**DocAssist (Building Intelligent Medical Decision Support System)**

**Capstone Project**

**Ritam Das**

**05/12/2023**

**Link:** <https://github.com/Ritam1D/KnowledgeHut_Capstone>

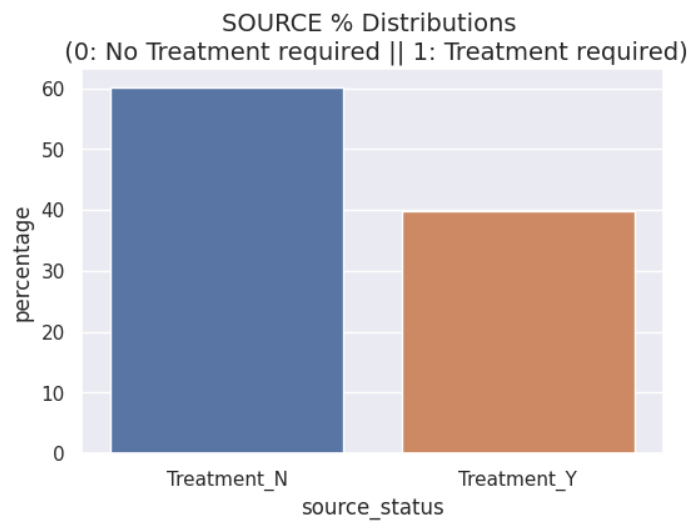
**Executive Summary:** In this project, we used machine learning approaches to analyzes patient data to assist doctors in making informed decisions about the best treatment options for individual patients.

**Problem Statement:**

The objective of this project is to develop an intelligent medical decision support system that analyzes patient data to assist doctors in making informed decisions about the best treatment options for individual patients. By leveraging machine learning and data analysis, the system will provide personalized treatment recommendations based on the patient's medical history, symptoms, lab results, and other relevant factors.  
  
**Objectives**:To Develop an Effective medical decision support system Model using, Machine Learning Models & EnhanceSecurity for Financial Transactions.

**Exploratory Data Analysis**:





* There are 2 classes available 0 stands for No Treatment required, whereas 1 stand for Treatment required.
* The dataset is **unbalanced**, the **positive class (**Treatment required**)** account for **39.8%** of all transactions.

**Handling Missing Values:** We began by identifying and addressing missing values in the dataset.NoMissing data points were found in any columns. So, no missing value treatment done.

**Dealing with Duplicates:** No Duplicates data points were found in the dataset which can distort model training. So, ensuring each data points entry is unique.

**Outliers Detection and Treatment:** NoOutliers, or extreme values in the data points found which can significantlyaffect model performance.

**Impact on the Model:** Proper data cleaning and outlier treatment are critical for model training.These steps help ensure that the model learns patterns and relationships in the data accurately, without being skewed by noise or inconsistencies. Inaccurate data pre-processing can lead to overfitting or underfitting, reducing the model's generalization performance on unseen data. These techniques enhance model performance by preventing certain features from dominating others during training. Outliers Detection and Treatment: Outliers, or extreme values in the data, can significantly affect model performance.

**Class Imbalance:** Class imbalance is a situation in which there are significantly more instances of oneclass than the other in the context of medical decision support system detection (or any other binary classification problem). For example, in the detection of assisting the doctor, the majority of No treatment required (Class 0), and minority are genuinely Treatment required (Class 1).

**Challenges with Imbalanced Data:** This class imbalance poses challenges during model training.Models tend to be biased towards the majority class, achieving high accuracy simply by predicting the majority class most of the time.

**Techniques to Balance Data:** Several techniques address this issue, broadly falling into two

categories: under sampling and oversampling.

1. **Under sampling**: Removing instances from the majority class to balance the class distribution.However, this can lead to a loss of valuable information.
2. **Oversampling**: Duplicating or generating new instances for the minority class to balance the classdistribution. Oversampling is often achieved through techniques like Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples based on existing minority samples.

**Why Used Oversampling:** Oversampling is a commonly used technique to handle imbalanced databecause it allows the model to learn from the existing minority class instances while generating new data points. By creating synthetic samples, the model can better understand the features and patterns associated with the minority class, improving its ability to predict the rare events accurately. This helps prevent the model from being biased towards the majority class and leads to a more balanced and effective machine learning model for medical decision support system detection.

**Encoding Techniques:** When working with categorical data in machine learning, it is essential to convert these variables into a numerical format that algorithms can understand. Two commonly used techniques for encoding categorical variables are **one-hot** and **label encoding**. Choosing the appropriate encoding method can significantly impact the performance of a ML model.

**Explanation of the chosen models:**

In this medical decision support system detection, selecting the right machine learning models was crucial for achieving accurate and reliable results. Here's an explanation of the chosen models and the reasons for their selection:

1. **Logistic Regression:** Logistic Regression is a simple yet effective linear classification algorithm.

It serves as a baseline model due to its simplicity and interpretability. It helps understand the importance of each feature in predicting.

1. **Random Forest:** Random Forest is an ensemble learning method that combines multiple decision

trees. Random Forest is robust, handles non-linearity, and can capture complex interactions between features. It also provides feature importance, aiding in understanding which features contribute most in detection.

1. **Grid Search CV**: Grid Search uses a different combination of all the specified hyperparameters and their values and calculates the performance for each combination and selects the best value for the hyperparameters. This makes the processing time-consuming and expensive based on the number of hyperparameters involved.
2. **K-Fold Cross validation (kf):** The dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time. Performance metrics from each fold are averaged to estimate the model’s generalization performance.

**Features Used for Data Splitting:** When splitting the data into training and test sets, all featuresexcept the "SOURCE” features are used. The " SOURCE " feature or target column, which indicates whether a patient is in requirement of treatment or not (0 for No treatment required, 1 for treatment required), is the target variable we want to predict.

**Model Training and Hyperparameter Tuning:** The models are trained using the train test split in 70:30 ratio in the dataset, which comprises a subset of the original data. During the training process, each model learns to identify patterns and relationships between the features and the target variable. Hyperparameter tuning is performed to optimize the model's parameters, ensuring it performs at its best.

**Tools and Technologies Used in the Project**:

**IDE (Integrated Development Environment):** Google Colab Used as the primary codeeditor for writing and managing the project code.

**Programming Language:** Python: The core programming language used for data analysis, machinelearning, and model development.

**Data Manipulation and Analysis Libraries:** Numpy: Used for numerical and mathematical operations  
on data arrays. Pandas: Used for data manipulation, data cleaning, and exploratory data analysis. Scipy: Used for advanced scientific and technical computing tasks.

**Machine Learning Libraries:** Scikit-learn (sklearn): Utilized for building, training, and evaluating machinelearning models.

**Data Visualization Libraries:** Matplotlib: Used for creating static, animated, and interactive visualizations   
in Python. Seaborn: Built on top of Matplotlib, used for creating aesthetically pleasing statistical graphics.

**Sampling Techniques:** Sampling: Refers to techniques used for handling imbalanced datasets. This mayinclude oversampling the minority class or undersampling the majority class to balance the dataset.   
  
**Microsoft Excel:** Used for data pre-processing tasks, such as initial data exploration and cleaning.

**Accuracy:** Accuracy measures the proportion of correctly classified transactions. However, accuracy may notbe the most appropriate metric for imbalanced datasets.

**Precision:** Precision calculates the ratio of true positive predictions to all positive predictions. It assesses themodel's ability to avoid false alarms (i.e., classifying no treatment required as treatment required).

**Recall (Sensitivity):** Recall calculates the ratio of true positive predictions to all actual positive cases. Itassesses the model's ability to detect all treatment required.

**F1-Score:** The F1-Score is the harmonic mean of precision and recall, providing a balance between the twometrics. It is useful when there's an uneven class distribution.

**Performance Metrics Used for Evaluation:** Several performance metrics are used to evaluatethe models' effectiveness in medical decision support system detection detection:

**The different classification reports for different machine learning models (Logistic Regression, Random Forest, K-Fold Cross validation , and Grid Search CV used   
here are the key findings:**

**Logistic Regression (logr)**:

* Accuracy score 0.69
* The F1-score 0.68

**GRID SEARCH CV (grd\_src):**

* The F1-score 0.67

**Random Forest (rfm)**:

* Accuracy score 0.73
* The F1-score 0.72

**k-Fold Cross validation (kf)**:

* The F1-score 0.67

**Decision Tree (tree):**

* F1 score: 0.67
* Accuracy: 0.67

**General Observations:**

* Comparison with different model Random Forest perform better than other model with

accuracy: 0.73 and F1-score: 0.72

* For capturing more treatment required and assist the doctor, Random Forest might be considered.

**Conclusion:**

We will go with Random Forest a sit performed better than other model.