Naïve Bayes Classifier Models for Predicting the Colon Cancer

Naïve Bayes Classifier Models for Predicting the Colon Cancer

Nafizatus Salmi[1] and Zuherman Rustam[1*]

[1]Department of Mathematics, University of Indonesia, Depok 16424, Indonesia

*Corresponding author email: rustam@ui.ac.id

Abstract. Cancer has been known as a disease consisting of several different types. Cancer is a life threatening disease in the world today. There are so many types of cancer in the world, one of which is colon cancer. Colon cancer is one of the number one killers in the world. However, because there isn't any obvious symptom of colon cancer at an early stage, people do not realize that they suffer from it. Even though cancer formation is different for each type of cancer, it is still a big challenge to make cancer classification with good accuracy. Many machine learning has been applied to the data of human's genes in order to get the most relevant genes in the classification of cancer. The author proposes the Naïve Bayes Classifier model as a classification method to show that the model has good accuracy, good precision, good recall, good $f_1 - score$ in classifying the data of patients suffering from colon cancer or not. In this proposed model, Naïve Bayes Classifier is a technique prediction based on simple probabilistic and on the application of the Bayes theorem (or Bayes rule) with a strong independence assumption. Therefore, this model is able to make higher classification accuracy with less complexity. In particular, it achieves up to 95.24% classification accuracy, thus this model can be an efficient analysis tool.

1. Introduction

In many countries, cancer ranks as the second most common cause of death after cardiovascular disease [1]. The number of cancer patients worldwide continues to increase significantly. The latest report released by the International Agency for Research on Cancer, the World Health Organization (WHO) estimates that there are 18.1 million new cancer cases and 9.6 million deaths that occurred in 2018.

About 70% of cancer deaths occur in low and middle income countries. About one third of cancer deaths are caused by five behavioral that is lack of physical activity, high body mass index, alcohol use, low fruit and vegetable intake, and tobacco use [2]. About 22% of tobacco users suffer from cancer and cause death, tobacco use is a factor that has the highest risk of cancer [3].

Cancer is a disease characterized by uncontrolled cell growth. Cancer can start anywhere in the body. Cancer cells can appear in one area, then spread to other areas in the body. Cancers are similar in some ways, but different in the way they grow and spread. This begins when cells grow out of control and the normal cell crowd. When these cells begin to grow in the body, the body will find it difficult to work normally. Cancer can occur in the breast, cervix, large intestine, lungs or even in the blood. Some types of cancer cause mashed cells to divide at a slower rate while others cause rapid cell growth [4].

There are so many types of cancer in the world, one of which is colon cancer, colon cancer is one of the number one killers in the world. However, this cancer is often not realized by patients because there are no obvious symptoms of colon cancer at an early stage, people do not realize that they suffer from it. Colon Cancer is one of the most common malignant tumor with high incidence rates in the age range of 40-50 and is a very serious threat to life and human health. In 2017, nearly 136,000 new cases of colon cancer were diagnosed. About 1 in 20 people will develop colon cancer during their lifetime [5].

Colon cancer is a type of cancer that attacks the large intestine or the last part of the human digestive system. Colon cancer occurs when there are genetic mutations, where DNA cells in certain areas of the body grow uncontrollably and are destructive. Most cases of colon cancer begin with the formation of small cell clots called adenoma polyps that are not cancerous. These clots then spread uncontrollably over time.

In colon cancer, abnormal growth of these cells begins in the lining of the inner intestine, then spreads and destroys other cells nearby, or even to several other areas of the body. Genetic mutations in colon cancer are of a dual nature. It means, someone who has a family member with colon cancer will be more at risk for suffering from this disease.

The exact cause of this cancer is unknown. However, understanding of genetic causes continues to increase. There are several factors that cause a person to develop colon cancer, such as bowel habits tend to change, experience constipation or diarrhea, and change the consistency of stool that lasts more than one month, abdominal discomfort such as cramps or pain, feeling that the intestine is not completely empty, weakness or fatigue, and weight loss that cannot be explained [6].

There are several stages to determine the severity of a person suffering from colon cancer. In stage 1, the cancer has not spread because it is still blocked by the intestinal wall, but the cancer has begun to grow in the large intestine. In stage 2, the cancer will spread throughout the large intestine wall and even penetrate the wall. In stage 3, the lymph nodes that are located adjacent to the large intestine have been eaten away by cancer. And in the final stage, cancer has spread further and attacks other organs, this stage is the most severe level of the spread of colon cancer.

We can prevent colon cancer by applying a healthy lifestyle. Like regular exercise, eating healthy foods, maintaining weight, stopping smoking and reducing or avoiding alcoholic beverages. The treatment can be done by chemotherapy, radiotherapy, and surgery. The chance to recover from the sufferer will depend on how severe the cancer has spread at the time of diagnosis.

In previous studies, colon cancer was used in *Classification tree analysis* [7], *Locality Sensitive Deep Learning*[8], *evolutionary neural network*[9] and *mechanism of tsoong inhibiting* [10]. In addition to colon cancer, brain cancer has also been applied in machine learning. The American Brain Tumor Association states that there are more than 120 types of brain cancer observed. Brain cancer has been the cause of cancer-related deaths for people under the age of 40 [11].

Classification is one of the learning processes where each data is labeled. This classification learning model was built using a collection of training data and a collection of testing data with the aim of predicting new classes by studying old class categories and labels [12]. In predicting a class, there are many data sets that will be used.

At the moment there is a lot of machine learning for data classification, such as disease data, stock data, or other data. Here the authors propose a classification method, namely the Naïve Bayes model in predicting the class of data of patients suffering from colon cancer or not. Naïve Bayes Classifier is a technique prediction based on simple probabilistic and on the application of the Bayes theorem (or Bayes rule) with a strong independence assumption. In this case, it is assumed that the absence or presence of a particular event from a group is not related or has nothing to do with the absence or presence of other events.

Naïve Bayes Classifiers have been used in multi-label learning, where training data sets consist of several instances where each label is linked between one label and another, and the task is to predict a collection of labels from invisible instances [13]. In the paper to overcome the problem of learning used a method called MLNB (Multi-Label Naïve Bayes) that adapts Naïve Bayes classification to handle multi-label instances. This method can handle data with high dimensions.

In addition to multi-label learning, Naïve Bayes has also been used for traffic classification schemes with little training data. By using the proposed scheme, information on can be stored into the classification process. Experiments carried out on the traffic data set show the effectiveness of the proposed scheme. Experimental results show that the method outperforms existing sophisticated methods with large limits [14]. In the paper provides a solution to achieve a classification of traffic with high performance without spending a lot of time.

For Naïve Bayes disease has been used to diagnose the origin of tumor and pancreatic tissue using RNA-Seq data obtained from The Cancer Genome Atlas (TCGA) and 1000 gene features for Bayes algorithm training data [15]. The accuracy in the model proposed in the paper is> 95% based on the validation of the TCGA. Based on the findings of the paper, the TOD-Bayes algorithm is a strong stone methodology for identifying the origin of unknown tumor cells from cancer using RNA-Seq data.

2. Methodology

The methodology used in this paper is the study of literature on Naïve Bayes Classifier.

2.1. Naïve Bayes

Naïve Bayes Classifier is a classification method based on the Bayes theorem. Naïve Bayes Classifier is known to be better than some other classification methods. Because first, the main characteristic of Naïve Bayes is a very strong (naive) assumption of independence from each condition or event. Second, its model is simple and easy to make. Third, the model can be implemented for large data sets. The basis of the Naïve Bayes theorem used is the Bayes formula as follows: (Han, Kamber, & Pei, 2012) [17].

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \tag{1}$$

Where:
X : attributes
C : class
$P(C|X)$: probability of even C given X has occured
$P(X|C)$: probability of even X given C has occurred
$P(C)$: probability of event C
$P(X)$: probability of event X

X can be written as follow:

$$X = (x_1, x_2, x_3, \ldots, x_n) \tag{2}$$

The relationship between C and X can be seen in the picture below:

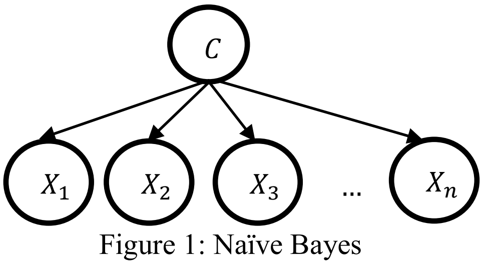

Figure 1: Naïve Bayes

With the substitution of X, the Bayes formula can be written as follows (H. Zhang, 2004) [18]:

$$P(C|x_1, x_2, \ldots, x_n) = \frac{P(C)P(x_1, x_2, \ldots, x_n|C)}{P(x_1, x_2, \ldots, x_n\ )} \tag{3}$$

Furthermore we can elaborate $P(C|x_1, x_2, \dots, x_n)$ such as:

$$\begin{aligned} P(C|x_1, x_2, \dots, x_n) &= P(C)P(x_1, x_2, \dots, x_n|C) \\ &= P(C)P(x_1\ |C)(x_2 \dots x_n\ |C, x_1) \\ &= P(C)P(x_1\ |C)P(x_2\ |C, x_1)(x_3 \dots x_n\ |C, x_1, x_2) \\ &= P(C)P(x_1\ |C)P(x_2\ |C, x_1)P(x_3\ |C, x_1, x_2) \dots P(x_n\ |C, x_1, x_2 \dots, x_{n-1}) \end{aligned} \tag{4}$$

It can be seen that Equation (4) contains complex factors of probability values which is almost impossible to analyze one by one. As a result, the calculation becomes difficult to do. At this point we need the assumption that each of the factor in Equation (4) is free from each other. With this assumption $P(C|x_1, x_2, \dots, x_n)$ can be written as follows:

$$P(C|x_1, x_2, \dots, x_n) = P(C) \prod_{i=1}^{n} P(\ x_i|C) \tag{5}$$

In Naïve Bayes Classifiers we need to maximize the probability value of each class, which is expressed as the Hypothesis Maximum aa Posteriori (HMAP):

$$H_{MAP} = arg\ max\ P(C\ |x_1, x_2, \dots, x_n) = arg\ max\ P(C) \prod_{i=1}^{n} P(\ x_i|C) \tag{6}$$

In Naïve Bayes Classifier, by means of Equation (6) we can predict which classes can be used in Naïve Bayes Model [16]. But, if the attribute X in Equation (6) has quantitative types then the probability will be very small such that the value $P(X|C)$ cannot be used to find the value of H_{MAP}. So we need to use other approach such as normal (Gaussian) distribution that is (Rohith, 2018) [19].

$$P = \left(X_i = x_i \middle| C = c_j\right) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} exp\left(-\frac{(X_i - \mu_{ij})^2}{2\sigma_{ij}^2}\right) \tag{7}$$

Where:
P : opportunity
X_i : the i^{th} attribute
x_i : the i^{th} attribute value
C : class
C_i : the i^{th} sub class
μ : the mean of all attributes
σ : standard deviation

2.2. *Confusion Matrix*

Confusion matrix is used to measure the performance of a classification algorithm. The terminology related to the confusion matrix can be rather confusing, but the matrix itself is simple to understand (See Table 1).

Table 1. Confusion Matrix

	Actual	
Predicted	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

Note that 'True' or 'False' in table 1 indicates if the class is correctly predicted or not, while 'Positive' or 'Negative' indicates the prediction of the class of people infected with cancer or not. From the confusion matrix, we can find accuracy, precision, recall and f_1 - score. Where the formula of each of these things are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$f_1 - score = 2 \cdot [\frac{precision \cdot recall}{precision + recall}]$$

3. Experiment

The following data from Al-Islam Hospital Bandung Indonesia can be used to establish Naïve Bayes Classifier model (See Table 2).

Table 2. Colon Cancer Data

Age	CEA	Haemoglobin (gram/dL)	Leukocytes (cell/mm^3)	Haematocrit (mL/L)	Thrombocyte (cell/mm^3)	Class
74	3.26	11.8	19400	37.3	341000	0
84	29.12	8	12400	26.6	465000	1
81	4.5	8.8	19900	26.2	468000	0
56	0.96	13.9	9400	41.5	260000	0
75	3.24	7.7	13500	22.5	377000	0
58	0.71	11	18200	34	259000	0
63	1.65	10.1	19900	32.1	151000	0
73	36.49	11.1	9700	33.4	267000	1
72	1.36	11.8	7600	36.7	582000	0
75	0.92	16.6	10100	49.3	285000	0
…	…	…	…	…	…	…
64	19.71	11.6	9000	34.7	529000	1
45	3158.73	10.1	20200	29.5	484000	1
76	0.96	12	12600	36.1	468000	0
73	1.61	12.4	8700	37.5	286000	0
75	1.14	10.1	16800	31.6	492000	0
76	0.5	12.8	7600	39.8	229000	0
52	0.75	13.6	9700	39.7	327000	0
72	3.55	10	7700	3.2	589000	0
36	3.72	7.5	9900	23.6	446000	0
45	4.93	14.1	15100	43.9	25200	0

This data consists of 7 columns and 209 lines. Each column, in succession, indicates age, CEA (Carcinoembryonic Antigen), Haemoglobin, Leukocytes, Haematocrit, Thrombocytes, and class are patients infected with cancer or not, where 1 indicates that person has infected with cancer and 0 indicates that the person has not been infected.

From the data on table 2, we define X as predictor those are data from column 1 up to column 6, and we define y as independent variable that is data from column 7. Then we split the data into training data and testing data. After that we input the data into the Naïve Bayes Classifier model, and make prediction and with confusion matrix we get the accuracy of the prediction.

4. Result and Discussion

By applying the algorithm Naïve Bayes discussed above and using python we can produce good accuracy (See Table 3).

Table 3. Accuracy of Naïve Bayes Classifier

% Data Training	% Accuracy
10	91.01
20	90.48
30	90.48
40	91.27
50	90.48
60	92.86
70	93.65
80	95.24
90	90.47

After several trials and some training data, we can see from table 3 that the best accuracy is 95.24% with 80% training data.

The performance of the Naïve Bayes classification system can be searched using confusion matrix (See Table 4).

Table 4. Confusion Matrix of Naïve Bayes Classifier with 80% Training Data

		Actual	
		Positive	Negative
Predicted	Positive	78.57	0
	Negative	4.76	16.67

From table 4, it can be seen that out of the 83.33% actual instances of positive (first column), the classifier predicted that 78.57% of them positive with cancer and 4.76% negative with cancer. Out of the 16.67% actual instances of negative (second column), the classifier predicted that 16.67% of them negative with cancer and 0% positive with cancer.

From the results of the confusion matrix, we can calculate other parameters such as accuracy, precision, recall, and f_1 - score (See Table 5).

Table 5. Matrix classification of Naïve Bayes Classifier with 80% Training Data

Accuracy	Precision	Recall	$f_1 - score$
0.95	1.00	0.94	0.96

Accuracy on table 5 shows how accurate is the Naïve Bayes Classifier in predicting whether the person infected with cancer or nor compared with actual data. On table 2, column 7 shows that are 2 classes of data, in this case the accuracy is about 95%. Precision shows how well is the performance of Naïve Bayes Classifier related to a specific class, if the precision is high is means the predicting is better. In this case the precision is 100%. Recall is a measure that shows how well is a classifier can be used to classify data into some specific classes based on actual positive data. Table 5 shows that the value of recall is 94%, that means the Naïve Bayes Classifier is good in predicting specific classes in actual positive data. F1-score is the harmonic mean from precision and recall. In this case the F1-score is 96%.

5. Conclusion

From the experiment conducted in part 4, we can see that when the training data is about 80% then the accuracy achieved is about 95.24%. From table 3, we can see that the accuracy is always high with its veracity is about 90% to 96%. The method used in this paper achieved result in classifying the data of patients suffering from colon cancer with high accuracy. The weakness of the Naïve Bayes Classifier is that the assumption of independences between attributes reduce the accuracy, because there is some part of the data where the attributes are related to each other.

Acknowledgment

This research was financially supported by University of Indonesia, with PIT.9 2019 research grant scheme (ID number NKB-0039/UN2.R3.1/HKP.05.00/2019).

References

[1] Xiaomei Ma, Herbert Yu, 2007. Global Burden of Cancer. NCBI [Internet]. [accesses on 2019]. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1994799/

[2] World Health Organization. 2018. Cancer. [Internet]. [accessed on 2019]. Available at https://www.who.int/news-room/fact-sheets/detail/cancer

[3] GBD 2015 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016 Oct; 388 (10053):1659-1724.

[4] Rachel Nall, 2018. What to know about cancer. Medical News Today [Internet]. [accessed on 2019]. Available at https://www.medicalnewstoday.com/articles/323648.php

[5] American Society of Colon and Rectal Surgeons. Colon Cancer. [Internet]. [accessed on 2019]. Available at https://www.fascrs.org/patients/disease-condition/colon-cancer

[6] Mayo Clinic. Colon Cancer. [Internet]. [accessed on 2019]. Available at https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669

[7] Camp, Nicola J. Slattery, Martha L. (2001). Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). Cancer Causes & Control; 13, 9; ProQuest pg. 813

[8] Sirinukunwattana, Korsuk. Raza, Shan E Ahmed. Tsang, Yee-Wash. Snead, David R J. Cree, Ian A. Rajpoot, Nasir M. (2016). Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. IEEE Transactions ON Medical Imaging Vol 35, No. 5.

[9] Kim, Kyung-Joong. Cho, Sung-Bae. (2004) . Prediction of colon cancer using an evolutionary neural network. Neurocomputing 61 361-379.

[10] Cao, Yong. Dai, Fei. Li, Yanmei. Jia, Lu. Luan, Yunpeng. Zhao, Youjie. (2018). The research on the mechanism of Tsoong inhibiting for colon cancer, (Saudi Journal of Bilogical Sciences).

[11] Panca, V. and Rustam, Z. (2017). Application of Machine Learning on Brain Cancer Multiclass Classification. AIP Conference Proceedings 1862, 030122.

[12] Nadira, T. and Rustam, Z. (2018). Classiification of Cancer Data Using Support Vector Machines with Features Selection Method Based on Global Artificial Bee Colony. AIP Conference Proceedings 2023, 020205.

[13] Zhang, Min-Ling. Peña, José M. Robles, Victor. (2009). Feature selection for multi-label naïve Bayes classification, Information Science 179 3218-3229

[14] Zhang, Jun. Chen, Chao. Xiang, Yang. Zhou, Wanlei. Xiang, Yong. (2013). Internet Trafiic Classification by Aggregating Correlated Naïve Bayes Predictions. IEEE Transactions On Information Forensics And Security, Vol 8, No. 1.

[15] Jaing,. Weiqin. (2015). A Naïve Bayes algorithm for tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system. International journal of cancer, Vol. 142, Issue. 2, Page 357-368

[16] Taheri, Sona. Mammadov, Musa. (2013). Learning The Naïve Bayes Classifier with Optimization Models. Int. J. Appl. Math. Comput. Sco., Vol. 23, No. 4, 787-795

[17] Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques (3rd ed). USA: Elsevier Inc.

[18] Zhang, Harry. (2004). The Optimality of Naïve Bayes. American Association for Artificial Intelligence.

[19] Gandhi, Rohith. (2018). Naïve Bayes Classifier. Towards Data Science. [Internet]. [accesses on 2019]. Available at https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c