Original Research Article

# PREDICTION OF LUNGS CANCER DISEASES DATASETS USING MACHINE LEARNING ALGORITHMS

***Abstract:*** *Lung cancer is the most common cause of mortality, and it is the only sort of cancer that affects both men and women globally. The primary goal of this paper is to creates a model for predicting lungs cancer using various machine learning classification algorithms like k Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Gaussian Naive Bayes (NB). Furthermore, assess and compare the performance of the varied classifiers using their accuracy in selecting the best algorithms. The lung cancer dataset is publicly available on the Kaggle Machine Learning Repository, thus the implementation phase dataset will be partitioned as 80% for the training phase and 20% for the testing phase before using machine learning methods. In all parameters, the support vector machine performed well.*
*Keywords: Breast Cancer, Machine Learning, Classification, Accuracy,* **support vector machine.**



## I. INTRODUCTION

Lung cancer, in other words, is the main global cause of death for both men and women (Brocken, et al (2012). Other research indicates that in 2015, lung cancer made up around 13% of all cancer diagnoses in the United States. According to the American Cancer Society Vivekanandan, lung cancer causes around 27% of all cancer-related deaths (2013). Lung nodules that are still developing need to be carefully checked and tracked. The ML and DL techniques for predicting cancer growth and progression were used by the researchers in this study to analyze cancer development and progression. The prediction models covered in this article are created by combining a range of supervised machine learning algorithms with different input and data samples. Images can be converted into arrays or images of integer labels, which are known as local binary patterns, using the image operator LBP. Further picture analysis, which is often displayed as a histogram, makes use of these designations. The LBP texture operator has been used in a variety of applications because of how specific it is and how simple it is to use (Vivekanandan, 2013).

These markers are then used by the histogram to analyze the image more thoroughly. For both men and women, cancer mortality from lung disease has remained higher over the past three years than cancer mortality from prostate or breast cancer (D'Cruz, 2016). This is largely attributable to the complex and systemic nature of the predictive models for breast and prostate cancer that have been created recently. To do this, an accurate early-stage lung cancer forecast model must be created as soon as possible (Shen, et al 2021). SVM, a powerful predictor in both

linear and nonlinear situations, is widely used in a variety of fields, including medicine (Abdullah, et al (2021). Even if SVM is an excellent approach to categorize items, cancer prognostic models are still created Jenipher, et al (2021). The results of a mutation test, which has become more significant in clinical trials, help establish the best therapy options for patients (Binson, et al 2021). Direct sequencing can be performed in addition to screening to find mutations that were overlooked during the screening procedure. The EGF receptor (EGFR) has a genetic mutation that can be used to identify genetic mutations in lung cancer. Artificial neural networks (ANN) and support vector machines (SVM) have been shown to perform better than their no ensemble equivalents (Xie, et al 2021). Misjudgments by the majority are more likely to occur than those by the minority because the majority's judgment has a greater weight than the minority's. The performance of classification algorithms that rely on conventional techniques falls short of what is possible (Miller, et al., 2021).

## II. MACHINE LEARNING ALGORITHMS

Programs that use machine learning algorithms are able to discover hidden patterns in data, forecast results, and enhance performance based on past performance. In machine learning, several algorithms can be employed for various tasks, such as basic linear regression for prediction issues like stock market forecasting and the KNN algorithm for categorization issues. Three general categories can be used to categorize machine learning algorithms:

1. Supervised Learning Algorithms
2. Unsupervised Learning Algorithms
3. Reinforcement Learning algorithm

a. Supervised Learning Algorithm

A type of machine learning called supervised learning requires outside supervision for the machine to learn. The labeled dataset is used to train the supervised learning models. After training and processing, the model is put to the test by being given a sample set of test data to see if it predicts the desired result.

In supervised learning, mapping input and output data is the main objective. It is the same as when a student is studying under the teacher's supervision because supervised learning is dependent on supervision. Spam filtering is a prime example of supervised learning.

Supervised learning can be divided further into two categories of problem:

Classification

Regression

b. Unsupervised Learning Algorithm

Unsupervised learning is a sort of machine learning where the computer can make its own decisions based solely on the data it is given. The algorithm needs to act on that data without any supervision, and the unsupervised models can be trained using the unlabelled dataset, which is neither classed nor categorized. In unsupervised learning, the model searches through the vast amount of data in search of meaningful insights rather than producing a predetermined result. These are employed in order to address the Association and Clustering issues. Consequently, it can be divided into two categories:

Clustering

c. reinforcement Learning

In reinforcement learning, an agent produces actions to interact with its environment and learns from feedback. The agent receives feedback in the form of rewards; for example, he receives a

positive reward for each good activity and a negative reward for each bad action. The agent is not under any oversight. Reinforcement learning makes use of the Q-Learning algorithm.
Association

**A. k Nearest Neighbor (KNN)**

In order to predict the estimations of the most recent data snippets, the k Nearest Neighbors algorithm makes use of "feature similarity." This further ensures that the new information point will be given a value based on how closely it resembles the points in the training set.

**B. Support Vector Machine (SVM)**

Support Vector Machine is one of the supervised machine learning characterisation techniques that is widely used in the field of determining and predicting cancerous growth. Support Vector Machine isolates the classes by developing a linear function that separates them as thoroughly as is practical using these help vectors, which are selected as basic examples from all classes and are referred to as help vectors. Accordingly, it is frequently stated that planning between an input vector and a high-dimensional space is framed using an SVM with the goal of finding the best plausible hyperplane to categorize the data set. By identifying the most suitable hyperplane, this linear classifier aims to increase the space between the decision hyperplane and consequently the nearest data, also known as the minimal distance.

**C. Logistic Regression (LR)**

A crucial machine-learning classification technique is logistic regression. Like polynomial and statistical regression, it fits in with the group of linear classifiers and is fairly useful. You can understand the results more easily and quickly by using logistic regression, which is rapid and straightforward. It is a path for binary classification, but it may also be used for multi-class problems. As statistical regression considers the forecast of consistent traits, this is typically not the same thing. The likelihood that a reaction fits into a particular classification is modelled using logistic regression. By using the Sigmoid function, a logistic regression model aids in the resolution of problems when the output can only take one of two values, either 0 or 1.

**D. Naive Bayes (NB)**

Naive Bayes is a classification technique that relies on the Bayes Theorem and presumes predictor independence. Simply put, a Naive Bayes classifier believes that the proximity of certain elements within a class has no bearing on the proximity of the other element. Although these characteristics depend on one another or on the absence of the opposite characteristics, they still freely increase the possibility of a class, which is why it is called "Naive." Because it assumes that estimating features are independent of one another, Naive Bayes (NB) is called "naive." Because it's (almost) never obvious, this is frequently naive. The naive Bayes model is simple to construct and is particularly useful for large data sets. In close proximity to simplicity, Naive Bayes is understood to

## III. ABOUT THE DATASET

This study is based on data from the US National Lung Screening Trial, which is publicly available (NLST). The data pertains to smokers who have smoked in the past and who have not, who were tracked for seven years while undergoing annual lung cancer screenings. There were no nonsmokers participated in the trial. The dataset's resulting attributes are as follows:

- PID - anonymous identifier of a person
- age - the age of a person at the start of the trial
- gender - Male/Female
- race - the race of a person

- smoker - Former/Current (Former is defined as quitting smoking in the last 15 years)

## IV. LITERATURE REVIEW

Predicting Cyber Security Incidents Using Machine Learning Algorithms: A Case Study of Korean SMEs is a 2019 study by Mohasseb et al. We examine a dataset from five small and medium-sized businesses in South Korea, which represents cyber security incidents and response measures, in their paper. We look into how the data illustrating various occurrences gathered from various businesses can help to increase classification accuracy and aid the classifiers in differentiating between various sorts of incidents. For the classification of incidents and the actions taken in response, a model using text mining techniques such as n-gram, bag-of-words, and machine learning algorithms has been built. The effectiveness of the classifiers for the prediction of various types of reactions and malware has been shown by experimental findings.

Based on the instruction sequences recovered from the file sample set, the authors (Fan et al., 2016) suggested a sequence mining approach to uncover dangerous sequential patterns. A Nearest-Neighbor (ANN) classifier was then built for malware identification based on the observed patterns. The proposed sequential pattern mining method and the ANN classifier make up the created data mining framework. Additionally, authors (Wang et al., 2006) suggested an integrated architecture to combat spy software and employed features taken from both static and dynamic analysis. These features were ranked based on the information gains they provided. A machine learning algorithm was also applied. Using Common N-Gram analysis (CNG), which depends on profiles for class representation, the authors of (Abou-Assaleh et al., 2004) presented a method based on byte n-gram analysis to identify malicious code. In order to detect undiscovered harmful code, the authors (Shabtai et al., 2012) collected OpCode n-gram patterns from the analyzed files after they were disassembled. The classification procedure makes use of the OpCode program patterns as characteristics.

## V. PROPOSED SYSTEM

With the breast cancer dataset imported and features needing to be retrieved, Figure 1 depicts the basic breast cancer classification model with machine learning calculations. The classification model is frequently trained and utilized for the prediction of benign and malignant breast cancer. Benign conditions are thought to be non-cancerous, which makes them safe. Unusual cell growth is the beginning of harmful cancer, which can quickly spread or assault normally dangerous tissue nearby.

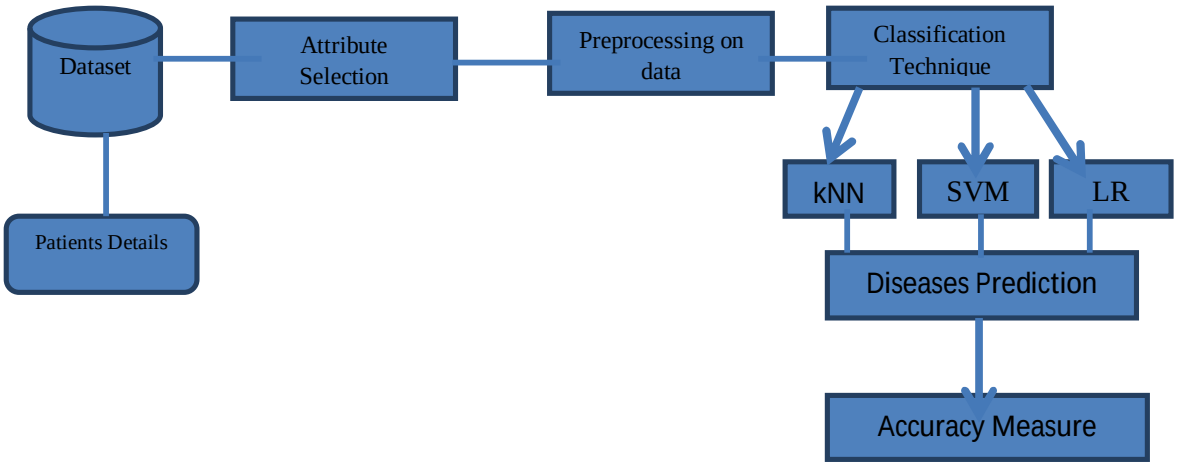

**Fig. 1. Bosom breast cancer classification model**

## VI. SOFTWARE USED

### 6.1 Python

An Anaconda-programmed web scraper was utilized to get the data. Python's syntax, according to Wikipedia, enables programmers to express concepts in less code. implementation of Python in December 1989. On October 16, 2000, Python 2.0 was released, then on December 3, 2008, Python 3.0.
Why use Python for web scraping rather than a different language? A module called "urllib2" is available in Python and provides the right functions for quickly opening webpages and extracting information. The web scraper that is in charge of gathering the weather data for the model is programmed in Python.

## VII. OBSERVATION

The confusion matrix is a table that is widely used to show how well a classification model performs when applied to a collection of test data for which truth values are known.

**TABLE 1.**

**CONFUSION MATRIX**

| | **Predicted Class** | | |
|---|---|---|---|
| **Actual Class** | | **Class = Yes** | **Class = No** |
| | **Class = Yes** | True Positive (TP) | False Negative (FN) |
| | **Class = No** | False Positive (FP) | True Negative (TN) |



The accurately predicted observations, designated as TP and FP in Table 2, are shaded in blue. We may want to completely reduce false positives and false negatives, which is why they are highlighted in red.
Classification accuracy:It is one of the crucial factors in figuring out how accurate a classification problem is. It specifies how frequently the model predicts the right result. The number of accurate predictions made by the classifier divided by the total number of predictions made by the classifiers can be used to compute it. The following is the formula:

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

Misclassification rate: Also known as the "error rate," this word describes the frequency with which the model makes incorrect predictions. The ratio of the number of inaccurate predictions to the total number of predictions made by the classifier can be used to compute error rate. The equation is shown below:

$$\text{Error Rate} = \frac{FP+FN}{TP+FP+FN+TN}$$

**Precision:** Precision can be characterized as the proportion of the model's outputs that were accurate, or as the proportion of the model's correctly anticipated positive classes that really occurred. Using the formula below, it can be calculated:

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

**Recall:** Recall it is defined as the percentage of total positive classes that our model properly predicted. There must be a significant recall.

**F-measure:** It is challenging to compare two models if one has a high recall and a poor precision. F-score can therefore be used for this purpose. This score enables us to simultaneously assess recall and precision. If the recall and precision are equal, the F-score is at its highest. Using the formula below, it can be calculated:

$$\textbf{F-measure} = \frac{2*Recall*Precision}{Recall+Precision}$$

**Null Error rate:** It specifies the proportion of times our model would be off if it consistently predicted the majority class. The best classifier, according to the accuracy paradox, "has a larger mistake rate than the null error rate."

**ROC Curve:** The performance of a classifier for each potential threshold is shown on a graph called the ROC. The real positive rate (on the Y-axis) and the false positive rate are shown on a graph (on the x-axis).

## VIII. ACCURACY

The accuracy of the classifier is a measure of how effectively it can anticipate situations into the appropriate classification. It is the number of accurate predictions divided by the total number of occurrences in the data collection. It's important to note that accuracy is highly dependent on the edge that the classifier chooses and may, thus, vary for different testing sets. Accordingly, comparing several classifiers is not the best technique, although it may provide a summary of the classification. As a result, accuracy is frequently determined using the equation:

## VIII. RESULT AND DISCUSSION

This initiative seeks to identify the patients' current smoking status (Elkan 1997). Training sets and test sets are created from the dataset's records. following data preparation. Support vector is a method of data classification. Table 2 displays the levels of accuracy for each of the four machine learning algorithms.

**TABLE 2. ACCURACY VALUES**

| S/N | Algorithms | Accuracy (decimal) | Percentage | RMSE | MAPE |
|---|---|---|---|---|---|
| 1 | K Neural Network (kNN) | 0.97 | 97% | 0.660859 | 0.675199 |
| 2 | Support Vector Machine (SVM) | 0.98 | 98% | 0.601131 | 0.668442 |
| 3 | Logistics Regression (LR) | 0.89 | 89% | 0.687485 | 0.774507 |
| 4 | Naïve Bayes (NB) | 0.77 | 77% | 0.611737 | 0.725908 |

**Data Source: R-Studio Output**

The critical performance in Table 2 is completed by the support vector machine with a 98 percent accuracy rate, while the second performance is completed by the k neural network with a 97 percent accuracy rate. The third performance is successfully completed by logistics regression with an accuracy of 89 percent, while the fourth performance is successfully completed by naive-Bayes with an accuracy of 77%. The support vector machine has the highest accuracy for the supplied dataset, which is in line with the findings of the predictions. This demonstrates that support vector machine consistently outperforms support vector machine, logistic regression, and naive bayes for the prediction of lung cancer. Support vector machines exhibit the lowest Mean Square Error, according to the RMSE.

## XIII. CONCLUSION

The classification of lung cancer caused by smoking using machine learning approaches is well explained in this work. In order to anticipate the type of treatment that can be given to patients, the job of the classifier is essential in the healthcare sector. For the purpose of identifying effective and precise methods, the existing methodologies are examined and contrasted. By identifying patients at an early stage of the disease and allowing them to get preventive treatment, machine learning algorithms greatly increase the accuracy of lung cancer prediction. In this study, we compared the classification parameters for four machine learning methods, including Naive Bayes, Support Vector Machine, k Nearest Neighbor, and Logistic Regression. This comparison analysis sought to identify the most precise machine learning algorithm that might be used to aid in the detection of lung cancer. According to the findings of the prediction, k Nearest Neighbor has the highest accuracy for the provided dataset. This demonstrates that k Nearest Neighbor consistently outperforms Support Vector Machine, Logistic Regression, and Naive Bayes for the prediction of lung cancer.

## XII. REFERENCE


Abdullah, D. M., Abdulazeez, A. M., & Sallow, A. B. (2021). Lung cancer prediction and classification based on correlation selection method using machine learning techniques. *Qubahan Academic Journal*, *1*(2), 141-149.

Aruna, S., Rajagopalan, S., & Nandakishore, L.(2011). "Knowledge-based analysis of various statistical tools in detecting breast cancer" *Computer Science Information Technology*.

Benbrahim, H., Hachimi, H., & Amine, A. (2020) *Springer, Comparative Study of Machine Learning Algorithms Using the Breast Cancer Dataset*.

Binson, V. A., Subramoniam, M., Sunny, Y., & Mathew, L. (2021). Prediction of pulmonary diseases with electronic nose using SVM and XGBoost. *IEEE Sensors Journal*, *21*(18), 20886-20895.

Brocken, P., Kiers, B. A., & Looijen-Salamon M. G., (2012). "Timeliness of lung cancer diagnosis and treatment in a rapid outpatient diagnostic program with combined $^{18}$FDG-PET and contrast enhanced CT scanning," *Lung Cancer*, vol. 75, no. 3, pp. 336–341.

Chaurasia, V., & Pal, S. (2014). "Data mining techniques: To predict and resolve breast cancer survivability" *Int. J.Computer Science*.

D'Cruz, J. Jadhav, A. Dighe, A. Chavan, V. and J. Chaudhari, (2016) "Detection of lung cancer using back propagation neural networks and genetic algorithm," *Computing Technologies and Applications*, vol. 6, pp. 823–827.

Deepika, V., & Nidhi., M. (2017). "Comparative analysis of breast cancer and hypothyroid dataset using data mining classification techniques" IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI).

Elkan C. (1997). "Naive Bayesian Learning, Technical Report CS97-557", Department of Computer Science and Engineering, *University of California, San Diego, USA*.

Jenipher, V. N., & Radhika, S. (2021, February). SVM Kernel Methods with Data Normalization for Lung Cancer Survivability Prediction Application. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 1294-1299). IEEE.

Mariam, A., Saliha, O., Ikram, G., &Tolga, E., (2018). "Breast cancer classification using machine learning" Electric Electronics, Computer Science, *Biomedical Engineerings, Meeting* (EBBT).

Miller, H. A., Yin, X., Smith, S. A., Hu, X., Zhang, X., Yan, J., ... & Frieboes, H. B. (2021). Evaluation of disease staging and chemotherapeutic response in non-small cell lung cancer from patient tumor-derived metabolomic data. *Lung Cancer*, *156*, 20-30.

Shen, J., Wu, J., Xu, M., Gan, D., An, B., & Liu, F. (2021). A Hybrid Method to Predict Postoperative Survival of Lung Cancer Using Improved SMOTE and Adaptive SVM. *Computational and mathematical methods in medicine*, *2021*.

Vivekanandan, P. (2013). "An efficient SVM based tumor classification with symmetry non-negative matrix factorization using gene expression data," in *In 2013 International conference on information communication and embedded systems (Icices)*, pp. 761–768, IEEE, USA.

Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., ... & Leung, E. L. H. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational oncology*, *14*(1), 100907.

Appendix

```
> head(cancer)
    pid age gender  race  smoker
1 100001  70   Male White Current
2 100002  66   Male White Current
3 100003  64   Male White Current
4 100004  60   Male White  Former
5 100005  64   Male White  Former
6 100006  56 Female White Current
> tail(cancer)
       pid age gender                 race  smoker
1187 101195  71 Female                White  Former
1188 101196  61 Female                White Current
1189 101197  70   Male More than one race  Former
1190 101198  60   Male                White Current
1191 101199  55 Female                White Current
1192 101200  69 Female Black or African-American Current
```

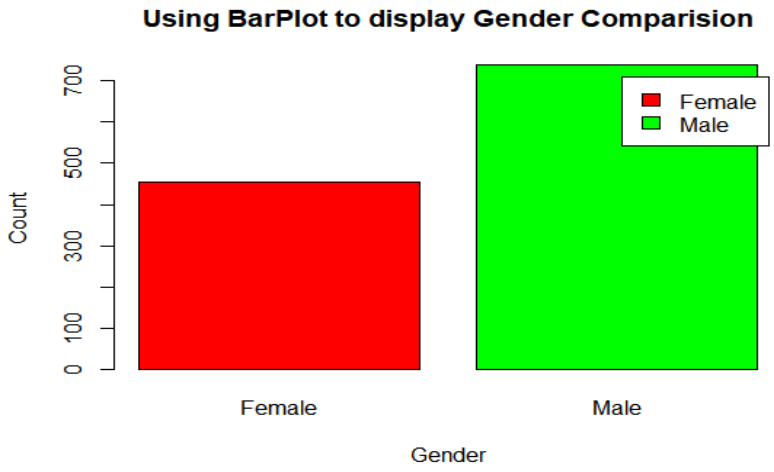


Fig 2: Using Bar plot to display Gender comparision

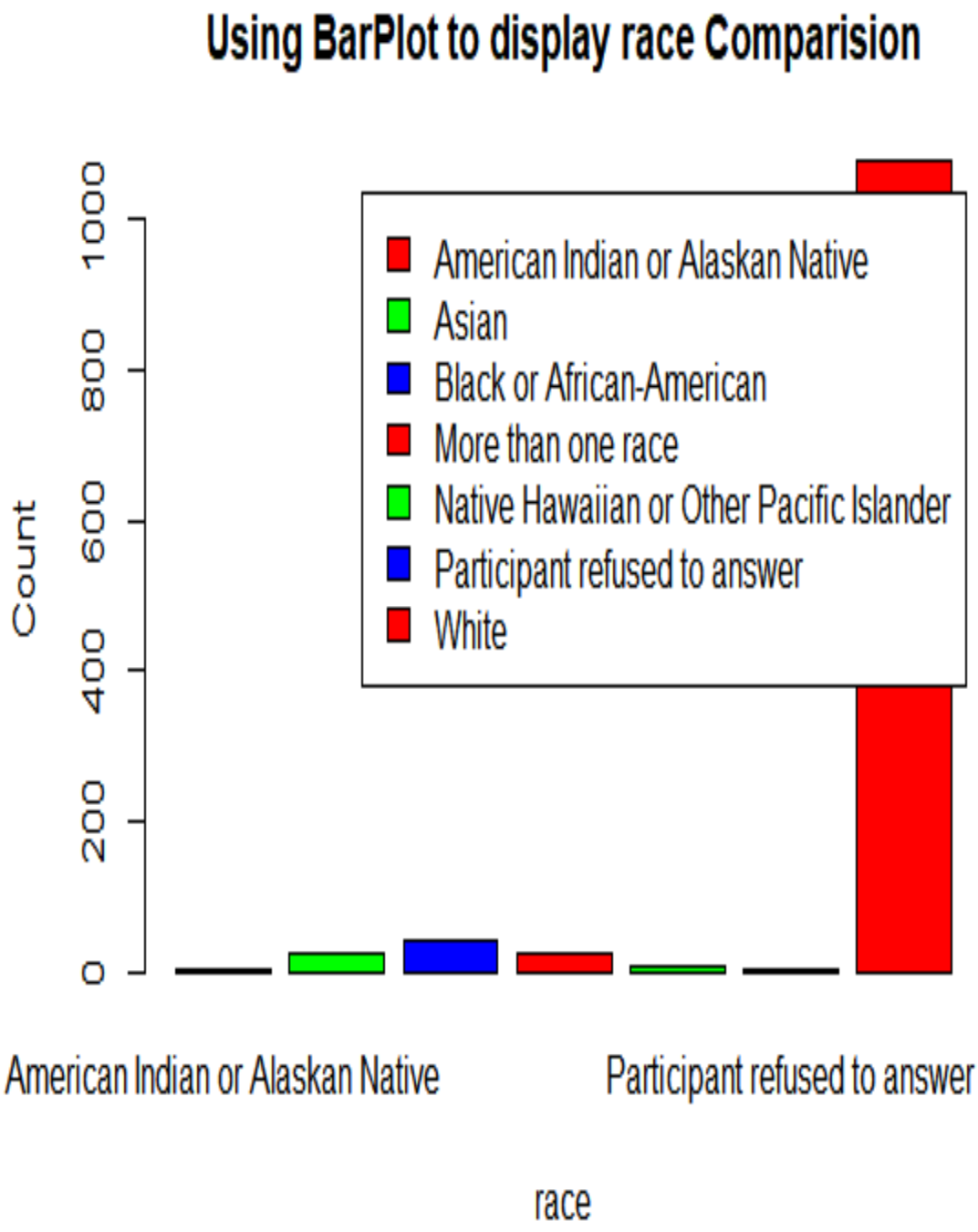


Fig 3:using Bar plot to display race comparision

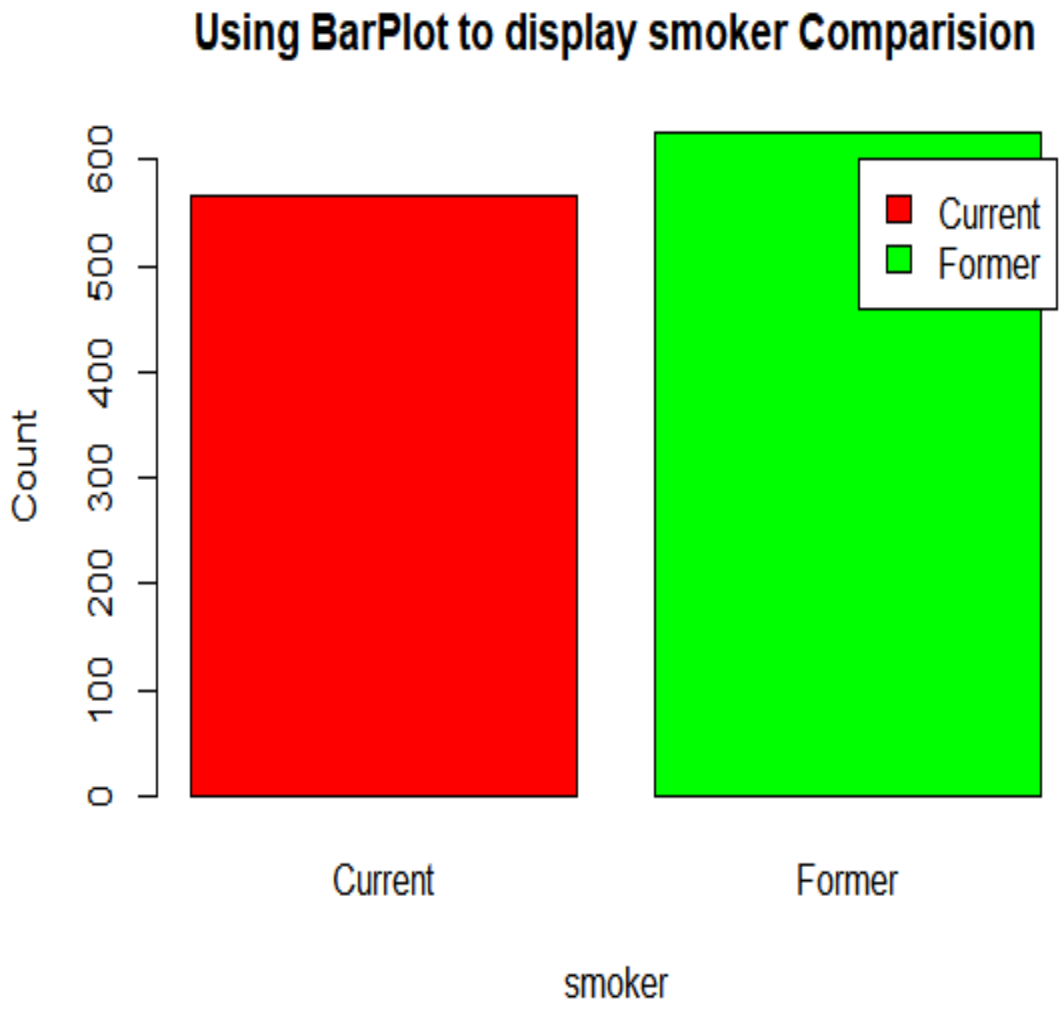


Fig 4: Using Bar plot to display smoker  comparision