a linear combination of variables that best separate two classes. The purpose of using LDA is for selecting the best features from the given set of attributes, which gives more accurate results. Linear Discriminant Analysis is done before applying dataset in the application. The large sets of attributes are reduced into fewer attributes by using LDA, so that it reduces complexity in training and testing data[13].

Fisher defined the following score function to capture the concept of separability. Hence the score function is $S(\alpha) = \frac{\alpha^T \mu_1 - \alpha^T \mu_2}{\alpha^T C \alpha}$ where μ_1 = Mean of first subset, μ_2 = Mean of second subset, T=targets, α= Linear model coefficients, C=Pooled Covariance matrix[14]. The use of function score is, to find the population means for each of given population μ_i. In this formula α is calculated as $\alpha = C^{-1}(\mu_1 - \mu_2)$ which is a linear coefficient which maximizes the score. In this function C represents pooled covariance matrix, μ_1 and μ_2 represents means of first and second subset. Since the LDA comes under the category of covariance, the covariance is calculated with the following equation, which gives the covariance matrix. Hence the equation is $C=\frac{1}{n_1+n_2}(n_1C_1+n_2C_2)$ is a pooled covariance matrix where α =Linear model coefficient, n_1=count of dataset belonging to class a, n_2=count of dataset belonging to class b C_1, C2 =Covariance matrices, μ_1, μ_2 =Mean vectors. After findingα, the values can be substituted in the scoring function $Z=x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$ and to find the effectiveness of discriminant, the Mahalanobis distance between two groups are calculated. $\Delta^2 = \alpha^T(\mu_1 - \mu_2)$ is the formula for Mahalanobis[19] distance where Δ= Mahalanobis distance between 2 groups, α=Linear coefficient model, μ_1 and μ_2.are means of first and second subsets, T=targets. The advantage of using Mahalanobis distance is, it provides a powerful method for calculating some set of conditions to an ideal set of conditions and this method was introduced by P.C.Mahalanobis[19] in the year 1936. To test whether the data belong to Class 1 or Class 2 the following function is used. $\alpha^T \left[x - \left[\frac{\mu_1+\mu_2}{2}\right]\right] > log\ \frac{P(1)}{P(2)}$ where α^T=Coefficient vector, x =Data vector, μ_1 & μ_2 are mean vector, p(1) & p(2) are class probability. The table 1 shows the features of dataset

.

Table 1. Features of dataset

Sno	Attributes	Range of Values
1	Clumpthickness	1-10
2	Uniformity cell size	1-10
3	Uniformity cell shape	1-10
4	Marginal Adhesion	1-10
5	Single Epithelian cell size	1-10
6	Bare Nuclei	1-10
7	Normal Nucleoli	1-10
8	Mitosis	1-10
9	Class	1-10

Table 2. Accuracy of classification for different set of attributes before normalization.

Attributes	Accuracy
Clumpthickness,Uniformity cell size, Uniformity cell shape, Marginal Adhesion	92.0%
Clumpthickness, Uniformity cell size,Uniformity cell shape,Single epithelian cell size.	93.7%
Clumpthickness,Uniformity cell size,Uniformity cell shape,Bland Chromatin	94%
Clumpthickness,Uniformity cell size,Uniformity cell shape,Normal Nucleoli	92%
Clumpthickness,Uniformity cell size,Uniformity cell shape,Mitosis	92.3%
Clumpthickness,Uniformity cell size, Uniformity cell shape, Bare Nuclei	92.7%
Bare nuclei, Marginal adhesion, Single epithelian cell size, Normal nucleoli	91.3%
Bland chromatin, Marginal adhesion, Single epithelian cell size, Normal Nucleoli	88.7%
Bland Chromatin, Marginal adhesion, Single epithelian cell size, mitosis.	88.7%

Table 3. Accuracy of classification for different set of attributes after normalization.

Attributes	Accuracy
Clumpthickness,Uniformity cell size, Uniformity cell shape, Marginal Adhesion	95.0%
Clumpthickness, Uniformity cell size, Uniformity cell shape, Single epithelian cell size.	94.3%
Clumpthickness,Uniformity cell size,Uniformity cell shape,Bland Chromatin	94.3%
Clumpthickness,Uniformity cell size,Uniformity cell shape,Normal Nucleoli	94.7%
Clumpthickness,Uniformity cell size,Uniformity cell shape,Mitosis	94.3%
Clumpthickness,Uniformity cell size,Uniformity cell shape, Bare Nuclei	94.3%
Bare nuclei, Marginal adhesion, Single epithelian cell size, Normal nucleoli	92.7%
Bland chromatin, Marginal adhesion, Single epithelian cell size, Normal Nucleoli	92.7%
Bland Chromatin, Marginal adhesion, Single epithelian cell size, mitosis.	91.3%

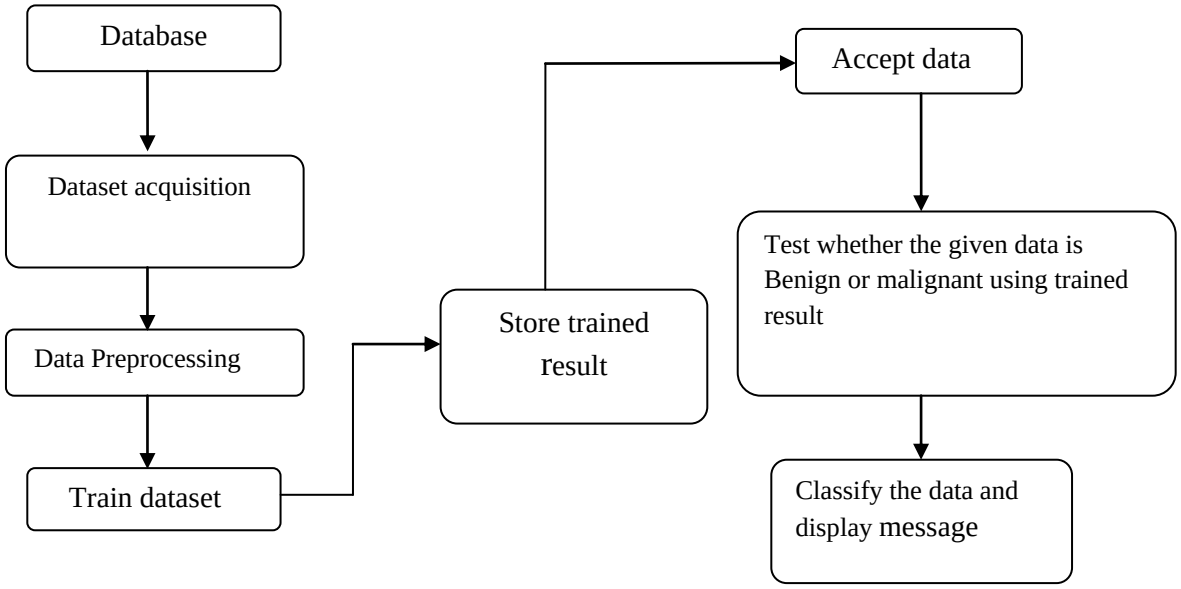

Figure 1 . Model Flow Chart

Since there is no much accuracy difference between all the nine attributes and four attributes, only four attributes are used in this work for classification. The Figure 1 shows the model flow chart of the proposed work.

3.2.2 Algorithm for performing LDA.

Step 1: Load dataset.

Step 2: Compute LDA for each and every attribute and select set of attributes.

Step 3: Check for the percentage of accuracy for each and every set of attributes (i.e., how effectively the attributes classify with fewer attributes.)

Step 4: If accuracy is better than previous set of attributes select those attributes for designing application.

Step5: Continue step 3 until step 4 is reached.

The following Figure 2 shows the result of LDA and the accuracy of each set of attributes in percentage before and after normalizing dataset by reducing size of attributes.

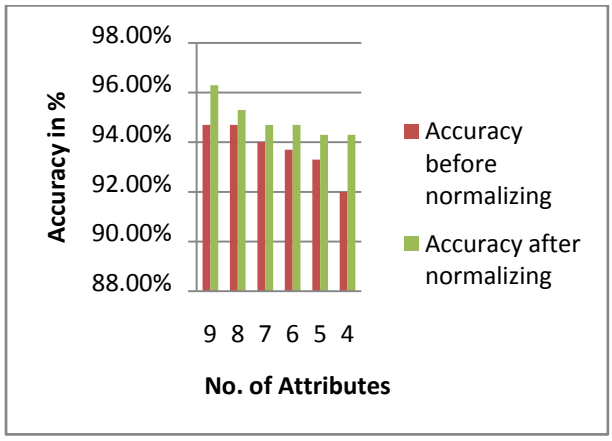

Figure 2. Accuracy of attributes before and after normalization

4. METHODOLOGY

The proposed algorithm used in this technique is as follows.

1. Load dataset.
2. Replace the missing values by using median method.
3. 50% of dataset is divided for training and remaining 50% of dataset is used for testing.
4. LDA is used for feature selection. Out of nine attributes only four attributes are selected for classification.
5. DA is used for feature selection. Out of nine attributes only four attributes are selected for classification.
6. Perform calculation using naive bayes classifier and the result is stored in a variable.
7. The given data is classified by using the stored result and pre classification rule. If both conditions are satisfied, the data is classified as Benign or malignant. Figure 3 shows the GUI interface design for classification.

Next, the following table 3 shows the accuracy of classification for different set of attributes after normalization. The accuracy is greater when the data is normalized. It is found that the accuracy for the combination of attributes Clumpthickness, Uniformity Cell Size, Uniformity Cell Shape and Marginal Adhesion is more (i.e., 95%) than other set of attributes and also without the Clumpthickness, Uniformity Cell Size, Uniformity Cell Shape the accuracy of classification is lesser(i.e.,92.7%).Usage of nine attributes gives the accuracy of up to 95.7%.

4.1 Implementation and result of this proposed work.

The proposed work runs on an Intel core i3 processor. This approach is applied in Wisconsin original dataset which consist of 699 samples of which 458 records are benign and 241 records are malignant. The dataset is divided into training and testing dataset. For classification, testing dataset of about 300 samples was used. Each and every data was entered manually, to classify the data whether it is benign or malignant and also to find out its accuracy. In order to evaluate, the performance of classifier, three main metrics have been computed. They are Sensitivity, Specificity and Accuracy.

Sensitivity or Recall rate =TP/ (TP+FN)

Precision Rate =TP/ (TP+FP)

Accuracy= (TP+TN)/ (TP+FP+FN+TN) where TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative. The table 4 shows the performance evaluation of the proposed classifier and the graphical representation is shown in figure 4

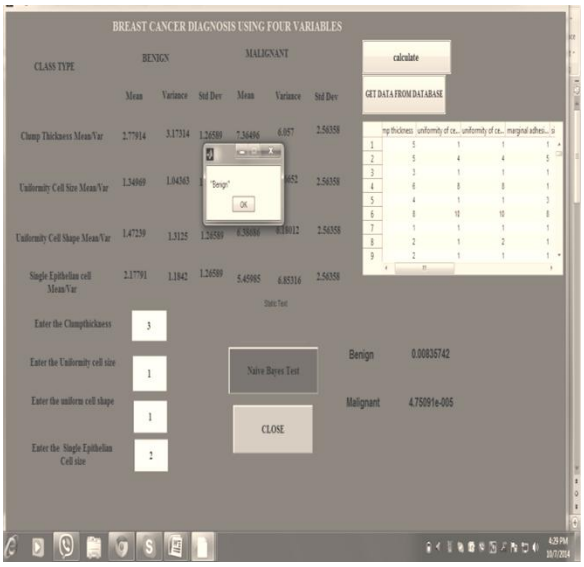

Figure 3: **GUI interface design for classification**

5. CONCLUSION

In this work the Gaussian Naïve bayes approach is used for predicting breast cancer, which analyses, with less attributes for large dataset. In this approach, the Wisconsin original dataset is used and GUI has been designed to enter the details of patient's, which predicts whether the given data is benign or malignant. This prediction gives the maximum accuracy of 96.6%. The purpose of this work is to demonstrate the ability of Naïve bayes classifier in predicting breast cancer with less attributes. In future the same methodology may be applied in other dataset, such as Wisconsin Prognostic Breast cancer dataset, Wisconsin Diagnosis Breast cancer dataset etc., by reducing the features, so that prediction time can be reduced and the treatment for the patients can be given as early as possible.

Table 4. Performance Evaluation of Gaussian Naïve Baye's Classifier

Algorithm	Dataset	Recall rate	Precision Rate	Accuracy
Naïve Bayes	300	94%	96%	96.6%

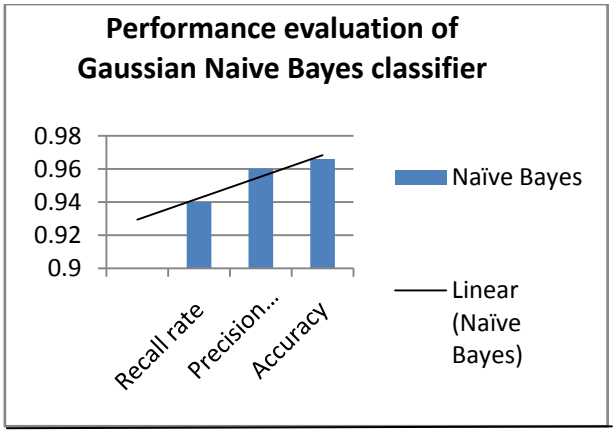

Figure 4: Performance evaluation of Proposed Classifier

6. ACKNOWLEDGMENTS

The breast cancer dataset used in this work is collected from UCI Machine learning repository [16].

7. REFERENCES

[1] Saravanakumar K.and Arthanariee A. M. Evaluate the multiple breast cancer factors and calculate the risk by software tool breast cancer risk evaluator, Indian Journal of Science and Technology, 3(Apr 2015),686-91.

[2] Vaidhehi K and Subhashini T.S, Breast Tissue Characterization using combined K-NN classifier, Indian Journal of Science and Technology,8 (Jan 2015),23-26.

[3] Aarthy S.L and Prabu.S, An Approach for Detecting Breast Cancer using Wavelet Transforms, Indian Journal of Science and Technology, 8 (Oct 2015),1-7

[4] Singh S, Dr.Gupta, P.R. and Sharma M.K, Breast Cancer Detection and Classification using Histopathological images, International Journal of Engineering Science and Technology, 3 (May 2011),4-9.

[5] Dumitru D. 2009 Prediction of recurrent events in breast cancer using the Naïve Bayesian classification, Annals University of Craiova, Mathematics and Computer Science Series.

[6] Nahar J, Chen Y P P and Ali S, Kernel Based Naïve Bayes Classifier for Breast Cancer prediction, Journal of Biological Systems,15 (Oct 2007),17-25.

[7] Kharya S, Agarwal S and Soni S, Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer. International Journal of Computer Applications ", 92 (Apr 2014), 26-31.

[8] Andrews J L and McNicholas P D, Variable selection for clustering and classification, Journal of Classification, 31 (Jul 2014), 136-153.

[9] You H and Rumbe G, Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data, International Journal of Interactive Multimedia and Artificial Intelligence, 1 (Dec 2010), 6-13.

[10] Güzel C., Mahmut Kaya M. and Yıldız O. 2013.Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with KNN Missing Data Imputation, 3rd World conference on innovation and Computer Sciences.

[11] Soria D, Garibaldi J M and Biganzoli E, 2008. A Comparison of Three Different Methods for Classification of Breast Cancer Data, Machine Learning and Applications, ICMLA '08. Seventh International Conference.

[12] Nezafat R, Tabesh A, Lucas C, Mohammed A and Zia M A. 1998. Feature Selection and Classification for Diagnosing Breast Cancer, "Proceedings of the IASTED International Conference of artificial intelligence and soft computing".

[13] Kitbumrungrat K, Comparison Logistic Regression and Discriminant Analysis in classification groups for Breast Cancer, International Journal of Computer Science and Network Security, 12 (May 2012), 111-15.

[14] Nanni L and Lumini A, Orthogonal linear discriminant analysis and feature selection for Micro-array data classification, Expert Systems with Applications, 37 (Oct 2010),7132-37.

[15] Nancy S G and Dr.Appavu alias Balamurugan S, A Comparative Study of Feature Selection Methods for cancer classification using Gene Expression Dataset, Journal of Computer Applications, 6 (Sep 2013), 78-84.

[16] Lichman,M.,'UCI Machine learning Repository', https://archive.ics.uci.edu/ml/datasets/BreastCancerWisconsin (Diagnostic),University of California, Irvine,CA,2013.

[17] Data preprocessing techniques for data mining-IASRI',http://www.slideshare.net/suganmca14/data-preprocessing-31610423, Feb 2014.

[18] 'Minitab-What is Mahalanobis distance' http://support.minitab.com/en-us/minitab/17/topic-Library/modeling-Statistics/multivariate/principal-components-and-factor-analysis/what-is-mahalanobis-Distance. Year-2015.

[19] 'Naive Bayes classification algorithm', http://software.ucv.rp/air/docs/naivebayes.pdf, Feb 2015.