**Exploratory Data Analysis for Heart Disease Dataset**

This document summarizes the exploratory data analysis (EDA) performed on a heart disease dataset. The main steps include:

**1. Data Loading and Cleaning**

* The heart disease dataset was loaded.
* The dataset was checked for missing values and duplicates.
* Duplicate rows were removed.
* The distribution of the target variable ('HeartDisease') was visualized.

**2. Data Visualization**

* Stacked bar charts were used to visualize the relationship between various features (e.g., 'SkinCancer', 'AlcoholDrinking', 'Stroke', etc.) and the target variable 'HeartDisease'. This helps in understanding how different features might influence the presence of heart disease.

**3. Handling Class Imbalance**

* The dataset likely had an imbalance in the target variable (more 'No Heart Disease' instances than 'Heart Disease').
* To address this, undersampling was applied to balance the class distribution. This ensures that the model is not biased towards the majority class.
* After undersampling, the feature relationships with the target variable were visualized again to see if there were any changes.

**4. Feature Selection**

* **One-Hot Encoding:** Categorical features were converted into numerical representations using one-hot encoding. The binary values had k-1 encoding (dummy encoding), where as the the other categorical values use one-hot encoding.
* **Correlation Analysis:**
  + **Point-Biserial Correlation:** Used to measure the correlation between continuous features (e.g., 'BMI', 'PhysicalHealth') and the binary target variable.
  + **Cramer's V:** Used to measure the association between categorical features and the target variable.
* The features were ranked based on their correlation with the target variable.
* Different feature sets were created by selecting the top N correlated features. This is for experimenting with different numbers of features to see how it affects model performance.We did top 10 features , top 20 features and finally all features

Websites used:

* <https://medium.com/analytics-vidhya/one-hot-encoding-method-of-feature-engineering-11cc76c4b627>

This EDA process helps in understanding the data, identifying potential issues, and preparing the data for further analysis and modeling.