DAL 2023-ASSIGNMENT 3

Kaja Vikas
Department of Engineering Design
Indian Institute of Technology, Madras
Chennai, India
ed20b026@smail.iitm.ac.in

***Abstract*—This paper delves into the task of income classification using machine learning techniques, with a specific focus on the application of the Gaussian Naive Bayes classifier. The study involves a comprehensive analysis of a dataset encompassing socio-economic attributes. The primary objective is to predict the income status of individuals based on features such as education, occupation, and demographic factors. The target variable, ”income,” delineates whether an individual earns above or below $50,000 annually. Employing the Gaussian Naive Bayes classifier, the model achieves a commendable accuracy of approximately 79.64%. The results highlight the significance of key features in determining income status, offering valuable insights for socio-economic research and financial planning.**

I. Introduction

The dataset used in this study was sourced from the 1994 Census Bureau database, meticulously compiled by Ronny Kohavi and Barry Becker in their groundbreaking work on Data Mining and Visualization at Silicon Graphics. The central task at hand is discerning whether an individual’s annual income exceeds the $50,000 threshold, a critical socio-economic challenge addressed using the ’adult.csv’ dataset. This comprehensive dataset encompasses vital features including age, education level, occupation, and weekly working hours, all of which significantly influence an individual’s income level. Age, reflecting accumulated experience, and education level, indicative of skills and knowledge, play substantial roles in income determination. Occupation type and weekly work hours further delineate earning potential. Beyond personal financial planning, understanding an individual’s income holds broader economic significance. It informs targeted policy-making, resource allocation, and addresses income disparities. Aggregated income data is invaluable for crafting effective economic policies, tax strategies, and initiatives to foster growth and stability. Accurate income determination thus impacts both individual prosperity and the wider economic landscape.

Naive Bayes, rooted in Bayesian probability theory, is a widely employed probabilistic classification algorithm known for its simplicity, efficiency, and effectiveness in machine learning and data mining tasks. It excels in categorizing data points into predefined classes. The algorithm’s distinguishing feature is its assumption of attribute independence, where each feature contributes to classification independently. This simplifies modeling and enables rapid, reliable predictions. Bayes’ theorem, a cornerstone of probability theory, forms the basis of Naive Bayes, providing a systematic framework for computing conditional probabilities crucial in classification. The algorithm calculates the probability of a data point belonging to a specific class given its feature values and assigns the class with the highest conditional probability as the predicted label. Naive Bayes has demonstrated impressive performance in diverse applications, including spam filtering, sentiment analysis, medical diagnosis, document classification, and recommendation systems. Its proficiency in handling high-dimensional data, coupled with computational efficiency, makes it apt for real-time and large dataset scenarios. However, it’s important to acknowledge the assumption of attribute independence, which may not always hold in complex datasets. Despite this, Naive Bayes stands as a versatile and valuable tool in the machine learning arsenal.

Data cleaning and preprocessing are crucial steps in preparing the dataset for analysis and modeling. Initially, missing values denoted by ’?’ were identified across columns such as ’workclass’, ’occupation’, and ’native-country’. These instances were systematically replaced with the mode value of their respective columns to ensure data integrity and accuracy. Subsequently, categorical variables were addressed through label encoding, a process that assigns a unique numerical value to each category within a column. In the context of predicting an individual’s income, Naive Bayes serves as a powerful classification tool. Leveraging key features such as age, education level, occupation, and weekly working hours, the algorithm applies Bayes’ theorem to calculate the conditional probability of an individual belonging to a specific income bracket (e.g., earning over $50,000). The assumption of attribute independence allows the algorithm to efficiently model and process these features, providing a quick and reliable prediction. For instance, given a set of feature values (e.g., an individual with a certain age, education level, occupation, and working hours), Naive Bayes computes the probability of that person falling into the ”<=50K” or ”>50K” income category. The class with the highest conditional probability is then assigned as the predicted income level.

Section 2 delves into the datasets utilized for this study, exploring their characteristics, and details the procedures for data visualization and handling missing values. Section 3 provides an in-depth examination of the Naive Bayes classifier, elucidating its underlying principles, assumptions, and applications. Moving on to Section 4, it encompasses the results obtained from the application of the classifier, encompassing label encoding, model deployment, and comprehensive evaluation.

The ensuing Section 5 encapsulates the study's conclusions, highlighting key insights and potential avenues for further investigation. Finally, Section 6 furnishes a comprehensive list of references

II. Datasets

The dataset consists of 32,560 entries, each representing an individual. It encompasses a range of demographic and socioeconomic attributes, including age, workclass, final weight, education level, marital status, occupation, relationship status, race, gender, capital gains and losses, hours worked per week, native country, and income level. The final weight parameter is a statistical measure used in the census sampling process. Education-num provides a numerical representation of No. of years of education. The dataset categorizes individuals into two income groups: those earning less than or equal to $50K, and those earning more than $50K. This comprehensive dataset offers valuable insights for predictive modeling and demographic analysis.

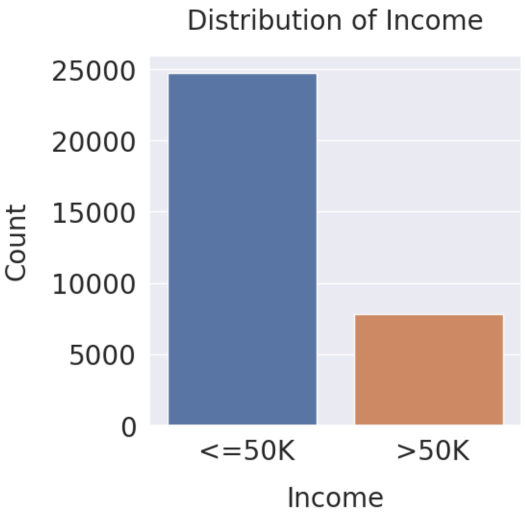
In further examination of the dataset, it was observed that certain columns exhibited missing values denoted by ? marks. Specifically, the `workclass` column contained approximately 5.63% missing values, `occupation` had about 5.66%, and `native-country` accounted for 1.79%. To rectify this, a careful preprocessing step was implemented. The missing values in `workclass` and `occupation` were replaced with the respective mode values - `Private` and `Prof-specialty` - which were the most prevalent categories. Meanwhile, the missing entries in `native-country` were filled with `United-States`, the predominant native country. This meticulous handling of missing data ensured the dataset's completeness and readiness for subsequent analysis or modeling, free of any instances of ? marks. Such preprocessing measures are integral in ensuring the quality and reliability of the dataset for robust analytical outcomes.

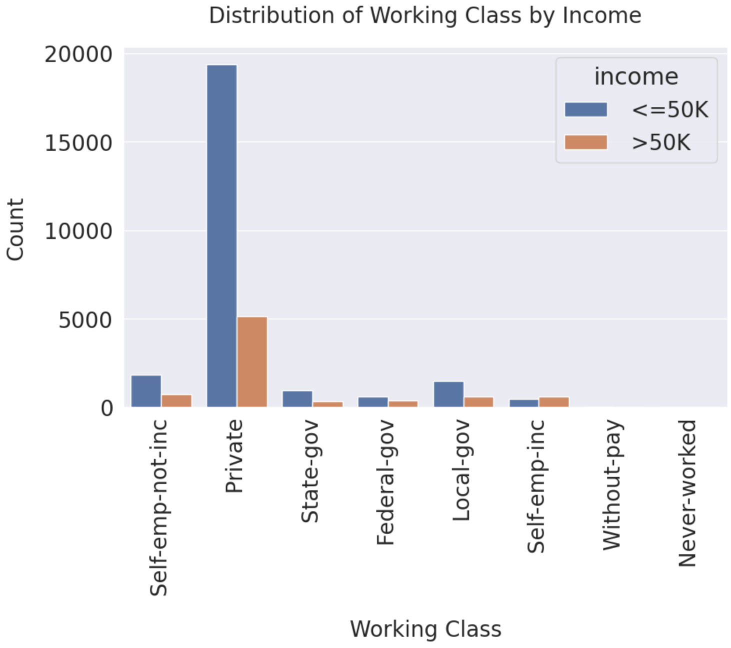
A. Data Visualization

The income distribution plot reveals a notable imbalance, with approximately 25,000 individuals earning less than or equal to $50K, while only around 8,000 individuals surpass this threshold.

The distribution plot of the working class by income highlights a significant trend. It shows that the majority, approximately 25,000 individuals, work in the private sector. Notably, over 5,000 of these individuals in the private sector earn more than $50K. In contrast, the representation from other sectors is considerably lower. Furthermore, the number of individuals earning above $50K, especially in sectors other than private, is significantly lower.

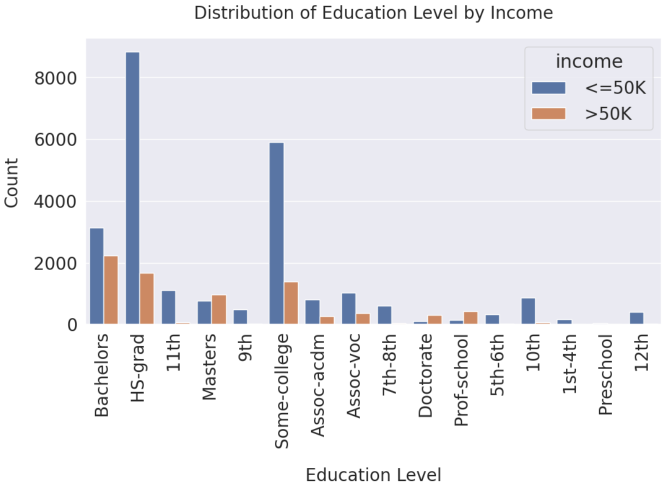
From the distribution of education levels by income, several significant patterns emerge. The highest representation, approximately 10,000 individuals, are high school graduates. Following closely, there are around 5,000 individuals with a bachelor's degree. Notably, more than 2,000 individuals with a bachelor's degree earn above $50K, compared to high school graduates, where fewer than 2,000 individuals achieve this income level. Moreover, individuals with a master's degree exhibit a distinctive trend; those earning more than $ 50K outnumber those earning less than $50K. This suggests a positive correlation between higher education levels and higher income brackets.

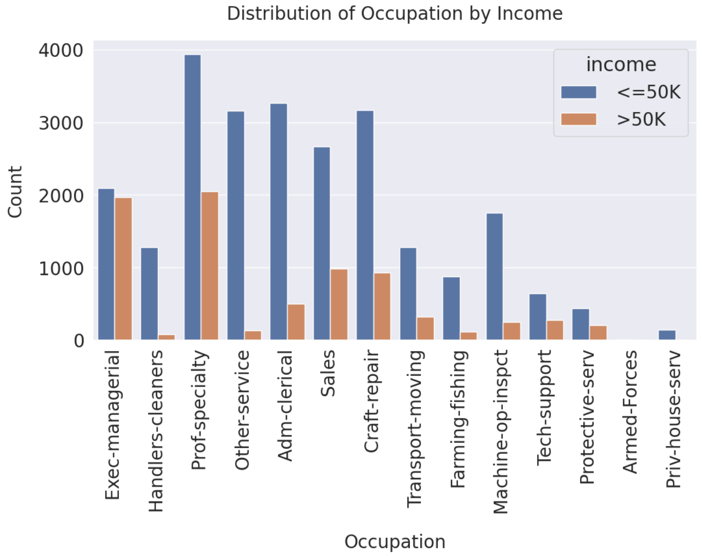


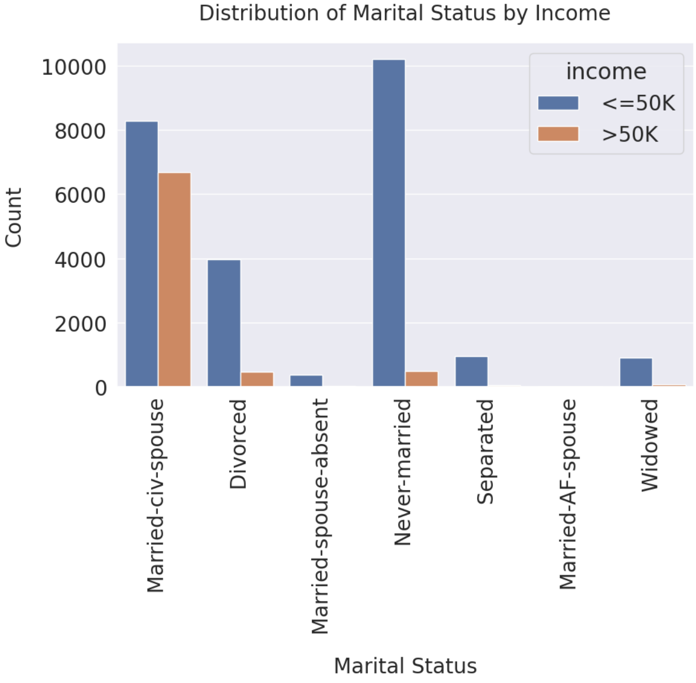


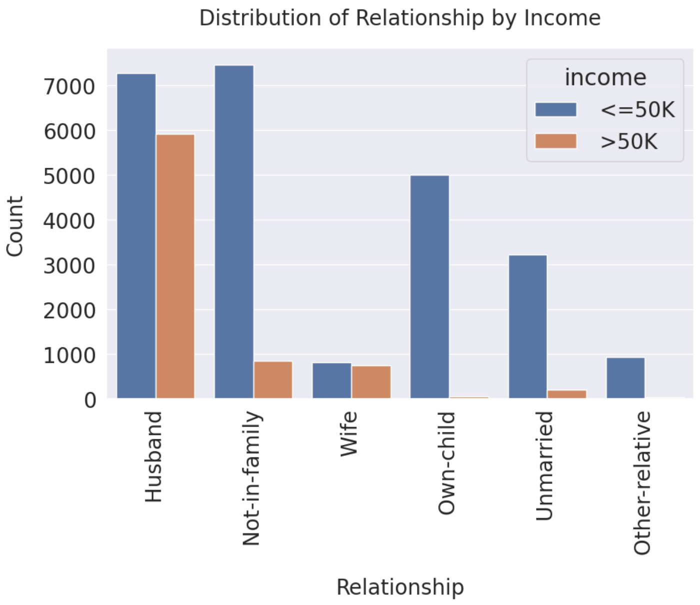
From the distribution of marital status by income, several notable trends emerge. The category Married - Civ Spouse stands out with a substantial representation of around 14,000 individuals. Among them, over 6,000 individuals earn more than $50K. Conversely, the groups Never Married and Divorced constitute a smaller proportion, with a majority earning below $50K. Notably, a lower percentage of individuals in these groups earn above $50K, underscoring the influence of marital status on income levels.

In the distribution of occupation by income, it's evident that in most occupations, the majority earn less than $50K.









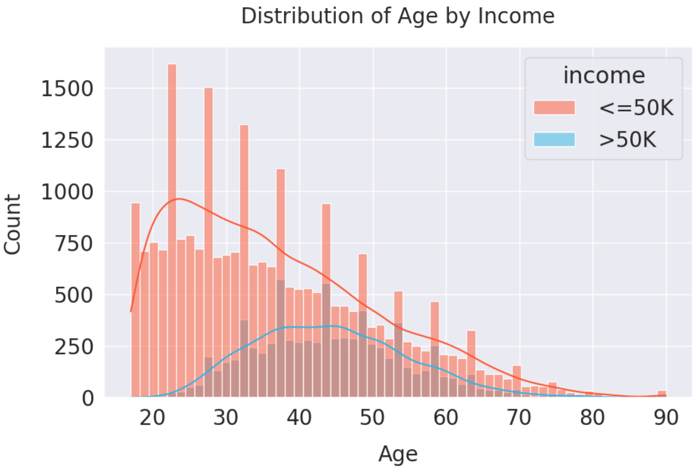
However, for ’Exec-managerial’ roles, the number of individuals earning both above and below $50K is relatively balanced. Notably, ’Prof-specialty’ occupations stand out with the highest number of individuals earning more than $50K compared to other occupation groups.

The distribution of relationships by income reveals that husbands constitute the majority in the dataset. Within this group, individuals earning more than $50K and less than $50K are roughly equal in number. On the other hand, individuals without a family tend to earn less than $50K in greater numbers. Additionally, although wives are fewer in number, they are fairly evenly distributed between those earning less than and more than $50K.

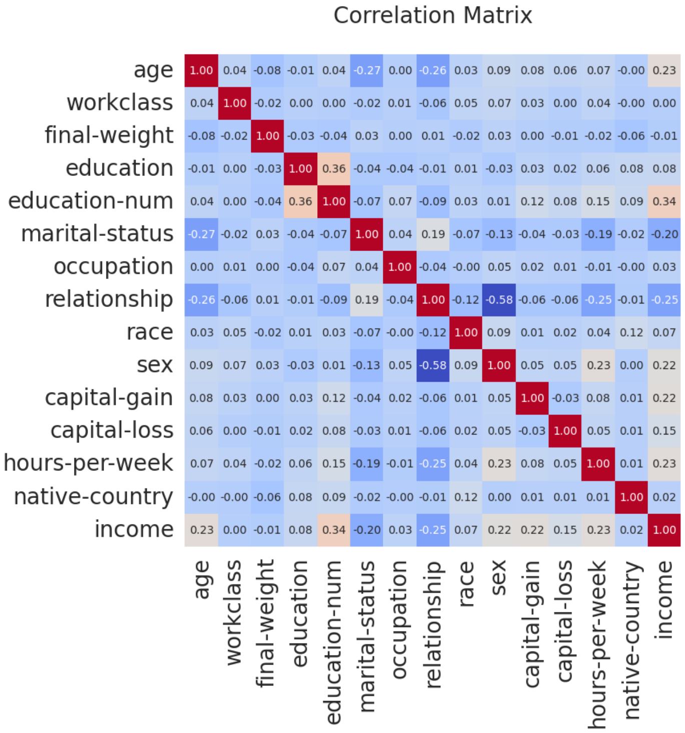
From the dataset, it’s evident that the majority of individuals are of White ethnicity. Notably, among all races, Whites have the highest number of individuals earning more than $50K. Additionally, males outnumber females in the dataset, and they also have a higher representation in the group earning more than $50K. Moreover, a significant portion of individuals’ native country is the United States, and interestingly, a higher proportion of those whose native country is the USA also fall into the category of earning more than $50K.

In the distribution of age by income, it’s evident that the majority of individuals earning less than $50K fall within the 20-30 year age range. Conversely, a significant portion of those earning above $50K are in the 40-50 age range. Interestingly, both distributions, for incomes less than and greater than $50K, exhibit a normal distribution pattern with respect to age.

The correlation heat map indicates that there is no significant correlation among the features, including those with the target variable. This suggests that all the features can be



considered for further data analysis without concerns about multicollinearity or strong linear relationships.



III. Naive Bayes Classifier

A. Introduction

Classification algorithms are essential in machine learning for predicting the class or category of a given observation. In Naive Bayes classifiers, we make a ”naive” assumption that the features are conditionally independent given the class label. This simplifies the computation and makes the algorithm efficient, although it may not always hold true in real-world scenarios. Naive Bayes classifiers, rooted in Bayes’ Theorem, offer a powerful approach to classification tasks.

B. Bayes’ Theorem

Bayes’ Theorem is a fundamental concept in probability theory, allowing us to update our beliefs about the probability of an event based on new evidence. It is expressed as:

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

Where:

- $P(A|B)$ is the posterior probability of event A given that event B has occurred.
- $P(B|A)$ is the likelihood or probability of observing event B given that event A has occurred.
- $P(A)$ is the prior probability of event A.
- $P(B)$ is the prior probability of event B.

C. Types of Naive Bayes Classifiers

Naive Bayes classifiers come in various flavors, each tailored to specific data characteristics and applications. Here, we delve deeper into the different types:

1) Gaussian Naive Bayes: The Gaussian Naive Bayes model assumes that numerical features in the dataset follow a Gaussian or normal distribution. This makes it particularly suitable for continuous data, where the values are distributed along a bell-shaped curve.

Formulation: The likelihood $P(x_i|y)$ for a given feature x_i in class y is modeled as a Gaussian distribution:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}}$$

2) Multinomial Naive Bayes: The Multinomial Naive Bayes model is primarily used for text classification tasks, where features represent the frequency of words or tokens. It assumes that features are generated from a multinomial distribution.

Formulation: The likelihood $P(x_i|y)$ for a given feature x_i in class y is calculated using the multinomial probability mass function:

$$P(x_i|y) = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

3) Bernoulli Naive Bayes: The Bernoulli Naive Bayes model is well-suited for binary feature data, where each feature can take on values of either 0 or 1. It models the presence or absence of features.

Formulation: The likelihood $P(x_i|y)$ for a given feature x_i in class y is calculated using the Bernoulli probability mass function:

$$P(x_i|y) = (1 - \theta_{yi})^{(1-x_i)} \cdot (\theta_{yi})^{x_i}$$

4) Complement Naive Bayes: The Complement Naive Bayes model is designed to handle imbalanced datasets, where one class significantly outnumbers the others. It adjusts the probabilities to give more weight to the under-represented class.

Formulation: The likelihood $P(x_i|y)$ in Complement Naive Bayes is defined as the complement of the likelihood in class y (inverse probability):

$$P(x_i|y) = 1 - P(x_i|\neg y)$$

5) Categorical Naive Bayes: The Categorical Naive Bayes model is used when features are categorical, representing distinct categories. It is well-suited for data with nominal attributes, such as color or country.

D. Applications

Naive Bayes classifiers find extensive applications in various domains, including text classification, medical diagnosis, recommendation systems, and image recognition. Their simplicity and efficiency make them valuable tools in the machine learning toolkit.

E. Assumptions and Limitations

While powerful, Naive Bayes classifiers come with assumptions and limitations:

- **Independence Assumption:** Assumes features are independent given the class label, which may not hold in complex datasets.
- **Sensitivity to Feature Distribution:** Gaussian Naive Bayes assumes Gaussian distribution of features; deviations may impact performance.
- **Limited Expressiveness:** May struggle to capture intricate feature interactions in complex data.

IV. Results

In this study, the target variable of interest is ”income,” indicating whether an individual earns more than $50,000 annually. The Gaussian Naive Bayes classifier was selected for its suitability to the dataset, assuming that numerical features follow a Gaussian distribution, which aligns with real-world scenarios.

The model was trained and tested on a preprocessed dataset, where categorical variables were encoded using a label encoder. Subsequently, the performance of the Naive Bayes model was evaluated.

A. Model Performance

The model demonstrated an accuracy of approximately 79.64% on the test set, signifying the proportion of correctly classified instances. The classification report further provides detailed metrics:

Class	Precision	Recall	F1-Score	Support
<=50K	0.81	0.95	0.88	4912
>50K	0.68	0.32	0.44	1600

These metrics offer insights into the model's performance, including false positives, false negatives, and overall accuracy.

B. Model Consideration

The Gaussian Naive Bayes model was chosen due to its alignment with the dataset's distribution assumptions. This choice facilitated effective training and yielded promising results.

Overall, the Naive Bayes model excels in accurately predicting individuals with an income below $50,000. However, it exhibits limitations in identifying individuals with higher incomes.

V. Conclusions

In this study, we examined the effectiveness of a Gaussian Naive Bayes classifier for income classification based on the provided dataset. The following key insights were obtained:

1) The Gaussian Naive Bayes model demonstrated a commendable accuracy of approximately 79.64% in predicting income status based on the provided features.
2) The model's performance varied across income categories, with higher accuracy in identifying individuals with an income below $50,000.
3) Features such as education level, working class, and occupation proved to be influential in determining income status.
4) The distribution of features closely aligned with the assumptions of the Gaussian Naive Bayes model, contributing to its effective performance.
5) Notably, individuals who remain unmarried are more likely to have an income below $50,000.

A. Further Avenues for Investigation

While this study provides valuable insights into income classification, there are several avenues for further investigation:

1) **Feature Engineering:** Exploring advanced feature engineering techniques could enhance the model's predictive capabilities.
2) **Model Selection:** Evaluating the performance of alternative classifiers such as Random Forest, Support Vector Machines, or Neural Networks could offer comparative insights.
3) **Ensemble Methods:** Investigating ensemble methods like Bagging or Boosting could potentially enhance the model's performance.
4) **Additional Data Sources:** Incorporating supplementary datasets, if available, might provide richer context and improve predictions.
5) **Temporal Analysis:** Assessing income trends over time may uncover valuable patterns and contribute to more accurate predictions.

Overall, this study establishes a foundation for future research endeavors in the field of income classification, with potential applications in various domains, including socioeconomic research and financial planning. .

References

[1] https://www.ibm.com/topics/logistic-regression: :text=Resources-,What

[2] https://www.geeksforgeeks.org/understanding-logistic-regression/.

[3] https://en.wikipedia.org/wiki/Logistic$_{r}egression$.

[4] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.