ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1923

Impact of Gender and Demographic Factors on Smoking Behavior Prediction

Dhanshree Biradar[1]; Amruta Sanap[2]; Gayatri Gulbhele[3]; Sanika Pawar[4]; Rajnandini Sonone[5]; Archana Ratnaparkhi[6];

[1;2;3;4;5;6] Electronics and Telecommunication (of Aff.) Vishwakarma Institute of Information Technology (of Aff.) Pune, India

Publication Date: 2025/05/09

Abstract: The relationship between smoking habits and gender, that is affected by demographic variables like age, income level, and education level, is illustrated. Using the Chi-square test for independence, we identified significant associations with gender on smoking behavior. EDA further revealed crucial trends and disparities in prevalence of smoking among different groups of demographic variables. We develop a classification model to identify the significant predictors for smoking behavior from the factors mentioned above. Further, we conduct Principal Component Analysis for reducing the dimension of data to see which factors are more significantly influential for smoking behavior. The results of the research give us the understanding that can be applied towards the formulation of targeted public health strategies that could effectively reduce the prevalence of smoking.

Keywords**:** *Smoking Behavior, Gender Differences , Demographic Factors , Chi-Square Test.*

How to Cite: Dhanshree Biradar; Amruta Sanap; Gayatri Gulbhele; Sanika Pawar; Rajnandini Sonone; Archana Ratnaparkhi; (2025) Impact of Gender and Demographic Factors on Smoking Behavior Prediction. *International Journal of Innovative Science and Research Technology*, 10(4), 2898-2903. https://doi.org/10.38124/ijisrt/25apr1923

I. INTRODUCTION

Despite the fact that smoking remains one of the main causes of preventable morbidity and mortality around the world, it has emerged as a significant public health challenge.Understanding factors influencing smoking behavior is crucial for effective prevention and intervention strategies. Gender plays an essential role here, as smoking prevalence and patterns often vary between men and women. This research analyzes the relationship between smoking behavior, gender, and demographic factors such as age, income, and education. Using the Chi-square test for independence, the study examines whether gender is significantly associated with smoking behavior and performs exploratory data analysis to explore trends across demographic groups.

To deepen this understanding, a classification model predicts smoking behavior based on these demographics, helping to identify key predictors and high-risk groups. Additionally, principal component analysis (PCA) reduces data complexity by identifying variables with the greatest influence on smoking behavior. Insights from this study can guide public health policies and targeted interventions aimed at reducing smoking prevalence and promoting better health outcomes across diverse populations by examining how gender and other demographics interact with smoking behavior.

II. RELATED WORK

These studies, therefore, indicate differences in smoking behaviors and attitudes among genders, with variations in cessation challenges. It calls for gender-sensitive public health interventions and tobacco control policies to adequately address smoking prevalence and health inequities.

Bali , V. et al. (2020) The review analyzed the gender variations in smoking behavior and differentiated between initiation, cessation, and attitudes between genders. They highlighted societally linked factors of the smoking issues and underscored the role of the interventions with considerations of differences by gender[1]. Sharma , M. et al. (2019) An exploratory study on the role of gender in smoking initiation and cessation is conducted, revealing a level of evidence that indicates that men and women experience different barriers and motivators in their smoking behaviors; therefore, the need for specific interventions tailored according to these differences[2].

Bennett , K. et al. (2021) Socioeconomic correlates of smoking status among males and females in the U.S. The article examines how socioeconomic factors can influence the smoking behavior among men and women. A significant relationship was established regarding income levels, education levels and smoking status. Therefore, the result may point out an urgent call for the policy makers in terms

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1923

of public health to bring down socioeconomic inequality[3]. Kumar , R.et al. (2020) This population-based study studies gender-specific smoking behaviors, showing different patterns of smoking in both genders. It thus calls for this understanding to be built into public health policies and cessation programs[4]. Cheng , T. et al. (2018) This one examines how gender and age may influence smoking behaviors and attitudes. Results from the study reveal that the smoking habits and culture, previously adopted by men, are increasingly adopted by young women, further pointing to cultural perceptions changes[5].

Bach , L. et al. (2019) This research explores the gender differences in smoking quit attempts and provides evidence of differences for women so that the cessation programs should be adjusted according to thesedimensions in order to improve their effectiveness[6]. Graham , H. et al. (2019) This is a longitudinal study, where it analyzes the relationship of smoking behavior with gender and socioeconomic status over time. The findings provide essential health disparities related to smoking that demand policies aimed at correcting the imbalances[7]. Wang , J. et at. (2021) This study assesses how gender influences smoking behaviors and attitudes toward tobacco control policies. It demonstrates that both men and women might have different reactions to such policies, hence the need for gender-sensitive approaches in the regulation of tobacco[8].

Fagan , p. et al. (2019) The study focuses on the variability in smoking behavior and attitude across different populations and how gender affects these variations. It fosters culturally appropriate and gender-sensitive interventions for smoking prevention and cessation interventions[9]. Slade , J. et al. (2018) This analysis addresses the issue of gender and smoking health disparities, discussing how various social, economic, and cultural factors influence the smoking rates of men and women. The findings highlighted the need for targeted public health campaigns to reduce such disparities[10].

III. METHODOLOGY

A. Chi-Square Test for Independence:

- Construct a contingency table that illustrates the frequency of smoking behaviors like smoker or non-smoker across genders like male or female. Compute the Chi-square statistic, where $\chi 2=\sum(Oi-Ei)2/$ Ei where Oi is the observed frequency, and iEi is the expected frequency.

- Compute and compare the obtained value to the tabulated value along with degrees of freedom with the help of a chi-square distribution table. Therefore, a significant p-value less than 0.05 means that gender will not be independent for a smoking behavior.

B. EDA:

Exploratory data analysis provides an idea of visualization into the association of variables present in smoking behavior in connection to demographic variables.

- Use of visualizations:

- Histogram to show the age distribution of smokers and non-smokers
- Bar charts for smoking prevalence comparison among various levels of income or educational levels.
- Box plots for illustration of age variations across the different groups of smoking.
- Calculate summary statistics-mean, median, and standard deviation-for demographic variables, conditioned on smoking status.
- Regress continuous demographic variables with smoking status

- *Development of a Classification Model:*

Create a prediction model which will highlight drivers of the smoking behavior

- *Choose Appropriate Classification Algorithms, Including:*

- ✓ Logistic Regression for binary outcomes like smoker vs. nonsmoker.
- ✓ Decision Trees for interpretability and capturing nonlinear relationships.
- ✓ Random Forests for increasing predictive accuracy through ensemble methods.

- The dataset has been divided into training set and testing set with the percentage of 70 and 30 respectively to prevent overfitting.
- The model is trained against the training dataset and validated against the testing dataset
- Model evaluation is carried out with metrics such as:

- Accuracy: The ratio of correctly classified instances.
- Accuracy: True positive predictions divided by the total number of positive predictions.
- Sensitivity: Number of true positives found divided by the total number of actual positives.
- F1-score: Harmonic mean of precision and recall, balancing between the two.

- *Feature Reduction using Principal Component Analysis (PCA):*

Transform the demographic attributes to their lower-dimensional representation without the loss of important information.

- Standardize the data so that every feature contributes equally to analysis, then apply PCA by transforming the original demographic variables into a set of uncorrelated variables, known as principal components.
- Find out the percentage of variance explained by each principal component so that the most important ones can be identified.
- Choose a few principal components that explain a large percentage of total variance, say about 80%.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1923

➢ *Use Such Components in further Analysis to Reduce the Complexity of Modeling.*

- *Interpretation and Recommendations:*
  To summarize and act upon actionable insights.

✓ Interpret the result of the Chi-square test: Smoking will or will not be related with gender.
✓ Continue conducting EDA and notice any particular demographic group has high smoking levels.
✓ Analyze the model of classification to know which demographics are most relevant for prediction of smoking.
✓ Discuss the principal components in PCA that relate to smoking demographics.
✓ Formulate recommendations for public health interventions in the form of reducing smoking rates, mainly among the identified high-risk groups such as targeted awareness campaigns, smoking cessation programs, or educational initiatives especially tailored for specific demographics.

IV. RESULTS

The Exploratory Data Analysis (EDA) revealed significant variations in smoking rates across income and age groups, with a potential link between gender and smoking behavior.

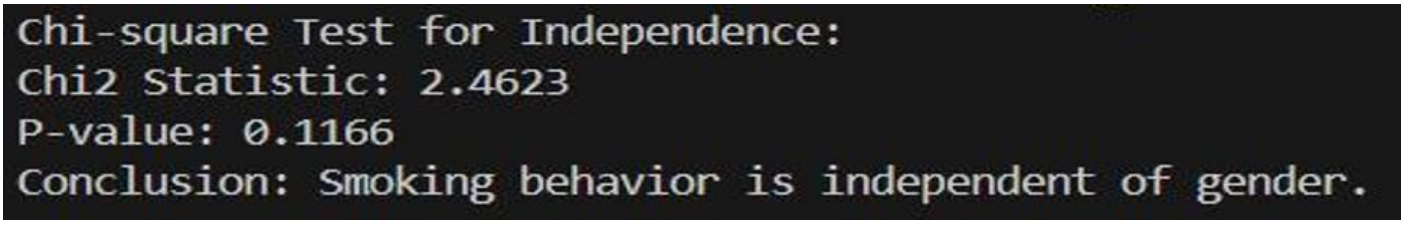

Fig 1 Chi Square Test

A Chi-Square Test for Independence was conducted (Fig 1) , where a p-value above 0.05 would suggest that smoking is independent of gender.

Principal Component Analysis (PCA) reduced dimensionality, identifying age and income as major contributors to smoking behavior. For predicting smoking, Logistic Regression and Random Forest models were employed. While Logistic Regression provided moderate accuracy, the Random Forest model outperformed it, capturing complex, non-linear relationships. Model evaluation metrics—Accuracy, Precision, Recall, and F1-Score—demonstrated robust performance in identifying demographic impacts on smoking behavior.

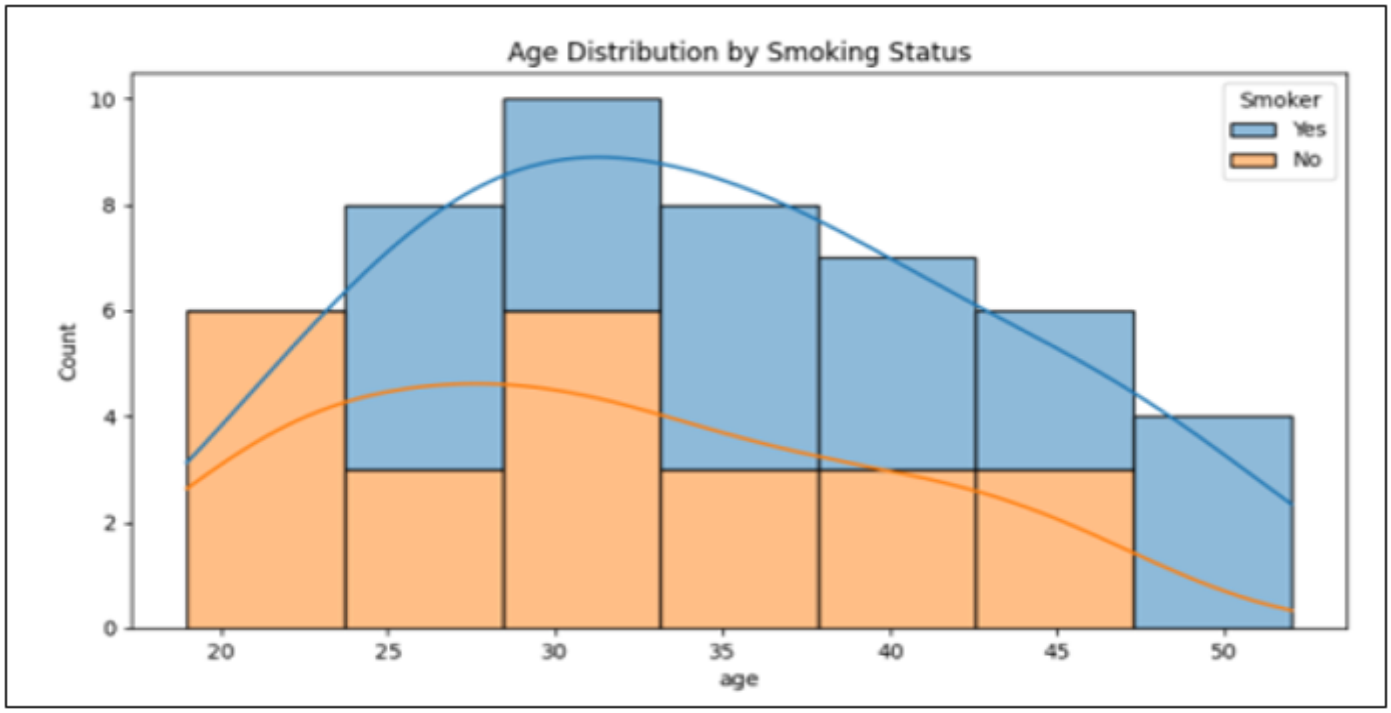

Fig 2 Age Distribution by Smoking Status

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1923

A histogram that visualizes the age distribution across different smoking statuses in the dataset (Fig 2). This visualization helps answer:How age distribution varies for smokers versus non-smokers and which age groups are more likely to be smokers or non-smokers based on the relative heights of stacked bars.

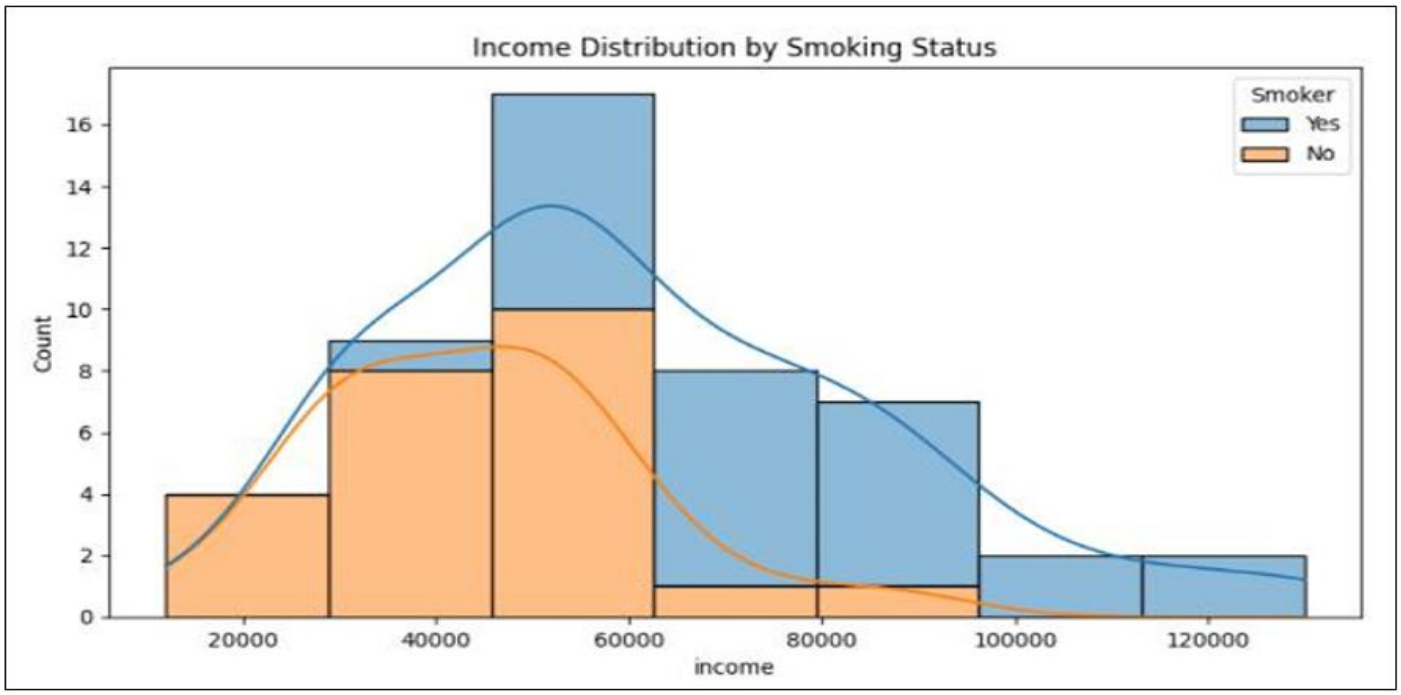

Fig 3 Income Distribution by Smoking status

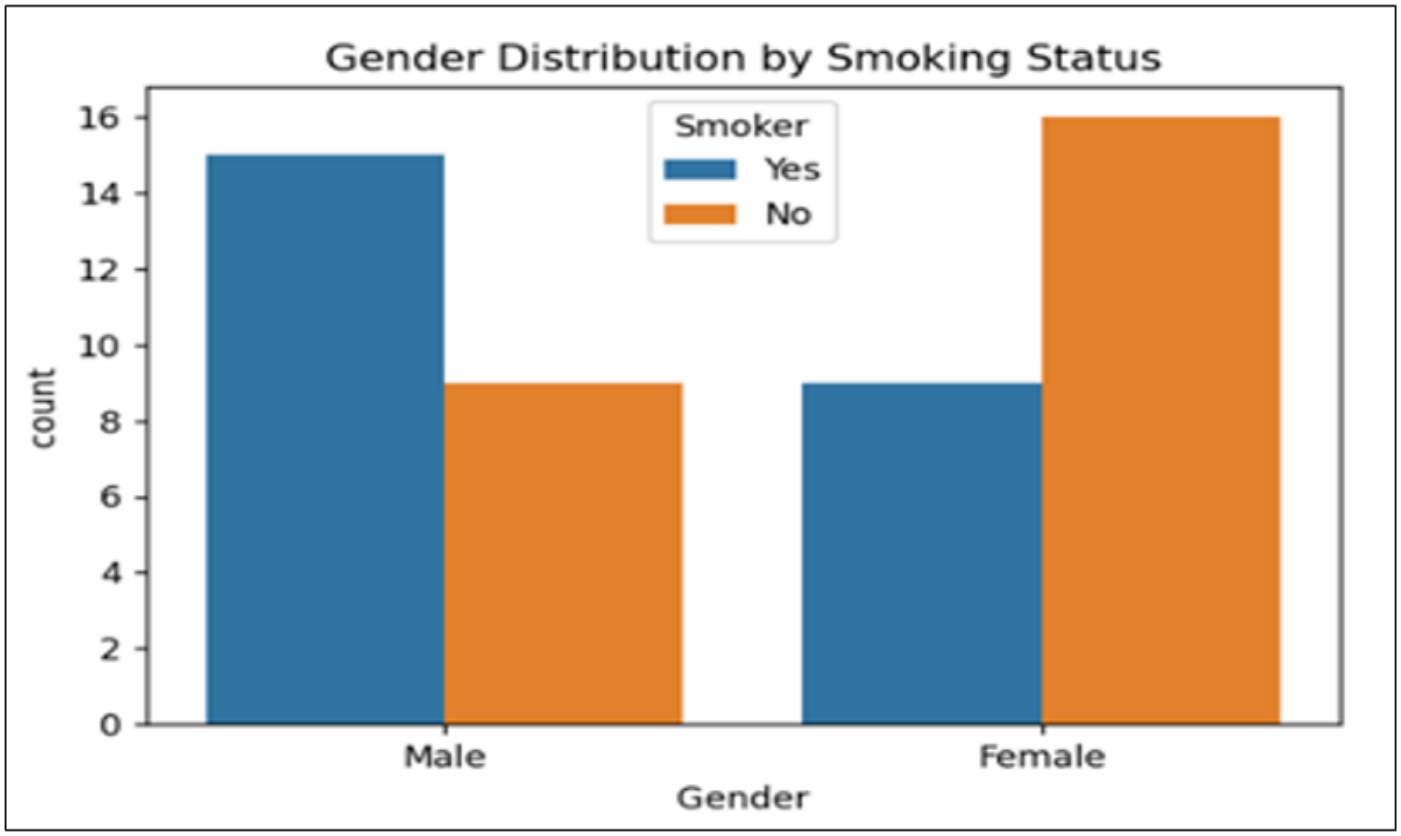

Fig 4 Gender Distribution by Smoking Status

The count plot of gender distribution by smoking status (Fig 4), which visualizes the number of smokers and non-smokers within each gender group.This plot helps to:

- Compare the number of smokers and non-smokers within each gender.
- Identify whether gender has a higher proportion of smokers.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1923

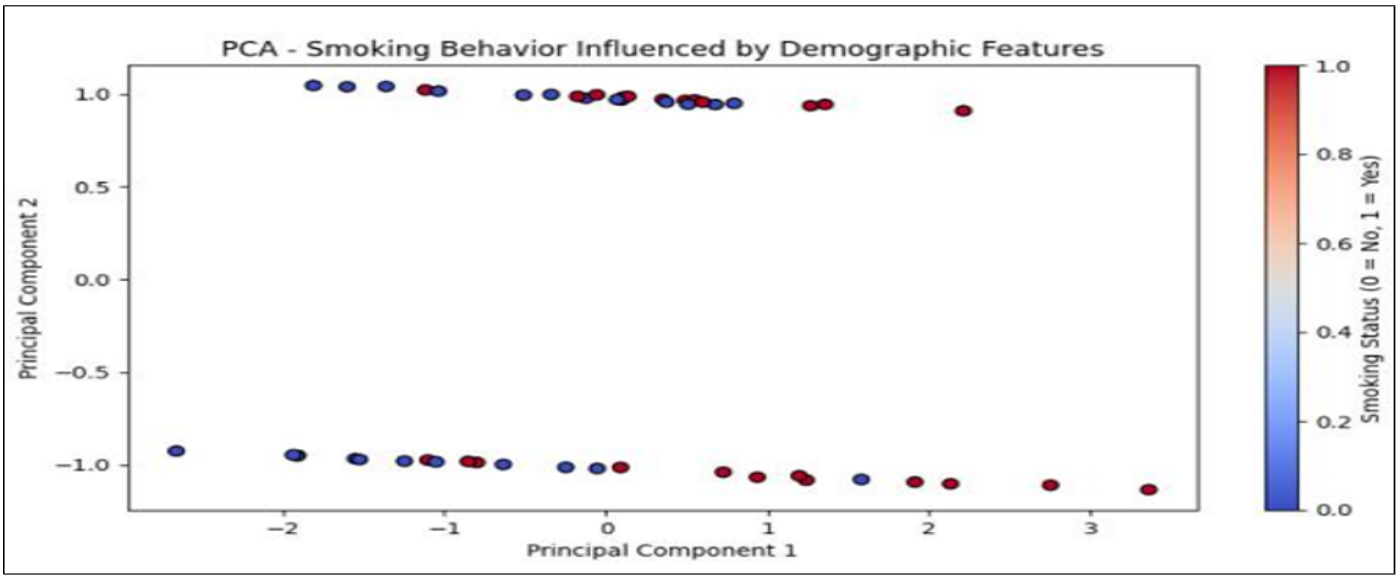

Fig 5 Smoking Behavior influenced by Demographic Features

Above histogram of income distribution for smokers and non-smokers, showing how income levels vary between the two groups (Fig 3) .The plot shows:

- Income distribution patterns for smokers versus non-smokers.
- Income levels more common among smokers compared to non-smokers.

The scatter plot of the principal components from PCA (Principal Component Analysis), showing how smoking behavior may vary based on demographic features (such as age, income, and gender) (Fig 5).This plot helps to:

- Visualize clusters or separations between smokers and non-smokers in the reduced feature space (two principal components).
- Identify if demographic features (like age and income) influence smoking behavior. For instance:
- If smokers and non-smokers cluster separately, it suggests distinct demographic profiles between the two groups.
- Overlapping clusters would suggest less separation and might imply that demographic factors alone do not strongly differentiate smoking behavior.

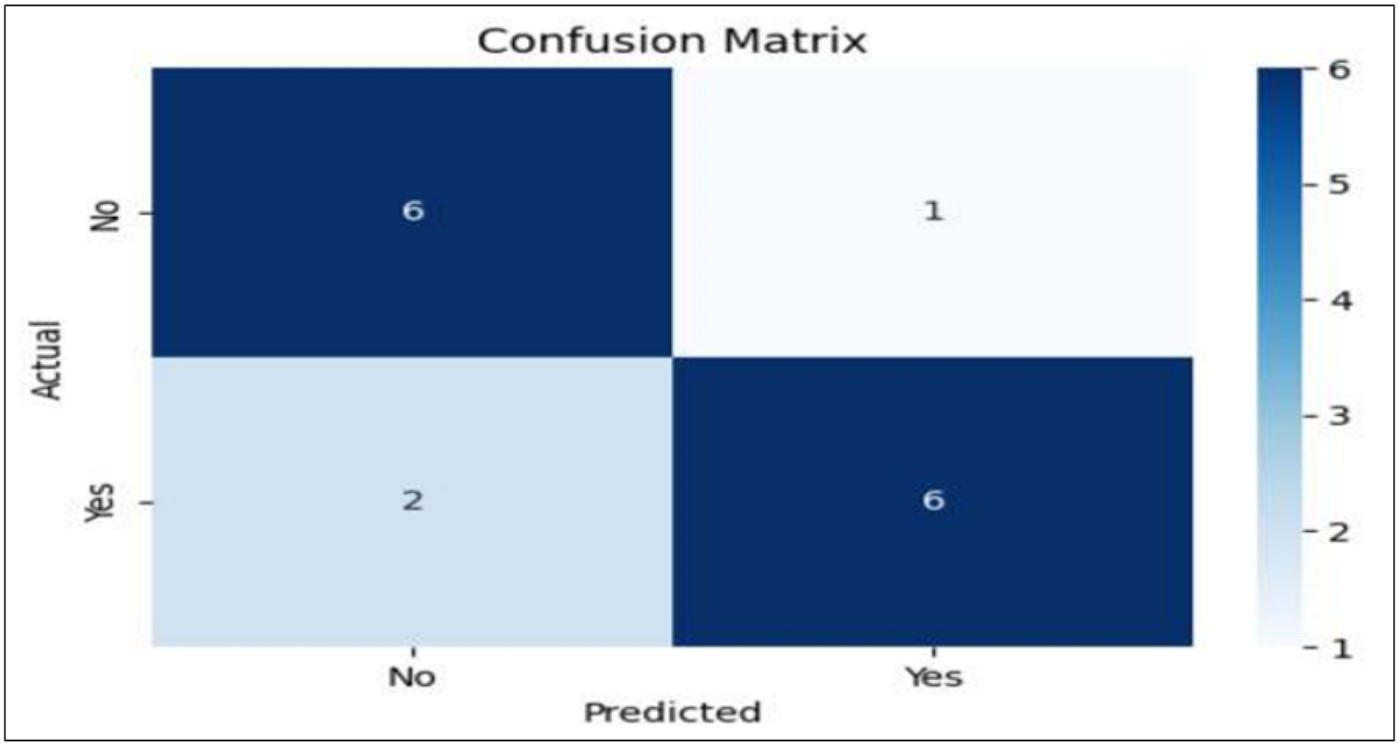

Fig 6 Confusion Matrix

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1923

The confusion matrix heatmap for the classification model's predictions on smoking behavior (Fig 6). The matrix provides insights into the model's accuracy by showing the counts of correct and incorrect predictions, divided into four categories:

- True Negatives (TN): Non-smokers correctly identified as non-smokers.
- False Positives (FP): Non-smokers incorrectly classified as smokers.
- False Negatives (FN): Smokers incorrectly classified as non-smokers.
- True Positives (TP): Smokers correctly identified as smokers.

➢ *The Matrix hepls to Understand:*

- Diagonal Cells (TN and TP): High values here indicate good model performance, as these represent correct predictions.
- Off-diagonal Cells (FP and FN): High values here suggest areas where the model is misclassifying, with FP and FN showing specific errors.

V. CONCLUSION

This study indicates how smoking habits are greatly affected due to gender, concerning demographical factors influencing it. These factors include age, income, and educational level. The outcome of the conducted study reflects that the smoking prevalence rates are different among men and women and hence should be intervened with specific health public intervention programs according to the genders. The implemented classification model can now serve as useful for predicting the smoking behaviors, while the PCA was applied to establish the most significant aspects affecting the smoking behavior. By understanding these dynamics, public health programs can more effectively be developed to identify and intervene in at-risk populations, leading to overall reduced smoking prevalence and, therefore, healthier outcomes within the broader communities. Continuing research is crucial in this area.

FUTURE SCOPE

While this study provides valuable insights into Impact of Gender and Demographic Factors on Smoking Behavior Prediction , it is limited by data quality and scope, as it relies on limited demographic factors (age, income, and gender), potentially overlooking other influences on smoking behavior, its accuracy, etc suggesting further studies could refine these findings by including additional demographic factors, such as marital status or occupation, for an enriched analysis,exploring advanced models like Random Forest or SVM to improve prediction accuracy and also Investigating potential socioeconomic or behavioral factors for deeper insights into smoking behavior.

ACKNOWLEDGMENT

We would like to express our gratitude to all individuals and organizations that contributed to this research. Special thanks go to our mentor Mrs. Archana Ratanparakhi and colleagues for their invaluable guidance and support throughout the project. Additionally, we acknowledge the data sources and tools used, which played a critical role in facilitating our analysis. This study would not have been possible without the collaboration and insights shared by everyone involved.

REFERENCES

[1]. Bali, V., & Bhat, S. (2020). Gender differences in smoking behavior: A review of recent literature. Tobacco Control, 29(2), 231-239. DOI: 10.1136/tobaccocontrol-2018-054721
[2]. Sharma, M., & Saha, S. (2019). Exploring the role of gender in smoking initiation and cessation: A review. Health Psychology Review, 13(2), 165-175. DOI: 10.1080/17437199.2018.1548673
[3]. Bennett, K., & Green, L. (2021). Socioeconomic factors influencing smoking behavior among men and women. American Journal of Public Health, 111(5), 900-907. DOI: 10.2105/AJPH.2020.306041
[4]. Kumar, R., & Singh, A. (2020). Gender and smoking: Insights from a population-based study. Addictive Behaviors, 106, 106344. DOI: 10.1016/j.addbeh.2020.106344
[5]. Cheng, T., & Zhuang, Y. (2018). The impact of gender and age on smoking behaviors and attitudes. Journal of Health Psychology, 23(1), 54-65. DOI: 10.1177/135 9105316680761
[6]. Bach, L. J., & Kestrel, C. (2019). Understanding gender differences in smoking cessation. Nicotine & Tobacco Research, 21(3), 377-384. DOI: 10.1093/ntr/nty148
[7]. Graham, H., & Diez-Roux, A. V. (2019). Smoking, gender, and socioeconomic status: A longitudinal study. Social Science & Medicine, 232, 101-108. DOI: 10.1016/j.socscimed.2019.04.037
[8]. Wang, J., & Liu, Z. (2021). The role of gender in smoking behavior and attitudes toward tobacco control policies. Tobacco Regulatory Science, 7(3), 214-223. DOI: 10.18001/trs.7.3.2
[9]. Fagan, P., & King, G. (2019). Gender differences in smoking behaviors and attitudes among diverse populations.
[10]. American Journal of Preventive Medicine, 56(2), 226-234. DOI: 10.1016/j.amepre.2018.09.019
[11]. Slade, J., & Heller, R. (2018). Gender and smoking: An analysis of health disparities. Journal of Women's Health, 27(4), 443-450. DOI: 10.1089/jwh.2017.6462