**OBJECTIVE:**

The goal is to utilize various machine learning models to predict if an individual possesses health insurance, determining which model yields the most accurate outcome. The target variable is referred to as the "Response."

**DATA AND DATABASE:**

The dataset, obtained from Kaggle, comprises 15 columns, including the target variable, and consists of 2974 records. To enhance data processing capabilities, I integrated Apache Spark with MongoDB.

**METHODOLOGY:**

I conducted exploratory data analysis (EDA) tasks such as analyzing feature distributions and creating a correlation heatmap of the predictors.

During the data preparation stage, I employed techniques like string indexing, one-hot encoding, vector assembly, label indexing, and standard scaling. I also made use of five transformers/estimators in a pipeline.

**MODELING:**

For the train-test split, 80% of the dataset was allocated to the train set, while the remaining 20% constituted the test set. I applied Logistic Regression, Decision Tree, Random Forest Classifier, and GBT Classifier on the processed dataset.

To evaluate model performance, I utilized a confusion matrix and a ROC-AUC curve. Logistic Regression achieved the highest performance with a score of 0.60.

**CONCLUSION:**

The objective was to predict whether individuals possess health insurance. The highest achieved score was 0.60, indicating the potential for further improvement through additional feature engineering, data transformations, and deeper analysis to gain insights and enhance the score.