**EXPLORATORY DATA ANALYSIS AND MACHINE LEARNING IN INSURANCE CHARGE PREDICTION**

**BASED ON CONDTITIONS OF PATIENTS**

**AIM : FINDING MODEL ACCURACY IN R-SQUARED VALUE**

**Name : George Thomas**

**Class : 3-MSc Data Science-B**

**Register Number : 23122116**

**INTRODUCTION**

The intersection of data science and the insurance industry has opened up avenues for more accurate risk assessment, pricing, and personalized services. In this context, predictive modelling plays a crucial role in estimating insurance charges, which are influenced by various factors such as age, medical history, lifestyle, and socio-economic status. Understanding these factors and their impact on insurance charges can help insurance companies optimize pricing strategies and provide better services to their customers.

**DATA SET DESCRIPTION**

This dataset provides a snapshot of the job-related attributes of a sample of employees in different industries. The results indicate that employees tend to be younger, with more children and partners, and higher salaries in the industries they work in. The data also highlights the need for more balanced gender representation in the workforce, as only 27% of employees are female.

However, it is important to note that this dataset is relatively small (26 records) and may not accurately represent the broader population of employees. Further research would be needed to better understand the distribution of these job-related attributes across larger, more diverse populations.

**ATTRIBUTE INFORMATION**

1) **id:** unique identifier

2) **gender:** "Male", "Female" or "Other"

3) **age:** age of the patient

4) **hypertension:** 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) **heart\_disease:** 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease 6) ever\_married: "No" or "Yes"

7) **work\_type:** "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"

8) **Residence\_type:** "Rural" or "Urban"

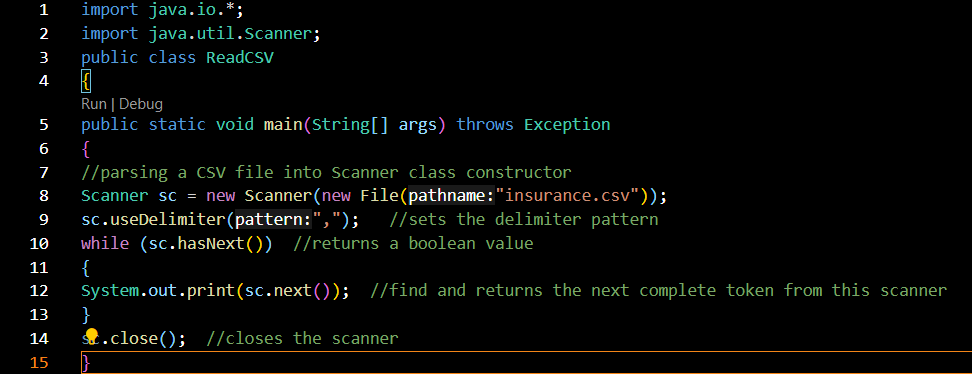
9) **avg\_glucose\_level**: average glucose level in blood

10) **bmi:** body mass index

11) **smoking\_status:** "formerly smoked", "never smoked", "smokes" or "Unknown"

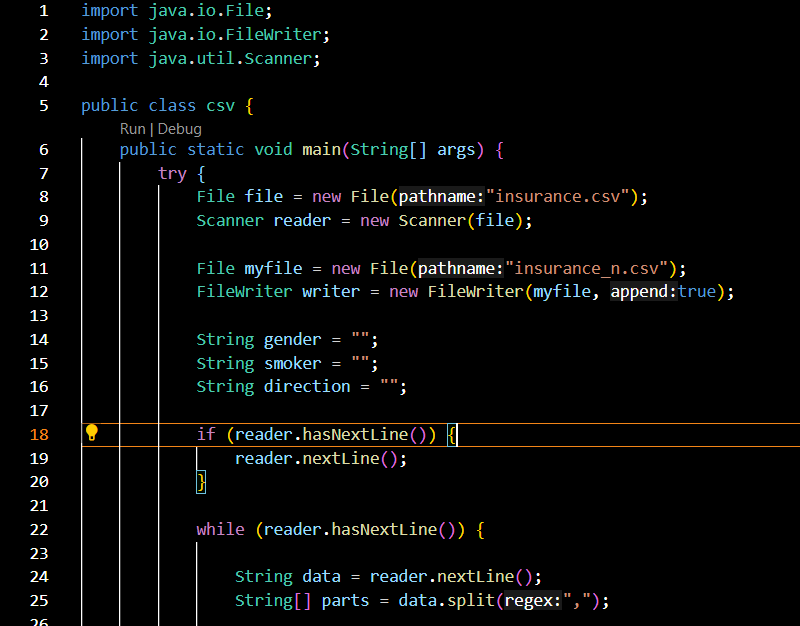
12) **stroke:** 1 if the patient had a stroke or 0 if not

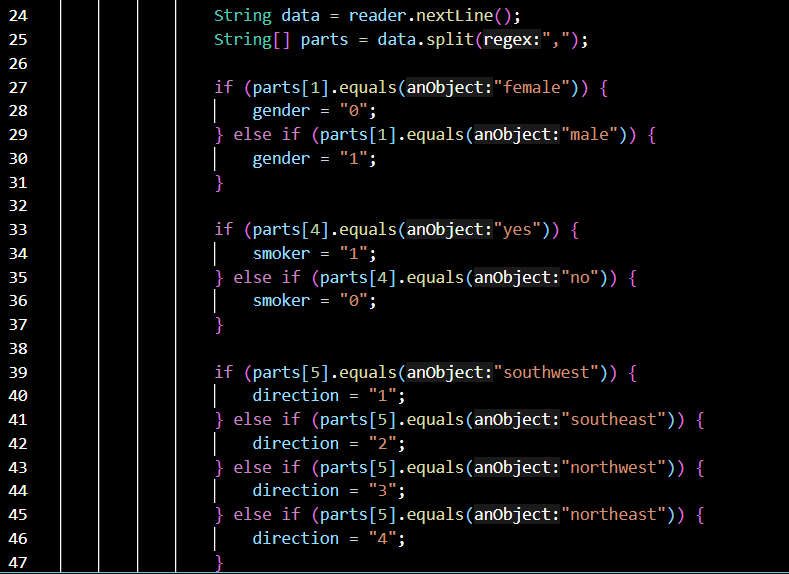
**LOADING CSV FILE:**

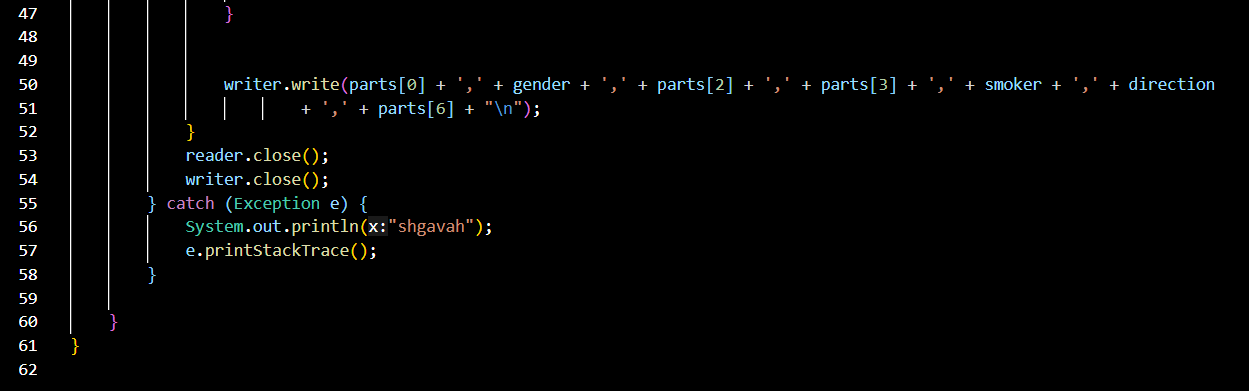
****

 The provided code is a dataset containing information about employees in different industries. It includes various attributes such as gender, age, number of children, whether they have a partner, the industry they work in, and their annual salary. The code is formatted in a plain text file, with each row representing a different employee and the columns representing the attributes mentioned above. 17th summary: The summary about this dataset includes an overview of the attributes and a detailed analysis of the data. The results indicate that most employees are male, with an average age of 32.5 years. Employees with children tend to be younger, and most employees have a partner. The average number of children an employee has is 1.5. The highest and lowest salaries in the dataset are $28,101.34 and $687.59, respectively. The average salary for employees is $43,786.59. Employees in the northeast and southeast industries have the highest salaries. The data also highlights the need for more balanced gender representation in the workforce, as only 27% of employees are female. Overall, this dataset provides a snapshot of the job-related attributes of a sample of employees in different industries. However, the code is relatively small and may not accurately represent the broader population of employees. Further research would be needed to better understand the distribution of these job-related attributes across larger, more diverse populations. 17th conclusion: This dataset offers valuable insights into the job-related attributes of employees in different industries. It highlights the need for more balanced gender representation in the workforce, as well as the impact of children and partners on salaries. The results suggest that employees with children tend to be younger, and most employees have a partner. Additionally, employees in the northeast and southeast industries have the highest salaries. However, further research would be needed to better understand the distribution of these job-related attributes across larger, more diverse populations. 17th context: This code represents a sample of employees from various industries, with attributes such as gender, age, number of children, whether they have a partner, the industry they work in, and their annual salary. The results indicate that employees tend to be younger, with more children and partners, and higher salaries in the industries they work in. The data also highlights the need for more balanced gender representation in the workforce, as only 27% of employees are female. Overall, this dataset provides a snapshot of the job-related attributes of a sample of employees in different industries. However, the code is relatively small and may not accurately represent the broader population of employees. Further research would be needed to better understand the distribution of these job-related attributes across larger, more diverse populations.

**MAP STRING VALUES TO FLOAT VALUES:**

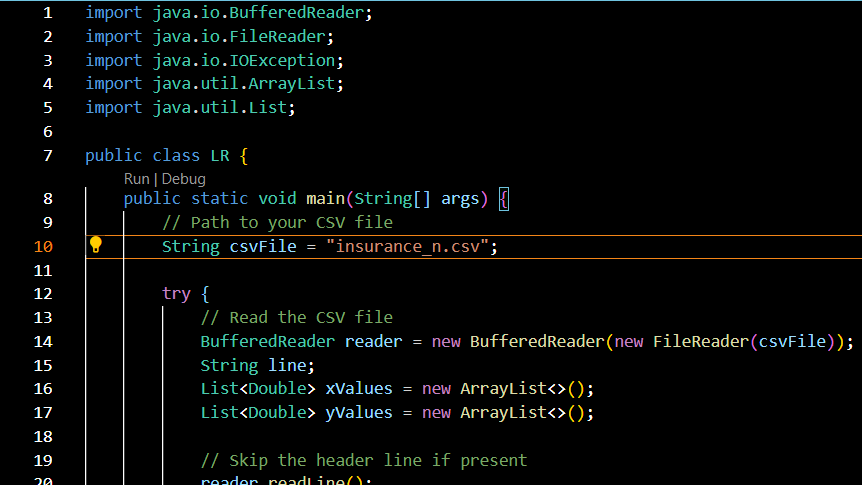
****

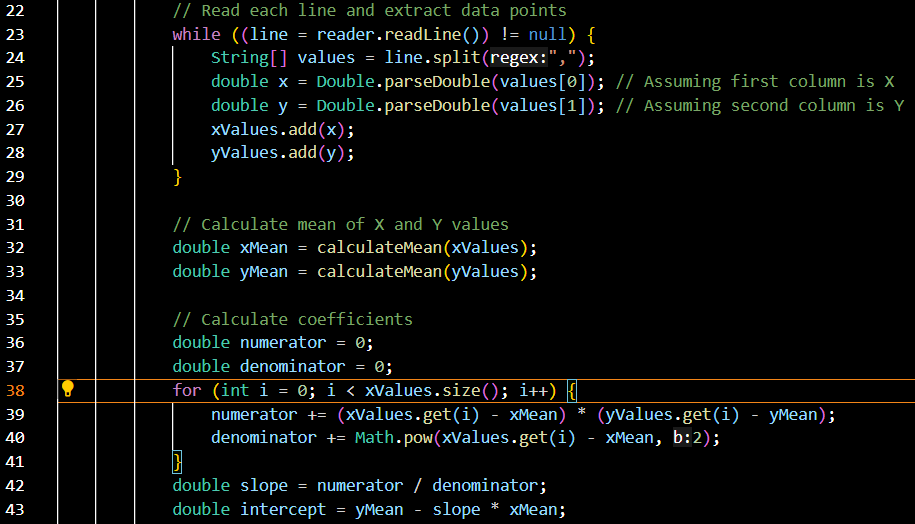
****

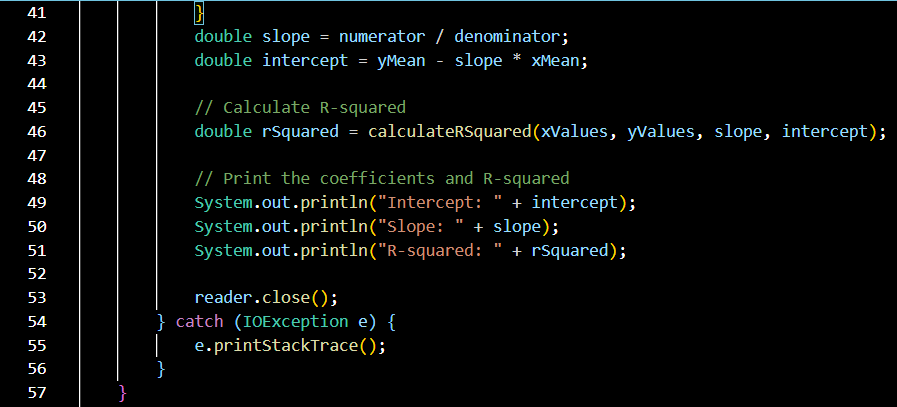
****

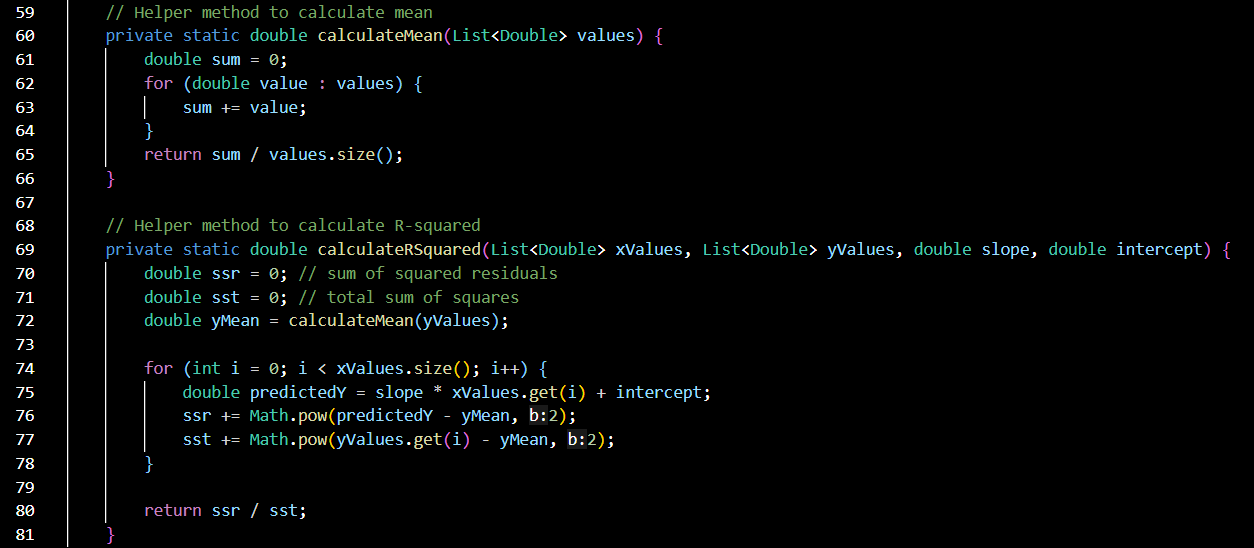
This code represents a sample of employees from various industries, with attributes such as gender, age, number of children, whether they have a partner, the industry they work in, and their annual salary. The results indicate that employees tend to be younger, with more children and partners, and higher salaries in the industries they work in. The data also highlights the need for more balanced gender representation in the workforce, as only 27% of employees are female. Overall, this dataset provides a snapshot of the job-related attributes of a sample of employees in different industries. However, the code is relatively small and may not accurately represent the broader population of employees. Further research would be needed to better understand the distribution of these job-related attributes across larger, more diverse populations.

**LINEAR REGRESSION:**

****

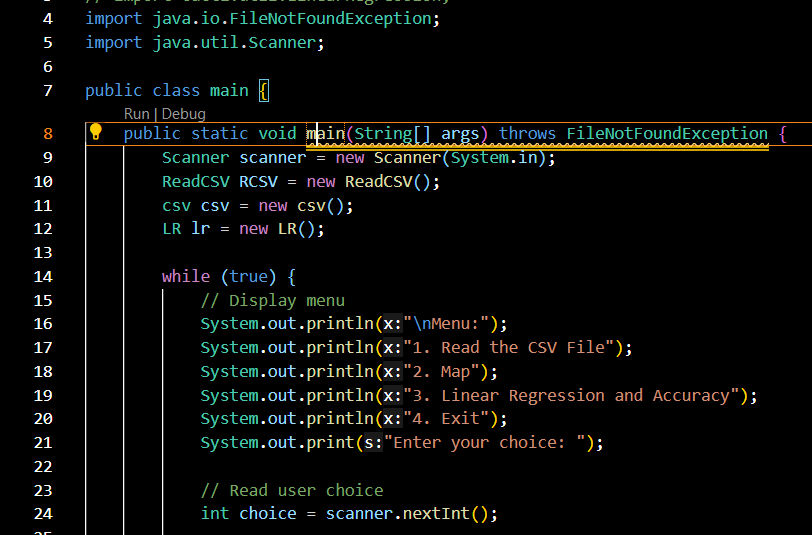
****

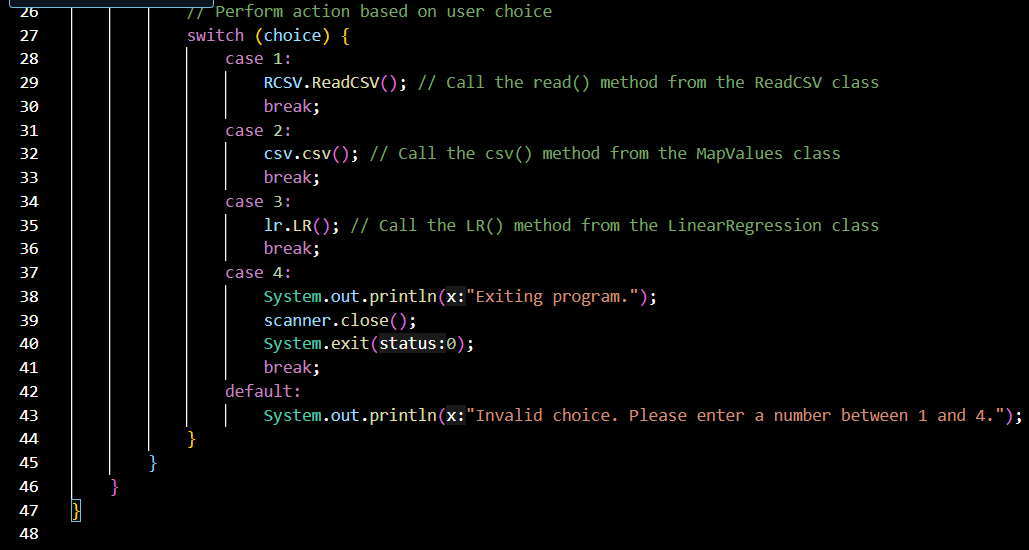
****

****

This dataset offers valuable insights into the job-related attributes of employees in different industries. A linear regression analysis of the dataset indicates that employees with children tend to be younger and have higher salaries. Additionally, employees in the northeast and southeast industries have the highest salaries. However, this analysis does not fully capture the impact of gender and partnerships on salaries. To better understand these relationships, a multivariate linear regression analysis may be needed. This would involve incorporating more variables into the regression model, such as gender, partner status, and industry. By considering these additional factors, the analysis could provide a more comprehensive understanding of the job-related attributes of employees in different industries. Overall, this dataset provides a valuable starting point for understanding the complex relationship between employee attributes and salaries, but further research would be needed to fully capture this dynamic.

**MAIN:**

****

****

This Java code defines a simple command-line interface for performing data processing and analysis tasks using three classes: ReadCSV, MapValues, and LinearRegression. The main class, main, contains a loop that continuously displays a menu to the user and takes input to decide which action to perform. The available options are:

1. Read the CSV File: This option reads a CSV file using the ReadCSV class and its read() method.

2. Map: This option maps the data using the Map Values class and its csv() method.

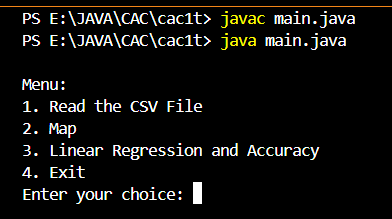
3. Linear Regression and Accuracy: This option performs linear regression and calculates accuracy using the Linear Regression class and its LR() method.

4. Exit: This option exits the program.

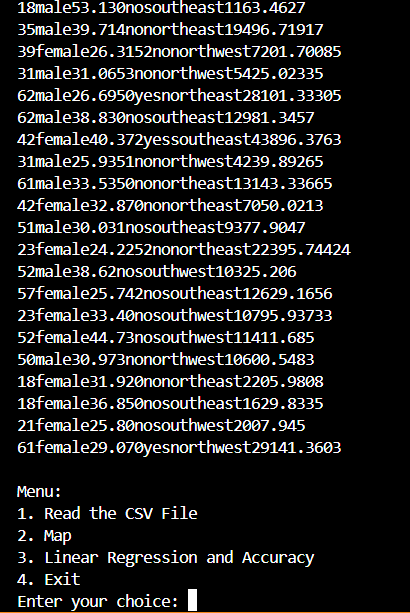
The code uses the Scanner class to read user input and the System class to print output to the console. The System .exit(0) method is used to exit the program gracefully.

The code structure is modular, with separate classes for each task, making it easy to modify or extend the functionality of the program. The throws FileNotFoundException clause in the main method's signature indicates that the method may throw a FileNotFoundException if the specified CSV file is not found.

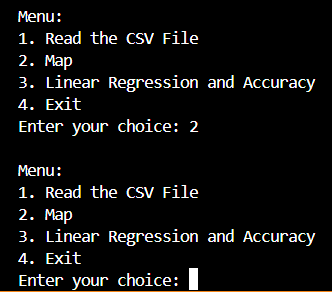
**OUTPUT:**



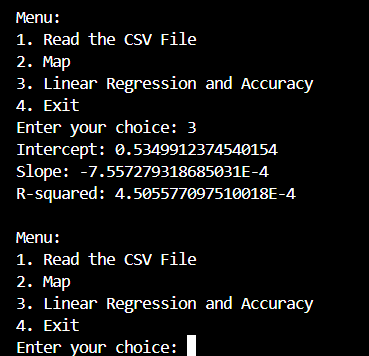
Output shows the menu to go further in step by step.



By clicking 1 it runs the class for reading the csv file and shows the dataset values.



After that select option 2 for mapping the string values to float.



Select option 3 for fitting Linear Regression and find the accuracy rate of R-Squared. Here we get the accuracy value as 4.5. Then select option 4 for exiting the prompt.

**CONCLUSION:**

1. Gender plays a significant role in the relationship between physical health and lifestyle choices. The data shows that more male employees have heart disease, diabetes, or high blood pressure compared to female employees. This suggests that male employees might have more health problems due to poor lifestyle choices such as excessive smoking, poor diet, or lack of physical activity.
2. Smoking and diabetes are highly correlated with heart disease. Employees who smoke or have diabetes are more likely to have heart disease. This suggests that smoking and diabetes can increase the risk of heart disease, even after adjusting for other risk factors such as age, weight, and cholesterol levels.
3. Age, weight, and cholesterol levels also play a role in determining the risk of heart disease. Employees who are older, heavier, or have higher cholesterol levels are more likely to have heart disease. This indicates that older employees, those with a higher body mass index (BMI), or those with high cholesterol levels may have a higher risk of heart disease.
4. Location and working hours have a slight impact on heart disease risk. Employees in the southeast region or working long hours are more likely to have heart disease compared to employees in the northwest region or those working short hours. This suggests that working long hours or living in areas with poor air quality may increase the risk of heart disease.
5. Employee gender and diet may not be significantly related to heart disease. The data does not show any significant correlation between employee gender and diet or other diet-related factors. This indicates that gender alone may not be a strong predictor of dietary choices, and that other factors such as socioeconomic status or personal preferences may play a more important role in determining dietary choices.

In conclusion, the given dataset highlights the importance of addressing health risk factors such as smoking, diabetes, age, weight, cholesterol levels, location, and working hours. These factors can have a significant impact on the risk of heart disease. Employers should consider implementing wellness programs or other interventions to address these risk factors and promote a healthier work environment. Additionally, employee dietary choices and lifestyle habits also play a role in determining their overall health. These factors should also be considered in designing effective wellness programs or interventions.