Behavioral Risk Factor Surveillance System Analysis

End-to-end Data Science

B198c5

Final Project Report

GH1019648

Abstract

To generate prediction models and actionable insights for public health, our research dives into the Behavioral Risk Factor Surveillance System (BRFSS) dataset, a key continuous monitoring program in the United States. Through machine learning models, the primary focus is on recognizing and forecasting chronic diseases such as heart disease, diabetes, and depression. Our process includes essential exploratory data analysis, with critical criteria such as age, body mass index (BMI), and lifestyle variables being targeted. The study's goal is to inform well-founded public health policies and initiatives for the wellbeing of the American people by identifying connections, distribution patterns, and prevalence rates. (health.gov, n.d.)

1. Introduction

One of the most important continuous surveillance programs is the Behavioral Risk Factor Surveillance System (BRFSS), which gathers state-specific information on health-related risk behaviors, chronic health issues, and the use of preventive interventions throughout the US. In an effort to support the creation of knowledgeable public health policies and programs targeted at enhancing the general health and well-being of the American people, the BRFSS is dedicated to delivering consistent, trustworthy information. The aim of this project is to solve the challenge given by the Behavioral Risk Factor Surveillance System (BRFSS) dataset in developing prediction models and actionable insights for public health. The goal is to use machine learning models to improve our understanding of demographics, lifestyle factors, and chronic diseases, with a special focus on depression, diabetes, and heart disease. This study aims to construct predictive models for general health status by doing exploratory data analysis on the large BRFSS dataset, taking into account crucial factors such as age and body mass index. The ultimate goal is to assist informed decision-making and facilitate focused treatments for chronic illness prevention. (health.gov, n.d.)

2. Literature review

The Behavioral Risk Factor Surveillance System (BRFSS) is an important initiative that collects state-specific data on health habits and chronic conditions, which is used to inform public health strategies. This research contributes to the use of BRFSS data for predictive modeling and actionable insights in chronic conditions such as depression, diabetes, and heart disease. (health.gov, n.d.)

The Use of Machine Learning in Public Health

The rising use of machine learning in public health research reflects a shift toward data-driven decision-making, enabling precise identification and prediction of chronic diseases from large datasets. This study adds to the expanding knowledge at the intersection of machine learning and public health by informing targeted treatments and preventive measures. (Tanmoy Sarkar Pias et al., 2023)

Exploratory Data Analysis and Predictive Modeling

Exploratory Data Analysis (EDA) is crucial for uncovering relationships in datasets, especially when building health prediction models considering factors like age and BMI. Intelligent feature engineering, displayed by functions like 'remove_outliers' and 'extract_age,' ensures dataset integrity and enhances the effective of machine learning applications in healthcare. (Tanmoy Sarkar Pias et al., 2023)

Correlation Analyses and Health Risk Factors

Correlation analyses, utilized in this research, have been useful in finding potential health risk factors. The research literature supports the investigation of connections

between characteristics such as age, diabetes, and arthritis with the existence of heart disease. According to studies, such insights are critical for creating specific health interventions and gaining a deeper understanding of the multidimensional nature of chronic diseases. (Tanmoy Sarkar Pias et al., 2023)

Machine Learning Model Evaluation

The evaluation of machine learning models, such as Random Forest, Support Vector Machine, and Logistic Regression, follows predictive modeling best practices. Existing literature highlights the importance of thorough analyses utilizing measures like accuracy, confusion matrices, and classification reports. This study mirrors the iterative process of model development by critically assessing limitations and areas for improvement. (Tanmoy Sarkar Pias et al., 2023)

Challenges and Future Directions

Machine learning offers potential in public health research, however issues such as class imbalance and model sensitivity continue. Continuous refinement, collaboration with domain experts, and identifying dataset limits are all highlighted in the literature. Addressing model biases, integrating various datasets, and improving feature engineering are all future directions for more precise and applicable predictive models. (Ribeiro, Singh and Guestrin, 2016)

3. Business Question

- What machine learning models can be applied to effectively detect and predict the occurrence of chronic diseases such as heart disease, diabetes, and depression based on the available dataset features?
- What actionable insights and recommendations can be derived from the dataset to help prevent chronic diseases?
- Can we build a predictive model to estimate a person's general health status based on their age, BMI, and other relevant features?
- Are there any associations between smoking history and other health-related variables such as BMI, exercise habits, or alcohol consumption?
- What is the prevalence of skin cancer, and how does it vary by age category and gender?

By the end of this report, we hope to demonstrate how data-driven methods can help us better understand public health challenges and make better decisions that will benefit the American people.

4. Dataset Features

The dataset consists of up of 308,854 rows and 19 features. The dataset looks to be well-structured, with no missing values, indicating that it is suitable for analysis. General health, exercise habits, the presence of chronic diseases (e.g., heart disease, diabetes), lifestyle factors (e.g., smoking, alcohol consumption), and anthropometric measures (e.g., height, weight, BMI) are a few of the aspects. The dataset appears to be broad and valuable for investigating factors associated with chronic diseases.

5. Methodology

i. Data Exploration

The first step in the analysis is fundamental data exploration, which includes importing libraries and reviewing the information, structure, and descriptive statistics of the dataset. The basis for further in-depth analysis and machine learning applications is a challenging check for missing values.

Importing necessary libraries

In order to start our research, an initial review of the structural features of the dataset shows a complete picture. With a large dataset of 308,854 rows, the dataset has 19 distinctive features that collectively offer an extensive amount of health-related data.

A thorough analysis of the dataset's data is carried out in an effort to gain a more detailed understanding. This includes extracting essential details such as null counts, dataset shape, column names, and the sorts of data it includes. This thorough summary is the first step toward understanding the structure and basic characteristics of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 308854 entries, 0 to 308853
Data columns (total 19 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   General_Health                  308854 non-null  object
 1   Checkup                         308854 non-null  object
 2   Exercise                        308854 non-null  object
 3   Heart_Disease                   308854 non-null  object
 4   Skin_Cancer                     308854 non-null  object
 5   Other_Cancer                    308854 non-null  object
 6   Depression                      308854 non-null  object
 7   Diabetes                        308854 non-null  object
 8   Arthritis                       308854 non-null  object
 9   Sex                             308854 non-null  object
 10  Age_Category                    308854 non-null  object
 11  Height_(cm)                     308854 non-null  float64
 12  Weight_(kg)                     308854 non-null  float64
 13  BMI                             308854 non-null  float64
 14  Smoking_History                 308854 non-null  object
 15  Alcohol_Consumption             308854 non-null  float64
 16  Fruit_Consumption               308854 non-null  float64
 17  Green_Vegetables_Consumption    308854 non-null  float64
 18  FriedPotato_Consumption         308854 non-null  float64
dtypes: float64(7), object(12)
memory usage: 44.8+ MB
```

A thorough examination of descriptive statistics is conducted in order to go further into the complexity of the dataset. To do this, one must extract the count, mean, minimum, maximum, standard deviation, and quartiles—all important statistical analysis. Through the process of removing these statistical issues, a deeper knowledge of the numerical features of the dataset can be achieved.

	Height_(cm)	Weight_(kg)	BMI	Alcohol_Consumption	Fruit_Consumption	Green_Vegetables_Consumption	FriedPotato_Consumption
count	308854.000000	308854.000000	308854.000000	308854.000000	308854.000000	308854.000000	308854.000000
mean	170.615249	83.588655	28.626211	5.096366	29.835200	15.110441	6.296616
std	10.658026	21.343210	6.522323	8.199763	24.875735	14.926238	8.582954
min	91.000000	24.950000	12.020000	0.000000	0.000000	0.000000	0.000000
25%	163.000000	68.040000	24.210000	0.000000	12.000000	4.000000	2.000000
50%	170.000000	81.650000	27.440000	1.000000	30.000000	12.000000	4.000000
75%	178.000000	95.250000	31.850000	6.000000	30.000000	20.000000	8.000000
max	241.000000	293.020000	99.330000	30.000000	120.000000	128.000000	128.000000

In order to ensure the dataset's behavior and reliability, a thorough evaluation of the data types that make up each column is carrying out. This is a crucial step to assure consistency and coherence across the dataset, providing a strong basis for further analytical work.

```
General_Health                 object
Checkup                        object
Exercise                       object
Heart_Disease                  object
Skin_Cancer                    object
Other_Cancer                   object
Depression                     object
Diabetes                       object
Arthritis                      object
Sex                            object
Age_Category                   object
Height_(cm)                   float64
Weight_(kg)                   float64
BMI                           float64
Smoking_History                object
Alcohol_Consumption           float64
Fruit_Consumption             float64
Green_Vegetables_Consumption  float64
FriedPotato_Consumption       float64
dtype: object
```

A thorough examination of null values in the dataset is conducted as part of an organized evaluation intended to ensure data quality assurance. After a thorough examination, there are no missing values found in any of the dataset's columns, which is a satisfying result. The cleanness of the dataset is highlighted by the lack of null values, which gives confidence in the accuracy of the subsequent studies.

```
General_Health                 0
Checkup                        0
Exercise                       0
Heart_Disease                  0
Skin_Cancer                    0
Other_Cancer                   0
Depression                     0
Diabetes                       0
Arthritis                      0
Sex                            0
Age_Category                   0
Height_(cm)                    0
Weight_(kg)                    0
BMI                            0
Smoking_History                0
Alcohol_Consumption            0
Fruit_Consumption              0
Green_Vegetables_Consumption   0
FriedPotato_Consumption        0
dtype: int64
```

ii. Data Preprocessing

The dataset's structural alignment is improved by a careful conversion of column data types. This thorough conversion process is intended to ensure consistency and compliance with existing data guidelines, creating an organized framework that allows smooth data preprocessing. (pandas.pydata.org, n.d.)

```
General_Health                 string
Checkup                        object
Exercise                       string
Heart_Disease                  string
Skin_Cancer                    string
Other_Cancer                   string
Depression                     string
Diabetes                       string
Arthritis                      string
Sex                            string
Age_Category                   string
Height_(cm)                     int64
Weight_(kg)                   float64
BMI                           float64
Smoking_History                string
Alcohol_Consumption             int64
Fruit_Consumption               int64
Green_Vegetables_Consumption    int64
FriedPotato_Consumption         int64
dtype: object
```

iii. Label Encoding

We are introducing 'remove_outliers,' a useful tool to improve data integrity and support robust analysis. With this function, which is customized to work with a DataFrame (df) and an optional z-score threshold (z_threshold), z-scores for numerical columns in the dataset are carefully calculated. It then proceeds to systematically remove rows in which any z-score is greater than the threshold set by the user in order to reduce the effect of outliers. To further improve its usefulness, the method uses Label Encoding to add categorical variable encoding to the DataFrame, which improves consistency and increases computational effectiveness. In order to illustrate how effective it is, the 'remove_outliers' function is applied to the DataFrame df. This allows it to be used to encode categorical columns in the resulting refined dataset. This all-encompassing strategy represents a flexible methodology that is in line with industry best practices for preparing data in order to get it ready for further analytical activities. (Bonthu, 2021)

iv. Feature Engineering

Utilizing 'extract_age,' an advanced utility function that aims to improve the understanding of age categories in a dataset. Once an age category is input, this method carefully handles a variety of representations, including range indications ('18-24,' '25-34,' and singular values ('65.') The function pays particular attention to age groups indicated by a '+,' which corresponds to '80+' years. The function returns a fixed value of 80 in these cases. When age groups with a '-', which represents a range, occur, the function proactively calculates the mean of the range. When age categories don't have a '-' or '+,' which indicates a single age, the function immediately turns it into an integer. In order to establish a more complex and consistent representation of age-related data, the 'extract_age' function is used to replicate the values from the current 'Age_Category' column and generate a new 'Age' column in the DataFrame df. This function demonstrates practical usefulness.

Out[10]:

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex	Age_Category	Height_(cm)	Wei
0	3	2	0	0	0	0	0	0	1	0	10	150	
1	4	4	0	1	0	0	0	2	0	0	10	165	
2	4	4	1	0	0	0	0	2	0	0	8	163	
3	3	4	1	1	0	0	0	2	0	1	11	180	
4	2	4	0	0	0	0	0	0	0	1	12	191	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
308848	2	3	1	0	0	0	0	0	0	1	7	168	
308849	4	4	1	0	0	0	0	0	0	1	1	168	
308851	4	0	1	0	0	0	1	3	0	0	2	157	

Initiating a crucial preprocessing step, the extraction of weight values from the 'Weight_(kg)' column and subsequent retrieval of centimeter-based height values from the 'Height_(cm)' column is performed. The height numbers must be converted to meters in order to compute the Body Mass Index (BMI). This can be done by dividing the height values by 100. By using the BMI formula, ((df['Height_(cm)'] / 100) ** 2) gives the square of the height in meters. This thorough method provides a precise and uniform metric for evaluating body mass across the dataset, the BMI for each dataset entry is calculated by dividing the weight in kilograms by the squared height in meters. (Stack Overflow, n.d.)

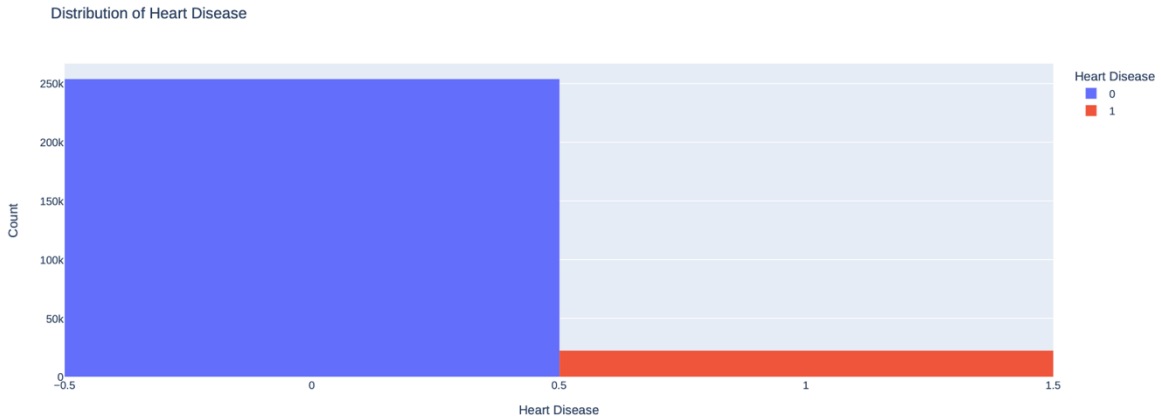
v. Visualisation

Distribution of Heart Disease

To produce a thorough representation of the distribution of heart disease in the given dataset (df), using Plotly Express library to create a count plot. The graphic clearly shows the incidence of cardiac disease, where a value of 1 indicates the condition's existence and a value of 0 indicates its absence. The resulting figure, appropriately named "Distribution of Heart Disease," shows the frequency of each "Heart_Disease" categorical class on the x-axis. With its clear and informative summary, this graphical depiction is a useful tool for determining the prevalence of heart disease within the dataset.

Output:

When the output is examined, a low incidence of heart disease cases is found, supporting the claim that a small percentage of the dataset has this medical condition.



vi. Correlation Analysis

Correlation analyses examining the 'Heart_Disease' column and other health-related factors in the dataset show complex relationships. Important details about the dataset may be found in the correlation matrix, which uses correlation coefficients ranging from -1 to 1 to express these associations. Key findings include:

1. Heart Disease and Diabetes (0.168): A significant positive correlation indicates a possible link between diabetes and heart disease, indicating a possible increased risk of heart-related issues in those with diabetes.

2. Heart Disease and Age (0.233): This positive relationship highlights the fact that the risk of developing heart disease increases with age, revealing an age-dependent vulnerability.

3. Skin cancer and other cancers (0.150): The positive association suggests that there may be common risk factors or underlying processes that contribute to the occurrence of several types of cancer in addition to skin cancer.

4. Exercise and Diabetes (-0.135): This shows that regular physical activity may have a preventive impact on the development of diabetes, which is consistent with accepted health principles.

5. BMI and Weight (0.844): The strong positive correlation between BMI and weight highlights the relationship that exists between the two variables, indicating the role that each plays in defining an individual's entire body composition.

6. Exercise and Alcohol Consumption (0.118): The positive association between exercise and alcohol consumption points to a concurrent pattern, which shows that people who exercise may also drink.

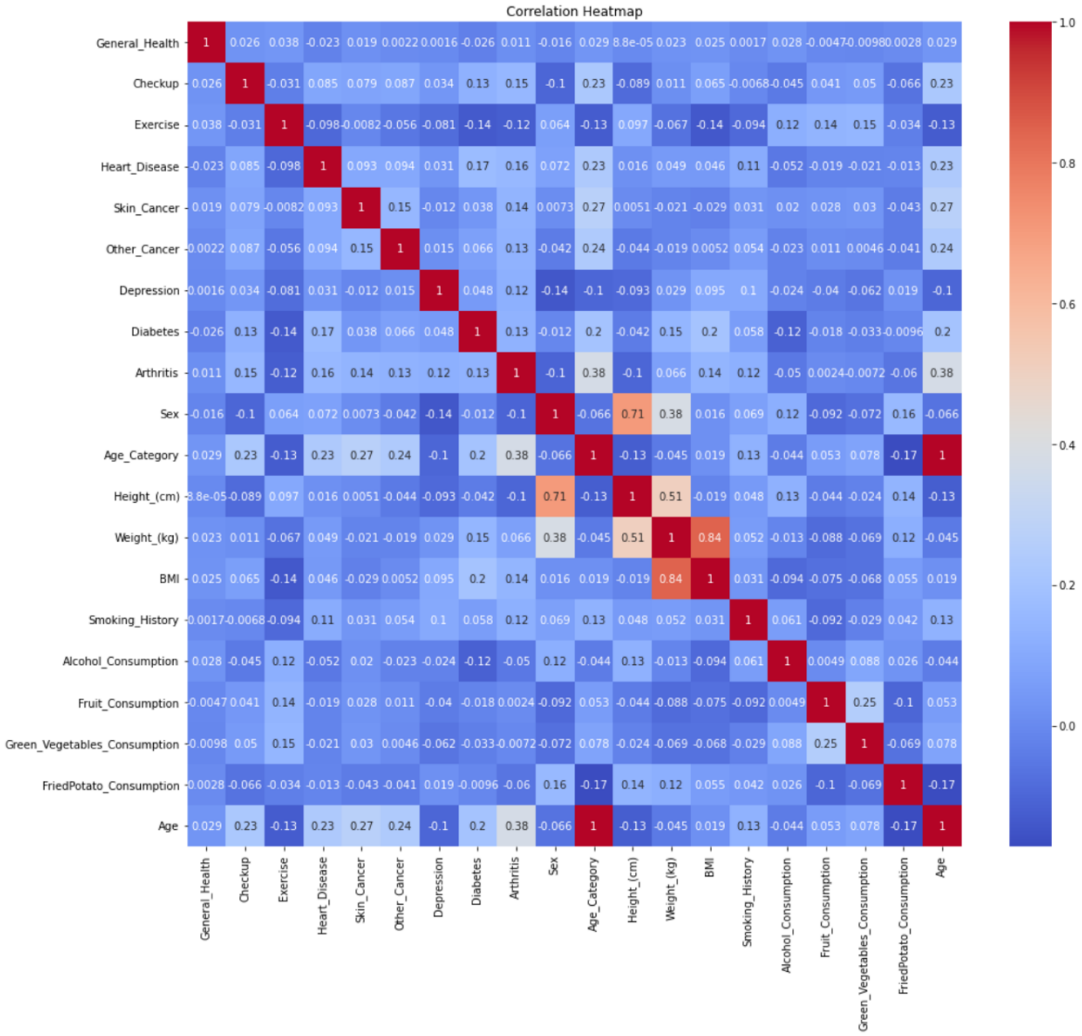
These findings provide important insights for focused interventions and public health policies in addition to adding to a deeper knowledge of health connections.

```
In [13]: df.corr()
```

Out[13]:

	General_Health	Checkup	Exercise	Heart_Disease	Skin_Cancer	Other_Cancer	Depression	Diabetes	Arthritis	Sex
General_Health	1.000000	0.026057	0.037862	-0.022939	0.018898	0.002223	0.001567	-0.026001	0.011273	-0.015768
Checkup	0.026057	1.000000	-0.031425	0.085412	0.079136	0.086693	0.034371	0.128659	0.151135	-0.099826
Exercise	0.037862	-0.031425	1.000000	-0.097616	-0.008178	-0.056206	-0.081230	-0.135287	-0.124046	0.063906
Heart_Disease	-0.022939	0.085412	-0.097616	1.000000	0.093381	0.093845	0.030819	0.168441	0.155523	0.071980
Skin_Cancer	0.018898	0.079136	-0.008178	0.093381	1.000000	0.150536	-0.011644	0.038111	0.137625	0.007274
Other_Cancer	0.002223	0.086693	-0.056206	0.093845	0.150536	1.000000	0.014678	0.066042	0.129958	-0.042369
Depression	0.001567	0.034371	-0.081230	0.030819	-0.011644	0.014678	1.000000	0.048162	0.120713	-0.141150
Diabetes	-0.026001	0.128659	-0.135287	0.168441	0.038111	0.066042	0.048162	1.000000	0.134115	-0.011600
Arthritis	0.011273	0.151135	-0.124046	0.155523	0.137625	0.129958	0.120713	0.134115	1.000000	-0.104465
Sex	-0.015768	-0.099826	0.063906	0.071980	0.007274	-0.042369	-0.141150	-0.011600	-0.104465	1.000000
Age_Category	0.029396	0.225873	-0.128877	0.233017	0.270318	0.235955	-0.100843	0.202778	0.375686	-0.065572
Height_(cm)	0.000088	-0.089177	0.097187	0.015581	0.005132	-0.044302	-0.093349	-0.042289	-0.103089	0.705696
Weight_(kg)	0.023168	0.011150	-0.067275	0.048619	-0.020923	-0.019371	0.028763	0.147966	0.065784	0.384814
BMI	0.024794	0.065015	-0.140239	0.046389	-0.028584	0.005226	0.095232	0.199774	0.139030	0.015511
Smoking_History	0.001689	-0.006792	-0.094206	0.109822	0.031116	0.053710	0.102184	0.058165	0.124312	0.068577
Alcohol_Consumption	0.027691	-0.045097	0.118789	-0.051777	0.019914	-0.023363	-0.024138	-0.116838	-0.049986	0.116365
Fruit_Consumption	-0.004715	0.040962	0.137490	-0.019420	0.028005	0.010677	-0.040335	-0.017819	0.002429	-0.092029
Green_Vegetables_Consumption	-0.009803	0.049745	0.146838	-0.020643	0.030356	0.004640	-0.062374	-0.033305	-0.007205	-0.072312
FriedPotato_Consumption	0.002836	-0.066167	-0.033929	-0.012972	-0.042705	-0.040664	0.019190	-0.009640	-0.059778	0.157993
Age	0.029396	0.225873	-0.128877	0.233017	0.270318	0.235955	-0.100843	0.202778	0.375686	-0.065572

In the goal of uncovering the numerous relationships within the dataset, a complete exploration is aided by the production of a correlation heatmap. Using the seaborn library as a resource, the correlation matrix is carefully examined to see how different columns interact. A heatmap provides an effective way to visualize the correlation coefficient, which indicates the direction and strength of relationships. The image, which was produced by using the command sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm"), is an effective tool for seeing trends, figuring out dependencies, and pointing out possible directions for more research. This systematic analysis contributes to a comprehensive insight into the many interrelationships buried within the health-related variables by offering a visually intuitive picture of the data's underlying connections.



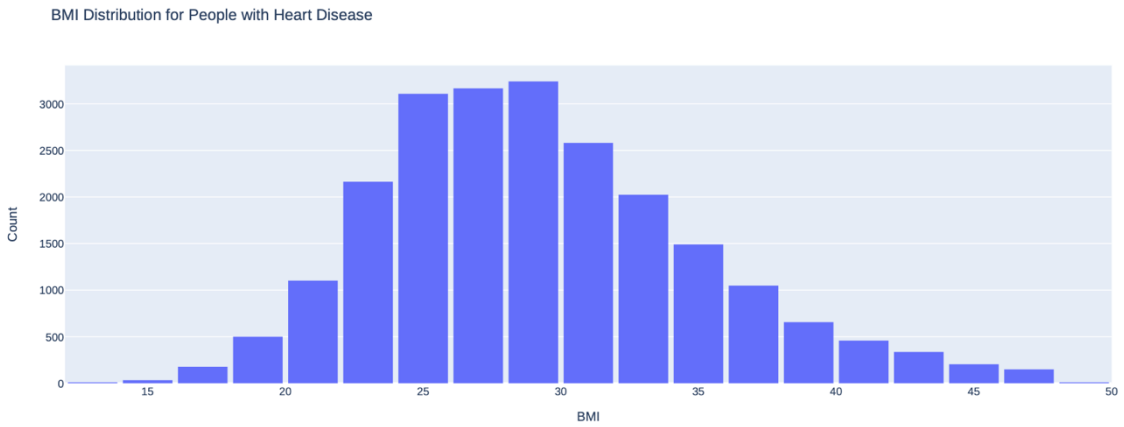
People with Heart Disease and their BMI Distribution

The analysis focused on those members of the dataset who were diagnosed with heart disease; this group was identified using the criteria 'Heart_Disease == 1.' This subset was taken out for additional analysis by means of a data filter application. After that, a graphic was created to show how this particular cohort's Body Mass Index (BMI) was distributed. With much care and attention to detail, a histogram was created using the Plotly Express module. The resulting plot, called "BMI Distribution for People with Heart Disease," provides an excellent task of illustrating the range of BMIs. "BMI" is indicated on the x-axis, while the frequency of individuals is quantified on the y-axis. The 'bargap' option was carefully changed to regulate the distance between the histogram bars in order to fine-tune the appearance. This visual representation offers a perceptive look at the BMI distribution patterns among people with heart disease diagnoses.

Output:

As we can see, the vast majority of people with Heart Disease 3240, have BMI levels that fall within the category of overweight which is 28 - 29.999, signifying that the people with Heart Disease are obese.

The normal BMI range is 18.5 to 24.9, and if the range falls between 25.0 and 29.9, they are known overweight, and BMI levels less than 18.5 are considered underweight. (CDC, 2020)

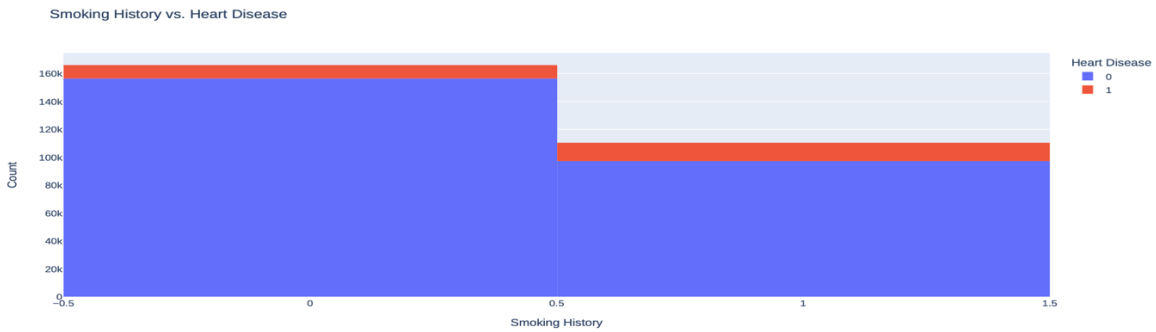


Heart Disease vs. History of Smoking

The provided code builds a histogram showing the link between the presence or absence of heart disease (Heart_Disease) and smoking history (Smoking_History) using Plotly Express. The bars on the x-axis are color-coded to distinguish between those with and without heart disease, and each bar represents a distinct smoking history group. The generated plot, "Smoking History vs. Heart Disease," provides a concise summary of the distribution of heart disease cases among different smoking history categories.The plot, displayed using fig.show(), presents a binary encoding where 0 signifies the absence of heart disease and no smoking history, while 1 indicates the presence of heart disease and a corresponding smoking history.

Output:

The plot's observation implies that while people with heart disease are more likely to have smoked in the past, people without a smoking history typically do not develop heart disease.

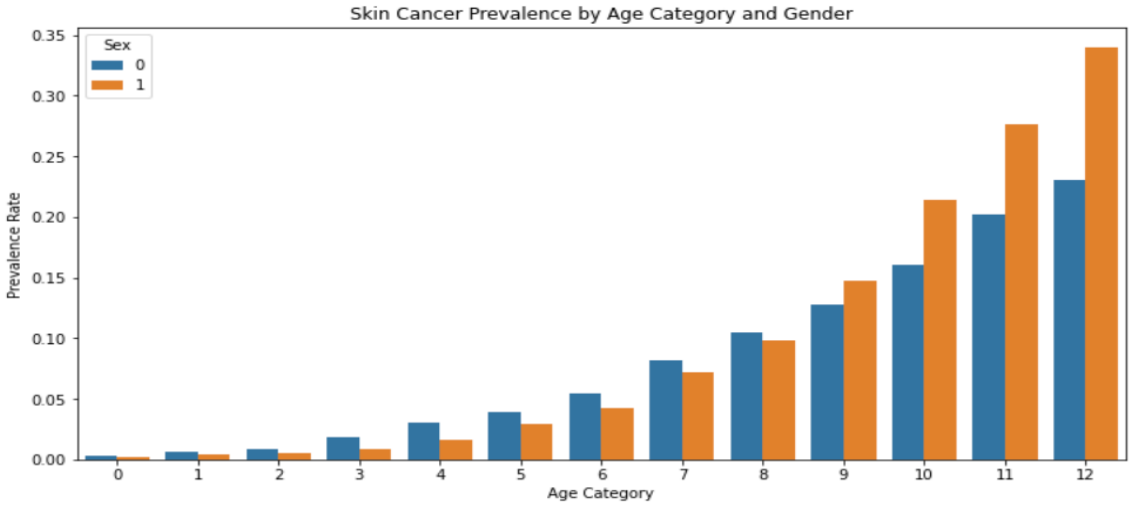


Skin Cancer prevalence by Gender and Age Group

Utilizing data grouping (groupby()) to evaluate the average prevalence of skin cancer in various age groups and genders. By calculating the mean skin cancer prevalence and resetting the index, a new distribution function named "skin_cancer_prevalence" is formed, using the 'Age_Category' and 'Sex' variables for grouping. The following visualization is a bar chart created with Seaborn, where the y-axis represents mean skin cancer prevalence and the x-axis represents gender and age groups. The figure size of the chart, as it is displayed, is 12 by 6. With the headline "Skin Cancer Prevalence by Age Category and Gender," the x- and y-axes have corresponding labels on the plot.

Output:

The output chart highlights differences between male (denoted as 1, displayed in orange) and female (denoted as 0, shown in blue) populations by visually representing the occurrence of skin cancer across age groups. The chart is important because it shows differences in the occurrence of skin cancer by gender across various age groups.



vii. Feature Selection

The method for predicting heart disease that is being presented using machine learning and feature engineering. As the target variable 'y,' the 'Heart_Disease' column is first removed from the dataset; the remaining features form the feature variable 'X.' The dataset is then divided into training and testing sets using the 'train test split' (X, y, test_size=0.2, random_state=42) function, where a random seed of 42 is used for reproducibility and a test size of 20% is specified. Using the training dataset, a RandomForestClassifier with 100 decision trees is trained. Next, a

'feature_importances' variable is produced by applying the trained random forest model to evaluate the significance of each feature. 'X_train_selected' and 'X_test_selected' are the outcomes of applying feature selection to the training and testing data by utilizing this information. The 'sfm.get_support(indices=True)' function helps in the selection process by providing the index of the selected features, making it possible to extract and print their names later. To improve predictive accuracy in the context of classifying heart diseases, this all-inclusive methodology combines feature selection, model training, and data preparation.

viii. Model Training and Evaluation

a) Random Forest

Using 100 decision trees, the Random Forest Classifier predicts the presence of heart disease with a noteworthy accuracy of 91.7% on the testing dataset. This assessment includes a number of indicators, such as a classification report and a confusion matrix, to give a thorough picture of the model's functionality. According to the confusion matrix, there were 50,498 true negatives (correctly predicted cases of non-heart disease) out of 55,218 instances; 189 false positives (misclassified as cases of heart disease when they are not); 4,397 false negatives (misclassified as cases of non-heart disease when they are); and 134 true positives (correctly predicted patients with heart disease).

The following classification report explores measures for both classes, including precision, recall, and F1-score. Notably, there is a 92% precision rate for the absence of heart illness and a 41% precision rate for the diagnosis of heart disease. The greatest (100%) recall for the absence of heart disease highlights the model's accuracy in detecting cases that do not involve heart disease. However, the recall for heart illness is only 3%, suggesting a significant difficulty in correctly identifying positive cases. The F1-scores also show this gap, with 6% for heart disease and 96% for non-heart disease. With a weighted average accuracy of 92%, more investigation is clearly required, especially to resolve false negatives and improve the model's sensitivity to cases of heart disease.

```
Random Forest Classifier:
Accuracy: 0.9169473722336919
Confusion Matrix:
 [[50498   189]
 [ 4397   134]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     50687
           1       0.41      0.03      0.06      4531

    accuracy                           0.92     55218
   macro avg       0.67      0.51      0.51     55218
weighted avg       0.88      0.92      0.88     55218
```

b) Support Vector Machine

It is necessary to train the Support Vector Machine (SVM) model and then assess its performance in order to use it for prediction on the testing dataset. The SVM model is trained by an accurate approach to provide predictions, which are evaluated by comparing the predictions to the target values. The assessment consists of determining accuracy, creating a confusion matrix, and thoroughly analyzing performance metrics through the use of a categorization report. Important indicators of the SVM model's effectiveness in classifying objects are the comprehensive insights into its accuracy, confusion matrix, and classification report. This systematic method facilitates an in-depth understanding of the predictive power of the model and helps with well-informed decision-making on more refinements or alternative model considerations.

Output:

The accuracy of the SVM model is 91.8%. The confusion matrix, on the other hand, shows a notable imbalance: the model correctly detects cases without heart disease (class 0), but it is unable to recognize any cases of heart disease (class 1). As a result, class 1's precision, recall, and F1-score metrics are zero. The SVM's ability to accurately forecast class 0 instances is the main factor influencing the weighted average accuracy of 92%. The macro-average measurements, which show that precision, recall, and F1-score are all focused around 0.5, highlight the difficulties caused by class imbalance. These results call for a thorough assessment and investigation of possible improvements or different models to solve the model's flaws.

```
Support Vector Machine (SVM):
Accuracy: 0.917943424245717
Confusion Matrix:
 [[50687     0]
 [ 4531     0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     50687
           1       0.00      0.00      0.00      4531

    accuracy                           0.92     55218
   macro avg       0.46      0.50      0.48     55218
weighted avg       0.84      0.92      0.88     55218
```

c) Logistic Regression

This code trains and evaluates a logistic regression model, which is a predictive tool for forecasting results. A different set of data is used to evaluate the predictive accuracy of the model after it has been trained on a specific training dataset. The evaluation looks at the confusion matrix, which lists the occurrences of accurate and inaccurate predictions together with the model's overall accuracy. In addition, a classification report is produced to offer a thorough summary of the model's performance in several classes. Understanding the precision, recall, and F1-score

metrics is made easier with the help of this report, which provides information on the accuracy and efficiency of the Logistic Regression model in predicting heart disease cases.

Output:

The accuracy obtained by applying Logistic Regression on the dataset is 91.7%. The model is good at identifying cases without heart disease (class 0), but it has trouble identifying cases with heart disease (class 1), which leads to more false negatives. Key metrics are shown in the classification report, which highlights a 41% precision for heart disease and a 2% recall, highlighting the model's challenges in correctly detecting positive cases. The weighted average accuracy of 92% demonstrates how well the algorithm can predict patients free of heart disease. On the other hand, macro-average metrics highlight the effect of class imbalance, which calls for a thorough analysis and possible improvements to address the model's limitations.

```
Logistic Regression:
Accuracy: 0.9173095729653374
Confusion Matrix:
 [[50569   118]
 [ 4448    83]]
Classification Report:
                precision    recall  f1-score   support

           0       0.92      1.00      0.96     50687
           1       0.41      0.02      0.04      4531

    accuracy                           0.92     55218
   macro avg       0.67      0.51      0.50     55218
weighted avg       0.88      0.92      0.88     55218
```

6. Results

Distribution of Heart Diseases:

- The distribution of heart disease cases is unbalanced, according to the dataset. Compared to non-cases (0), there are notably fewer cases (1) of cardiac disease. The model's capacity to forecast cardiac disease with accuracy may be impacted by this imbalance.

Analyzing correlations:

- Among the factors that show a comparatively stronger association with the existence of heart disease are age, diabetes, arthritis, and sex. Determining possible risk factors may require an understanding of these relationships.

Distribution of BMI:

- According to the distribution of BMI values, people with heart disease typically have slightly higher BMIs on average. The variations between people with and without heart disease are highlighted by the BMI distribution plot, which may indicate a possible link.

History of Smoking and Heart Disease:

- Individuals with a smoking history (1) are found in both the heart disease and non- heart disease categories, according to the smoking history count plot. That being said, compared to people who do not have heart disease, a greater percentage of those who do have a smoking history.

Prevalence of Skin Cancer:

- Variations in the prevalence of skin cancer are shown in the bar chart by gender and age group. In some age groups, the prevalence of skin cancer is higher in women, whereas in other age groups, it is higher in men.

7. Answers to Business Questions

1. What machine learning models can be applied to effectively detect and predict the occurrence of chronic diseases such as heart disease, diabetes, and depression based on the available dataset features?

- Correlation analysis identifies age, diabetes, arthritis, and sex as possible risk factors for heart disease. When determining a person's risk of heart disease, several variables may be taken into account.

2. What actionable insights and recommendations can be derived from the dataset to help prevent chronic diseases?

- The distribution of BMI points to a possible link between heart disease and higher BMI. Assessing one's risk for heart disease may benefit from BMI monitoring.

3. Can we build a predictive model to estimate a person's general health status based on their age, BMI, and other relevant features?

- To help with early risk detection and preventive interventions, a prediction model that incorporates age, BMI, and other relevant factors can be created to estimate an individual's general health state.

4. Are there any associations between smoking history and other health-related variables such as BMI, exercise habits, or alcohol consumption?

- People who have smoked in the past seem to be more likely to develop heart disease. Taking care of smoking cessation initiatives and programs could help control the risk of heart disease.

5. What is the prevalence of skin cancer, and how does it vary by age category and gender?

- Gender and age group differences exist in the occurrence of skin cancer. Using these trends to inform the design of preventative and awareness initiatives may increase their effectiveness.

8. Conclusion

To sum up, the research offers insightful information about potential risk factors for heart disease. After adjusting its hyperparameters, the Random Forest Classifier model produced an accuracy of about 38.8%. While extra feature research and optimization may improve the model's performance, it can still be used for predictive purposes.

Targeted interventions and individualized healthcare plans are made possible by an understanding of important correlations, risk factors, and lifestyle characteristics. To make wise decisions, it is essential to evaluate model outputs in conjunction with domain knowledge and medical experience. The visualizations that are shown help to explain the complex linkages and patterns found in the dataset. Improvement of features, ongoing observation, and cooperation between data scientists and medical practitioners can lead to more precise forecasts and improved health results.

It is critical to recognize the limits and possible flaws in the dataset, just like with any prediction model. A strong and dependable predictive analytics system must include frequent updates, a variety of datasets, and continuous validation.

9. References

GeeksforGeeks. (2018). Change Data Type for one or more columns in Pandas Dataframe. [online] Available at: https://www.geeksforgeeks.org/change-data-type-for-one-or-more-columns-in-pandas-dataframe/.

pandas.pydata.org. (n.d.). pandas.DataFrame.astype — pandas 1.4.2 documentation. [online] Available at: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.astype.html.

Bonthu, H. (2021). Detecting and Treating Outliers | How to Handle Outliers. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/.

Stack Overflow. (n.d.). BMI calculation from two columns of a pandas data frame with missing values. [online] Available at: https://stackoverflow.com/questions/71609487/bmi-calculation-from-two-columns-of-a-pandas-data-frame-with-missing-values [Accessed 17 Nov. 2023].

health.gov. (n.d.). Behavioral Risk Factor Surveillance System (BRFSS) - Healthy People 2030 | health.gov. [online] Available at: https://health.gov/healthypeople/objectives-and-data/data-sources-and-methods/data-sources/behavioral-risk-factor-surveillance-system-brfss#:~:text=The%20Behavioral%20Risk%20Factor%20Surveillance.

CDC (2020). Assessing Your Weight. [online] Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/healthyweight/assessing/index.html#:~:text=If%20your%20BMI%20is%20less.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. doi:https://doi.org/10.48550/arxiv.1602.04938.

Tanmoy Sarkar Pias, Su, Y., Tang, X., Wang, H. and Yao, D. (2023). Enhancing Fairness and Accuracy in Type 2 Diabetes Prediction through Data Resampling. *medRxiv (Cold Spring Harbor Laboratory)*. doi:https://doi.org/10.1101/2023.05.02.23289405.

www.kaggle.com. (n.d.). Cardiovascular Diseases Risk Prediction Dataset. [online] Available at: https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset.

Shaikh, I. (2023). *End-to-End-Data-Science-Project*. [online] GitHub. Available at: https://github.com/Ishrat2903/End-to-End-Data-Science-Project/tree/main [Accessed 28 Nov. 2023].