Assignment 3: a mathematical essay on naive bayes classifier

Krishna Sah Teli
Cyber-Physical Systems (I2MP)
Indian Institute of Technology
Madras - 600036, India
ge22m018@smail.iitm.ac.in

***Abstract*—In this paper, naive bayes has been explored and different plots with the help of visualizing tools such as matplotlib, and seaborn, have been generated to depict the correlation among the different variables as well as to extract the insight from the data. It illustrates how the income of people is related to their education level, race, marital status, age, sex, native country, etc. Interestingly, it has been found out that about 95% workers are from United State and out of which white race workers are dominant with 85.97%. It also illustrates that the ratio of male and female workers is double. However, education level plays vital role in getting wages more than 50K as the workers holding degree of bachelor have equal chance to getting $>$ 50K as well as $<=$ 50K wages and holding higher degree have about 80% chance of getting wages more than 50K. Finally, a classification model based on naive bayes has been built and has been analysed its accuracy, precision and over-fitting. As I went through this paper, I found its overall flow and content is meaningful. In this revised paper, I have made improvement in Introduction, and Dataset and analytics section.**

***Index Terms*—Naive bayes, Matplotlib, Seaborn, Classification Model, Over-fitting**

I. Introduction

Everyone has their own skills regarding selection of best out of number of choices. It depends on the experiences, knowledge of domain, and other related parameters. Generally, We do correlate the related parameters along with knowledge and experiences connected to the target and finally select the best as per our knowledge. Basically, we start classify the data into different categories to make it easy to analyse and understand. More priority is given to those parameters which are more related or close to the target as well as we do consider the parameters having high negative influences. Similarly, to help in making decision, different algorithms have been developed such as decision tree classifier, SVM classifier, KNN-classifier, etc. The classification algorithms have changed our ability of making decision and one can easily come up with better decision within short period of time provided related data and information. It has brought revolutionary changes in medical field by helping in classifying the different stages, kinds of diseases, sickness, etc. [1], [6], [11] as well as in recommending the medicine and checkup [16], [17]. It has great application in weather estimation, spam and intrusion detection, sentiment analysis [9], analyzing the performance and match [14], etc.

The main focus in this paper is to demonstrate how the various parameters influence the income of workers, by plotting different graphs followed by visualization and establishing correlation with each others. It also shows ground truth such as working environment is dominated by white race and male workers dominant as 85.97% of the workers are white people and male is more than double of female population.

Naive bayes has been an easy classification technique which discriminates the object based on certain features or property. It builds a probabilistic model which relates the dependent and non-dependent variables and predict or classify the target. It has been used widely in various fields such as e-commerce, IT [9], banking, scientific studies, forecasting, biology [10], etc. The same model has been used for carrying out the classification task and predicting whether one can make income more than 50 thousands or not.

The rest of the paper has been organized as follows: section II describes the data set related to our problem, section III explains the mathematics behind the naive bayes classifier and tools used for developing our model. section IV Modelling and section V concludes the paper with key contributions made and directions for future works.

II. Data set

The source of data which has been used for visualization, analysis and developing classifier model is kaggle which is famous for the data science and storage of data. This data has totally 15 columns namely age, workclass, fnlwgt, education, education_num, marital_status, occupation, relationship, race, sex, capital_gain, capital_loss, hours_per_week, native_country, income. Out of 15 data columns, only 9 columns are selected for further preprocessing and building a classification model. The 9 important data columns are workclass, sex, age, race, relationship, education, occupation, hours_per_week and income. Here "income" column is the target variable (dependent features) which is comprised of two classes i.e $<=$ 50K and $>$ 50K and whereas the other remaining variables are taken as independent variables which help classifier to make decision by providing evidences.

With the data analysis and data visualization, the correlation among the data features has been studied and correlated with the income of workers. It has been found out that income has positive correlation with hours_per_week variable, education,

race, and relationship whereas has negative correlation with sex, marital_status and workclass variables. As we can see in fig.1, the number of workers who work at wages less than 50 thousand is exponentially higher than whom work for more than 50 thousands. There can be different reasons for working at less wages such as lacking of higher education level, less working hours, etc. As we can see in fig. 2 more number of workers work in private companies. so, it is obvious that they will get salary comparatively less despite of long working hours.

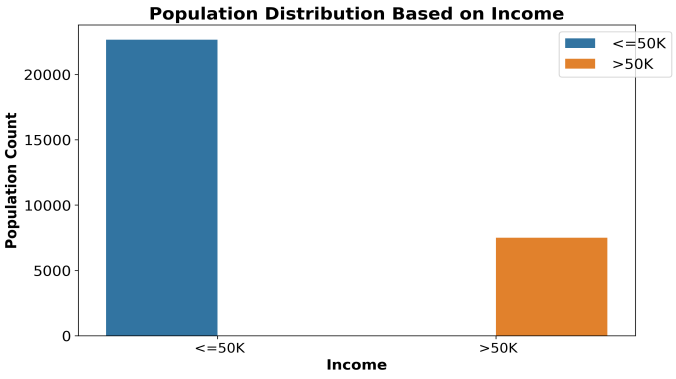

Fig. 1. Count plot showing workers count against income

Similarly, from the fig.2 we can analyze that the most of the workers work in private industries. It may be because government jobs are limited and more competent. It may be that salary is not satisfactory. There are tremendously increasing number of startup companies and they recruit more number of workers.

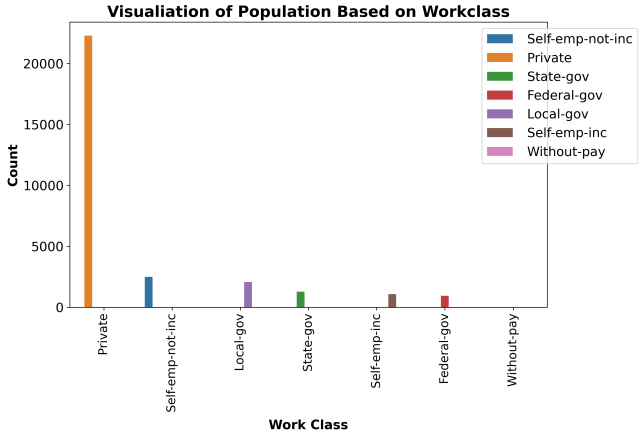

Fig. 2. Plot showing the population density based on workclass

Interestingly, if we see in fig.3, it says that private companies have provided more jobs but about 70% of the workers wages are less than 50 thousands. However, fig. 4 illustrates that the child workers (under the 20 age group) are paid less than 50 thousands for sure. It may be because they are not capable for doing heavy work. However, they are given job but paid less. It also depicts that the density of population working for less than 50 thousands wages increases upto the age of 25 and start decreasing. It may be because some of them may be doing part-time job along with higher studies and completion of studies, they transfer into high paid job. Thus the population density for more than 50 thousands wages is increasing for 20 and above age group.

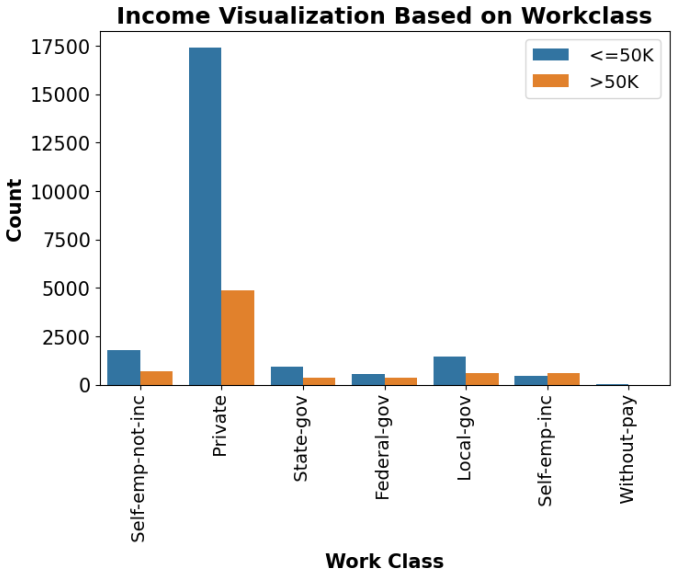

Fig. 3. Plot showing count of population based on workclass and income

similarly, if we talk about education then most of employee hold higher secondary degree followed by college degree, master degree. Interestingly, as the education level raises, the probability of getting high paid salary also increases as shown in fig. 5.

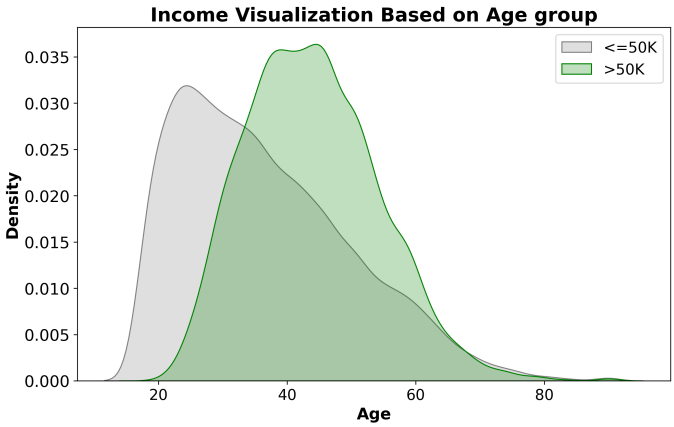

Fig. 4. Count plot for population based on ages and income

Interestingly, the workers population density based on sex is really self speaking as the number of male workers are almost double to the female workers. Similarly the number of male workers for less wages are also higher than high paid as their count is more. However, if we see in fig. 6 which tells about the white race people dominance in the working environment.

White race workers hold about 95 % job opportunities. If we see it country-wise then about 85% of them are from United State and remaining are from rest of the world. It may be because United States is the hub of technical company and thereby creates more number of jobs.

Similarly, if we count the workers based on their working occupation then we will find most of them are working in prof-specialty followed by craft-repair and executive managerial, adm-clerical and sales. These are famous occupation where about more than 50% of workers work.

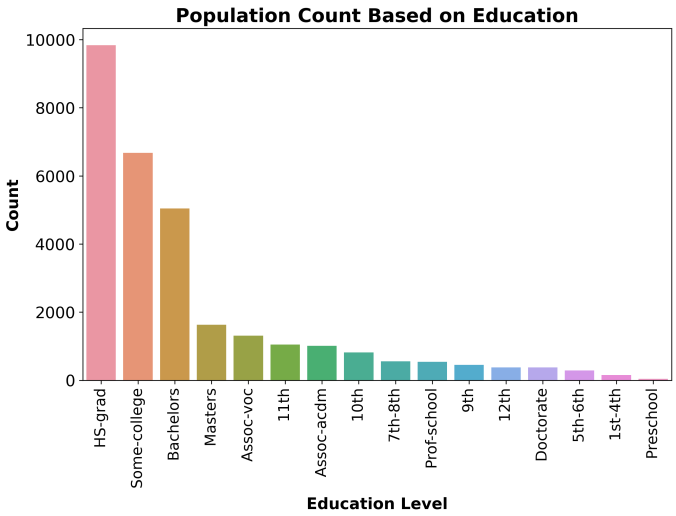

Fig. 5. Plot showing density of population based on education level

If we considering the marital status then the couple (ie. married) workers population is more and for them there is equal probability of working for less or more than 50 thousands wages where as others do not bother about the salary because they just want to some money to survive.

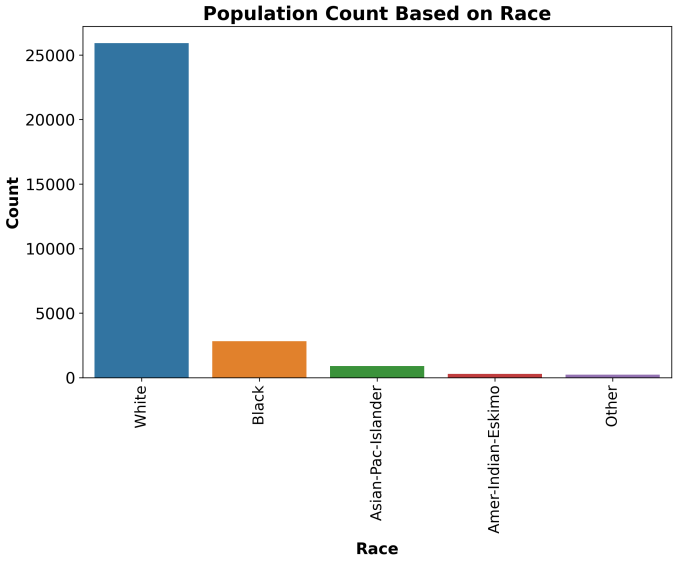

Fig. 6. Plot showing density of population based on race

Similarly, working hours per week has great impact on the salary as shown in the fig.7 which illustrates that the average working hours per week is about 45 hours with less flexibility for the workers getting more than 50 thousand salary whereas the there is enough flexibility with average 45 hours in a week as working hours for workers who are paid less than 50 thousand wages.

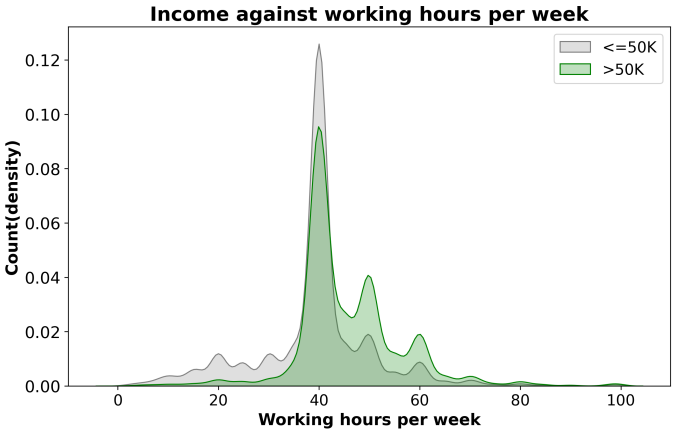

Fig. 7. Distribution of population based working hours per week

III. Naive Bayes

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class.

Bayesian classifier is based on Bayes' theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered "naive".

A. Bayes's Theorem

Let $X = x1, x2, ..., xn$ be a sample, whose components represent values made on a set of n attributes. In Bayesian terms, X is considered "evidence". Let H be some hypothesis, such as that the data X belongs to a specific class C. For classification problems, our goal is to determine $P(H|X)$, the probability that the hypothesis H holds given the "evidence", (i.e. the observed data sample X). In other words, we are looking for the probability that sample X belongs to class C, given that we know the attribute description of X.

According to Bayes' theorem, the probability that we want to compute $P(H|X)$ can be expressed in terms of probabilities $P(H)$, $P(X|H)$, and $P(X)$ as

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

and these probabilities may be estimated from the given data.

B. Naive Bayesian Classifier

The naive Bayesian classifier works as follows: 1. Let T be a training set of samples, each with their class labels. There are K classes, $C1, C2, ..., Ck$. Each sample is represented by an n-dimensional vector, $X = x1, x2, ..., xn$, depicting n measured values of the n attributes, $A1, A2, ..., An$, respectively. 2. Given a sample X, the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X. That is X is predicted to belong to the class Ci if and only if

$P(Ci|X) > P(Cj|X) for 1jm, j6 = i$

Thus we find the class that maximizes $P(Ci|X)$. The class Ci for which$P(Ci|X)$is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

3. As$P(X)$ is the same for all classes, only $P(X|Ci)P(Ci)$ need be maximized. If the class a priori probabilities, P(Ci), are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C1) = P(C2) = ... = P(Ck)$, and we would therefore maximize $P(X|Ci)$. Otherwise we maximize $P(X|Ci)P(Ci)$. Note that the class a priori probabilities may be estimated by P$(Ci) = freq(Ci, T)/|T|$.

4. Given data sets with many attributes, it would be computationally expensive to compute$P(X|Ci)$. In order to reduce computation in evaluating $P(X|Ci)P(Ci)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample. Mathematically this means that

$$P(Ci|X) = \prod_{i=1}^{n} P(xk|Ci)$$

The probabilities $P(x1|Ci), P(x2|Ci), ..., P(xn|Ci)$ can easily be estimated from the training set. Recall that here xk refers to the value of attribute Ak for sample X.

(a) If $Ak$is categorical, then $P(xk|Ci)$ is the number of samples of class Ci in T having the value xk for attribute Ak, divided by $freq(Ci, T)$, the number of sample of class Ci in$T$.

(b) If $Ak$is continuous-valued, then we typically assume that the values have a Gaussian distribution with a mean and standard deviation . so that $p(xk|Ci) = g(xk, Ci, Ci)$. We need to compute $Ci$and Ci, which are the mean and standard deviation of values of attribute Ak for training samples of class Ci.

5. In order to predict the class label of $X, P(X|Ci)P(Ci)$is evaluated for each class Ci. The classifier predicts that the class label of$X$ is Ci if and only if it is the class that maximizes $P(X|Ci)P(Ci)$.

IV. Modelling

Naive Bayesian Classifier is based on conditional probability. It predicts the output based on the given input. It uses a conditional probability based on Baye's theorem to classify outcome into discrete i.e 0 and 1. In our model, it relates the data features with some weight and finds the probability of income. If the probability given input is more than 0.5 then income $>$ 50K otherwise it is $<=$ 50K. Since sex has more correlation with income features, hence sex has more weight while finding probability. To build a Naive bayes model, from sklearn regression model is imported and GaussianNB as solver is used with l1 penality. For evaluating the model, accuracy_error and mean_absolute_error is used. To develop a perform report, classification_report is used. For this model which has been trained using titanic data set, its accuracy is 81.38 % followed by recall_score 77.27% and precision 71.8%.

Acknowledgement

It is my pleasure to express with deep sense of gratitude to Professor Gaurav Raina, for his guidance, continual encouragement, understanding.

I would like to extend my gratitude to IIT Madras, for providing with an environment to work in and for his inspiration during the tenure of the course.

Conclusion

In this paper, Naive Bayes has been used to get insight from the given data set. Data analysis and visualization have been applied in order to find out the correlation among the variables and cause-effect. It has been depicted from the dataset is that more male population is dominant in the working environment than female population. It is more frequent in the population under the 20 years old are obvious given less salary and education plays major role to improve the salary of one's. So in order to boost the econcomy of the country then government has to improve the illiteracy rate and provide more jobs for the people. This paper also guides future data analytic enthusiastic to get insight in naive bayes classification methodology.

References

[1] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," J Inf Sci, vol. 44, no. 1, 2018, doi: 10.1177/0165551516677946.

[2] S. Taheri and M. Mammadov, "Learning the naive bayes classifier with optimization models," International Journal of Applied Mathematics and Computer Science, vol. 23, no. 4, 2013, doi: 10.2478/amcs-2013-0059.

[3] A. Khajenezhad, M. A. Bashiri, and H. Beigy, "A distributed density estimation algorithm and its application to naive Bayes classification," Appl Soft Comput, vol. 98, 2021, doi: 10.1016/j.asoc.2020.106837.

[4] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," Knowl Based Syst, vol. 192, 2020, doi: 10.1016/j.knosys.2019.105361.

[5] J. Ababneh, "Application of Naïve Bayes, Decision Tree, and K-Nearest Neighbors for Automated Text Classification," Mod Appl Sci, vol. 13, no. 11, 2019, doi: 10.5539/mas.v13n11p31.

[6] M. Abbas, K. Ali Memon, and A. Aleem Jamali, "Multinomial Naive Bayes Classification Model for Sentiment Analysis," IJCSNS International Journal of Computer Science and Network Security, vol. 19, no. 3, 2019.

[7] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes," Information (Switzerland), vol. 12, no. 5, 2021, doi: 10.3390/info12050204.

[8] B. Lakshmi Devi, V. Varaswathi Bai, S. Ramasubbareddy, and K. Govinda, “Sentiment analysis on movie reviews,” in Advances in Intelligent Systems and Computing, 2020, vol. 1054. doi: 10.1007/978-981-15-0135-7_31.

[9] A. J. Meerja, A. Ashu, and A. Rajani Kanth, “Gaussian naïve bayes based intrusion detection system,” in Advances in Intelligent Systems and Computing, 2021, vol. 1182 AISC. doi: 10.1007/978-3-030-49345-5_16.

[10] M. Singh, “Classification of spam email using intelligent water drops algorithm with Naïve bayes classifier,” in Advances in Intelligent Systems and Computing, 2019, vol. 714. doi: 10.1007/978-981-13-0224-4_13.

[11] P. K. Dubey, H. Suri, and S. Gupta, “Naïve Bayes Algorithm Based Match Winner Prediction Model for T20 Cricket,” in Advances in Intelligent Systems and Computing, 2021, vol. 1172. doi: 10.1007/978-981-15-5566-4_38.

[12] A. Ali, A. Khairan, F. Tempola, and A. Fuad, “Application Of Naïve Bayes to Predict the Potential of Rain in Ternate City,” E3S Web of Conferences, vol. 328, 2021, doi: 10.1051/e3sconf/202132804011.

[13] M. Ilić, Z. Srdjević, and B. Srdjević, “Water quality prediction based on Naïve Bayes algorithm,” Water Science and Technology, vol. 85, no. 4, 2022, doi: 10.2166/wst.2022.006.

[14] R. Jayadi, H. M. Firmantyo, M. T. J. Dzaka, M. F. Suaidy, and A. M. Putra, “Employee performance prediction using naïve bayes,” International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 6, 2019, doi: 10.30534/ijatcse/2019/59862019.

[15] Y. Tan, H. Chen, J. Zhang, R. Tang, and P. Liu, “Early Risk Prediction of Diabetes Based on GA-Stacking,” Applied Sciences (Switzerland), vol. 12, no. 2, 2022, doi: 10.3390/app12020632.

[16] S. Josephine Theresa and D. J. Evangeline, “Classification of diabetes milletus using naive bayes algorithm,” in Advances in Intelligent Systems and Computing, 2021, vol. 1167. doi: 10.1007/978-981-15-5285-4_40.

[17] N. A. Mansour, A. I. Saleh, M. Badawy, and H. A. Ali, “Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy,” J Ambient Intell Humaniz Comput, vol. 13, no. 1, 2022, doi: 10.1007/s12652-020-02883-2.

[18] D. Setsirichok et al., “Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening,” Biomed Signal Process Control, vol. 7, no. 2, 2012, doi: 10.1016/j.bspc.2011.03.007.

[19] G. Kaur and A. Oberoi, “Novel Approach for Brain Tumor Detection Based on Naïve Bayes Classification,” in Advances in Intelligent Systems and Computing, 2020, vol. 1042. doi: 10.1007/978-981-32-9949-8_31.

Source code link: https://drive.google.com/drive/folders/10abm1W0_-4Hf0HsIZfYKP9gQbWfOoiJ9?usp=share_link