

## Medical datamining with probabilistic classifiers

Ranjit Abraham<sup>1</sup>, Jay B.Simha<sup>2</sup>, Iyengar S.S<sup>3</sup>

<sup>1</sup>*Ejyothi Services Pvt. Ltd,  
Kurisupally Road, Cochin.*

[ranjit.abraham@ascellatech.com](mailto:ranjit.abraham@ascellatech.com)

<sup>2</sup>*Abiba Systems, Bangalore, INDIA.*

[jbsimha@gmail.com](mailto:jbsimha@gmail.com)

<sup>3</sup>*Department of Computer Science,  
Louisiana State University, Baton Rouge, USA*

[iyengar@bit.cse.lsu.edu](mailto:iyengar@bit.cse.lsu.edu)

**Abstract:** - *Statistical classifiers typically build (parametric) probabilistic models of the training data, and compute the probability that an unknown sample belongs to each of the possible classes using these models. We utilize two established measures to compare the performance of statistical classifiers namely; classification accuracy (or error rate) and the area under ROC. Naïve Bayes has obtained much relevance in data classification for machine learning and datamining. In our work, a comparative analysis of the accuracy performance of statistical classifiers namely Naïve Bayes (NB), MDL discretized NB, 4 different variants of NB and 8 popular non-NB classifiers was carried out on 21 medical datasets using classification accuracy and true positive rate. Our results indicate that the classification accuracy of Naïve Bayes (MDL discretized) on the average is the best performer. The significance of this work through the results of the comparative analysis, we are of the opinion that medical datamining with generative methods like Naïve Bayes is computationally simple yet effective and are to be used whenever possible as the benchmark for statistical classifiers.*

**Keywords:** Bayesian networks, Naïve Bayes classifier, discretization, Minimum Description Length (MDL)

### 1. Introduction

In the last few years, the digital revolution has provided relatively inexpensive and available means to collect and store large amounts of patient data in databases containing rich medical information and made available through the Internet for Health services globally. For a Physician who is guided by empirical observation and clinical trials, this data becomes appropriate if it is provided in terms of generalized knowledge such as information pertaining to patient history, diseases, medications, and clinical reports. Several computer programs have been developed to carry out optimal management of data for extraction of knowledge or patterns contained in the data. One such program approach has been data classification utilizing statistical algorithms with the goal of providing information such as if the patient is suffering from the illness or not from a case or collection of symptoms.

Naïve Bayes (NB) is a simple yet consistently performing probabilistic model based on the theory of Bayesian networks. Data classification with naïve Bayes is the task of predicting the class of an instance from a set of attributes describing that instance and assumes that all the attributes are conditionally independent given the class. This assumption grossly violates real-world problems and much effort has been focused in the name of naïve Bayes variants by relaxing the independence assumptions to improve classification accuracy. Research shows Naïve Bayes still performs well in spite of strong dependencies among attributes [19].

Research work show that naïve Bayes classification works best for discretized attributes and the application of Fayyad & Irani's Minimum discretization length (MDL) discretization gives on the average best classification accuracy performance [22].

In this paper we compare the accuracy performance of non-discretized NB with MDL discretized NB, popular variants of NB and with state-of-the-art classifiers such as k-nearest neighbor, Decision trees, Logistic regression, Neural networks, Support vector machines, RIPPER, RIDOR and Decision Tables.

## 2. Naïve Bayes (NB)

Naïve Bayes (NB), a special form of Bayesian network (BN) has been widely used for data classification in that its predictive performance is competitive with state-of-the-art classifiers such as C4.5 [26]. NB is best understood from the perspective of Bayesian networks. Bayesian networks graphically represent the joint probability distribution of a set of random variables. A BN is an annotated directed acyclic graph that encodes a joint probability distribution over a set of attributes  $X$ . Formally a BN for  $X$  is a pair  $B = \langle G, \theta \rangle$ , where  $G$  represents the directed acyclic graph whose nodes represent the attributes  $X_1, X_2, \dots, X_n$  and whose edges represent direct dependencies between the attributes. The BN can be used to compute the conditional probability of a node given values assigned to the other nodes. The Bayesian Network can be used as a classifier where the learner attempts to construct a classifier from a given set of training examples with class labels. Here nodes represent dataset attributes.

Assuming that  $X_1, X_2, \dots, X_n$  are the  $n$  attributes corresponding to the nodes of the BN and say an example  $E$  is represented by a vector  $x_1, x_2, \dots, x_n$  where  $x_1$  is the value of the attribute  $X_1$ . Let  $C$  represent the class variable and  $c$  its value corresponding to the class node in the BN, then the class  $c$  of the example  $E$  ( $c(E)$ ) can be represented as a classifier by the BN [11] as

$$c(E) = \arg \max_{c \in C} p(c) p(x_1, x_2, \dots, x_n | c) \tag{1}$$

Although BN can represent arbitrary dependencies it is intractable to learn it from data. Hence learning restricted structures such as Naïve Bayes is more practical. The NB classifier represented as a BN has the simplest structure. Here the assumption made is that all attributes are independent given the class and equation 1 takes the form

$$c(E) = \arg \max_{c \in C} p(c) \prod_{i=1}^n p(x_i | c) \tag{2}$$

The structure of NB is graphically shown in Figure 1.

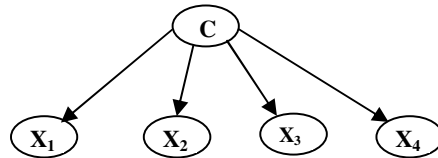


Figure 1. Structure of Naïve Bayes

Accordingly each attribute has a class node as its parent only. The most likely class of a test example can be easily estimated and surprisingly effective [6]. Comparing NB to BN, a much more powerful and flexible representation of probabilistic dependence generally did not lead to improvements in accuracy and in some cases reduced accuracy for some domains [19].

## 3. MDL Discretized Naïve Bayes

Discretization is the process of transforming data containing a quantitative attribute so that the attribute in question is replaced by a qualitative attribute [25]. Data attributes are either numeric or categorical. While categorical attributes are discrete, numerical attributes are either discrete or continuous. Research study shows that naïve Bayes classification works best for discretized attributes and discretization effectively approximates a continuous variable [2].

The Minimum Description Length (MDL) discretization is Entropy based heuristic given by Fayyad and Irani [9]. The technique evaluates a candidate cut point between each successive pair of sorted values. For each candidate cut point, the data are discretized into two intervals and the class information entropy is

calculated. The candidate cut point that provides the minimum entropy is chosen as the cut point. The technique is applied recursively to the two sub-intervals until the criteria of the Minimum Description Length (MDL). MDL discretized datasets show good classification accuracy performance with naïve Bayes [22].

#### 4. Variants of Naïve Bayes Classifier

The Tree Augmented Naïve Bayes (TAN) is an extended NB [10] where with a less restricted structure in which the class node directly points to all attribute nodes and an attribute node can have only one parent attribute node. TAN is a special case of Augmented Naïve Bayes (ANB), which is equivalent to learning an optimal BN, which is N-P hard. TAN has shown to maintain NB robustness and computational complexity and at the same time displaying better accuracy. The structure of TAN is shown in Figure 2.

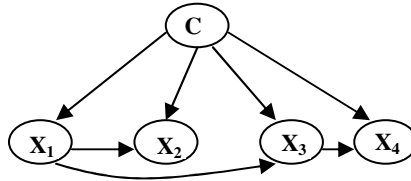


Figure 2. Structural representation of Tree Augmented Naïve Bayes (TAN)

Boosting involves learning a series of classifiers, where each classifier in the series learns more attention to the examples that have been misclassified by its predecessors. Hence each next classifier learns from the reweighed examples. The final boosted classifier outputs a weighted sum of the outputs of each individual classifier series with each weighted according to its accuracy on its training set. Boosting requires only linear time and constant space and hidden nodes are learned incrementally starting with the most important [8]. A graphical representation for Boosted Naïve Bayes (BAN) is shown in Figure 3. The hidden nodes  $\psi$  correspond to the outputs of the NB classifier after each iteration of boosting. With sample datasets BAN shows comparable accuracy with TAN.

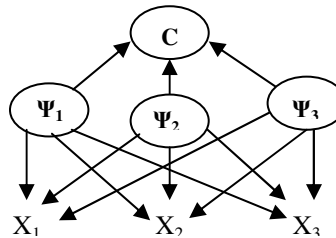


Figure 3. Structural representation for the Boosted Augmented Naïve Bayes (BAN)

The Forest augmented Naïve Bayes (FAN) represents an Augmented Bayes Network defined by a Class variable as parent to every attribute and an attribute can have at most one other attribute as its parent [11][24]. By applying the algorithm [12] incorporating Kruskal’s Maximum Spanning Tree algorithms an optimal Augmented Bayes Network can be found. A graphical structural representation for the Forest augmented NB is shown in Figure 4.

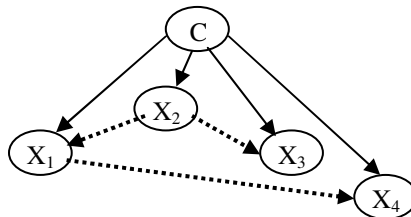


Figure 4. Structural representation for Forest augmented Naïve Bayes (FAN)

The Selective Naïve Bayesian classifier (SNB) uses only a subset of the given attributes in making the prediction [17]. The model enables to exclude redundant, irrelevant variables so that they do not reflect any differences for classification purposes. Experiments with sample datasets reveal that SNB appears to overcome the weakness of NB classifier. An example structural representation for SNB is shown in Figure 5.

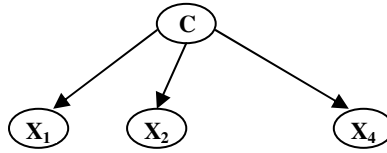


Figure 5. Structural representation for Selective Naïve Bayes (SNB)

For the above given model, and an example given by  $E = \langle x_1, x_2, x_3, x_4 \rangle$ , will be assigned to the class

$$c(E) = \arg \max_{c \in C} p(c) p(x_1 | c) P(x_2 | c) P(x_4 | c)$$

## 5. Popular non-NB statistical classifiers

The idea of a Decision Tree (DT) [21] is to partition the input space into small segments, and label these small segments with one of the various output categories. A DT is a k-ary tree where each of the internal nodes specifies a test on some attributes from the input feature set used to represent the data. Each branch descending from a node corresponds to one of the possible values of the feature specified at that node. Each test results in branches, which represent different outcomes of the test. The basic algorithm for DT induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The class probability of an example is estimated by the proportion of the examples of that class in the leaf into which the example falls.

k-NN is a supervised learning algorithm where the result of new instance query is classified based on majority of K-nearest neighbor category [6]. The purpose of this algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and only based on memory. Given a query point, we find K number of objects or (training points) closest to the query point. The classification is using majority vote among the classification of the K objects. Any ties can be broken at random. k-NN algorithm uses neighborhood classification as the prediction value of the new query instance. K-nearest neighborhood may be influenced by the density of the neighboring data points.

Logistic regression (LR) is part of a category of statistical models called generalized linear models. LR allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these [17]. LR is often referred to as a discriminative classifier unlike NB which is referred to as a generative classifier.

Artificial neural networks (NN) are relatively crude electronic networks of "neurons" based on the neural structure of the brain. They process records one at a time, and "learn" by comparing their classification of the record (which, at the outset, is largely arbitrary) with the known actual classification of the record. The errors from the initial classification of the first record is fed back into the network, and used to modify the networks algorithm the second time around, and so on for many iterations [14].

The Support Vector Machine (SVM) classification is based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. This approach constructs hyperplanes in a multidimensional space that separates cases of different class labels. SVM can handle multiple continuous and categorical variables [5].

The Repeated Incremental Pruning to Produce Error Reduction (RIPPER) is a decision-tree learning algorithm developed by William Cohen of AT&T Laboratories. This method offers modifications to IREP, C4.5, and C4.5 rules yielding faster training and lower error rates [3].

A Decision Table (DTab) is essentially a hierarchical table in which each entry in a higher level table gets broken down by the values of a pair of additional attributes to form another table. For an unlabelled instance, a decision table classifier searches for exact matches in the decision table. If no instances are found, then the majority class from the decision table is returned, otherwise the majority class matching all the instances is returned. [16].

The Ripple Down Rule learner (RIDOR) is an approach to building knowledge based systems (KBS) incrementally, while the KBS is in routine use. Here a default rule is generated and then the exceptions to the default rule (least weighted error rate). The best exception rules are generated iteratively to predict classes other than the default. [4]

## 6. Experimental Evaluation

Table 1 provides the specification for the 21 medical datasets used for the experimental evaluation. We have used 10-fold cross validation test method to all the Medical datasets [15]. The dataset was divided into 10 parts of which 9 parts were used as training sets and the remaining one part as the testing set. The classification accuracy was taken as the average of the 10 predictive accuracy values.

Table 1: Specifications for the Medical datasets

SL. No.	Medical Dataset	Total Instances	Total attributes	Number of Classes	Missing attr. status	Noisy attr. status
1	Wisconsin Breast Cancer [1]	699	10	2	Yes	No
2	Pima Diabetes [1]	768	9	2	No	No
3	Bupa Liver Disorders [1]	345	7	2	No	No
4	Cleveland Heart Disease [1]	303	14	2	Yes	No
5	Hepatitis [1]	155	20	2	Yes	No
6	Thyroid (ann-train) [1]	3772	22	3	No	No
7	Statlog- heart [1]	270	14	2	No	No
8	Hepatobiliary disorders [13]	536	10	4	No	No
9	Appendicitis [23]6	106	9	2	Yes	No
10	Stover Audiology [20]	1848	6 (5)	2	No	No
11	Leisenring neo audiology [20]	3152	8 (7)	2	No	No
12	Norton neonatal audiology [20]	5058	9 (7)	2	Yes	No
13	CARET PSA [20]	683	6 (5)	2	No	No
14	Ultrasound hepatic mets [20]	96	3	2	No	No
15	Pancreatic Ca biomarkers [20]	141	3	2	No	No
16	Laryngeal 1 [18]	213	17	2	No	No
17	Laryngeal 2 [18]	692	17	2	No	No
18	Laryngeal 3 [18]	353	17	3	No	No
19	RDS [18]	85	18	2	No	No
20	Voice_3 [18]	238	11	3	No	No
21	Voice_9 [18]	428	11	9 (2)	No	No

We have used two established measures of classifier performance in our experiments. The first one is the typical one namely Classification accuracy. The second measure is based on true positive rate which is used in Receiver Operator Characteristic Curve (ROC). Accordingly the area under the curve (AUROC) becomes a single-number performance measure for comparing classifiers.

Table 2 shows the accuracy results for non-discretized NB, MDL discretized NB and variants of NB. The 4 variants of naïve Bayes chosen for our experiments are Selective naïve Bayes (SNB), Boosted naïve Bayes (BNB), Tree augmented naïve Bayes (TAN) and Forest augmented naïve Bayes (FAN). Table 3 shows the accuracy performance with non-discretized NB, MDL discretized NB and some popular non-NB classifiers. The 8 popular non-NB statistical classifiers are Decision Tree (DT), k -Nearest Neighbor (k-NN), Logistic Regression (LR), Neural Network (NN), Support Vector Machine (SVM), Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Decision Table (DTab) and Ripple Down Rules Learner (RIDOR). The wins at the bottom of Table 2 and Table 3 provides the ratio of medical datasets where the accuracy is highest among the considered classifiers to the total number of datasets used for our experiments. Clearly the MDL discretized NB classifier on the average is the best performer compared to the other variants of NB as well as the 8 popular non-NB statistical classifiers that have been considered.

To further substantiate the results obtained in Table 2 and 3, we have tabled the results for the Area under the Receiver Operator Characteristics (AUROC) for the above mentioned statistical classifiers. Clearly the wins obtained by MDL discretized NB classifier proves that it is the best performer.

Table 2: Classification Accuracy with Naïve Bayes (NB), MDL discretized NB and variants of NB.

SI No.	Medical Dataset	NB	NB (MDL)	Variants of NB			
				SNB	BNB	TAN	FAN
1	Wisconsin Breast Cancer	95.9943	96.9957	96.7096	95.5651	96.7096	95.5651
2	Pima Diabetes	76.3021	77.8646	77.0833	74.349	74.6094	73.9583
3	Bupa Liver Disorders	55.3623	63.1884	61.4493	66.087	56.2319	68.9855
4	Cleveland Heart Disease	83.8284	83.8284	84.4884	83.4983	83.4983	83.4983
5	Hepatitis	84.5161	84.5161	87.0968	82.5806	83.2258	83.2258
6	Thyroid (ann-train)	95.5196	98.807	95.6257	93.0806	99.3107	99.3637
7	Statlog- heart	84.8148	83.3333	84.8148	80.7407	80.7407	80.3704
8	Hepatobiliary disorders	47.9478	68.4701	49.0672	45.1493	65.4851	79.1045
9	Appendicitis	84.9057	88.6792	88.6792	84.9057	87.7358	86.7925
10	Stover Audiology	96.5368	100	100	100	100	100
11	Leisenring neo audiology	100	100	100	100	100	100
12	Norton neonatal audiology	96.0854	96.6983	96.0854	97.0542	96.9751	96.9751
13	CARET PSA	74.3777	85.3587	76.1347	84.1874	84.041	85.0659
14	Ultrasound hepatic mets	82.2917	84.375	84.375	81.25	81.25	80.2083
15	Pancreatic Ca biomarkers	73.7589	78.7234	73.7589	70.922	70.922	72.3404
16	Laryngeal 1	75.5869	86.8545	85.446	82.1596	84.0376	84.0376
17	Laryngeal 2	84.6821	94.2197	95.9538	94.5087	95.9538	96.5318
18	Laryngeal 3	64.3059	70.8215	73.3711	68.5552	71.3881	71.6714
19	RDS	89.4118	95.2941	91.7647	89.4118	90.5882	92.9412
20	Voice_3	69.7479	76.8908	80.2521	71.8487	74.3697	73.1092
21	Voice_9	78.7383	84.5794	89.2523	86.6822	88.0841	94.8598
Wins		2/21	10/21	9/21	3/21	2/21	7/21

**Abbreviations Used:** **NB**- Naïve Bayes, **NB (MDL)** – Naïve Bayes with MDL discretization, **SNB** – Selective Naïve Bayes, **BNB**- Boosted Naïve Bayes, **TAN**- Tree Augmented Naïve Bayes, **FAN** – Forest Augmented Naïve Bayes

Table 3: Classification Accuracy with Naïve Bayes (NB), MDL discretized NB and non-NB classifiers

SI No.	Medical Dataset	NB	NB (MDL)	Popular non-NB Classifiers							
				DT	k-NN	LR	NN	SVM	RIPPER	DTAB	RIDOR
1	Wisconsin Breast Cancer	95.9943	96.9957	94.5637	94.9928	96.5665	95.279	96.9957	95.7082	95.422	95.8512
2	Pima Diabetes	76.3021	77.8646	73.8281	70.1823	77.2135	75.1302	77.3438	75.1302	72.2656	75
3	Bupa Liver Disorders	55.3623	63.1884	68.6957	62.8986	68.1159	71.5942	58.2609	64.6377	57.6812	63.1884
4	Cleveland Heart Disease	83.8284	83.8284	75.9076	75.9076	84.8185	80.8581	85.1485	74.9175	76.2376	72.2772
5	Hepatitis	84.5161	84.5161	83.871	80.6452	82.5806	81.9355	85.1613	80.6452	81.2903	77.4194
6	Thyroid (ann-train)	95.5196	98.807	99.7084	92.1262	96.8717	96.2354	93.7964	99.6819	99.6819	99.5493
7	Statlog- heart	84.8148	83.3333	76.2963	75.5556	83.3333	83.3333	82.963	77.4074	80.7407	77.4074
8	Hepatobiliary disorders	47.9478	68.4701	71.0821	73.3209	59.3284	60.8209	42.3507	67.7239	63.4328	67.7239
9	Appendicitis	84.9057	88.6792	86.7925	83.0189	87.7358	87.7358	86.7925	82.0755	85.8491	86.7925
10	Stover Audiology	96.5368	100	100	99.513	100	100	99.6212	100	100	100
11	Leisenring neo audiology	100	100	100	100	100	100	100	100	100	100
12	Norton neonatal audiology	96.0854	96.6983	97.0739	94.6619	97.0542	96.9553	97.0542	96.9751	97.0344	97.0146
13	CARET PSA	74.3777	85.3587	83.4553	77.306	79.063	84.6266	72.4744	84.9195	83.6018	83.4553
14	Ultrasound hepatic mets	82.2917	84.375	81.25	79.1667	82.2917	80.2083	84.375	81.25	81.25	78.125
15	Pancreatic Ca biomarkers	73.7589	78.7234	73.0496	72.3404	80.1418	65.9574	63.8298	80.1418	68.7943	78.7234
16	Laryngeal 1	75.5869	86.8545	78.4038	79.8122	84.507	82.1596	84.0376	81.2207	80.7512	80.2817
17	Laryngeal 2	84.6821	94.2197	94.9422	95.0867	97.2543	96.2428	96.0983	95.0867	95.5202	95.3757
18	Laryngeal 3	64.3059	70.8215	65.7224	66.5722	75.0708	68.8385	72.8045	71.6714	69.4051	70.5382
19	RDS	89.4118	95.2941	84.7095	82.3529	87.0588	87.0588	88.2353	90.5882	87.0588	85.8824
20	Voice_3	69.7479	76.8908	74.7899	71.8487	78.1513	76.4706	78.5714	76.8908	75.2101	75.6303
21	Voice_9	78.7383	84.5794	91.1215	94.8598	87.1495	89.486	85.2804	87.8505	91.8224	88.7850
Wins		2/21	9/21	4/21	3/21	4/21	3/21	6/21	3/21	2/21	2/21

**Abbreviations Used:** **NB**- Naïve Bayes, **NB (MDL)** – Naïve Bayes with MDL discretization, **DT** – Decision Tree, **k-NN**- k -Nearest Neighbor, **LR**- Logistic Regression, **NN**-Neural Network, **SVM** – Support Vector Machine, **RIPPER**- Repeated Incremental Pruning to Produce Error Reduction, **DTAB**- Decision Table, **RIDOR**- Ripple Down Rules Learner

Table 4: AUROC (in percentage) with Naïve Bayes (NB), MDL discretized NB and variants of NB.

SI No.	Medical Dataset	NB	NB (MDL)	Variants of NB			
				SNB	BNB	TAN	FAN
1	Wisconsin Breast Cancer	98.75	99.2	99.11	97.62	98.94	98.68
2	Pima Diabetes	81.86	84.64	82.79	80.08	80.75	78.42
3	Bupa Liver Disorders	64.01	55.95	61.78	68.41	51.36	73.67
4	Cleveland Heart Disease	90.71	91.27	88.46	89.44	90.92	90.92
5	Hepatitis	85.95	86.94	85.82	85.15	87.70	86.43
6	Thyroid (ann-train)	99.73	99.94	99.72	99.88	99.95	99.97
7	Statlog- heart	90.83	91	87.99	87.86	90.20	87.11
8	Hepatobiliary disorders	74.35	87.48	71.34	54.85	85.53	91.14
9	Appendicitis	79.33	79.94	78.38	80.81	78.85	85.60
10	Stover Audiology	99.65	100	100	100	100	100
11	Leisenring neo audiology	100	100	100	100	100	100
12	Norton neonatal audiology	60.91	58.44	60.91	61.04	57.83	56.01
13	CARET PSA	84.97	91.47	85.98	90.02	90.60	91.20
14	Ultrasound hepatic mets	81.51	77.66	80.04	76.48	77.22	75.66
15	Pancreatic Ca biomarkers	83.15	84.11	83.15	82.07	79.02	78.31
16	Laryngeal 1	90.24	93.62	91.02	89.82	89.97	91.19
17	Laryngeal 2	96.95	98.82	97.70	96.06	97.84	95.60
18	Laryngeal 3	96.67	98.6	97.06	85.69	96.64	96.47
19	RDS	95.39	98.78	90.78	96.11	97.33	96.22
20	Voice_3	89.9	95.01	90.67	82.95	92.17	88.87
21	Voice_9	90.91	95.36	91.75	92.28	94.50	98.29
Wins		2/21	14/21	2/21	3/21	3/21	6/21

**Abbreviations Used:** **NB**- Naïve Bayes, **NB (MDL)** – Naïve Bayes with MDL discretization, **SNB** – Selective Naïve Bayes, **BNB**- Boosted Naïve Bayes, **TAN**- Tree Augmented Naïve Bayes, **FAN** – Forest Augmented Naïve Bayes

Table 5: AUROC (in percentage) with Naïve Bayes (NB), MDL discretized NB and non-NB classifiers

SI No.	Medical Dataset	NB	NB (MDL)	Popular non-NB Classifiers							
				DT	k-NN	LR	NN	SVM	RIPPER	DTab	RIDOR
1	Wisconsin Breast Cancer	98.75	99.2	95.47	97.31	99.33	98.61	96.82	96.49	95.41	95.65
2	Pima Diabetes	81.86	84.64	75.14	65.01	83.18	79.09	71.95	73.43	76.27	70.15
3	Bupa Liver Disorders	64.01	55.95	66.5	62.96	71.76	74.16	50.34	65.28	54.18	61.23
4	Cleveland Heart Disease	90.71	91.27	78.28	78.28	90.86	88.30	84.75	75.54	80.36	72.09
5	Hepatitis	85.95	86.94	70.82	65.35	80.26	81.63	75.62	62.54	70.83	56.87
6	Thyroid (ann-train)	99.73	99.94	99.97	82.50	97.56	99.24	84.99	99.24	97.85	98.32
7	Statlog- heart	90.83	91	76.73	75.42	90.44	88.55	82.33	79.09	81.16	76.75
8	Hepatobiliary disorders	74.35	87.48	78.95	80.07	77.37	80.69	75.90	78.31	78.53	73.05
9	Appendicitis	79.33	79.94	63.81	78.54	79.10	77.20	72.04	63.84	74.12	72.04
10	Stover Audiology	99.65	100	100	99.46	100	100	99.48	100	100	100
11	Leisenring neo audiology	100	100	100	100	100	100	100	100	100	100
12	Norton neonatal audiology	60.91	58.44	51.68	51.41	62.50	59.12	50.00	50.99	51.16	49.98
13	CARET PSA	84.97	91.47	88.93	75.19	89.20	91.36	59.38	83.54	90.87	78.68
14	Ultrasound hepatic mets	81.51	77.66	72.75	76.58	80.21	79.53	80.16	73.91	72.75	74.68
15	Pancreatic Ca biomarkers	83.15	84.11	77.03	70.17	87.60	73.51	50.00	79.01	76.67	78.66
16	Laryngeal 1	90.24	93.62	75.23	77.84	90.83	87.07	83.54	83.66	81.43	79.56
17	Laryngeal 2	96.95	98.82	86.45	84.54	97.89	98.21	81.45	80.68	84.96	80.19
18	Laryngeal 3	96.67	98.6	91.53	87.48	96.75	94.50	93.94	89.43	83.82	88.68
19	RDS	95.39	98.78	89.28	79.44	92.31	94.56	88.06	95.44	91.28	85.69
20	Voice_3	89.9	95.01	68.46	77.87	90.42	90.78	84.39	75.33	84.20	77.16
21	Voice_9	90.91	95.36	89.75	92.63	91.74	93.11	77.54	84.08	90.61	83.30
Wins		2/21	15/21	3/21	1/21	5/21	3/21	1/21	2/21	2/21	2/21

**Abbreviations Used:** **NB**- Naïve Bayes, **NB (MDL)** – Naïve Bayes with MDL discretization, **DT** – Decision Tree, **k-NN**- k -Nearest Neighbor, **LR**- Logistic Regression, **NN**-Neural Network, **SVM** – Support Vector Machine, **RIPPER**- Repeated Incremental Pruning to Produce Error Reduction, **DTab**- Decision Table, **RIDOR**- Ripple Down Rules Learner

In Fig. 6, we show 4 medical datasets used in our experiments where MDL discretized Naïve Bayes classification accuracy provides the best results compared to all other statistical classifiers. Fig 7 shows the same 4 medical datasets that gave the best results for the AUROC compared to all the other statistical classifiers.

Fig. 6: Classification accuracy of Statistical classifiers

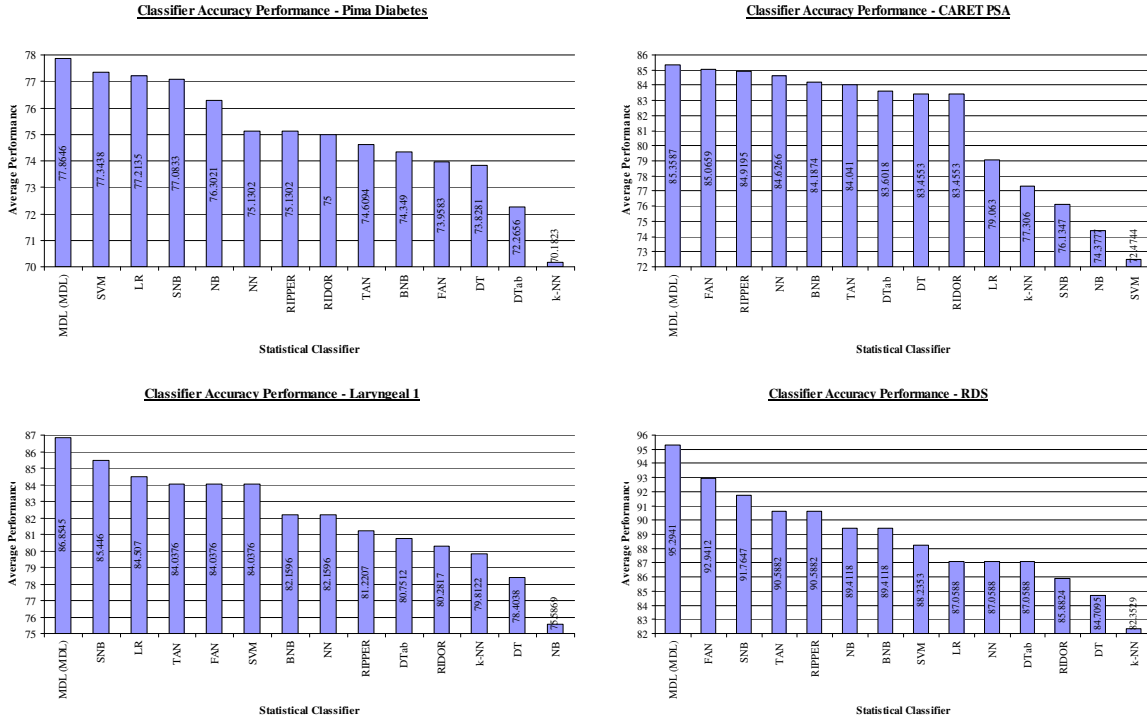
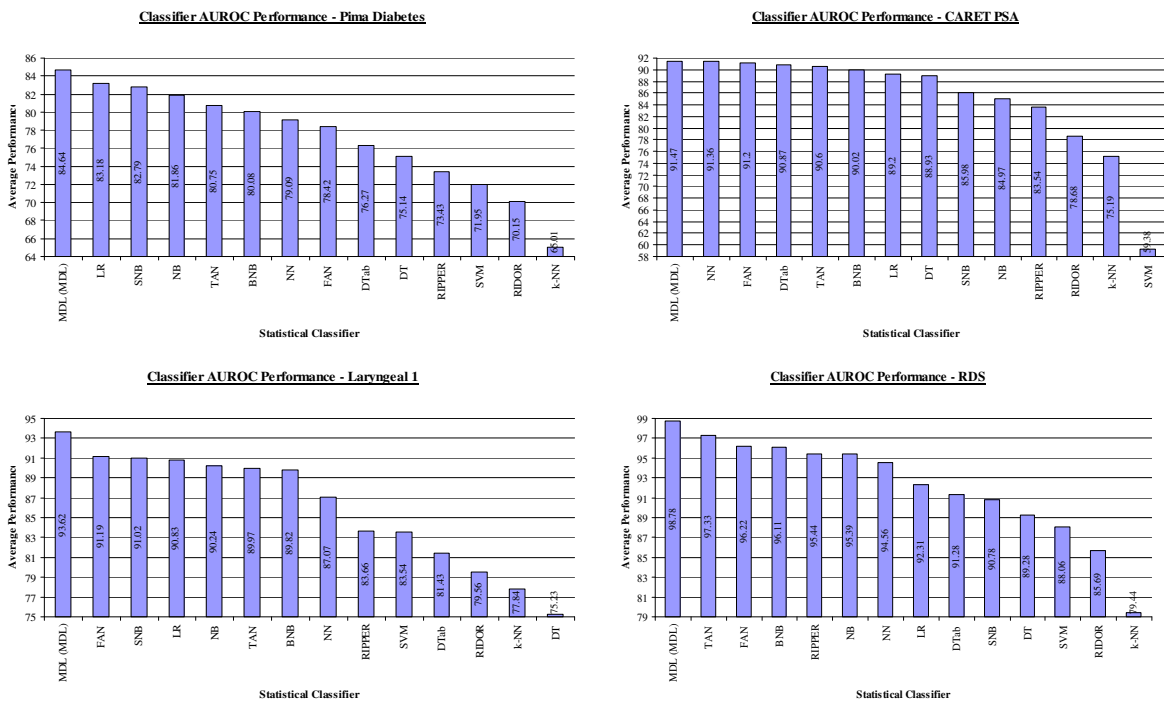


Fig. 7: Classification AUROC (in percentage) of Statistical classifiers



## 7. Conclusions

In this research work an attempt was made to evaluate various probabilistic classifiers that could be used for medical datamining. We show using two established measures for comparing performance of statistical classifiers, on an average, naïve Minimum Description Length (MDL) discretization seems to be the best performer compared to the considered various naïve Bayes and non-naïve Bayes classifiers. Hence forth we are of the opinion that generative methods like naïve Bayes discretized with MDL is simple yet effective and are to be used whenever possible to set the benchmark for other statistical classifiers. The work is presently under progress to explore feature selection methods in achieving better naïve Bayes classification performance.

## References

- [1] Blake C.L, Merz C.J., “*UCI repository of machine learning databases*”. [<http://www.ics.uci.edu/~mlern/MLRepository.html>], Department of Information and Computer Science, University of California, Irvine.
- [2] Chun-Nan Hsu, Hung-Ju Huang, Tsu-Tsung Wong, “*Why Discretization works for Naïve Bayesian Classifiers*”, 17th ICML, pp 309-406, 2000.
- [3] Cohen W, “*Fast Effective Rule Induction*”, In Machine Learning: Proceedings of the Twelfth International Conference, Lake Tahoe, California, 1995.
- [4] Compton P., Peters L., Edwards G., Lavers T.G., “*Experience with Ripple-Down Rules*”, Knowledge-Based Systems, Volume 19, Issue 5, pp. 356-362, 2006
- [5] Cortes C., Vapnik V., “*Support Vector Networks*”, Machine Learning, 20(3), pp. 273-297, 1995
- [6] David W. Aha, Dennis Kibler, Mark C Albert, “*Instance-Based learning algorithms*”, Machine Learning, 6, pp. 37-66, 1991.
- [7] Domingos P., Pazzani M., “*Beyond independence: Conditions for the optimality of the simple Bayesian classifier*”, Machine Learning, 29:103—130. 1997.
- [8] Elkan C., “*Boosting and Naïve Bayesian learning*”, (Technical Report) University of California, San Diego, 1997.
- [9] Fayyad U. M., Irani K. B., “*Multi-interval discretization of continuous-valued attributes for classification learning*”, In Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022–1027, 1993.
- [10] Friedman N., Geiger D., Goldszmidt M., “*Bayesian network classifiers*”, Machine Learning, vol. 29, pages 131-163, 1997.
- [11] Keogh E.J., Pazzani M.J., “*Learning Augmented Bayesian Classifiers: A Comparison of Distribution-based and Classification-based Approaches*”, Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics: 225-230, 1999.
- [12] Hamine V., Helman P., “*Learning Optimal Augmented Bayes Networks*”, Tech Report TR-CS-2004-11, Computer Science Department, University of New Mexico, 2004.
- [13] Hayashi Y., “*Neural expert system using fuzzy teaching input and its application to medical diagnosis*”, *Information Sciences Applications*, Vol. 1, pp. 47-58, 1994.
- [14] Herve Abdi, “*A Neural Network Primer*”, Journal of Biological Systems, Vol 2(3), pp. 247-283, 1994.
- [15] Kohavi R., “*A study of cross-validation and bootstrap for accuracy estimation and model selection*”. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 1137–1145, 1995.
- [16] Kohavi R., “*The Power of Decision Tables*”, Machine Learning: ECML-95: 8th European Conference on Machine Learning, Heraclion, Crete, Greece, 1995.
- [17] le Cessie S., van Houwelingen J., “*Ridge estimators in logistic regression*”, Applied Statistics, Vol 41, no 1, pp. 191-201, 1992.
- [18] Ludmila I. Kuncheva, - School of Informatics, University of Wales, Bangor, Dean Street, Bangor Gwynedd LL57 1UT, UK. [http://www.informatics.bangor.ac.uk/~kuncheva/activities/real\\_data\\_full\\_set.htm](http://www.informatics.bangor.ac.uk/~kuncheva/activities/real_data_full_set.htm)
- [19] Pearl J, “*Probabilistic Reasoning in Intelligent Systems*”, Morgan Kaufmann Publishers, 1988.
- [20] Pepe M.S., “*The Statistical Evaluation of Medical Tests for Classification and Prediction*”, <http://www.fhrc.org/science/labs/pepe/book/>, Oxford Statistical Science Series, Oxford University Press. 2003.
- [21] Quinlan, J. R., “*C4.5, Programs for Machine Learning*”, Morgan Kaufmann, San Mateo, Ca, 1993.
- [22] Ranjit Abraham, Jay B.Simha, Iyengar S.S., “*A comparative analysis of discretization methods for Medical datamining with Naïve Bayesian classifiers*”, 9<sup>th</sup> International Conference for Information Technology (ICIT2006), Bhubaneshwar, India, Dec 18-21, 2006.
- [23] Shalom M. Weiss, (for the Medical dataset on Appendicitis).
- [24] Saha, ‘[http://jbnf.sourceforge.net/JP\\_Sacha\\_PhD\\_dissertation.pdf](http://jbnf.sourceforge.net/JP_Sacha_PhD_dissertation.pdf)’.
- [25] Ying Yang, Geoffrey I Webb, “*A Comparative Study of Discretization Methods for Naïve Bayes Classifiers*”, In Proceedings of PKAW, Japan pp 159-173, 2002.
- [26] Zhang H., Jiang L., Su J. , “*Hidden Naïve Bayes*” , Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05) , AAAI Press, 2005.