On the use of m-probability-estimation and imprecise probabilities in the naïve Bayes classifier

Javier G. Castellano, Serafín Moral-García, Carlos J. Mantas and Joaquín Abellán
Department of Computer Science and Artificial Intelligence
University of Granada, Granada, Spain
{*fjgc,seramoral,cmantas,jabellan*}*@decsai.ugr.es*

Received (received date)
Revised (revised date)

Within the field of supervised classification, the naïve Bayes (NB) classifier is a very simple and fast classification method that obtains good results, being even comparable with much more complex models. It has been proved that the NB model is strongly dependent on the estimation of conditional probabilities. In the literature, it had been shown that the classical and Laplace estimations of probabilities have some drawbacks and it was proposed a NB model that takes into account the a priori probabilities in order to estimate the conditional probabilities, which was called m-probability-estimation. With a very scarce experimentation, this approximation based on m-probability-estimation demonstrated to provide better results than NB with classical and Laplace estimations of probabilities. In this research, a new naïve Bayes variation is proposed, which is based on the m-probability-estimation version and takes into account imprecise probabilities in order to calculate the a priori probabilities. An exhaustive experimental research is carried out, with a large number of data sets and different levels of class noise. From this experimentation, we can conclude that the proposed NB model and the m-probability-estimation approach provide better results than NB with classical and Laplace estimation of probabilities. It will be also shown that the proposed NB implies an improvement over the m-probability-estimation model, especially when there is some class noise.

Keywords: supervised learning; naïve Bayes; m-estimate; m-probability-estimation; imprecise probabilities; noisy data.

1. Introduction

Supervised learning or classification[1] has been considered as a crucial task in data mining. This machine learning task starts with a data set of observations, each one with an assigned value of a class variable, which is the variable under study. Each observation is described via a set of attributes. The goal of classification is to extract useful knowledge from the data in order to predict the value of the class label when a new case or instance appears.

In the literature, a large number of techniques has been applied to carry out this labor. Among these approaches stand out classical statistical methods[2], Decision Trees (DT)[3], Bayesian Networks (BN)[4] or Artificial Neural Networks (ANN)[5]. For each one of these techniques, a lot of algorithms have been developed in order to

improve the results of previous models. For instance, within DTs, ensemble models such as bagging[6], boosting[7], or Random Forest[8] emerged, obtaining better results than classical DT algorithms, like ID3[9] or C4.5[3].

Among Bayesian networks models, the naïve Bayes (NB) algorithm[10] is quite popular. NB assumes that all attributes are conditionally independent given the value of the class variable and all of the variables have the same influence on the class. It is obvious that this independence condition is unsatisfied in many cases. Despite this unrealistic assumption, NB shows remarkable results in terms of accuracy and it has been applied successfully in practice, often being comparable with other far more complex models, especially when the attributes are not strongly correlated[11,12,13]. In fact, NB has been effectively applied in a lot of practical applications, such as systems performance management[14], text classification[15] or gene expression analysis[16]. Moreover, the NB model, as a consequence of its independence assumption, is much faster than other more sophisticated models and the required computational cost is significantly lower. Therefore, we can say that the key to the success of the NB is its simplicity: no Bayesian network structure learning algorithm is required because its structure is fixed, the parameters of the model need only to be estimated from the data set using only bi-dimensional statistics for the class and each attribute and the classification process is very efficient[17].

The NB classifier, like Bayesian classifiers, is based on Bayes formula. In this particular case this formula is used naïvely, i.e assuming the independence condition. In (Cestnik[18], 1990) it is shown that the evaluation of the naïve Bayesian formula is pretty influenced by the estimation of conditional probabilities. Concretely, in that work it is shown that the classical probability estimation through relative frequencies has important issues. Furthermore, it is also illustrated that the Laplace estimation[19], which is used to solve those issues, has some drawbacks too. For this reason, precisely in (Cestnik[18], 1990), a new probability estimation is introduced, that is called m-probability-estimation in contrast with Laplace-probability-estimation and it is also known as m-estimate[a]. This estimation consists of taking into account the a priori probabilities of the class variable when we estimate the conditional probability of the class variable given the value of an attribute. The m-probability-estimation technique has a parameter m as input which is related to the amount of noise in the data. It was shown empirically[18] that m-probability-estimation provides better results than the previous approaches of conditional probability estimation, although the experimentation carried out was very scarce, using only four databases without added noise. In addition, it is experimentally shown[20] that this probability estimation in tree pruning improves the results of classical standard methods.

In spite of the improvement of the performance of the NB model with this way of

[a]The reader should be aware that, from a statistical perspective, the term m-estimate (coined by Cestnik[20]) is an unfortunate choice, since M-estimation is already a well-known term in Statistics and it refers to an entirely different approach.

estimating probabilities, the algorithm is still quite sensitive to noise. This happens because the estimation of the a priori probability is still done by the classical method of computing relative frequencies, which is clearly deteriorated with the presence of noise. For this reason, in this paper, we also use the Imprecise Dirichlet Model (IDM)[21] in order to estimate the a priori probability of the class variable. The use of this imprecise model has been tested to be useful in order to improve the performance of standard models when there is noise in the data. An example can be found in the Credal-C4.5 algorithm[22] which is based on the IDM and provides better results than the classical C4.5, especially with the presence of class noise[23].

In this work, we propose a new naïve Bayes algorithm, the Imprecise m-probability-estimation naïve Bayes (ImNB), which combines the m-probability-estimation with the IDM in order to obtain a classifier which is less sensitive to class noise. An extensive experimentation is carried out where our new NB approach is compared with the NB using the m-probability-estimation and the Laplace and classical probability estimations. The mentioned algorithms are applied to a set of data sets without noise and with different levels of added class noise. Another contribution of this research is that the experimentation shows that Cestnik proposal obtains much better results than NB with Laplace and classical estimations of probabilities, with a much more exhaustive experimentation than in (Cestnik[18], 1990). Besides, the experimental study shows that the new method performs better than the Cestnik model, where the differences are significant when there is class noise in the data.

The rest of this paper is structured as follows: Section 2 contains all the theoretical background that is necessary to know in order to understand the proposed model. In section 3 our new naïve Bayes approach is explained. In section 4 our experimentation is described and the obtained results are commented. Finally, section 5 is devoted to the concluding remarks.

2. Previous knowledge

2.1. *Preliminaries: Conditional Probability, Independence and Bayes Theorem*

Let X_1 and X_2 be two random variables. The conditional probability of X_1 given X_2 is defined as follows:

$$P(X_1|X_2) = \frac{P(X_1, X_2)}{P(X_2)} \tag{1}$$

It is said that X_1 and X_2 are independent if it is verified that

$$P(X_1, X_2) = P(X_1)P(X_2) \tag{2}$$

As can be observed from (1) and (2), the variables X_1 and X_2 are independent if and only if $P(X_1|X_2) = P(X_1)$ and $P(X_2|X_1) = P(X_2)$.

Now let X_3 be another variable. It is said that X_1 and X_2 are conditionally independent given X_3 if:

$$P(X_1, X_2|X_3) = P(X_1|X_3)P(X_2|X_3) \tag{3}$$

The previous definitions of independence and conditional independence can be extended to a general set of variables. Let $X_1, \ldots, X_n$ be a set of variables, $n \in \mathbb{N}$. We say that $X_1, \ldots, X_{n-1}$ are conditionally independent given X_n if:

$$P(X_1, \ldots, X_{n-1}|X_n) = \prod_{i=1}^{n-1} P(X_i|X_n) \tag{4}$$

Once the preliminary definitions have been exposed the most important theorem upon which Bayesian classifiers are based, the Bayes' theorem can be shown.

Theorem 1.

Let $\mathbf{X}$ *and* $\mathbf{Y}$ *be two set of variables. It is verified that*

$$P(\mathbf{X}|\mathbf{Y}) = \frac{P(\mathbf{Y}|\mathbf{X})P(\mathbf{X})}{P(\mathbf{Y})} \tag{5}$$

2.2. *The naïve Bayes model*

We recall that the aim of supervised learning is to predict the value of the class variable C from an instance given a set of attribute variables $X_1, \ldots, X_n$, and their respective correspondent values $x^1_{t_1}, \ldots, x^n_{t_n}$, $n \in \mathbb{N}$ of that instance and being T_i the number of possible values of X_i. Suppose that the possible values of C are $\{c_1, \ldots, c_k\}$.

The NB classifier is a Bayesian classifier; these kind of classifiers predict the class label by maximizing $P(C|X_1, \ldots, X_n)$ over C, i.e, they assign the value c of C that verifies

$$c = \arg \max_{c_j, 1 \leq j \leq k} P(C = c_j | X_1 = x^1_{t_1}, \ldots, X_n = x^n_{t_n}) \tag{6}$$

According to Theorem 1:

$$P(C = c_j | X_1 = x^1_{t_1}, \ldots, X_n = x^n_{t_n}) = \frac{P(C = c_j)P(X_1 = x^1_{t_1}, \ldots, X_n = x^n_{t_n}|C = c_j)}{P(X_1 = x^1_{t_1}, \ldots, X_n = x^n_{t_n})}, \quad \forall j \tag{7}$$

Thus, it is easy to observe that maximizing $P(C = c_j | X_1 = x^1_{t_1}, \ldots, X_n = x^n_{t_n})$ is equivalent to maximize $P(C = c_j)P(X_1 = x^1_{t_1}, \ldots, X_n = x^n_{t_n}|C = c_j)$, $\forall j$.

The NB model assumes that the attributes are conditionally independent given the class value. Hence,

$$P(X_1 = x^1_{t_1}, \ldots, X_n = x^n_{t_n}|C = c_j) = \prod_i P(X_i = x^i_{t_i}|C = c_j), \forall j \tag{8}$$

In consequence, the NB algorithm assigns the class label c that verifies:

$$c = \arg \max_{c_j, j=1,\ldots,k} P(C=c_j) \prod_i P(X_i=x^i_{t_i}|C=c_j) \tag{9}$$

Furthermore, according to Theorem 1:

$$P(X_i=x^i_{t_i}|C=c_j) = \frac{P(C=c_j|X_i=x^i_{t_i})P(X_i=x^i_{t_i})}{P(C=c_j)}, \forall i \tag{10}$$

Therefore, the NB classifier assigns the value c of C that verifies

$$c = \arg \max_{c_j, 1\leq j\leq k} P(C=c_j) \prod_i h_{jt_i} \tag{11}$$

where

$$h_{jt_i} = \frac{P(C=c_j|X_i=x^i_{t_i})}{P(C=c_j)}, \forall i \tag{12}$$

2.3. *Problems of classical and Laplace estimations of Probabilities*

Let N_{t_i} be the number of cases of our data set in which $X_i=x^i_{t_i}$, let N be the total number of cases in the data set, N_{c_j} the instances in the data set which verifies that $C=c_j$ and let N_{t_i,c_j} be the number of cases in which $C=c_j$ and $X_i=x^i_{t_i}$. The classical estimations of $\hat{P}(C=c_j|X_i=x^i_{t_i})$ and $\hat{P}(C=c_j)$ are given by:

$$\hat{P}(C=c_j) = \frac{N_{c_j}}{N} \tag{13}$$

$$\hat{P}(C=c_j|X_i=x^i_{t_i}) = \frac{N_{t_i,c_j}}{N_{t_i}} \tag{14}$$

The main problem arises when $N_{t_i,c_j}=0$. In this case $\hat{P}(C=c_j|X_i=x^i_{t_i})=0$ and, therefore, $h_{jt_i}=0$ and $\hat{P}(C=c_j|X_1=x^1_{t_1},\ldots,X_n=x^n_{t_n})=0$. Consequently, only the value of one attribute can affect drastically the value of the probability of the class label. In fact, this probability becomes zero when $\hat{P}(C=c_j|X_1=x^1_{t_1},\ldots,X_{i-1}=x^{i-1}_{t_{i-1}},X_{i+1}=x^{i+1}_{t_{i+1}},\ldots,X_n=x^n_{t_n})$ may be pretty high. In addition, it may happen that if N_{t_i} is very small, the estimation of $\hat{P}(C=c_j|X_i=x^i_{t_i})$ could be an unstable estimation.

The Laplace's law of succession[19] was introduced to estimate probabilities when there are few observations, or no observations at all (zero-probability events). With that in mind, the Laplace's law of succession was used for the estimation of the probabilities in the NB with the aim of solving this problem. It assumes a uniform distribution for all classes. Hence, it states that the probability that a new instance with the value x_{t_i} for the attribute X_i has the value c_j for C is given by:

$$\hat{P}(C=c_j|X_i=x_{t_i}^i)=\frac{N_{t_i,c_j}+1}{N_{t_i}+k} \tag{15}$$

and the a priori probability $\hat{P}(C=c_j)$:

$$\hat{P}(C=c_j)=\frac{N_{c_j}+1}{N+k} \tag{16}$$

then

$$h_{jt_i}=\frac{N_{t_i,c_j}+1}{(N_{t_i}+k)\hat{P}(C=c_j)} \tag{17}$$

where $\hat{P}(C=c_j)$ is determined by (16). This way of estimating probabilities using the Laplace's law of succession is called Laplace smoothing or additive smoothing.

Let us analyze the cases in which $N_{t_i}=0$ or $N_{t_i,c_j}=0$:

- When $N_{t_i}=0$, we obtain $h_{jt_i}=\frac{1}{k\hat{P}(C=c_j)}$; therefore h_{jt_i} increases when $\hat{P}(C=c_j)$ decreases and viceversa.
- When $N_{t_i,c_j}=0$, we obtain $h_{jt_i}=\frac{1}{(k+N_{t_i})\hat{P}(C=c_j)}$; consequently, we have that h_{jt_ii} is inversely proportional to $\hat{P}(C=c_j)$. This strange behavior is due to the assumption of uniformity for the distribution of the class values.

2.4. *The m-probability-estimation model*

In order to correct the questionable behavior of the Laplace's law of succession in some cases, a more appropriate and flexible class of initial probabilities is proposed[24]. According to this, after r success in N trials, the probability of getting a success in a next trial is of the form:

$$\hat{P}_s(r,N)=\frac{r+a}{N+a+b} \tag{18}$$

where $a>0$ and $b>0$[b]. In (Cestnik[18], 1990) the choice of parameters a and b was as follows: $a=m\hat{P}(C=c_j)$ and $b=m-a$, where m is a parameter of this model, and is called m-probability-estimation. In this way, the conditional probability is estimated by:

$$\hat{P}(C=c_j|X=x_{t_i}^i)=\frac{N_{t_i,c_j}+m\hat{P}(C=c_j)}{N_{t_i}+m} \tag{19}$$

$\forall j, \forall i$. As can be observed, if the value of an attribute is known, this model also takes into consideration the a priori probability of the class $\hat{P}(C=c_j)$ in the calculation

[b]The Laplace estimation of probabilities is a particular case of Equation 18 where $a=1$ and $b=k$.

of $\hat{P}(C=c_j|X_i=x_{t_i}^i)$. This a priori probability of the class is estimated using the Laplace's law of succession.

According to Cestnik, the higher the level of noise, the higher the value of the parameter m should be.

2.5. *The Imprecise Dirichlet Model*

The Imprecise Dirichlet Model (IDM)[21] is a specific type of reachable probabilities intervals, a formal mathematical theory based on imprecise probabilities[21], that also represents a belief function[25] that is employed to estimate probability intervals for each value of a certain variable X_i in a data set. In concrete, the IDM establishes that the probabilities that X_i has the value $x_{t_i}^i$ are within the interval

$$P(X_i=x_{t_i}^i) \in \left[\frac{N_{t_i}}{N+s}, \frac{N_{t_i}+s}{N+s}\right], \forall i \tag{20}$$

where s is a parameter of the model. It is easy to observe that the interval is getting wider when the value of s increases and the interval is getting narrower when the sample size is larger and the parameter s remains unchanged, that is, the parameter s determines how quickly the lower and upper probabilities converge as more data become available. Two values for the parameter s are suggested[21]: $s=1$ and $s=2$.

Through these probability intervals a convex set of probabilities on the variable X_i, $K(X_i)$, can be extracted[25].

$$K(X_i)=\left\{P|P(X_i=x_{t_i}^i) \in \left[\frac{N_{t_i}}{N+s}, \frac{N_{t_i}+s}{N+s}\right], \sum_{t_i} P(X_i=x_{t_i}^i)=1, \forall i\right\} \tag{21}$$

3. Imprecise m-probability-estimation naïve Bayes

We want to remark that the new NB approach, which we call Imprecise m-probability-estimation naïve Bayes (ImNB), as a variation of the NB model, assigns to a new instance the class label c_j, $\forall j$ that maximizes $\hat{P}(C=c_j|X_1=x_{t_1}^1,\ldots,X_n=x_{t_n}^n)$, being X_i the attributes of the instance and $x_{t_i}^i$ their possible values,$\forall i$. It has been proved in section 2.2 that it is equivalent to maximize Equation 11.

The difference between the ImNB and the rest of the NB algorithms resides in the estimation of the probabilities $\hat{P}(C=c_j)$ and $\hat{P}(C=c_j|X_i=x_{t_i}^i)$.

The estimation of the a priori probability $\hat{P}(C=c_j)$, which we call $\hat{P}_j$, is based on the credal set given in eq (21). We need to choose one probability distribution of this set. We choose those distribution that maximizes the Shannon entropy[26]. It is shown[22,27] that the use of this probability distribution provides good results when it is employed in C4.5, specially when there is noise in the data.

We can use Algorithm 1 to calculate the estimated probability $\hat{P}_j$.

```
Procedure CalculateAPriori(s)
s' = s;
for j = 1 to k do
    N'_{c_j} = N_{c_j};
end
while s' > 0 do
    s'' = min{s', 1};
    Let A = {c_j | N'_{c_j} = min_{i=1,...,k} N'_{c_i}};
    for j = 1 to k do
        if c_j ∈ A then
            N'_{c_j} = N'_{c_j} + s''/|A|;
        end
    end
    s' = s' - 1;
end
for j = 1 to k do
    P̂_j = N'_{c_j} / (N + s);
end
return P̂_j
```

Algorithm 1: Procedure of calculating the a priori probability of $\hat{P}(C=c_j)$, which is called $\hat{P}_j$.

In order to calculate the conditional probability $\hat{P}(C=c_j|X_i=x_{t_i}^i)$, $\forall j$, $\forall i$ the m-probability-estimation model is taken as a reference, i.e. the a priori probabilities are taken into account. However, these a priori probabilities are now the estimated through Algorithm 1. According to (Cestnik[18], 1990) the parameter m should be greater with more amount of noise in the data, as happens with the parameter s of the IDM[27]. For this reason the same values for s and m have been chosen.

Thus, the estimation of the conditional probability $\hat{P}(C=c_j|X_i=x_{t_i}^i)$ is given by:

$$\hat{P}(C=c_j|X_i=x_{t_i}^i)=\frac{N_{t_i,c_j}+s\hat{P}_j}{N_{t_i}+s} \tag{22}$$

where $\hat{P}_j$ is the estimated a priori probability of $\hat{P}(C=c_j)$ obtained from Algorithm 1. Therefore, ImNB, for an instance, chooses the class value c that verifies:

$$c=\arg\max_{c_j,j=1,\ldots,k}\hat{P}_j\prod_i\frac{\hat{P}(C=c_j|X_i=x_{t_i}^i)}{\hat{P}_j} \tag{23}$$

where $\hat{P}(C=c_j|X_i=x_{t_i}^i)$ is given by (22), $\forall j$, $\forall i$.

4. Experimental study

4.1. *Description of the experiments and results*

For the experiments, we have selected 75 well-known data sets, obtained from the *UCI repository of machine learning*[28]. All these data sets have been widely used in the specialized literature for comparing supervised learning algorithms. Table 1 shows the most relevant characteristics of each data set. As can be observed, these data sets are different with respect to the number of instances, the number of features, the number of states of the class variable, the type of the features (if they are discrete or continuous) and the number of states of discrete variables.

Since the new algorithm only works with discrete variables the databases have been discretized previously. For this purpose, the minimum description length principle criterion[29] has been employed.

In this experimentation, four algorithms have been compared. The first of them is the NB with classical estimation of probabilities, called classical NB. The second one is Laplace NB, i.e the naïve Bayes using the Laplace smoothing to estimate the probabilities. The third algorithm is the naïve Bayes with m-probability-estimation (which will be noted as mNB), explained in section 2.4. The fourth model is our proposal, called ImNB, which was explained in section 3.

For the experimental research, the Weka software[30] has been used. The four algorithms considered in the experiments have been implemented in this software taking the implementation of the standard NB as a reference. For the ImNB model, in preliminary experiments, it has been noted that the method has a good performance with $m = 4$ as the default value. Obviously, the results will improve if we tune the value of the parameter m to the level of noise of the data, however, its default value is used for the experiments. As regard to the mNB model, Cestnik did not make any recommendation for the m value, we only know that it is related to data noise, so a wide range of values for the parameter m have been considered. In concrete, twenty values have been tested for mNB ($m = 1, \ldots, m = 20$) in order to obtain the most appropriate value for each noise level.

Within the experiments, for each algorithm, four noise levels have been considered: 0%, 10%, 20%, and 30%. In our case only class noise has been taken into account. The noise had been added only to training sets. Weka's filters have been used to add the noise in the different cases. The corresponding noise has been added in each case by using a random procedure: Given a percentage x, an x% of the instances are selected randomly of the training data set and the class label is changed randomly to another possible class value. The instances belonging to the test data set were left unmodified. To compare the results of the classifiers, 10 times a 10-fold cross validation procedure was repeated for each data set and for each level of noise and for each algorithm.

In order to compare the results the following procedure has been carried out for each level of noise: The best value of the parameter m has been selected for the mNB model. All the algorithms have been compared, all of them with their default

Table 1. Data set description. Column 'N' is the number of instances in the data sets, column 'Feat' is the number of features or attribute variables, column 'Num' is the number of numerical variables, column 'Nom' is the number of nominal variables, column 'k' is the number of cases or states of the class variable (always a nominal variable) and column 'Range' is the range of states of the nominal variables of each data set.

Dataset	N	Feat	Num	Nom	k	Range
acute-infl-nephritis	120	6	1	5	2	2
anneal	898	38	6	32	6	2-10
appendicitis	106	7	7	0	2	-
arrhythmia	452	279	206	73	16	2
audiology	226	69	0	69	24	2-6
autos	205	25	15	10	7	2-22
balance-scale	625	4	4	0	3	-
bank-marketing	4521	16	7	9	2	2-12
banknote-auth	1372	4	4	0	2	-
breast-cancer	286	9	0	9	2	2-13
breast-cancer-wisconsin	699	9	9	0	2	-
bridges-version1	107	11	3	8	6	2-54
bridges-version2	107	11	0	11	6	2-54
bupa	345	6	6	9	2	-
car	1728	6	0	6	4	3-4
cmc	1473	9	2	7	3	2-4
horse-colic	368	22	7	15	2	2-6
credit-rating-australian	690	15	6	9	2	2-14
credit-rating-german	1000	20	7	13	2	2-11
dermatology	366	34	1	33	6	2-4
diabetes-pima	768	8	8	0	2	-
dresses-sales	500	12	1	11	2	5-25
ecoli	366	7	7	0	7	-
fertility-diagnosis	100	9	9	0	2	-
flags	194	29	2	27	8	4-194
glass	214	9	9	0	7	-
glioma16	50	16	16	0	2	-
haberman	306	3	2	1	2	12
heart-disease-cleveland	303	13	6	7	5	2-14
heart-disease-hungarian	294	13	6	7	5	2-14
heart-statlog	270	13	13	0	2	-
hepatitis	155	19	4	15	2	2
hypothyroid	3772	30	7	23	4	2-4
ionosphere	351	35	35	0	2	-
iris	150	4	4	0	3	-
japanese-crx	690	15	6	9	2	2-14
kr-vs-kp	3196	36	0	36	2	2-3
letter	20000	16	16	0	26	-
liver-disorders	345	6	6	0	2	-
lsvt-voice-rehab	126	310	310	0	2	-
lymphography	146	18	3	15	4	2-8
mfeat-pixel	2000	240	0	240	10	4-6
mol-splice-junction	3190	60	0	60	3	4-5
nursery	12960	8	0	8	4	2-4
optdigits	5620	64	64	0	10	-
page-blocks	5473	10	10	0	5	-
parkinsons	195	22	22	0	2	-
pendigits	10992	16	16	0	10	-
postoperative-patient	90	8	8	0	3	2-4
primary-tumor	339	17	0	17	21	2-3
qsar-biodegradation	1055	41	41	0	2	-
qualitative-bankruptcy	250	6	0	6	2	3
saheart	462	9	8	1	2	2
segment	2310	19	16	0	7	-
seismic-bumps	2584	18	14	4	2	2-3
sick	3772	29	7	22	2	2
solar-flare2	1066	12	0	6	3	2-8
sonar	208	60	60	0	2	-
soybean	683	35	0	35	19	2-7
spambase	4601	57	57	0	2	-
spect	267	22	0	22	2	2
spectf	349	44	44	0	2	-
spectrometer	531	101	100	1	48	4
splice	3190	60	0	60	3	4-6
sponge	76	44	0	44	3	2-9
tae	151	5	3	2	3	2
thoracic-surgery	470	16	3	13	2	2-7
tic-tac-toe	958	9	0	9	2	3
turkiye-student	5820	32	32	0	13	-
vehicle	946	18	18	0	4	-
vote	435	16	0	16	2	2
vowel	990	11	10	1	11	2
waveform	5000	40	40	0	3	-
wine	178	13	13	0	3	-
zoo	101	16	1	16	7	2

parameter values and mNB with its selected value of m.

Following the recommendation of Demšar[31], we have used the Friedman test[32,33] for the algorithms comparisons. It is a non-parametric test, which ranks the classification algorithms for each data set separately, according to their performance, in ascending order (from the best-performing algorithm to the worst-performing one). The null hypothesis is that all algorithms are equivalent. If this hypothesis is rejected, then we compare the algorithms using the Holm's procedure[34]. The level of significance used is $\alpha = 0.05$.

To make the selection of the best value for the parameter m used in the mNB algorithm, the Friedman ranking has been used. In this way, the chosen value for the parameter m is the one which obtains a better value for the rank. For each noise level, the results of these ranks can be observed in Table 2 and the best value of the parameter m for each noise level is emphasized using bold fonts. For the sake of clarity and simplification, the accuracy results obtained for the computation of the Friedman test used to obtain the best value of the parameter m for mNB have not been included.

With the selected values for the parameter m obtained in Table 2 for mNB, all the studied approaches were used to classify the data sets. Table 3 presents the average accuracy results of the methods used in the experimentation: Classical NB, Laplace NB, mNB and ImNB. In this table, the best algorithm for each added noise level is emphasized using bold fonts, the second best is marked with italic fonts. Tables that present the detailed accuracy results of the different algorithms for different percentages of class noise are described in Appendix A.1.

Table 2. Friedman ranks about the accuracy for the mNB approach with different values of m and different levels of class noise.

m	0% Noise	10% Noise	20% Noise	30% Noise
1	**7.93**	9.63	11.75	12.49
2	8.14	**8.95**	10.74	11.92
3	7.98	8.99	10.41	11.51
4	8.11	9.35	10.30	11.07
5	8.65	9.59	10.21	11.04
6	9.05	9.67	10.19	10.66
7	9.39	9.39	10.71	10.11
8	9.23	9.28	**9.99**	10.27
9	9.55	9.68	10.15	10.29
10	10.25	10.26	10.15	10.26
11	10.49	10.71	10.05	10.48
12	10.87	11.11	10.20	10.23
13	11.67	11.59	10.17	**9.52**
14	11.95	11.58	10.17	9.73
15	12.15	11.81	10.18	9.77
16	12.01	11.44	10.64	9.81
17	12.65	11.49	11.07	9.77
18	12.93	11.74	11.03	9.88
19	13.35	11.94	10.81	10.45
20	13.63	11.80	11.09	10.75

Table 4 shows Friedman's ranks about the accuracy of the studied NB variations when they are applied on data sets with and without class noise. The best approach

for each noise level is noted using bold fonts, the second best method is emphasized with italic fonts.

Table 3. Average accuracy results of the NB variations when they are built from data sets with and without added class noise.

Algorithm	noise 0%	noise 10%	noise 20%	noise 30%
Classical NB	77.05	74.69	73.08	71.08
Laplace NB	77.33	75.44	74.09	72.19
mNB$_{BESTm}$	*79.60*	*77.74*	*76.11*	*73.54*
ImNB	**79.88**	**78.67**	**77.17**	**74.64**

Table 4. Friedman's ranks about the accuracy of the algorithms when they are built from datasets with different percentages of added noise.

Algorithm	noise 0%	noise 10%	noise 20%	noise 30%
Classical NB	2.76	3.12	3.05	2.97
Laplace NB	2.97	2.81	2.77	2.66
mNB$_{BESTm}$	*2.21*	*2.35*	*2.37*	*2.36*
ImNB	**2.07**	**1.73**	**1.81**	**2.01**

Tables 5, 6, 7 and 8 show the p-values of the Holm test on the pairs of comparisons. In the event that there is a significative difference, the best algorithm is marked with bold fonts. The figures 1, 2, 3 and 4 show the same information graphically. These figures are critical difference (CD) diagrams, where average ranks of examined methods are presented. Bold lines indicate groups of classifiers which are not significantly different (their average ranks differ by less than critical difference value).

Table 5. p-values of the Holm test about the accuracy on data sets without added class noise. Holm test rejects the hypotheses that the methods are equivalent if the corresponding p-value is ≤ 0.025.

i	algorithms	p
6	Laplace NB vs. **ImNB**	0.000020
5	Laplace NB vs. **mNB**$_{m=1}$	0.000312
4	Classical NB vs. **ImNB**	0.001006
3	Classical NB vs. **mNB**$_{m=1}$	0.008673
2	Classical NB vs. Laplace NB	0.326930
1	mNB$_{m=1}$vs. ImNB	0.506640

Table 6. p-values of the Holm test about the accuracy on datasets with 10% of added class noise. Holm test rejects the hypotheses that the methods are equivalent if the corresponding p-value is ≤ 0.025.

i	algorithms	p
6	Classical NB vs. **ImNB**	0.000000
5	Laplace NB vs. **ImNB**	0.000000
4	Classical NB vs. **mNB**$_{m=2}$	0.000244
3	mNB$_{m=2}$ vs. **ImNB**	0.003272
2	Laplace NB vs. mNB$_{m=2}$	0.029112
1	Classical NB vs. Laplace NB	0.137208

Since we are using data sets with label noise, it would be desirable to use a robustness metric to establish the expected behavior of a classifier against noisy data. A measure which considers performance and robustness individually for each

Table 7. p-values of the Holm test about the accuracy on datasets with 20% of added class noise. Holm test rejects the hypotheses that the methods are equivalent if the corresponding p-value is ≤ 0.025.

i	algorithms	p
6	Classical NB vs. **ImNB**	0.000000
5	Laplace NB vs. **ImNB**	0.000005
4	Classical NB vs. **mNB**$_{m=8}$	0.001125
3	mNB$_{m=8}$ vs. **ImNB**	0.007900
2	Laplace NB vs. mNB$_{m=8}$	0.053732
1	Classical NB vs. Laplace NB	0.184126

Table 8. p-values of the Holm test about the accuracy on datasets with 30% of added class noise. Holm test rejects the hypotheses that the methods are equivalent if the corresponding p-value is ≤ 0.016667.

i	algorithms	p
6	Classical NB vs. **ImNB**	0.000005
5	Laplace NB vs. **ImNB**	0.001942
4	Classical NB vs. **mNB**$_{m=13}$	0.003622
3	mNB$_{m=13}$ vs. ImNB	0.093737
2	Laplace NB vs. mNB$_{m=13}$	0.137208
1	Classical NB vs. Laplace NB	0.154729

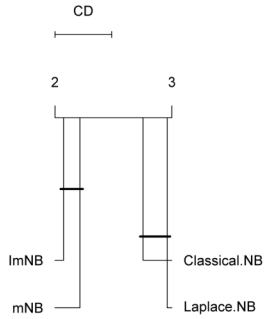

Fig. 1. Critical difference diagram about the accuracy on data sets without added class noise.

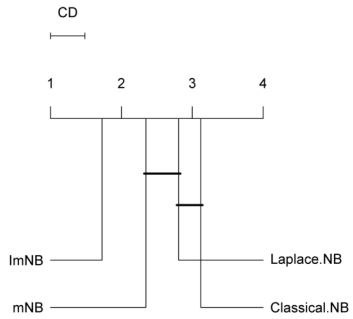

Fig. 2. Critical difference diagram about the accuracy on data sets with 10% of added class noise.

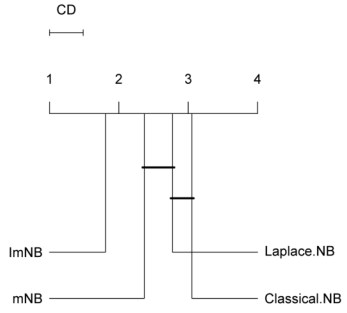

Fig. 3. Critical difference diagram about the accuracy on data sets with 20% of added class noise.

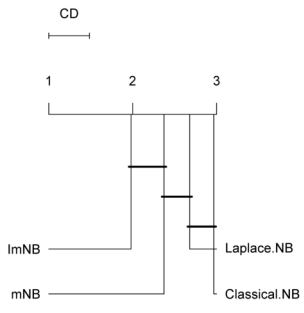

Fig. 4. Critical difference diagram about the accuracy on data sets with 30% of added class noise.

classifier is the *Equalized Loss of Accuracy* (ELA)[35]. This measure computes the performance without added noise taking into account which algorithm is more ap-

propriate to deal with noisy data. The approach with the lowest value for ELA will be the most robust. In Table 9, it can be observed the obtained results of the ELA metric for each NB variation (in bold it is marked the best one and in italic the second best for each noise level).

Table 9. Values of the ELA measure obtained for each noise level.

Algorithm	noise 10%	noise 20%	noise 30%
Classical NB	0.3285	0.3494	0.3753
Laplace NB	0.3176	0.3351	0.3596
mNB_{BESTm}	*0.2796*	*0.3001*	*0.3324*
ImNB	**0.2670**	**0.2858**	**0.3175**

4.2. *Analysis of the results*

From a general point of view, we can state that the two approaches in which our work is focused, i.e. mNB (tuned) and ImNB, have a better performance than the naïve Bayes models used as reference (NB with and without Laplace smoothing) on data sets with and without label noise. The improvement is not only with respect to the classifier accuracy, via the tests of Friedman and Holm carried out, but also in terms of robustness, via ELA measure. Furthermore, we can note that our contribution obtains better results than Cestnik's approach being this improvement statistically significant in some cases.

Next, we are going to analyze in detail the experimental results with respect to the level of noise, taking into account the following aspects: Average accuracy, Friedman's ranking, Holm test, and robustness:

- **Average accuracy**: According to this aspect, the methods based on m-probability-estimation, i.e. mNB and ImNB, attains constantly the best average accuracy, regardless of the level of added noise. The best result has always been for our proposed method (ImNB) and the second best result have been invariably obtained by the Cestnik's approach (mNB) with tuned parameter. In this regard, the worst results are invariably obtained by the NB without Laplace smoothing (Classical NB). We want to emphasize that this ordering from the best NB variation (ImNB) to the worst (Classical NB) occurs independently of the percentage of added noise.
- **Friedman's ranking and Holm test**: The outcomes of the set of statistical tests carried out, reinforce the comments made about the average accuracy. The tuned mNB and the ImNB variations obtain the best Friedman's rank for all levels of added noise, and always the finest classifier in the ranking is our proposal (ImNB). According to the Friedman's ranking, the worst NB variation for data sets without added noise is, interestingly, the NB with Laplace smoothing (Laplace NB). Notwithstanding, the NB

without Laplace smoothing (Classical NB) obtain the worst score in the rank for any level of added noise (10%, 20% or 30%).
With respect to the Holm test, our approach is the only one that outperforms always the NB classifiers with or without Laplace smoothing, regardless of the level of added class noise. The Cestnik proposal shows a less consistent behavior, significantly outperforming the NB without Laplace smoothing for all noise levels but it is only statistically better than NB with Laplace smoothing when the level of noise is 0%. The differences between ImNB and tuned mNB are not always statistically significant in accordance with the Holm test, but the ImNB is statistically better than tuned mNB when the level of noise is 10% or 20%.

- **ELA measure**: In relation to this metric, there is no doubt about what methods of the comparison are the most robust to label noise. The ordering presented in the average accuracy measure and the Friedman's ranking is the same that the sequence obtained by the ELA measure. Therefore, our proposal is in the first place, the tuned mNB classifier is the second, the NB with the Laplace smoothing is in the third place and the worst result is obtained by the NB without Laplace smoothing (classical NB). This outcome occurs independently of the percentage of added noise.

With the above analysis of the results we can conclude that the methods based on m-probability-estimation outperform the conventional approaches of the NB, that is, NB with and without Laplace smoothing. These outcomes are presented consistently, regardless of whether the data sets suffer or not of class noise. Taking into account differences statistically significant, we notice that the better outcomes are achieved by our proposal.

5. Conclusions

In this research, two naïve Bayes approaches are the subject of our study: the first one is the naïve Bayes using m-probability-estimation (mNB) proposed by Cestnik[18] and the second one, our proposal, a naïve Bayes classifier using m-probability-estimation and imprecise probabilities (ImNB). We believe that the Cestnik approach has not received sufficient attention from the scientific community and we think this is why we have not found an extensive comparison where this model evidences its worth. The second classifier in this study, our contribution, is a variation of the Cestnik approach which achieves better results without parameter tuning.

We also consider that this paper studies different ways of estimating probabilities in the naïve Bayes algorithm. In concrete, we have started with the classical estimations and the Laplace smoothing, recalling their drawbacks. We have also considered the m-probability-estimation model presented in (Cestnik[18], 1990), the NB model that estimates conditional probabilities taking into account the a priori probabilities. Moreover, we have proposed a new way of estimating probabilities in NB based on imprecise probabilities and the m-probability-estimations, our ImNB

algorithm.

A wide experimentation has been carried out , considering 75 data sets with different characteristics and different levels of noise. This experimental study has indicated four things:

- The best choice of the parameter m in m-probability-estimation depends on the level of noise in the data, being generally higher the more there is noise in the data.
- tuned mNB and ImNB are clearly better than the naïve Bayes estimating probabilities in a classical way and with Laplace smoothing, regardless of the level of noise in the data.
- ImNB used with its default m value is always better than m-probability-estimation even with parameter tuning, although the differences are not statistically significant in some cases.
- ImNB is always statistically better than NB with and without Laplace smoothing.

Hence, in this work, it has been demonstrated, with a far more exhaustive experimentation than in (Cestnik[18], 1990), that the m-probability-estimation model supposes a very considerable improvement over Laplace and classical estimations of probabilities in naïve Bayes. Furthermore, we have presented a new way of estimating probabilities in NB based on m-probability-estimation and imprecise probabilities that involves an improvement.

Acknowledgements

This work has been supported by the Spanish "Ministerio de Economía y Competitividad" and by "Fondo Europeo de Desarrollo Regional" (FEDER) under Project TEC2015-69496-R.

References

1. D. J. Hand, *Construction and Assessment of Classification Rules*. John Wiley and Sons, New York, 1997.
2. D. Hand, *Discrimination and Classification*. John Wiley, 1981.
3. J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.
4. J. Pearl, *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Mateo, CA, 1988.
5. G. Romero, M. G. Arenas, J. G. Castellano, P. A. Castillo, J. Carpio, J. J. Merelo, A. Prieto, and V. Rivas, "Evolutionary computation visualization: Application to g-prop," in *Parallel Problem Solving from Nature PPSN VI*, (Berlin, Heidelberg), pp. 902–912, Springer Berlin Heidelberg, 2000.
6. L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

7. Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)* (L. Saitta, ed.), pp. 148–156, Morgan Kaufmann, 1996.
8. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
9. J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
10. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: John Wiley and Sons, 1973.
11. P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2, pp. 103–130, 1997.
12. N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2, pp. 131–163, 1997.
13. P. Langley, W. Iba, and K. Thompson, "An analysis of bayesian classifiers," in *Proceedings of the 10th national conference on Artificial intelligence (AAAI'92)*, pp. 223–228, MIT Press, 1992.
14. J. L. Hellerstein, T. S. Jayram, and I. Rish, "Recognizing end-user transactions in performance management," in *Proceedings of the 7th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pp. 596–602, 2000.
15. H. Zhang and D. Li, "Naïve bayes text classifier," in *IEEE International Conference on Granular Computing (GRC 2007)*, pp. 708–708, 2007.
16. L. M. de Campos, A. Cano, J. G. Castellano, and S. Moral, "Bayesian networks classifiers for gene-expression data," in *11th International Conference on Intelligent Systems Design and Applications*, pp. 1200–1206, 2011.
17. J. Abellán and J. G. Castellano, "Improving the naive bayes classifier via a quick variable selection method using maximum of entropy," *Entropy*, vol. 19, no. 6, 2017.
18. B. Cestnik, "Estimating probabilities: A crucial task in machine learning," in *Proceedings of the 9th European Conference on Artificial Intelligence (ECAI'90)*, pp. 147–149, London Pitman Publishing, 1990.
19. I. J. Good, "Probability and the weighing of evidence," *British Journal of Social Medicine*, vol. 4, no. 3, pp. 170–171, 1950.
20. B. Cestnik and I. Bratko, "On estimating probabilities in tree pruning," in *European Working Session on Learning (EWSL-91)*, vol. 482 of *Lecture Notes in Computer Science (LNCS)*, pp. 138–150, Springer Berlin Heidelberg, 1991.
21. P. Walley, "Inferences from multinomial data; learning about a bag of marbles (with discussion).," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 3–57, 1996.
22. C. J. Mantas and J. Abellán, "Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data," *Expert Systems with Applications*, vol. 41, no. 10, pp. 4625 – 4637, 2014.
23. C. J. Mantas, J. Abellán, and J. G. Castellano, "Analysis of Credal-C4.5 for classification in noisy domains," *Expert Systems with Applications*, vol. 61, pp. 314 – 326, 2016.
24. I. J. Good, *The Estimation of Probabilities. An essay on modern Bayesian Methods*, vol. 30 of *Volume 30 of MIT Research monograph*. The M. I. T. Press Cambridge, 1965.
25. J. Abellán, "Uncertainty measures on probability intervals from the imprecise dirichlet model," *International Journal of General Systems*, vol. 35, no. 5, pp. 509–528, 2006.
26. C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

27. J. Abellán, C. J. Mantas, and J. G. Castellano, "AdaptativeCC4.5: Credal C4.5 with a rough class noise estimator," *Expert Systems with Applications*, vol. 92, no. Supplement C, pp. 363 – 379, 2018.
28. M. Lichman, "UCI machine learning repository," 2013.
29. U. Fayyad and K. Irani, "Multi-valued interval discretization of continuous-valued attributes for classification learning," in *Proceeding of the 13th International joint Conference on Artificial Inteligence*, pp. 1022–1027, Morgan Kaufmann, 1993.
30. I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., second ed., 2005.
31. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
32. M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, pp. 675–701, 1937.
33. M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.
34. S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.
35. J. A. Sáez, J. Luengo, and F. Herrera, "Evaluating the classifier behavior with noisy data considering performance and robustness: the equalized loss of accuracy measure," *Neurocomputing*, vol. 176, pp. 26 – 35, 2014. Recent Advancements in Hybrid Artificial Intelligence Systems and its Application to Real-World Problems, selected papers from the {HAIS} 2013 conference.

Appendix A. Appendices

A.1. *Tables about accuracy results*

Tables 10, 11, 12 and 13 show the accuracy results obtained by the different NB methods when they classify data sets with different percentage of added noise. The best algorithm for each data set is emphasized using bold fonts.

Table 10. Accuracy results of Classical NB, Laplace NB, mNB and ImNB when they are used on data sets without added class noise.

Dataset	Classical NB	Laplace NB	$mNB_{m=1}$	$ImNB_{m=4}$
acute-infl-nephritis	95.17	95.00	**100.00**	**100.00**
anneal	93.44	86.59	**97.26**	96.50
appendicitis	85.21	85.21	85.25	**86.20**
arrhythmia	61.51	62.40	68.85	**71.68**
audiology	**78.78**	72.64	77.53	77.72
autos	60.11	57.41	**69.30**	65.41
balance-scale	**90.53**	**90.53**	71.56	71.56
bank-marketing	86.72	86.77	87.86	**88.01**
banknote-auth	84.01	84.01	91.98	**92.03**
breast-cancer	71.93	72.70	72.17	**72.91**
breast-cancer-wisconsin	96.07	96.07	**97.18**	97.11
bridges-version1	60.51	**69.33**	62.98	65.25
bridges-version2	60.59	**67.25**	63.64	66.25
bupa	55.03	54.89	**56.85**	**56.85**
car	85.86	85.46	**85.90**	84.80
cmc	**50.51**	50.48	50.01	50.15
horse-colic	78.24	78.70	79.98	**81.19**
credit-rating-australian	77.90	77.86	**86.19**	85.93
credit-rating-german	75.25	75.16	74.95	**75.29**
dermatology	97.73	97.43	**98.23**	97.93
diabetes-pima	**75.77**	75.75	75.25	75.51
dresses-sales	58.22	61.64	60.90	**62.94**
ecoli	**85.59**	85.50	81.39	81.31
fertility-diagnosis	86.50	86.50	**88.00**	**88.00**
flags	52.54	52.49	51.65	**57.32**
glass	49.54	49.45	72.75	**73.09**
glioma16	82.40	82.20	**87.20**	82.40
haberman	75.29	**75.36**	71.97	72.07
heart-disease-cleveland	83.11	83.34	**83.47**	**83.47**
heart-disease-hungarian	**84.84**	83.95	84.13	84.50
heart-statlog	**83.59**	**83.59**	82.56	82.70
hepatitis	84.38	83.81	**85.42**	85.23
hypothyroid	95.59	95.30	**98.90**	98.77
ionosphere	82.17	82.17	89.49	**89.66**
iris	**95.53**	**95.53**	93.20	93.33
japanese-crx	77.90	77.86	**86.19**	85.93
kr-vs-kp	**87.81**	87.79	**87.81**	87.24
letter	64.07	64.07	**74.79**	74.64
liver-disorders	55.03	54.89	**56.85**	**56.85**
lsvt-voice-rehab	54.33	54.33	76.45	**81.37**
lymphography	81.42	83.13	**85.72**	83.70
mfeat-pixel	75.55	93.36	**93.74**	93.58
mol-splice-junction	95.48	95.42	**95.57**	95.38
nursery	90.31	90.30	**90.32**	**90.32**
optdigits	91.39	91.39	**92.24**	92.21
page-blocks	90.03	90.01	**93.88**	93.82
parkinsons	70.14	70.14	79.53	**81.01**
pendigits	85.76	85.76	**88.14**	87.91
postoperative-patient	63.33	68.11	63.56	**69.78**
primary-tumor	47.41	49.71	48.92	**49.83**
qsar-biodegradation	75.88	75.89	80.99	**81.00**
qualitative-bankruptcy	**99.60**	99.32	**99.60**	99.24
saheart	**71.10**	71.05	67.68	68.65
segment	80.17	80.17	**91.89**	91.42
seismic-bumps	86.68	**86.72**	81.58	82.48
sick	92.59	92.75	96.75	**97.09**
solar-flare2	**98.05**	97.56	97.09	97.57
sonar	67.71	67.71	**76.80**	76.07
soybean	**94.67**	92.94	91.60	90.82
spambase	79.56	79.56	**89.85**	89.80
spect	80.29	78.68	79.66	**82.92**
spectf	71.72	71.75	79.18	**79.81**
spectrometer	41.96	42.06	45.08	**45.16**
splice	95.48	95.42	**95.57**	95.38
sponge	**95.00**	92.11	90.70	92.50
tae	**55.12**	54.01	46.32	42.59
thoracic-surgery	77.13	77.74	81.94	**84.87**
tic-tac-toe	69.57	69.64	69.57	**71.15**
turkiye-student	**25.78**	25.77	25.53	25.41
vehicle	44.68	44.68	61.04	**61.16**
vote	**90.09**	90.02	**90.09**	89.98
vowel	62.06	**62.90**	59.09	58.82
waveform	80.01	80.01	79.96	**80.10**
wine	97.46	97.46	98.37	**98.54**
zoo	**96.35**	95.07	96.15	91.80
Average	77.05	77.33	79.60	**79.88**

Table 11. Accuracy results of Classical NB, Laplace NB, mNB and ImNB when they are used on data sets with a percentage of added class noise equal to 10%.

Dataset	Classical NB	Laplace NB	$mNB_{m=2}$	$ImNB_{m=4}$
acute-infl-nephritis	97.00	97.00	**99.83**	**99.83**
anneal	82.35	83.37	90.30	**91.98**
appendicitis	**85.26**	**85.26**	83.49	83.67
arrhythmia	55.40	58.26	70.37	**72.35**
audiology	70.54	70.61	71.78	**73.25**
autos	52.94	50.39	58.84	**60.97**
balance-scale	**89.51**	**89.51**	74.53	74.42
bank-marketing	86.32	86.33	87.59	**87.82**
banknote-auth	83.96	83.95	**90.59**	90.55
breast-cancer	71.45	72.29	71.83	**72.50**
breast-cancer-wisconsin	96.25	96.25	97.21	**97.24**
bridges-version1	52.55	**65.67**	58.35	60.01
bridges-version2	52.99	**66.29**	61.32	64.14
bupa	55.65	55.71	57.89	**58.03**
car	**83.44**	83.26	83.43	83.06
cmc	49.80	49.89	50.26	**50.33**
horse-colic	77.48	78.05	79.51	**80.74**
credit-rating-australian	78.12	78.17	85.38	**85.46**
credit-rating-german	**75.07**	74.91	73.79	74.17
dermatology	95.20	97.60	**97.82**	97.66
diabetes-pima	**75.17**	**75.17**	73.41	73.24
dresses-sales	57.62	60.02	59.94	**61.24**
ecoli	85.47	**85.62**	81.67	81.76
fertility-diagnosis	84.30	84.30	**88.00**	**88.00**
flags	47.79	49.53	50.00	**55.08**
glass	45.35	45.26	62.73	**63.15**
glioma16	80.00	80.00	**82.60**	80.20
haberman	**75.16**	**75.16**	71.52	71.96
heart-disease-cleveland	81.36	81.86	83.11	**83.67**
heart-disease-hungarian	83.23	82.32	83.07	**83.41**
heart-statlog	**84.07**	84.04	82.33	82.93
hepatitis	82.70	82.12	82.61	**83.47**
hypothyroid	94.14	94.07	97.21	**97.22**
ionosphere	81.97	81.94	88.63	**88.94**
iris	93.00	93.07	**94.00**	**94.00**
japanese-crx	78.12	78.17	85.38	**85.46**
kr-vs-kp	**86.34**	86.31	86.32	85.77
letter	62.24	62.24	**72.86**	72.84
liver-disorders	55.65	55.71	57.89	**58.03**
lsvt-voice-rehab	50.72	50.72	75.84	**76.85**
lymphography	80.02	**82.24**	80.20	81.17
mfeat-pixel	74.88	92.93	**93.06**	92.85
mol-splice-junction	93.36	93.45	93.45	**93.62**
nursery	90.39	**90.40**	**90.40**	**90.40**
optdigits	91.07	91.07	**91.36**	91.28
page-blocks	88.93	88.93	92.32	**92.38**
parkinsons	65.06	65.11	78.43	**78.97**
pendigits	85.43	85.43	**86.24**	86.18
postoperative-patient	59.78	64.11	60.44	**64.33**
primary-tumor	45.25	**47.17**	45.72	46.61
qsar-biodegradation	70.71	70.71	78.50	**78.84**
qualitative-bankruptcy	**99.20**	**99.20**	**99.20**	**99.20**
saheart	**70.68**	70.64	67.20	68.72
segment	72.41	72.40	**89.36**	89.23
seismic-bumps	85.80	85.83	85.46	**86.61**
sick	91.82	92.06	95.71	**96.38**
solar-flare2	98.30	97.51	98.31	**98.53**
sonar	66.23	66.23	**73.34**	71.67
soybean	**92.59**	91.92	90.07	89.92
spambase	71.30	71.30	**89.12**	**89.12**
spect	74.36	74.09	73.94	**81.64**
spectf	57.76	57.76	75.10	**78.35**
spectrometer	39.00	38.96	43.22	**43.33**
splice	93.36	93.45	93.45	**93.62**
sponge	89.16	79.63	63.29	**91.07**
tae	**52.15**	51.88	45.64	45.84
thoracic-surgery	76.32	76.79	82.43	**84.91**
tic-tac-toe	69.82	69.84	69.82	**71.20**
turkiye-student	24.37	24.35	26.96	**26.97**
vehicle	44.97	44.97	**58.96**	58.75
vote	**89.97**	89.90	89.81	89.88
vowel	57.30	**58.28**	54.05	54.30
waveform	78.24	78.24	79.05	**79.24**
wine	95.89	95.89	96.96	**97.07**
zoo	89.92	94.96	**96.45**	92.79
Average	74.69	75.44	77.74	**78.67**

Table 12. Accuracy results of Classical NB, Laplace NB, mNB and ImNB when they are used on data sets with a percentage of added class noise equal to 20%.

Dataset	Classical NB	Laplace NB	$mNB_{m=8}$	$ImNB_{m=4}$
acute-infl-nephritis	97.83	97.67	98.58	**98.75**
anneal	79.26	81.82	88.65	**89.56**
appendicitis	**83.75**	83.65	80.28	80.76
arrhythmia	43.95	49.03	66.63	**67.67**
audiology	66.29	**69.76**	65.86	67.31
autos	49.79	48.89	56.25	**59.02**
balance-scale	**88.55**	**88.55**	74.22	74.22
bank-marketing	86.04	86.03	88.45	**88.54**
banknote-auth	82.57	82.57	**88.22**	88.02
breast-cancer	69.83	70.77	71.23	**71.72**
breast-cancer-wisconsin	96.27	96.27	96.72	**96.78**
bridges-version1	46.35	**64.16**	55.82	57.47
bridges-version2	47.88	**64.33**	59.55	63.06
bupa	54.63	54.66	57.86	**58.01**
car	**82.49**	82.45	82.44	82.38
cmc	49.55	49.58	50.03	**50.12**
horse-colic	76.31	76.85	77.91	**79.57**
credit-rating-australian	78.22	78.36	84.54	**84.58**
credit-rating-german	**74.19**	74.16	72.83	73.50
dermatology	94.71	97.43	**97.63**	97.30
diabetes-pima	75.07	**75.09**	73.35	73.37
dresses-sales	56.38	58.64	59.62	**59.88**
ecoli	**83.59**	83.53	79.81	80.11
fertility-diagnosis	81.90	81.70	**88.00**	**88.00**
flags	43.71	46.63	40.18	**51.57**
glass	42.34	42.06	**56.79**	56.47
glioma16	**77.40**	**77.40**	73.80	69.20
haberman	73.72	**73.88**	71.63	71.44
heart-disease-cleveland	80.74	81.43	83.37	**83.40**
heart-disease-hungarian	81.87	81.23	82.87	**83.37**
heart-statlog	**83.74**	**83.74**	78.93	78.85
hepatitis	**82.24**	81.93	82.03	82.22
hypothyroid	93.78	93.72	**96.56**	96.46
ionosphere	81.85	81.85	88.75	**88.84**
iris	90.60	90.67	93.53	**93.73**
japanese-crx	78.22	78.36	84.54	**84.58**
kr-vs-kp	85.53	85.54	**85.55**	84.98
letter	60.63	60.63	71.41	**71.43**
liver-disorders	54.63	54.66	57.86	**58.01**
lsvt-voice-rehab	49.49	49.49	**72.06**	69.53
lymphography	78.01	**80.77**	78.74	79.72
mfeat-pixel	75.72	92.55	**92.59**	92.58
mol-splice-junction	91.59	91.68	91.51	**92.13**
nursery	90.55	**90.56**	**90.56**	**90.56**
optdigits	90.02	90.02	**90.64**	90.56
page-blocks	88.51	88.51	**91.64**	91.60
parkinsons	63.85	63.85	77.05	**79.46**
pendigits	84.16	84.16	84.88	**84.92**
postoperative-patient	57.67	61.22	61.89	**63.44**
primary-tumor	42.98	**46.08**	44.69	45.87
qsar-biodegradation	68.01	68.01	75.20	**76.40**
qualitative-bankruptcy	99.16	99.16	99.16	**99.20**
saheart	**70.42**	70.40	67.51	68.12
segment	69.52	69.52	87.48	**87.59**
seismic-bumps	85.55	85.55	92.35	**92.71**
sick	91.59	91.73	96.41	**96.59**
solar-flare2	98.28	97.71	98.71	**98.89**
sonar	65.94	65.94	**70.56**	62.08
soybean	**91.32**	91.17	88.72	88.81
spambase	72.65	72.64	88.63	**88.66**
spect	72.70	72.74	71.69	**81.83**
spectf	54.50	54.50	72.05	**74.92**
spectrometer	36.03	35.73	35.85	**37.08**
splice	91.59	91.68	91.51	**92.13**
sponge	80.84	72.70	44.13	**88.16**
tae	**48.15**	47.82	44.64	43.52
thoracic-surgery	75.51	75.96	82.74	**84.30**
tic-tac-toe	69.25	69.30	69.36	**70.52**
turkiye-student	23.72	23.70	**34.03**	**34.03**
vehicle	44.14	44.13	**56.72**	56.59
vote	89.45	89.45	89.49	**89.56**
vowel	52.19	**53.51**	46.98	46.59
waveform	77.30	77.30	78.33	**78.55**
wine	**94.77**	**94.77**	93.82	94.37
zoo	89.31	93.20	**93.65**	92.20
Average	73.08	74.09	76.11	**77.17**

Table 13. Accuracy results of Classical NB, Laplace NB, mNB and ImNB when they are used on data sets with a percentage of added class noise equal to 30%.

Dataset	Classical NB	Laplace NB	$mNB_{m=13}$	$ImNB_{m=4}$
acute-infl-nephritis	**96.17**	**96.17**	91.00	89.25
anneal	78.11	80.07	**87.84**	87.43
appendicitis	78.58	78.58	80.26	**80.56**
arrhythmia	39.29	43.52	58.67	**59.25**
audiology	61.12	**67.42**	58.06	64.82
autos	46.64	46.74	53.20	**54.90**
balance-scale	**87.97**	**87.97**	68.55	68.53
bank-marketing	84.84	84.86	88.69	**88.74**
banknote-auth	81.68	81.68	87.12	**87.16**
breast-cancer	66.50	67.49	69.02	**70.42**
breast-cancer-wisconsin	95.99	95.99	**96.72**	96.67
bridges-version1	43.70	**63.99**	54.15	55.30
bridges-version2	44.47	**62.81**	59.45	61.35
bupa	53.12	53.06	57.63	**57.75**
car	**81.90**	81.87	81.44	81.79
cmc	48.72	**48.79**	47.72	48.01
horse-colic	74.90	75.48	76.42	**78.10**
credit-rating-australian	75.87	75.96	**83.26**	83.14
credit-rating-german	71.36	71.41	70.56	**71.53**
dermatology	93.42	96.97	**97.11**	96.61
diabetes-pima	73.87	**73.89**	69.87	70.58
dresses-sales	53.68	54.84	56.82	**57.24**
ecoli	**83.59**	**83.59**	78.48	78.61
fertility-diagnosis	74.00	73.70	**88.00**	**88.00**
flags	40.88	42.73	36.72	**48.81**
glass	43.69	43.60	**51.26**	50.13
glioma16	**69.40**	**69.40**	63.60	62.20
haberman	69.47	69.93	**70.29**	68.67
heart-disease-cleveland	79.71	80.11	**82.10**	82.06
heart-disease-hungarian	80.72	80.55	82.66	**83.27**
heart-statlog	**81.07**	**81.07**	62.74	62.67
hepatitis	79.50	79.49	79.24	**81.06**
hypothyroid	93.33	93.28	**96.19**	96.01
ionosphere	82.32	82.32	**84.52**	80.99
iris	88.20	88.07	**91.80**	91.73
japanese-crx	75.87	75.96	**83.26**	83.14
kr-vs-kp	84.39	**84.40**	84.32	83.46
letter	59.23	59.22	70.17	**70.21**
liver-disorders	53.12	53.06	57.63	**57.75**
lsvt-voice-rehab	50.02	50.02	64.28	**67.63**
lymphography	76.01	**78.38**	75.97	77.20
mfeat-pixel	77.47	92.11	**92.12**	92.06
mol-splice-junction	90.03	90.16	89.89	**90.70**
nursery	90.58	90.59	**90.60**	90.59
optdigits	88.88	88.88	**89.84**	89.69
page-blocks	88.26	88.26	**91.19**	91.01
parkinsons	62.05	61.94	76.48	**76.54**
pendigits	82.64	82.64	83.57	**83.61**
postoperative-patient	56.00	59.67	60.78	**62.00**
primary-tumor	42.04	**45.29**	42.72	43.95
qsar-biodegradation	65.68	65.68	73.66	**75.01**
qualitative-bankruptcy	98.56	98.52	98.48	**98.60**
saheart	**68.97**	**68.97**	64.89	65.51
segment	67.28	67.28	**85.92**	85.87
seismic-bumps	81.54	81.53	**93.19**	93.17
sick	88.24	88.33	96.37	**96.51**
solar-flare2	97.58	97.07	**98.62**	98.54
sonar	**65.35**	**65.35**	60.09	55.21
soybean	**90.18**	**90.18**	86.91	87.41
spambase	72.50	72.50	88.31	**88.40**
spect	70.37	70.18	69.40	**80.32**
spectf	52.84	52.84	69.53	**72.98**
spectrometer	**32.52**	32.34	16.01	17.17
splice	90.03	90.16	89.89	**90.70**
sponge	68.39	64.16	39.32	**80.84**
tae	**45.43**	45.10	42.98	40.53
thoracic-surgery	72.43	73.34	81.68	**83.68**
tic-tac-toe	67.95	67.91	68.11	**68.69**
turkiye-student	22.91	22.91	**35.19**	**35.19**
vehicle	43.44	43.44	**53.67**	53.58
vote	**89.40**	**89.40**	89.38	89.36
vowel	47.56	**49.03**	38.70	37.77
waveform	76.73	76.73	77.90	**78.10**
wine	**93.14**	**93.14**	92.58	**93.14**
zoo	87.44	90.42	**90.60**	88.55
Average	71.08	72.19	73.54	**74.64**