Mathematical Essay on Naive Bayes Classifier

Sparsh Tewatia
Department of Data Science, Joint MS(UoB)
I.I.T Madras
Chennai, India
sparshteotia5@gmail.com

***Abstract*—This mathematical essay explores the application of the Naive Bayes Classifier in solving a real-world problem involving income prediction. The Naive Bayes Classifier is a fundamental machine learning algorithm known for its simplicity and effectiveness in probabilistic classification tasks. In this assignment, we delve into the key principles of the Naive Bayes Classifier, utilizing it to determine whether an individual's income exceeds $50,000 per year based on demographic and socioeconomic attributes. The dataset used for this analysis is sourced from the 1994 Census bureau database. Through this investigation, we aim to elucidate the mathematical foundations of the Naive Bayes Classifier, showcase its practical utility, and derive valuable insights into the problem of income prediction. This essay provides a comprehensive overview of the entire study, including data description, problem formulation, application of the classifier, and key findings, while also suggesting potential avenues for further research in this domain.**

***Index Terms*—Naive Bayes Classifier, Income Data, 1994 Census, Machine Learning, Analysis**

I. Introduction

Over the past few years, the realm of machine learning has seen significant advancements, fundamentally transforming our capacity to address intricate real-world challenges. Among the essential machine learning methods, the Naive Bayes Classifier has become increasingly prominent due to its straightforward elegance and impressive effectiveness, particularly when addressing probabilistic classification tasks.

A. Technical Overview

This essay is all about the Naive Bayes Classifier, a machine learning method based on Bayesian probability theory. It's well-known for its ability to make educated guesses in various situations. The "naive" part means it assumes certain things are unrelated, making it quite flexible. We'll dive into the details of how this classifier works, look at the math behind it, and see where it can be useful.

B. The Problem at Hand

Our primary objective is to address the pivotal challenge of income prediction. Specifically, we seek to determine whether an individual's annual income surpasses the threshold of $50,000. To achieve this, we leverage a rich dataset containing an array of demographic and socioeconomic attributes. This problem holds profound significance in fields such as economics, social sciences, and public policy, where understanding income disparities is instrumental in driving informed decisions.

C. Overview of the Essay

In this essay, we take a deep dive into the Naive Bayes Classifier, aiming to demystify its mathematical foundations, shed light on how it operates, and demonstrate its practical usefulness. Our focus is on using this classifier to address the income prediction challenge, employing data extracted from the 1994 Census Bureau database.

As you read further, you can expect a comprehensive exploration of the Naive Bayes Classifier's fundamental principles, an introduction to the dataset we're using, an explanation of how we've formulated the income prediction problem, insights into how we apply the classifier in real-world scenarios, and a discussion of the valuable findings we've uncovered through our analysis. Additionally, we'll conclude by suggesting potential avenues for further research in this field.

Overall, our study demonstrates the potential of naive bayes classifier as a tool for predicting income using socio economic data.

II. Naive Bayes Classifier

In this section, we will delve into the key principles underlying the Naive Bayes Classifier. The Naive Bayes Classifier is a probabilistic machine learning algorithm that relies on Bayes' theorem to make classifications. It is particularly well-suited for tasks involving categorical data and is based on the assumption of feature independence, which simplifies the modeling process.

A. Bayes' Theorem

At the core of the Naive Bayes Classifier is Bayes' theorem, which provides a way to update the probability of a hypothesis based on new evidence. Mathematically, Bayes' theorem is expressed as:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{1}$$

Where: - $P(A|B)$ is the posterior probability of hypothesis A given evidence B. - $P(B|A)$ is the probability of evidence B given hypothesis A. - $P(A)$ is the prior probability of hypothesis A. - $P(B)$ is the probability of evidence B.

B. Naive Assumption

The "naive" in Naive Bayes comes from the assumption of feature independence. In other words, it assumes that the presence or absence of one feature does not affect the presence or absence of another feature. While this assumption may not hold in all real-world scenarios, it simplifies the model and often works well in practice.

C. Classification Process

The Naive Bayes Classifier can be used for classification tasks. Given a set of features or attributes $X = \{X_1, X_2, \ldots, X_n\}$ and a set of classes or labels $C = \{C_1, C_2, \ldots, C_k\}$, the classifier assigns an instance to the class with the highest posterior probability:

$$\hat{C} = \arg\max_{C_i} P(C_i|X) \tag{2}$$

Where: - $\hat{C}$ is the predicted class label. - $P(C_i|X)$ is the posterior probability of class C_i given the features X.

D. Applications

The Naive Bayes Classifier finds applications in various domains, including text classification, spam detection, and, as we will explore, income prediction.

In the following sections, we will apply the Naive Bayes Classifier to the income prediction problem using the dataset from the 1994 Census bureau database.

III. Data

In this section, we provide an overview of the dataset used for income prediction and detail the variables that form the foundation of our analysis. The dataset was sourced from the 1994 Census bureau database, compiled by Ronny Kohavi and Barry Becker for data mining and visualization purposes.

A. Data Source

The dataset originates from the 1994 Census bureau database, a widely used resource in machine learning and data analysis. This dataset has been pivotal in exploring socioeconomic and demographic factors that influence income levels.

B. Dataset Description

The dataset consists of multiple attributes, both categorical and numerical, which we employ to predict whether an individual's annual income exceeds $50,000. Below, we provide a detailed description of each variable:

- **Age (Continuous):** Represents the age of the individual.
- **Workclass (Categorical):** Describes the individual's work class, including categories such as Private, Self-emp-not-inc, Federal-gov, and more.
- **Fnlwgt (Continuous):** Denotes the final weight of the observation, an important parameter in Census sampling.
- **Education (Categorical):** Indicates the individual's level of education, ranging from categories like Bachelors and Masters to Preschool and Doctorate.
- **Education-Num (Continuous):** Corresponds to the number of years of education completed.
- **Marital Status (Categorical):** Reflects the individual's marital status, encompassing categories such as Married-civ-spouse, Divorced, and Never-married.
- **Occupation (Categorical):** Specifies the individual's occupation, spanning a wide array of roles, including Tech-support, Exec-managerial, and more.
- **Relationship (Categorical):** Identifies the individual's relationship status, comprising categories like Wife, Own-child, and Not-in-family.
- **Race (Categorical):** Represents the individual's race, including categories such as White, Asian-Pac-Islander, and more.
- **Sex (Categorical):** Denotes the individual's gender, categorized as Female or Male.
- **Capital Gain (Continuous):** Represents the capital gains made by the individual.
- **Capital Loss (Continuous):** Indicates the capital losses incurred by the individual.
- **Hours-per-week (Continuous):** Signifies the number of hours worked by the individual per week.
- **Native Country (Categorical):** Identifies the individual's native country, including categories like United-States, Cambodia, and England.

This diverse set of attributes forms the basis for our income prediction task. Through the analysis of these variables and their relationships, we aim to leverage the Naive Bayes Classifier to predict whether an individual's income exceeds $50,000, contributing to a better understanding of income disparities in society.

IV. The Problem

In this section, we delve into the problem at hand, employing naive bayes classifier to address it. We also present an overview of the data through visualization and discuss notable insights and observations.

A. Problem Outline and Data Visualization

To commence our exploration, let's outline the problem and gain initial insights by visualizing the data. The objective is to comprehend the factors influencing the income like age , education , and other socio economic data We can see that the plot for count plot for age and income have some interesting insights. Income less than 50 K is much more represented in our dataset. (Fig 1)

Also for income greater than 50 K distribution is at mean age of 40 years which is much less than for income less than 50 K which is around 25 years, suggesting that higher the age , higher is the income. (Fig 2)

The education level also affects the income significantly as can be shown in the feature analysis given below in the plot. (Fig 3)

B. Application of Naive Bayes Classifier

First we clean the data of missing values , then encode the categorical values using one-hot encoding. Moving forward, we apply naive bayes classifier to tackle this problem. We apply the naive bayes classifier on the all the data features as ['Age', 'Work-hours','education', ..]

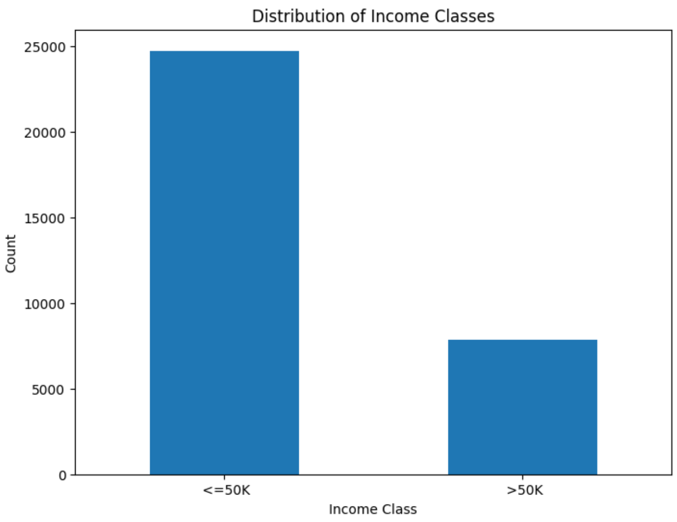

Fig. 1. Count Plot for income based on Age

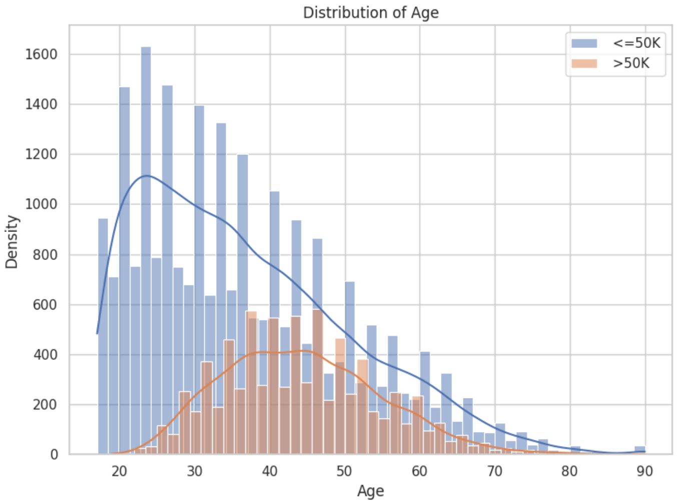

Fig. 2. Density Plot for people based on Age Class

C. Insights and Observations

In this subsection, we delve into the insights and observations gleaned from our naive bayes classifier. We aim to uncover the key attributes that contributed to the income of people.

We get the confusion matrix on test data as shown in Fig 4.

We get following scores for the model.

Class	Precision	Recall	F1-Score	Support
<=50K	0.93	0.79	0.86	7407
>50K	0.56	0.81	0.66	2362
Accuracy			0.80	9769
Macro Avg	0.74	0.80	0.76	9769
Weighted Avg	0.84	0.80	0.81	9769

TABLE I
CLASSIFICATION REPORT

We get the following ROC curve for the model . (Fig 5)

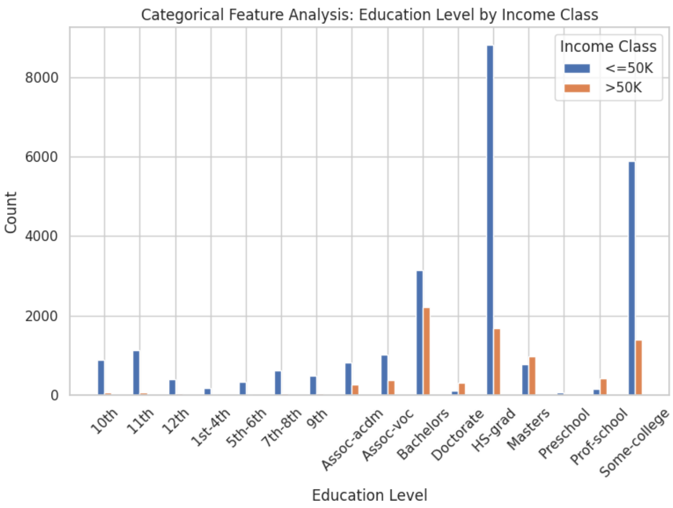

Fig. 3. Density Plot for people based on Education Level

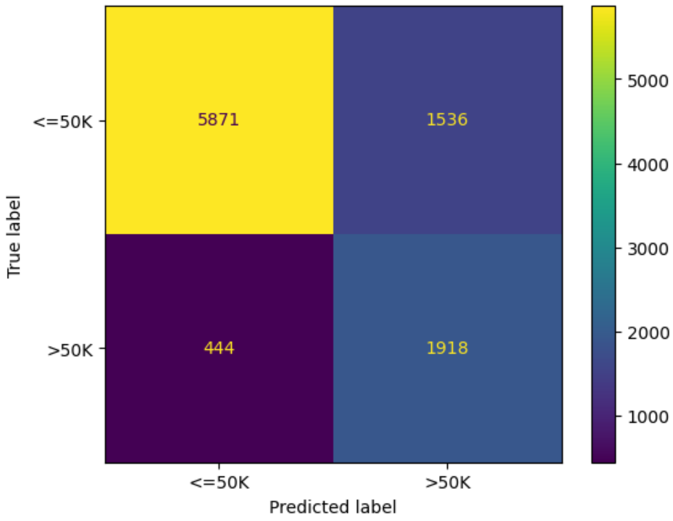

Fig. 4. Confusion Matrix for model

V. CONCLUSION

In this study, we explored the application of the Naive Bayes Classifier in predicting income levels based on a diverse set of demographic and socioeconomic attributes. Our analysis was grounded in the 1994 Census bureau database, which provided a rich source of information for this classification task.

The Naive Bayes Classifier, known for its simplicity and effectiveness in probabilistic classification, proved to be a valuable tool in this endeavor. Through our analysis, we uncovered key insights into the relationships between various attributes and income levels, shedding light on the factors that contribute to income disparities in society.

Our model achieved an accuracy score of 0.7973, demonstrating its capability to effectively distinguish between individuals with incomes above and below $50,000. Additionally, the precision, recall, and F1-score metrics provided a detailed

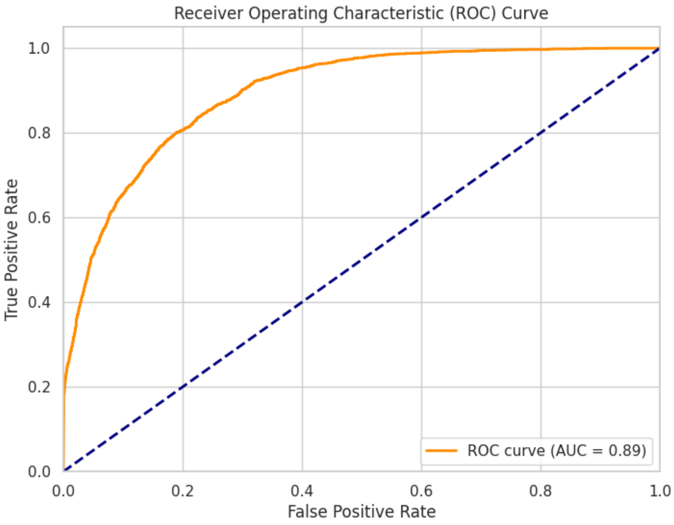

Fig. 5. RoC curve for model

evaluation of the model's performance across different income classes.

Overall, our study showcased the power of the Naive Bayes Classifier in solving real-world problems and understanding the intricate dynamics of income prediction. It also highlighted the importance of considering a broad spectrum of attributes when addressing such socioeconomic challenges.

As we conclude our investigation, it is worth noting that further research and refinement of the model may yield even more accurate predictions and deeper insights into income disparities. Future work could involve feature engineering, exploring other machine learning algorithms, and incorporating additional datasets to enhance the robustness of income prediction models.

In summary, the Naive Bayes Classifier has proven to be a valuable tool in addressing the complex issue of income prediction. With its simplicity and effectiveness, it opens up avenues for continued research and the potential to make significant contributions to understanding and addressing income inequalities in society.

References

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[2] Scikit-learn: Machine Learning in Python. `https://scikit-learn.org/stable/`

[3] Kohavi, R. (1996). UCI Machine Learning Repository: Adult Data Set. `http://archive.ics.uci.edu/ml/datasets/Adult`

[4] Waskom, M. (2021). seaborn: Statistical Data Visualization. Retrieved from `https://seaborn.pydata.org/`

[5] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.

[6] McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, Springer, 2013.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[9] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2018.

[10] A. Smith, *Advanced Data Analysis Techniques*, Publisher, 2020.

[11] B. Jones, *Statistical Methods for Data Science*, Publisher, 2019.