1 Chapter 1

1.1 Introduction

Smoking has been proven to negatively affect your health in a multitude of ways. Smoking and secondhand smoke can magnify current harmful health conditions, and has been linked as the cause for others. Smoking and secondhand smoke often trigger asthma attacks for persons suffering from Asthma, and almost every case of Buerger's disease has been linked to some form of tobacco exposure. Various forms of cancer are caused by smoking, secondhand smoke, and other tobacco products. In addition to being deemed the cause of certain cancers, most commonly known for causing lung and gum cancer, smoking and secondhand smoke also prevents the human body from fighting against cancer. Gum disease is often caused by chewing tobacco products, but continuing to smoke after gum damage can inhibit the body from repairing itself, including the gums. Smoking, secondhand smoke, and tobacco products are included in creating and preventing the recovery of the following additional diseases or health conditions: chronic obstructive pulmonary disease (COPD), diabetes, heart disease, stroke, HIV, mental health conditions such as depression and anxiety, pregnancy, and vision loss or blindness. The lungs are significantly impacted by smoking. A third of all cancer cases are brought on by smoking. For instance, it may have an impact on respiration, resulting in coughing and shortness of breath. Additionally, it raises the possibility of respiratory infections, which ultimately lowers quality of life. The consequences of smoking vary from person to person depending on how exposed they are to the smoking components. Smoking is not only a problem for public health, but it also costs countries a lot of money. Because of the severe harm that smoking causes to one's body, it should be prohibited in all public areas. Smoking can quickly render a body frail and result in a slow and painful demise. Smoking has an effect on a person's well being in addition to these grave health repercussions. The senses of taste and smell are affected. Additionally, it hinders one's capacity for physical activity.It also impairs your outward look by causing things like yellow teeth and wrinkled skin. You also have a higher chance of

developing anxiety or depression. Additionally, smoking has an impact on our relationships with friends, family, and coworkers. Most significantly, it is a costly habit. In other words, it comes at a high cost. Some people spend their little resources on smoking despite not having enough money to get by.

On the world smoking map, India occupies a very special place. As the second most popular country in the world, India's share of the global burden of smoking-induced disease and death is substantial. As the second-largest producer and consumer of tobacco in the world, the complex interplay of economic interests and public health commitments becomes particularly prominent in the Indian context. There is, therefore, an even greater need to examine the case for a comprehensive tobacco control program in such a setting.
India's tobacco and smoking problem is more complex than probably that of any other country in the world, with a large consequential burden of tobacco related disease and death. The prevalence of tobacco use among men has been reported to be high from almost all parts of India (more in rural than in urban areas). Women from most parts of India report smokeless tobacco use and the prevalence varies between 15 percent and 60 percent. Among 13 to 15-year-old school-going children, the current use of any tobacco product varies from 3.3 percent in Goa to 62.8 percent in Nagaland. In the late 1980s, the number of tobacco-attributable deaths in India was estimated as 630,000 per year. On conservative estimates, the tobacco-attributable deaths currently range between 800,000 and 900,000 per year. The cost of the tobacco -attributable burden of just three groups of diseases cancer, heart disease and lung disease was estimated as Rs 277.611 billion 1999. This increased to Rs 308.33 billion in the year 2002-2003.

According to estimates made by the WHO , Currently about 5 million people die prematurely Every year in the world due to the use of tobacco, Mostly cigarette smoking. These deaths are Currently divided somewhat evenly between Developed and developing countries. More Important is the fact that this epidemic of Disease and death caused by tobacco is Increasing very rapidly. By 2030, it is estimated That the number of premature deaths attribute- Able to tobacco would double to 10 million Deaths every year, with about 7 million of the Deaths taking place in developing countries. Among people alive today in the world, about 500 million would die prematurely due to Tobacco use and smoking; most of these are children and Young adults of today.

1.2 Motivation

Smoking is a widespread habit with severe health consequences that pose significant challenges to public health. Despite the known harmful effects, millions of people worldwide continue to smoke, leading to numerous preventable diseases and premature deaths. This project aims to explore in detail how smoking impacts the human body. The motivation behind conducting this project work lies in the urgent need to address the escalating health risks associated with smoking. Smoking is a leading cause of preventable death and disability globally, responsible for an alarming number of deaths each year due to cardiovascular diseases, respiratory disorders, and cancer. Understanding the specific physiological changes induced by smoking is vital for raising awareness and designing effective interventions to curb the smoking epidemic. This research serves as a powerful tool to motivate individuals to quit smoking, promote healthy lifestyles, and reduce the burden of smoking-related diseases on society.

1.3 Literature Review

In 2018, Charles Frank,Asmail Habach,Raed Seetan,Abdullah Wahbeh conducted survey on "Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis". The main objective is to investigate the viability and effectiveness of some machine learning algorithms for predicting the smoking status of patients based on their blood tests and vital readings results. They uses One-way ANOVA analysis with SAS tool to show the statistically significant difference in blood test readings between smokers and non-smokers. The results show that the difference in INR, which measures the effectiveness of anticoagulants, was significant in favor of non-smokers which further confirms the health risks associated with smoking and also they use five machine learning algorithms: Naive Bayes, MLP, Logistic regression classifier, J48 and Decision Table to predict the smoking status of patients. To compare the effectiveness of these algorithms we use: Precision, Recall, F-measure and Accuracy measures. The results show that the Logistic algorithm outperformed the four other algorithms with Precision, Recall, F-Measure, and Accuracy of 83 percent, 83.4 percent, 83.2 percent, 83.44 percent respectively.

In 2018, Abdel-Salam G et.al conducted survey on "Statistical analysis for the impact of smoking on the behavior and health of Qatari adolescents". This study focuses on exploring the patterns of tobacco use and its impacts on the adolescents by conducting a survey in different schools across Qatar. The questionnaire was administered in five schools, selected by proportional random sampling. The responses were recorded from the sample for general questions regarding interest in physical activities, relationship with family and friends, mental satisfaction, health, academics and access to cigarettes. To reach the aims of this research, several statistical techniques were used. They done descriptive analysis and testing for association to the regression analysis. The descriptive analysis was performed to obtain an overview of the responses recorded. Chi-squared (2)-tests were carried out to test the association between smoking status and student grades, family history and feeling lonely. Factor analysis was carried out prior to the logistic regression analysis, to avoid possible multi-collinearity problem; as a 50-item questionnaire was used for survey.

In 2011, D.K.Gautam et.al conducted a study on "Effect of cigarette smoking on the periodontal health status: A comparative, cross sectional study". The objective of the study was to evaluate the periodontal health status among cigarette smokers and non cigarette smokers, and oral hygiene measures. This study included 400 male

(200 cigarette smokers and 200 non smokers) aged 18-65 years. The subjects were randomly selected from the patients attending dental out-patient department of civil hospital and Himachal Dental College, Sundernagar. Community Periodontal Index (CPI) score was recorded for each patient and a questionnaire was completed by each patient. They used Chi square and t-test for statistical analysis. From the study they concluded that positive association was observed between periodontal disease and cigarette smoking. It was found that cigarette smoking was associated with lesser gingival bleeding and deeper pockets as compared to non-smokers.

In 2017, Maja Malenica et.al conducted study on "Effect of Cigarette Smoking on Haematological Parameters in Healthy Population". The objective of study was to assess the extent of adverse effects of cigarette smoking on biochemical characteristics in healthy smokers.Statistical analysis is conducted using SPSS version 20.0 (SPSS Inc.). Before statistical analysis, normal distribution and homogeneity of the variances were tested using Kolmogorov-Smirnov test respectively. Groups were compared using Student's unpaired t test for parameters with normal distribution or Mann–Whitney test for parameters with non-normal distribution. Correlations between parameters were analyzed using the Pearson R test for variables with normal distribution and the Spearman test for variables with non-normal distribution. Data are expressed as mean ± standard deviation or medians (interquartile range). P is less than 0.05 was considered significant. They concluded that continuous cigarette smoking has severe adverse effects on haematological parameters (e.g., hemoglobin, white blood cells count, mean corpuscular volume, mean corpuscular hemoglobin concentration, red blood cells count, hematocrit) and these alterations might be associated with a greater risk for developing atherosclerosis, polycythemia vera, chronic obstructive pulmonary disease and/or cardiovascular diseases.

1.4 Objectives

- To study the features of smokers and non smokers.
- To understand the association of smoking on blood sugar level and dental caries.
- To determine whether smokers cholesterol level influencing on blood pressure.
- To build a predictive model which identify diabetic category using smokers and non smokers body signal.

1.5 Scope of the study

Smoking has been proven to negatively affect health in a multitude of ways. Smoking has been found to harm nearly every organ of the body, cause many diseases, as well as reducing the life expectancy of smokers in general.
The main aim of the study is to make a meaningful contribution to the existing body of knowledge, with the potential to improve public health outcomes and contribute to the well-being of individuals and communities. It is essential to communicate the research effectively to both the scientific community and the general public to maximize its impact on health awareness and policy changes.

2 Chapter 2

2.1 Materials and Methods

The secondary data is selected from Kaggle. It has 55692 rows and 26 columns. The data providing information about both smokers and non-smokers, including anthropometric measurements, lipid panels, liver function tests, and dental conditions. By examining the changes in health conditions due to smoking, we aim to shed light on the detrimental effects of tobacco use on the human body.

2.2 About the Data:

ID : serial number given to a person.
Gender: Gender of person.
Age: 5-years gap age groups.
Height:Height of the person (in cm).
Weight: Weight of the person (in kg).
Waist : Waist of the person in circumference length.
Eyesight(left): Visual acuity of left eye .
Eyesight(right): Visual acuity of right eye.
Hearing(left): Auditory perception of the person.
Hearing(right): Auditory perception of the person.
Systolic: The blood pressure when the heart is contracting.
Relaxation: The blood pressure when the heart is relaxing.
Fasting blood sugar: Blood sugar level before meals.
Cholesterol: A compound of the sterol type found in most body tissues.
Triglyceride: Type of fat (lipid) found in your blood.
HDL: Type of Cholesterol.
LDL: Type of cholesterol.
Hemoglobin: A red protein responsible for transporting oxygen in the blood.
Serum creatinine: Blood test that measures the amount of creatinine in the blood.
AST: Enzyme mostly found in the liver.
ALT: Enzyme mostly found in the liver.
Gtp: Energy transfer within the cell.
Oral: Oral Examination status.
Dental caries: Tooth decay or dental cavities.
Tartar: Dental plaque that can form on the teeth.

2.3 Statistical Methods:

The Python programming languages has been used to carry out the analysis.

2.3.1 Python:

It is the programming language used for analysis and building model.
Python libraries:
Numpy:
NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting,discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
Matplotlib:
Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.
Pandas:
Pandas is a Python package to work with structured and time series data. The data from various file formats such as csv, sql etc can be imported using Pandas. It is a powerful open source tool used for data analysis and data manipulation operations such as data cleaning, merging, selecting. Pandas mainly used for machine learning in from of dataframes. Pandas allows various data manipulation operations such as groupby, join, merge, data cleaning.
Seaborn:
Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on top matplotlib library and is also closely integrated with the data structures from pandas.It aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs so that we can switch between different visual representations for the same variables for a better understanding of the dataset.

Python Plotly:
Python Plotly Library is an open-source library that can be used for data visualization and understanding data simply and easily. Plotly supports various types of plots like line charts, scatter plots, histograms, cox plots, etc.

2.3.2 Graphical techniques:

A graph is basically an illustration of how two variables relate to one another. Some simple popular graphical techniques are pie chart, bar plot, count-plot. These are very informative, simple to understand and interpret.

- **Bar-Graph:**
  Bar graphs are the pictorial representation of data (generally grouped), in the form of vertical or horizontal rectangular bars, where the length of bars are proportional to the measure of data. They are also known as bar charts. It is one of the means of data handling in statistics.

- **Boxplot:**
  A box plot is a visual representation of the distribution of a dataset, showing the median, quartiles, and outliers. The box represents the interquartile range (IQR), the middle line is the median, and whiskers extend to the minimum and maximum values within 1.5 times the IQR. It is a concise way to understand the spread, central tendency, and presence of extreme values in the data. Box plots are commonly used in data analysis and exploration to compare multiple datasets and identify key features of their distributions.

- **Line chart:**
  A line chart or line graph, also known as curve chart, is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields. It is similar to a scatter plot except that the measurement points are ordered (typically by their x-axis value) and joined with straight line segments. A line

chart is often used to visualize a trend in data over intervals of time .

- **Pie chart:**
  A pie chart (or a circle chart) is a circular statistical graphic which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area) is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented..

2.3.3 Two sample T test:

The t test is a statistical hypothesis test used to determine whether there is a significant difference between the means of two groups or conditions.
The hypothesis under consideration are given as follows :

H_0 :There is no significant difference between the means of the two groups.
H_1 :There is a significant difference between the means of the two groups.

The test statistic is given as follows :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where,
$\bar{x}_1$ and $\bar{x}_2$ are the means of group 1 and group 2, respectively.
s_p is the pooled standard deviation, calculated as:

$$s_p = \sqrt{\frac{s_1^2 + s_2^2}{n_1 + n_2 - 2}}$$

s_1 and s_2 are the sample standard deviations of group 1 and group 2, respectively.
n_1 and n_2 are the sample sizes of group 1 and group 2, respectively.

If the calculated t-statistic is greater than the critical t-value (in absolute value), reject the null hypothesis. There is a significant difference between the means of the two groups.

If the calculated t-statistic is less than or equal to the critical t-value (in absolute value), fail to reject the null hypothesis. There is no significant difference between the means of the two groups.

2.3.4 Chi-square test for independence of attributes:

The chi-square test for independence is a statistical test used to determine whether there is a significant association between two categorical variables in a contingency table. It is commonly applied to test whether the occurrence of one categorical variable is related to the occurrence of another categorical variable. The test assesses whether the observed frequencies in the contingency table differ significantly from the frequencies that would be expected if the two variables were independent of each other.
The hypothesis under consideration are given as follows :
H_0 : There is no association between the two categorical variables.
H_1 : There is association between the two categorical variables.

The test statistic is given as follows :

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

where,
n_{ij} is the observed frequency in cell (i, j) of the contingency table.
e_{ij} is the expected frequency in cell (i, j) under the assumption of independence, calculates as:

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

If the calculated χ^2 value is greater than the critical value, reject the null hypothesis, and conclude that there is a significant association between the two categorical variables.
If the calculated χ^2 value is less than or equal to the critical value, fail to reject the null hypothesis, and conclude that there is no significant association between the two categorical variables.

2.3.5 Fisher exact test of independence:

Fisher's exact test for independence is a statistical test used to determine whether there is a significant association between two categorical variables. It is often employed when dealing with small sample sizes or when the assumptions of other tests, like the chi-square test, are not met. The test is based on a contingency table, which shows the counts of different combinations of the two categorical variables.

It calculates the probability of obtaining the observed distribution of counts or a more extreme distribution, assuming that the two variables are independent (i.e., not related).
Fisher's exact test starts with a null hypothesis, assuming that there is no association between the two categorical variables. It suggests that any observed differences are due to random chance.If the p-value is greater than the chosen significance level (often 0.05), we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest a significant association between the two categorical variables.

2.3.6 One-way ANOVA:

One-Way Analysis of Variance (ANOVA) is a statistical test used to compare means of three or more groups to determine if there are significant differences between them. It assesses whether the variation between group means is greater than the variation within each group. The null hypothesis assumes that all group means are equal, while the alternative hypothesis suggests at least one group mean differs significantly from the others. The test statistic is given as follows :

$$F = \frac{MS_{within}}{MS_{between}}$$

where,
F is the F-statistic, which follows an F-distribution.
MS_{within} is the Mean Square Between Groups, calculated as the sum of squares between groups divided by the degrees of freedom between groups.
MS_{within} is the Mean Square Within Groups, calculated as the sum of squares within groups divided by the degrees of freedom within groups.

2.3.7 Random Forest Classifier:

Random Forest Classifier is an ensemble learning algorithm used for classification tasks in machine learning. It constructs multiple decision trees during training, where each tree is trained on a different subset of the data created through bootstrapping. Additionally, it introduces randomness by selecting only a subset of features at each node in every tree. During prediction, the algorithm combines the outputs of these individual trees through voting, where the class with the most

votes becomes the final predicted class. Random Forest excels in handling diverse data types, large datasets, and noisy data. It mitigates overfitting by leveraging the collective decisions of multiple trees. Furthermore, it provides valuable insights into feature importance, indicating which features contribute most significantly to the model's predictions. Due to its accuracy, robustness, and versatility, Random Forest is widely used in various fields and applications.

2.3.8 Decision Tree Classifier:

A Decision Tree Classifier is a supervised machine learning algorithm used for both classification and regression tasks. It constructs a tree-like model by recursively splitting the dataset into subsets based on the most significant attributes, ultimately leading to the prediction of the target variable. The decision tree operates through a top-down approach, starting at the root node, where the entire dataset is considered. At each step, the algorithm selects the attribute that best separates the data based on certain criteria, such as Gini impurity or information gain. The selected attribute becomes the decision node, and the data is divided into subsets (branches) based on its values. The process continues iteratively, creating child nodes that represent further attribute splits until a stopping condition is met. This condition could be a predefined depth limit, minimum samples per leaf, or when no significant improvement in impurity/gain is achievable. To make predictions, a new data point traverses the decision tree from the root node down to a leaf node, following the path based on attribute values. The majority class in the leaf node is then assigned as the predicted class for the data point (in the case of classification tasks). Decision Tree classifiers are known for their simplicity, interpretability, and ability to handle both numerical and categorical data. However, they can be prone to overfitting, especially on noisy datasets. Ensemble methods like Random Forest and boosting algorithms like AdaBoost are often employed to improve the generalization and accuracy of Decision Tree classifiers. Overall, Decision Trees are valuable tools in various domains, including finance, medicine, and natural language processing.

2.3.9 XGBoost Classifier:

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm widely used for both regression and classification tasks. It belongs to the ensemble learning family and combines the predictions of multiple weak learners (typically decision trees) to create a robust and accurate model. The algorithm works by it-

eratively building and optimizing decision trees. In each iteration, the model places more emphasis on the data points that were misclassified in the previous iteration, leading to a strong ensemble of trees. To prevent overfitting, regularization terms are incorporated into the objective function during model training. XGBoost's key features include handling missing data, automatic feature selection, and its ability to handle large datasets efficiently. The algorithm's speed and performance make it a popular choice for various data science competitions and real-world applications, achieving state-of-the-art results with relatively less hyperparameter tuning compared to other algorithms.

2.3.10 CatBoost classifier:

CatBoost classifier is a powerful gradient boosting algorithm designed to handle categorical features effectively. It leverages ordered boosting and Bayesian priors to combat overfitting and improve generalization. CatBoost automatically handles categorical variables, avoiding the need for explicit encoding. It uses "Ordered Target Encoding" to convert categorical features into numerical representations, considering the target variable. This technique helps maintain the integrity of the data and improves model performance. The algorithm is efficient, scalable, and has a user-friendly API. CatBoost's ability to handle mixed data types, its excellent performance, and built-in visualizations make it a popular choice for various classification tasks. It excels in real-world applications, data science competitions, and scenarios where feature interpretation and accuracy are crucial.

3 Chapter 3

3.1 Results and discussion: Graphical Analysis

3.1.1 Smoking status distribution:

Table 1: Smoking status distribution

Smoking Status	Count
Yes	20455
No	35237

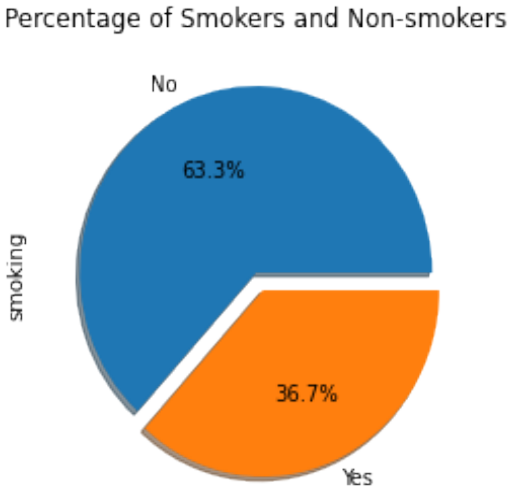

Figure 1: Pie chart for smoking status distribution.

From the pie chart we observe that 36.7% of persons are smokers and 63.3% of persons are non smokers

3.1.2 Smoking habits:Male vs Female

Table 2: Number of male and female smokers

Gender	Count
Male	19596
Female	859

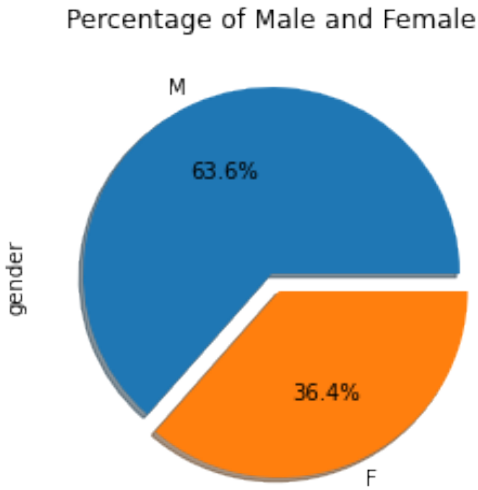

Figure 2: Pie chart of male and female smokers

From the pie chart graph we observe that 95.8% of persons are male and 4.2% of persons are Female.

3.1.3 Demographic overview of male and female smokers:

	Male	Female
Age	41	46
Height (in cm)	169	156
Weight(in kg)	71	56

Table 3: Demographic overview of male and female smokers

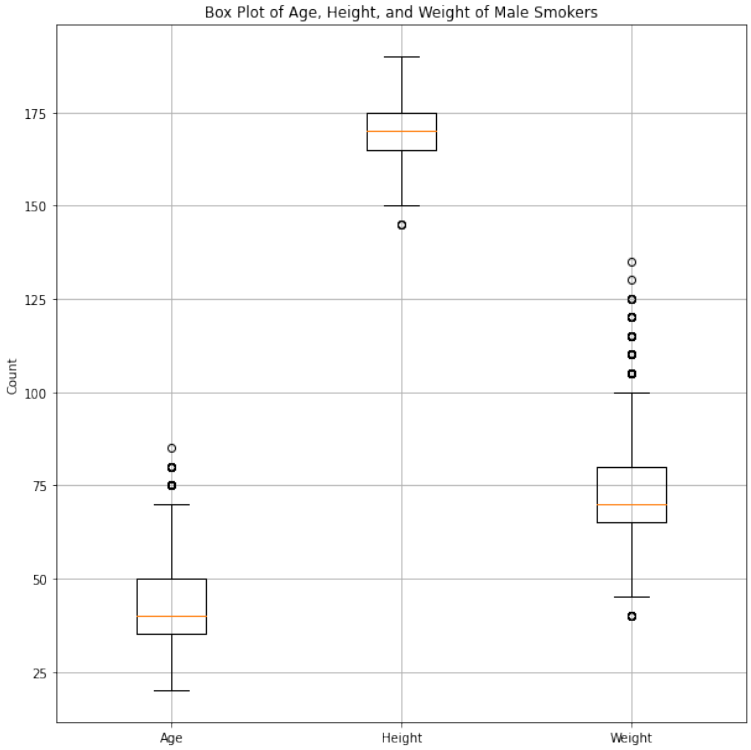

Figure 3: Box plot of age, height, weight of male smokers.

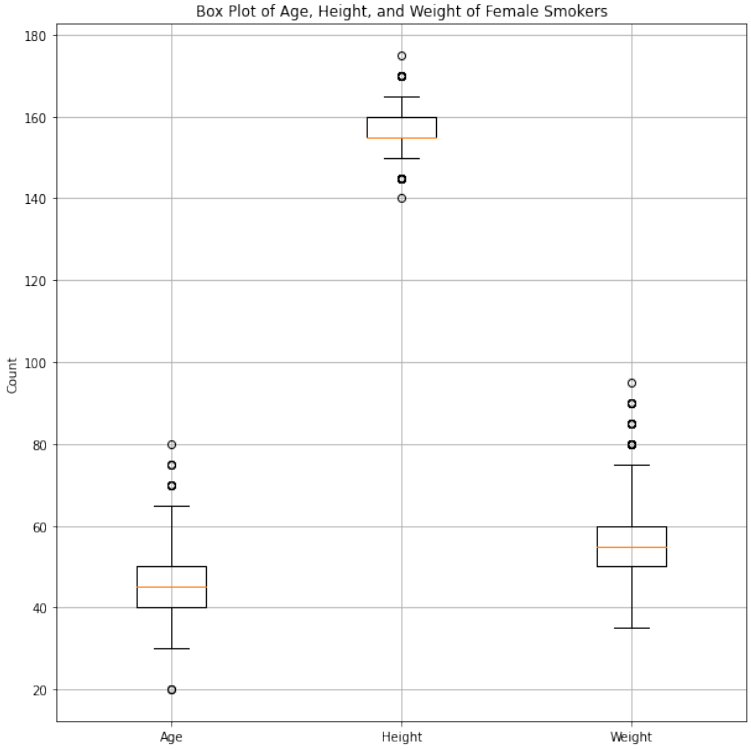

Figure 4: Box plot of age, height, weight of Female smokers.

The above boxplots shows that average age, height and weight of male smokers are 41, 169cm and 71kg respectively. Similarly average age, height and weight of female smokers are 46, 156cm and 56kg respectively.

3.1.4 Diabetic categories among Smokers and Non-Smokers:

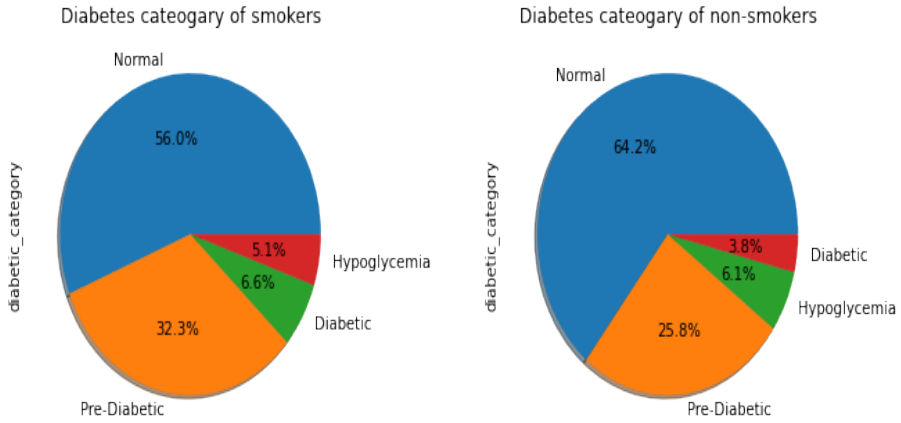

Figure 5: Pie Chart of Diabetic category in Smokers and Non-Smokers.

The above pie charts shows distribution of diabetic category of smokers and non smokers. As we can see that in diabetic category of smokers, 56% of peoples have Normal diabetics, 32.3% of peoples have Pre-diabetic level, 6.6% peoples have Diabetes and 5.5% people have Hypoglycemia. In diabetic category of non-smokers, 64.2% peoples have nomal diabetics, 25.8% peoples have Pre-diabetics, 6.1% peoples have Hypoglycemia and 3.8% peoples have diabetes. The graphs demonstrates that more number of non smokers have normal diabetes as compared to smokers.

3.1.5 Comparison of systolic blood pressure categories: Smokers vs Non-Smokers

Blood pressure category	Smoker	Non smoker
Hypotension	0.4%	1.0%
Normal	47.3%	53%
Elevated	27.6%	24%
Hypertension stage 1	18.4%	15.6%
Hypertension stage 2	4.1%	4.2%
Hypertension Crisis	2.1%	2.2%

Table 4: Systolic blood pressure of smokers and non smokers

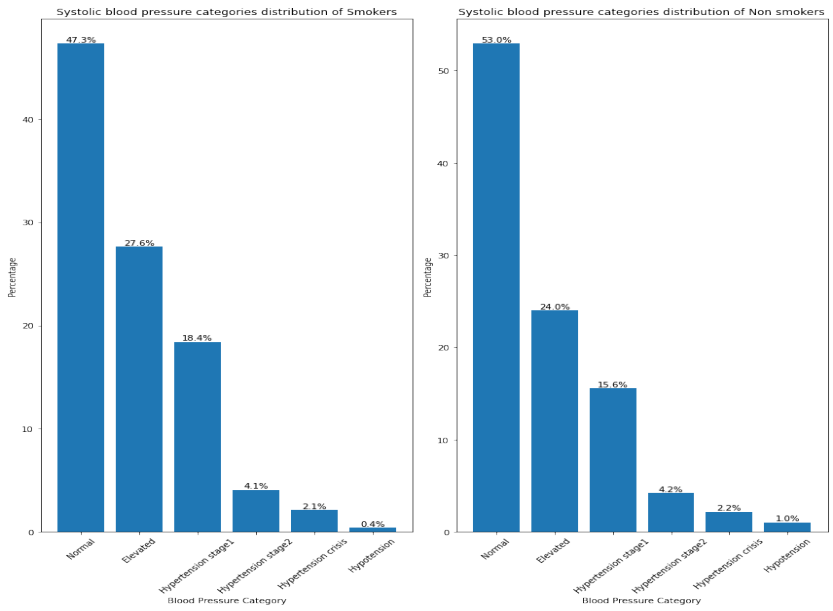

Figure 6: Bar graph of systolic blood pressure category of smokers and non smokers.

From smokers systolic blood pressure bar graph we observe that 47.3% of persons have normal blood pressure, 27.6% of persons have elevated blood pressure, 18.4% of persons have hypertension stage1, 4.1% of persons have hypertension stage 2 and 2.1% of persons have hypertension. Also from non smokers bar graph we observe that 53% of persons has normal blood pressure, 24% of persons have elevated blood pressure, 15.6% of persons have hypertension stage1, 4.2% of persons have hypertension stage2 and 2.2% of persons have hypertension crisis.

3.1.6 Comparison of Relaxation blood pressure categories: Smokers vs Non-Smokers

Blood pressure category	Smoker	Non smoker
Hypotension	0.1%	0.0%
Normal	64.2%	67.4%
Elevated	25.7%	20.8%
Hypertension stage 1	4.4%	7.2%
Hypertension stage 2	4.3%	3.8%
Hypertension Crisis	1.3%	0.8%

Table 5: Relaxation blood pressure of smokers and non smokers

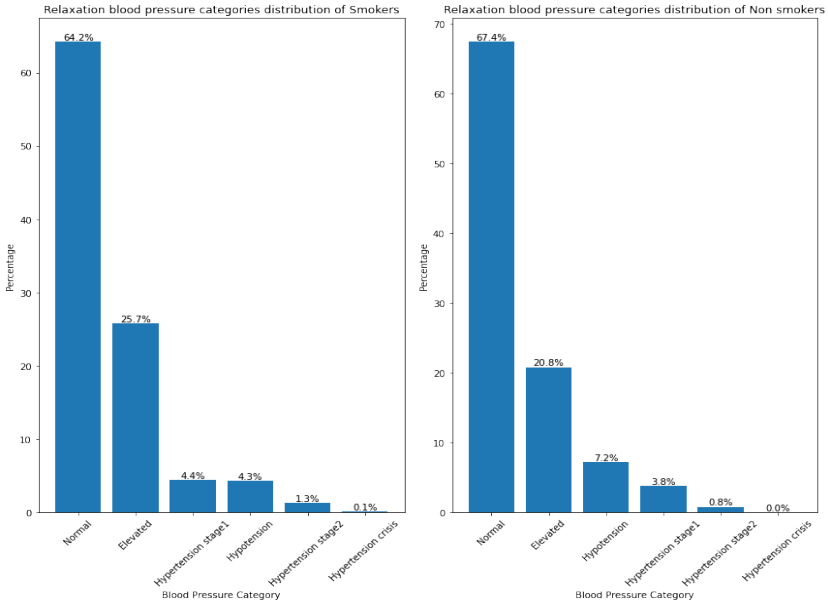

Figure 7: Bar graph of relaxation blood pressure category of smokers and non smokers.

From smokers relaxation blood pressure bar graph we observe that that 64.2% of persons have normal blood pressure, 25.7% of persons have elevated blood pressure, 4.4% of persons have hypertension stage1, 4.3% of persons have hypertension stage2, 1.3% of persons have hypertension crisis and 0.1% pf persons have hypotension who smokes. Also from non smokers bar graph we observe 67.4% of persons has normal blood pressure, 20.8% of persons have elevated blood pressure ,7.2% of persons have hypertension stage1, 3.8% of persons have hypertension stage2, 0.8% of persons have hypertension crisis and nobody have hypotension.

3.2 Results and discussion: Statistical Tests

3.2.1 To test the association between smoking status and dental caries using Chi-square test.

The following table shows the smokers and non smokers who have dental caries.

Table 6: Relationship between smoker status and Dental caries.

Smoking	Dental caries	
	Yes	No
Yes	28862	6375
No	14949	5506

The Hypothesis are as follows:

H_0 : The smoking status and dental caries are independent.
H_1 : There smoking status and dental caries are dependent.

The obtained values are as follows:

- Chi-Square test statistic :600.1881467493236
- Degrees of freedom : 1
- p-value : 1.5236163632659182e-132
- Cramer V value = 0.103857

Here we observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that there is relationship between the Smoking status and Dental caries.
From the Cramer V value, we can say that the strength of association between smoking status and dental caries is 0.10, which indicates that week association between these two variables.

3.2.2 To test whether smoking influencing on systolic blood pressure using Mann Whitney U test.

Assumption of t test, i.e normality and homogeneity of variance does not met. Hence we can use Mann Whitney U test.

To check for normality, Histogram is applied. The graphs are as follows:

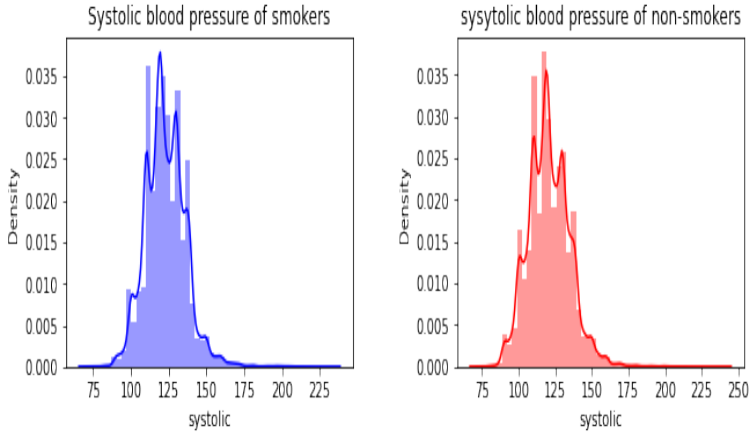

Figure 8: Histogram to check normality of systolic blood pressure level.

The following graph shows that normality does not exist.

To check the homogeneity of variance, Levene's test is used. The hypothesis are as follows:

H_0 : The variance of the groups are equal.

H_1 : The variance of the groups are not equal.

The obtained values are :

- Statistic= 83.1821
- p-value= 7.724588794620917e-20

We observe that the p-value is less than 0.05.Thus, we reject the null hypothesis. Hence we conclude that the variance of the groups are not equal. Hence we can use non parametric alternative, Wilcoxon Mann-Whitney U test to carryout the analysis.

The hypotheses of Mann Whitney U test are as follows:

H_0: Median systolic blood pressure of smoker and non smokers are same
H_1: Median systolic blood pressure of smoker and non smokers are not same.

The obtained values are as follows:

- W-Statistics = 393712768.0
- p-value = 2.900950687428121e-74

We observe that the p-value is less than 0.05. Thus, we reject the null hypothesis. Hence we conclude that smoking influencing on systolic blood pressure.

3.2.3 To test whether smoking influencing on relaxation blood pressure using Mann Whitney U test.

To check for normality, Histogram is applied. The graphs are as follows:

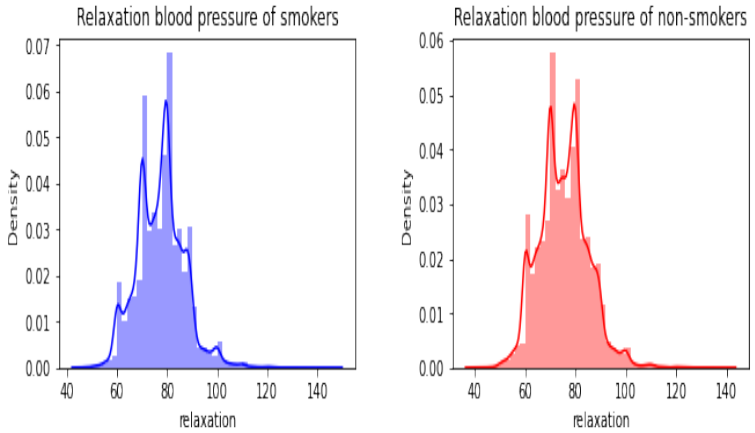

Figure 9: Histogram to check normality of relaxation blood pressure level.

The following graph shows that normality does not exist.

To check the homogeneity of variance, Levene's test is used.The hypothesis are as follows:

H_0 : The variance of the groups are equal.
H_1 : The variance of the groups are not equal.

The obtained values are :

- Statistic= 83.18215611787471
- p-value= 7.724588794620917e-20

We observe that the p-value is less than 0.05. Thus, We do reject the null hypothesis. Hence we conclude that the variance of the groups are not equal. Hence we can use non parametric alternative, Wilcoxon Mann-Whitney U test to carryout the analysis.

The hypotheses of Mann Whitney U test are as follows:

H_0: Median relaxation blood pressure of smoker and non smokers are same.
H_1: Median reelaxation blood pressure of smoker and non smokers are not same.

The obtained values are as follows:

- Statistic= 407760182.5
- p-value= 2.7663275867933146e-148

We observe that the p-value is less than 0.05. Thus, we reject the null hypothesis. Hence we conclude that smoking influencing on relaxation blood pressure of smoker.

3.2.4 Analysis of association between the smokers systolic blood pressure and fasting blood sugar level using Chi square test.

The following table shows the systolic blood pressure level and diabetes level of smokers.

	Diabetic category			
Blood pressure level	Hyperglycemia	Normal	Pre-diabetic	Diabetic
Hypotension	6	58	20	7
Normal	635	5908	2659	480
Eleveted	243	3120	1910	376
Hypertension stage1	138	1862	1440	326
Hypertension stage2	16	347	378	94
Hypertension crisis	6	58	20	7

Table 7: Smokers systolic blood pressure level and diabetes level.

The Hypothesis are as follows:

H_0 : The smokers systolic blood pressure and diabetes level are independent.
H_1 : The smokers systolic blood pressure and diabetes level are dependent.

The obtained values are as follows:

- Chi-Square test statistic: 533.069901464023
- Degrees of freedom: 15
- p-value: 5.640166840898203e-104
- Cramer V value: 0.0932033

Here we can observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that there is relationship between smokers systolic blood pressure level and diabetes level.
From the Cramer V value, we can say that the strength of association between smokers systolic blood pressure level and diabetes level is 0.09, which indicates that week association between these two variables.

3.2.5 Analysis of association between the smokers relaxation blood pressure level and diabetes level using Fisher's exact test of independence.

The following table shows the relaxation blood pressure level and diabetes level of smokers.

	Diabetic category			
Blood pressure level	Hyperglycemia	Normal	Pre-diabetic	Diabetic
Hypotension	71	570	197	35
Normal	798	7620	3966	755
Eleveted	147	2757	1930	432
Hypertension stage1	24	415	376	88
Hypertension stage2	9	94	118	36
Hypertension crisis	0	3	13	1

Table 8: Smokers relaxation blood pressure level and diabetes level.

The Hypothesis are as follows:

H_0: The smokers relaxation blood pressure and diabetes level are not dependent.
H_1: The smokers relaxation blood pressure and diabetes level are dependent.

The obtained values are as follows:

- p-value: 0.0004998
- Cramer V value: 0.0790805

Here we can observe that the obtained p-value is less than 0.05. Hence we reject the null hypothesis. And we conclude that there is relationship between smokers relaxation blood pressure level and diabetes level.
From the Cramer V value, we can say that the strength of association between smokers systolic blood pressure level and diabetes level is 0.07, which indicates that week association between these two variables.

3.2.6 To test the smokers diabetes level influencing cholesterol level using One way ANOVA.

To check the assumption of one way ANOVA, Histogram is used.The graph are as follows:

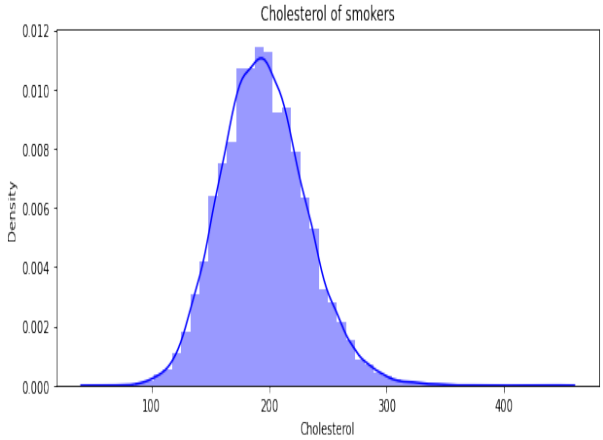

Figure 10: Histogram to check normality of Cholesterol.

The following graph shows the variable is approximately normal. Hence One way ANOVA is used.

The hypotheses of One way anova as follows:

H_0: Smokers diabetes level not influencing cholesterol level.
H_1: Smokers diabetes level influencing cholesterol level.

The obtained values are as follows:

- Test statistic: 56.1143
- p-value: 4.075895721427204e-36

Here p value is less than 0.05 hence we reject null hypothesis. Therefore smokers diabetes level influencing cholesterol level.

The following table obtained by Tukey's Kramer test which shows significant difference among the diabetes level.

Group1	Group2	Adjusted p-value	Decision
Diabetic	Hypoglycemia	0.001	Reject H_0
Diabetic	Normal	0.001	Reject H_0
Diabetic	Pre-diabetic	0.001	Reject H_0
Hypoglycemia	Normal	0.001	Reject H_0
Hypoglycemia	Pre-diabetic	0.001	Reject H_0
Normal	Pre-diabetic	0.001	Reject H_0

Table 9: Specific differences between diabetes level.

We observed that p-value of factors Hypoglycemia, Normal, Pre-diabetic and Diabetic is less than 0.05. Thus, we reject the null hypothesis. Hence we conclude that each diabetes level i.e Hypoglycemia, Normal, Pre-diabetic and Diabetic influencing cholestrol level in smokers.

3.3 Results and discussion: Machine Learning Models

In this section we discuss about various machine learning that build to predict diabetic level.
Data Splitting The data is splitted into train and test in the ratio 80:20. The shape of the data for train and test after splitting is as shown below:
x train shape: (44553, 20)
y train shape: (44553,)
x test shape: (11139, 20)
y test shape: (11139,)

Data transformation Diabetic category is the target variable which need to be predicted. Below transformation should be performed before fitting the model.

1. Standardising numeric data
2. Encoding categorical data

3.3.1 Model Building

Decision Tree Classifier:
From the model below observation are found

- Training set Accuracy: 0.6831
- Training set F1 score: 0.6792
- Training set Precision: 0.6721
- Training set Recall: 0.6866
- Testing set Accuracy: 0.6542
- Testing set F1 score: 0.6562
- Testing set Precision: 0.6584
- Testing set Recall: 0.6542

Here both test and train accuracy are low anf F1 score is also ranges betweeen 0 and 1, so we try some more models.

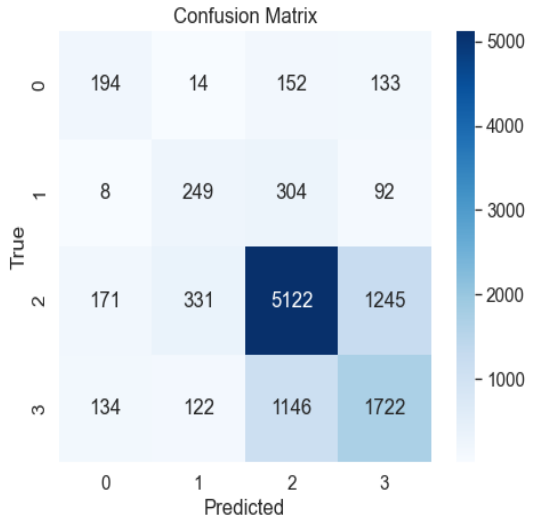

Figure 11: Confusion matrix for Decision Tree Classifier.

Random forest Classifier:
From the model below observation are found

- Training set Accuracy: 0.7661
- Training set F1 score: 0.7478
- Training set Precision: 0.7657
- Training set Recall: 0.7510
- Testing set Accuracy: 0.7455
- Testing set F1 score: 0.7207
- Testing set Precision: 0.7594
- Testing set Recall: 0.7455

From the above model performance, we conclude that train and test performance is good compare to Decision Tree Classifier. F1-Score is also very close to 1, so the model is better. There is no underfit and overfit present in the data.

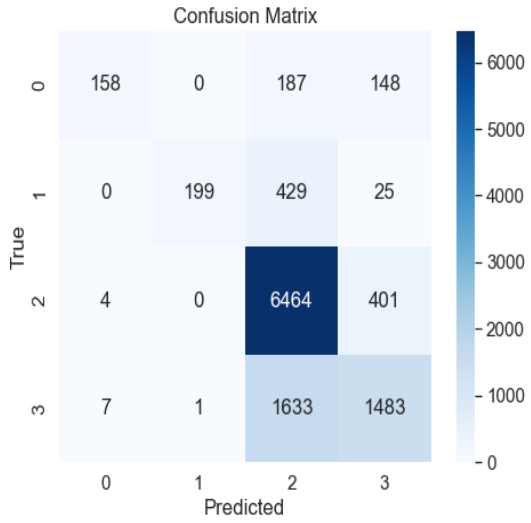

Figure 12: Confusion matrix for Random Forest Classifier.

Xgboost Classifier:
From the model below observation are found

- Training set Accuracy: 0.7346
- Training set F1 score: 0.700302
- Training set Precision: 0.760770
- Training set Recall: 0.734608
- Testing set Accuracy: 0.7034
- Testing set F1 score: 0.6983
- Testing set Precision: 0.7134
- Testing set Recall: 0.7064

Above model shows a decent level of performance with reasonably high accuracy, F1 score, precision, and recall on both the training and testing datasets. However, the slight drop in performance on the testing set compared to the training set indicate some overfitting.

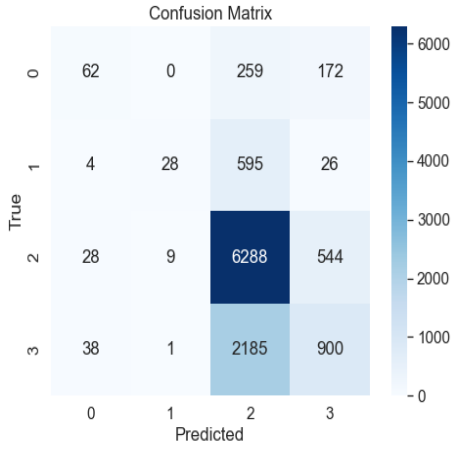

Figure 13: Confusion matrix for XGBoost Classifier.

Catboost Classifier:
From the model below observation are found Training set:

- Training set Accuracy: 0.7191
- Training set F1 score: 0.675938
- Training set Precision: 0.748221
- Training set Recall: 0.719076
- Testing set Accuracy: 0.6985
- Testing set F1 score: 0.6887
- Testing set Precision: 0.6882
- Testing set Recall: 0.6785

The Catboost classifier model demonstrates a reasonable level of performance, with accuracy, F1 score, precision, and recall being moderately high on both the training and testing datasets. However, the slight drop in performance on the testing set compared to the training set suggests that the model somewhat overfitting to the training data.

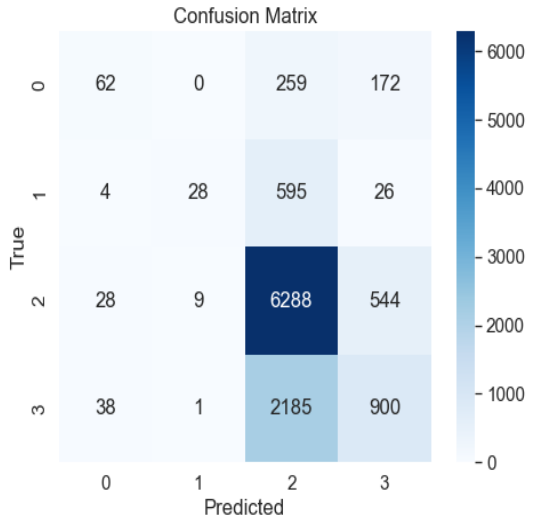

Figure 14: Confusion matrix for CatBoost Classifier.

3.3.2 Model comparison

The below table shows the comparison of all models.

Models	Train Accuracy	Test Accuracy	F1-score
Decision Tree	68.3%	65.4%	0.6562
Random Forest	76.6%	74.5%	0.7207
XGBoost	73.4%	70.3%	0.6983
CatBoost	71.9%	69.8%	0.6882

Table 10: Compression of all models.

Overall Random Forest model is the best-performing model among the four. It provides the highest test accuracy and F1-score, indicating better overall predictive performance and a good balance between precision and recall.

4 Chapter 4

4.1 Conclusion and Summary

4.1.1 Conclusion

1. Total 36.7% of persons are smokers and 63.3% of persons are non smokers. Among smokers, the majority (95.8%) are male smokers, while only a smaller percentage (4.2%) are female smokers.

2. The average age of male smokers (41 years) is lower compared to the average age of female smokers (46 years).

3. Among smokers, 56% have normal blood glucose levels, 32.3% are in the pre-diabetic category, 6.6% have diabetes, and 5.5% have hypoglycemia.

4. Among non-smokers, 64.2% have normal blood glucose levels, 25.8% are in the pre-diabetic category, 6.1% have hypoglycemia, and 3.8% have diabetes.

5. Total 47.3% of smokers have normal systolic blood pressure level and 53% of non smokers have normal systolic blood pressure level. So as compared to the smokers, more number of non smokers have normal systolic blood pressure level. Then 18.4% of smokers and 15.6% of non smokers have hypertension stage1. We say thet smokers have more chance of getting hypertension stage1.

6. Total 63.2% of smokers have normal relaxation blood pressure level and 67.4% of non smokers have normal relaxation blood pressure level. So as compared to the smokers, more number of non smokers have normal relaxation blood pressure level and smokers have more chance of having elevated, hypertension stage2, hypertension crisis and hypotension.

7. There is a relation between smoking status and dental caries. Also we say that smokers diabetes level effects blood pressure level and cholesterol level.

8. The performance of the Decision Tree model is good to predict diabetes category of smokers and non smokers.

4.1.2 Summary

This research studies the impact of smoking on health and how smoking influencing on metabolic health. From the study we got know that majority of smokers are male as compared with female.The finding that the average age of male smokers is lower compared to the average age of female smokers.The lower average age of male smokers suggests that males tend to start smoking at a younger age compared to females. This could be influenced by social, cultural, and environmental factors, and it raises concerns about early smoking initiation among males. The majority of smokers have glucose levels within a healthy range. However, the percentages of individuals in the pre-diabetic (32.3%) and diabetic (6.6%) categories are noteworthy, suggesting that a considerable proportion of smokers are at risk of developing diabetes or have already been diagnosed with the condition. Among non-smokers, the higher percentage (64.2%) with normal blood glucose levels indicates a healthier distribution compared to smokers. However, it is still concerning that a significant proportion of non-smokers are in the pre-diabetic (25.8%) and diabetic (3.8%) categories, highlighting the importance of diabetes prevention strategies even among non-smokers.

The analysis of blood pressure levels among smokers indicates a mix of both favorable and concerning findings. While a significant percentage of smokers have normal blood pressure, there is a notable presence of elevated blood pressure and hypertension. These results highlight the need for comprehensive health strategies, including smoking cessation efforts and blood pressure management, to reduce the cardiovascular risks associated with smoking. The proportion of non-smokers have normal systolic and relaxation blood pressure levels. This indicates that a majority of non-smokers have healthy blood pressure levels, which is a positive finding. This highlights the importance of maintaining healthy blood pressure levels in the population. It emphasizes the need for effective interventions to address elevated blood pressure and hypertension, which are crucial steps in reducing the risk of cardiovascular diseases and improving overall health outcomes for non-smokers.

The dependence between smoking status and dental caries underscores the adverse impact of smoking on oral health. Addressing this association is vital in designing effective preventive measures and promoting overall oral health and well-being among individuals who smoke. It also highlights the importance of comprehensive oral care and smoking cessation programs to reduce the prevalence of dental caries and its consequences in the smoking population.

The dependence between smokers' systolic blood pressure and relaxation blood pressure on their diabetes level. The relationship underscores the importance of diabetes

management and the significance of smoking cessation in improving blood pressure control and reducing the risk of cardiovascular complications in smokers with diabetes. The smokers' diabetes level can influence cholesterol levels, and this relationship underscores the importance of managing diabetes and smoking cessation in improving lipid profiles and reducing the risk of cardiovascular complications. A comprehensive approach to diabetes care, including attention to cholesterol levels, is essential to promote better cardiovascular health in this population. Encouraging smoking cessation and providing individualized care are essential components of cardiovascular risk reduction in individuals with diabetes who smoke.
Random Forest shows great promise as a robust and accurate model for diabetes prediction, which could potentially aid in early diagnosis and improve healthcare outcomes for individuals at risk of diabetes.

5 Chapter 5

5.1 Bibliography

1. Abdel-Salam. (2018). Statistical analysis for the impact of smoking on the behavior and health of Qatari adolescents-*International Journal of Adolescent Medicine and Health.*

2. DK Gautam. (2011). Effect of cigarette smoking on the periodontal health status: A comparative, cross sectional study-*Journal of Indian society of Periodontology*. 15(4), 383.

   Calsina. (2002). Effect of smoking on periodontal tissues-*Journal of clinical periodontalogy* .29(8),771-776.

3. Deepa M Gopal. (2012). Cigarette smoking exposure and heart failure risk in older adults: the health, Aging and Body Composition Study -*American heart journal* .164(2),236-242.

4. Robert West. (2017). Tobacco smoking: Health impact, prevalence, correlates and interventions -*Department of Behavioral Science and Health, University College London, London, UK.*

5. https://www.cdc.gov/tobacco/basic$_information/health_effects/index.htm$

6. https://www.health.gov.au/

7. https://www.degruyter.com/document/doi/10.1515/ijamh-2018-0045/html?lang=en

8. https://www.pythonfordatascience.org/anova-python/

9. https://www.analyticsvidhya.com/blog/2021/07/t-test-performing-hypothesis-testing-with-python/

10. https://www.statology.org/levenes-test-python/

11. https://towardsdatascience.com/methods-for-normality-test-with-application-in-python-bb91b49ed0f5

12. https://www.datacamp.com/tutorial/random-forests-classifier-python

13. https://www.section.io/engineering-education/machine-learning-with-xgboost-and-scikit-learn/

14. https://catboost.ai/en/docs/concepts/python-usages-examples

15. http://scikit-learn.org/stable/modules/naive$_b ayes.html$

6 Chapter 6

6.1 Appendix

Python Codes:

```
Importing various libraries required for analysis}
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import classification_report
from sklearn.metrics import precision_score
from sklearn.metrics import PrecisionRecallDisplay
from sklearn.metrics import accuracy_score
from sklearn.metrics import f1_score
from sklearn.metrics import PrecisionRecallDisplay
from sklearn.metrics import recall_score
from sklearn.metrics import confusion_matrix
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
import xgboost as xgb
from catboost import CatBoostClassifier
from sklearn.model_selection import RandomizedSearchC


#Splitting data to test and train set

from sklearn.model_selection import train_test_split
xtrain,xtest,ytrain,ytest=train_test_split(x,y,test_size=0.20,random_state=1)



#Function to analyse out come of different machine learning models

import seaborn as sns
def.analysis(ytrain,ypred)
```

```
model_test_accuracy = accuracy_score(ytest, ypred)
model_test_f1 = f1_score(ytest, ypred, average='weighted')
model_test_precision = precision_score(ytest, ypred , average='weighted')
model_test_recall  = recall_score(ytest, ypred,average='weighted')

model_train_accuracy = accuracy_score(ytrain, ypredd)
model_train_f1 = f1_score(ytrain, ypredd, average= 'weighted')
model_train_precision = precision_score(ytrain, ypredd,average='weighted')
model_train_recall = recall_score(ytrain, ypredd,average='weighted')

print('Model performance for Training set')
print("- Accuracy: {:.4f}".format(model_train_accuracy))
print('- F1 score: {:4f}'.format(model_train_f1))
print('- Precision: {:4f}'.format(model_train_precision))
print('- Recall: {:4f}'.format(model_train_recall))

print('Model performance for Test set')
print('- Accuracy: {:.4f}'.format(model_test_accuracy) )
print('- Fl score: {:.4f}'.format(model_test_f1))
print('- Precision: {:.4f}'.format(model_test_precision))
print('- Recall: {:.4f}'.format(model_test_recall))

conf_matrix1 = confusion_matrix(ytest, ypred)
plt.figure(figsize=(8, 6))
sns.set(font_scale=1.4)
sns.heatmap(conf_matrix1, annot=True, fmt="d", square=True)
plt.xlabel("Predicted")
plt.ylabel("True")
plt.title("Confusion Matrix")
plt.show()\\\end{flushleft}


#Decision Tree Classifier

# Fitting training and testing of Decision Tree Classifier
#define model
dtree=DecisionTreeClassifier()
dtree.fit(xtrain,ytrain)
```

```
ypredd4=dtree.predict(xtrain)
ypred4=dtree.predict(xtest)

#Random Forest Classifier

# Fitting training and testing of Random Forest Classifier
from sklearn.ensemble import RandomForestClassifier
rf=RandomForestClassifier()
rf.fit(xtrain,ytrain)
ypred5=rf.predict(xtest)
ypredd5=rf.predict(xtrain)


#XGBoost Classifier

# Fitting training and testing of XGBoost Classifier
pip install xgboost
import xgboost as xgb
model.fit(xtrain,ytrain)

y_train_pred = model.predict(xtrain)
y_test_pred =  model.predict(xtest)


#CatBoost Classifier

# Fitting training and testing of CatBoost Classifier
pip install catboost
from catboost import CatBoostClassifier
catboost_model = CatBoostClassifier(iterations=1000, depth=6, random_seed=42)
catboost_model.fit(xtrain, ytrain)
```