**Aim:** To apply Logistic Regression for binary classification problems using machine learning, and assess model performance through appropriate evaluation metrics.

**Software Used:** Google Python Colab

**Theory:**

**Logistic Regression** is a fundamental **supervised machine learning** algorithm used for **classification tasks**, particularly **binary classification**, where the target variable has only two possible outcomes (e.g., Yes/No, Spam/Not Spam, Positive/Negative). Unlike **Linear Regression**, which predicts continuous values, Logistic Regression predicts the **probability** that an instance belongs to a particular class.

Unlike Linear Regression, which predicts continuous values, Logistic Regression predicts the **probability** that a given input belongs to a particular class. To ensure the predicted values lie between **0 and 1**, it uses the **sigmoid (logistic) function**, which is defined as:

S(z)=11+e−zS(z) = \frac{1}{1 + e^{-z}}S(z)=1+e−z1​

Where:

* **S(z)** is the sigmoid output (probability),
* **z** is the linear combination of input features and weights,
* **e** is the base of the natural logarithm.

The decision boundary for classification is determined by setting a **threshold** (commonly 0.5). If the predicted probability is greater than or equal to the threshold, the model assigns one class; otherwise, it assigns the other class.

The key idea behind Logistic Regression is the use of the **logistic (sigmoid) function** to map any real-valued number into a range between **0 and 1**, which can be interpreted as a probability. The sigmoid function is given by:

**P(y=1|x) = 1 / (1 + e^(-z))**

Where:

* P(y=1|x) is the probability that the dependent variable y belongs to class 1, given input x
* z = β₀ + β₁x₁ + β₂x₂ + ... + βₙxₙ is the linear combination of input features
* β₀ is the intercept, β₁, β₂, …, βₙ are the model coefficients

If the probability **P** is greater than a chosen **threshold** (commonly 0.5), the model predicts class 1; otherwise, it predicts class 0.

### **Types of Logistic Regression:**

1. **Binary Logistic Regression** – for two possible outcomes (Yes/No)
2. **Multinomial Logistic Regression** – for more than two unordered categories
3. **Ordinal Logistic Regression** – for ordered categories

### **Model Training:**

Logistic Regression estimates the model parameters using **Maximum Likelihood Estimation (MLE)**, which finds the parameter values that maximize the probability of correctly classifying the observed data.

### **Model Evaluation Metrics:**

* **Accuracy** – proportion of correctly classified instances
* **Precision** – proportion of positive predictions that are actually correct
* **Recall (Sensitivity)** – proportion of actual positives that are correctly predicted
* **F1 Score** – harmonic mean of Precision and Recall, useful for imbalanced datasets
* **ROC Curve & AUC** – measure the trade-off between sensitivity and specificity

### **Advantages:**

* Simple and easy to implement
* Works well for linearly separable data
* Outputs probabilities, which are useful in decision-making

**Understanding Sigmoid Function**

1. The sigmoid function is an important part of logistic regression which is used to convert the raw output of the model into a probability value between 0 and 1.

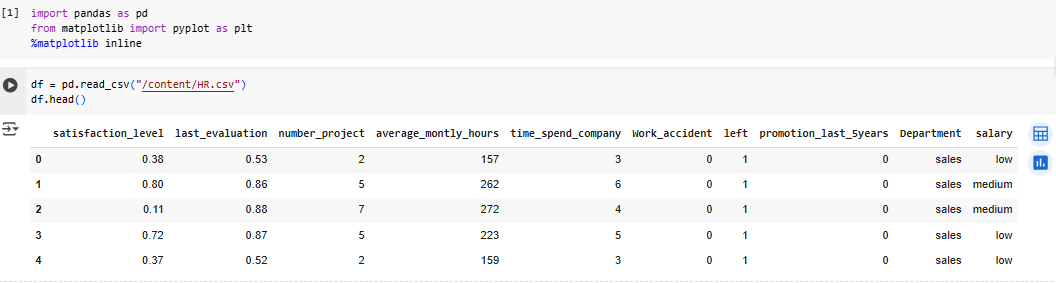
2. This function takes any real number and maps it into the range 0 to 1 forming an "S" shaped curve called the sigmoid curve or logistic curve. Because probabilities must lie between 0 and 1, the sigmoid function is perfect for this purpose.

3. In logistic regression, we use a threshold value usually 0.5 to decide the class label.

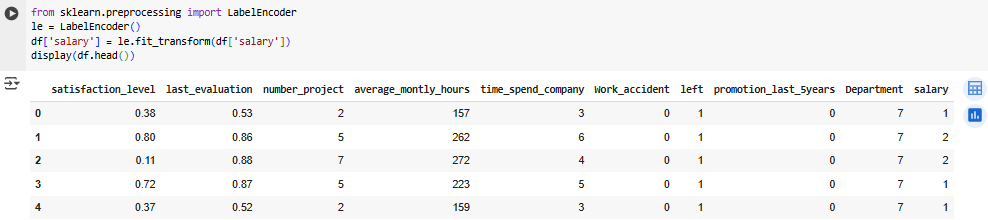
* If the sigmoid output is the same or above the threshold, the input is classified as Class 1.
* If it is below the threshold, the input is classified as Class 0.

**Output:**

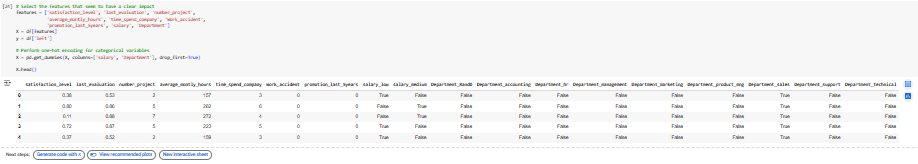
1. **HR**

****

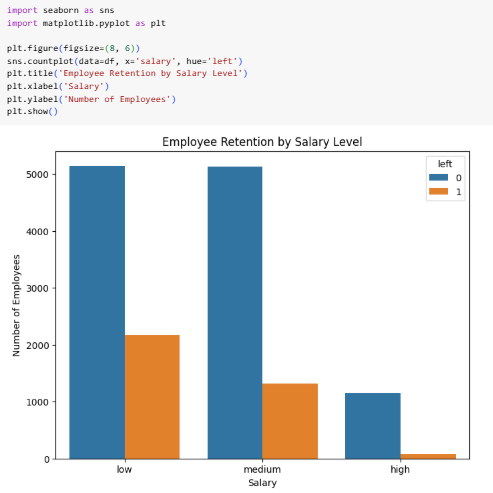
Preprocessed Data



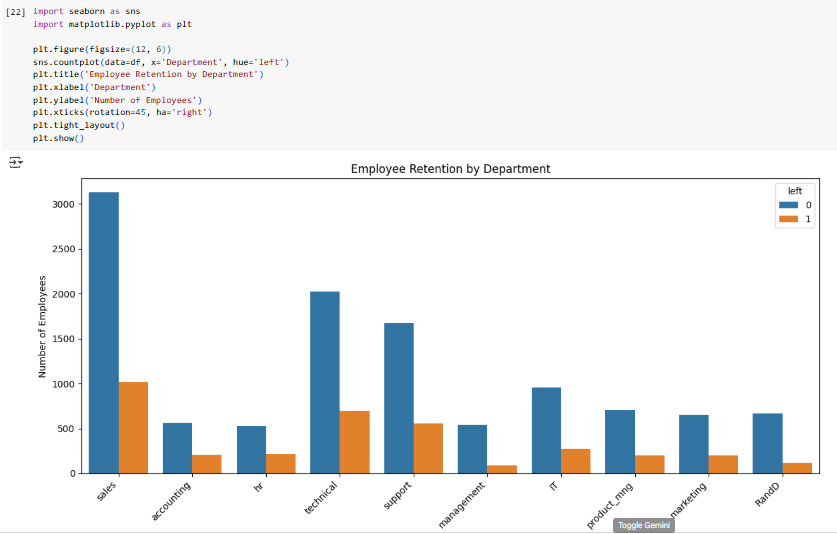
Exploratory data analysis to figure out which variables have direct and clear impact on employee retention (i.e. whether they leave the company or continue to work)

****

Plot bar charts showing impact of employee salaries on retention



Plot bar charts showing correlation between department and employee retention



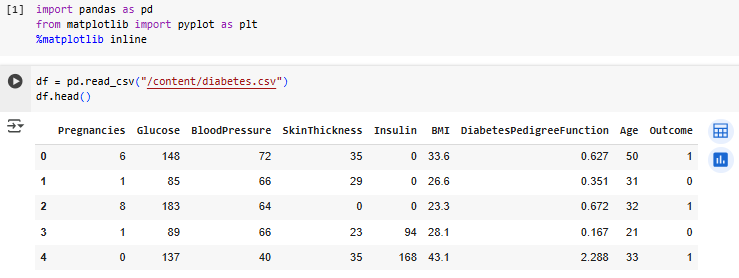
Build logistic regression model using variables that were narrowed down in step 1



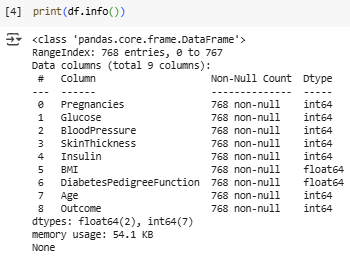
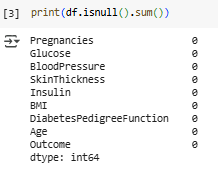
Measure the accuracy of the model



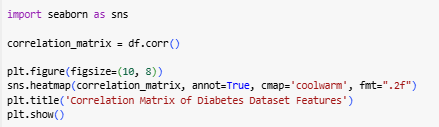
1. **Diabetes**

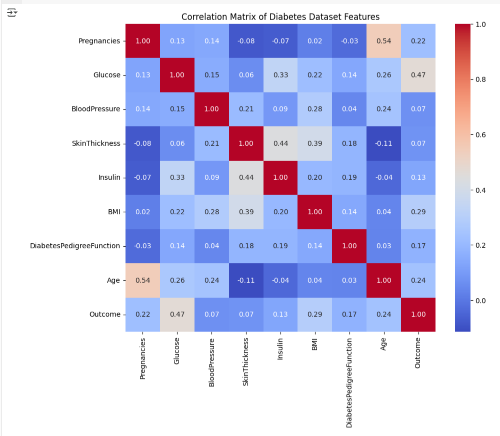
****

Preprocessing steps:

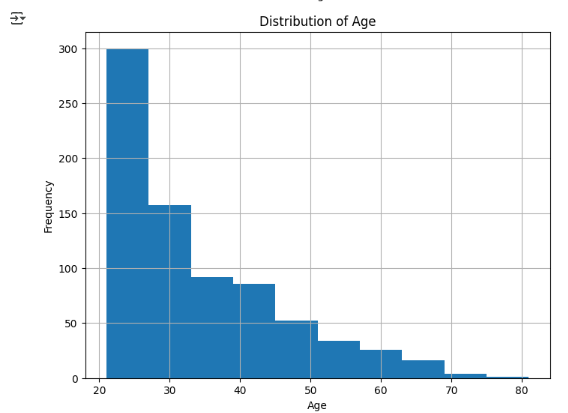
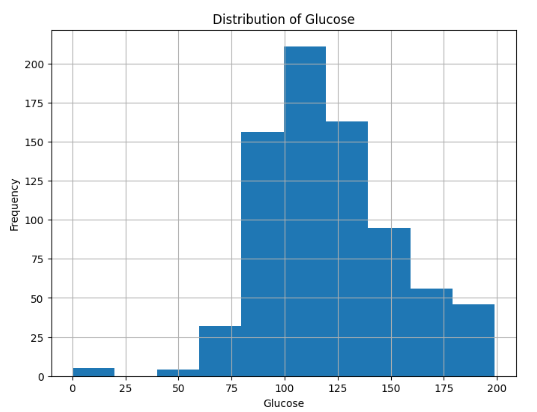


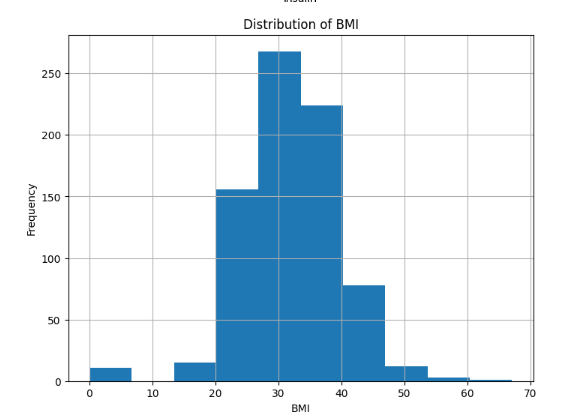
Perform **exploratory data analysis (EDA)** to understand the distribution of features and identify any strong indicators of diabetes.



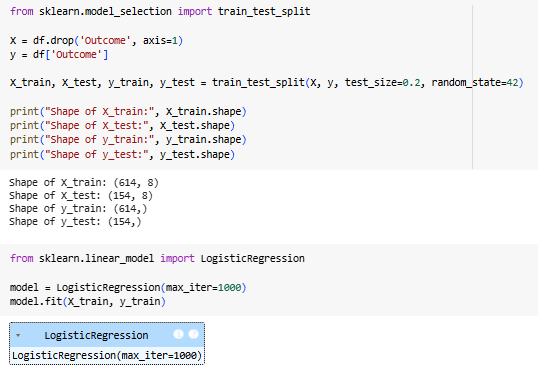


Visualize the relationship between selected features (like *glucose*, *age*, *BMI*) and the outcome using appropriate **bar charts or boxplots**.



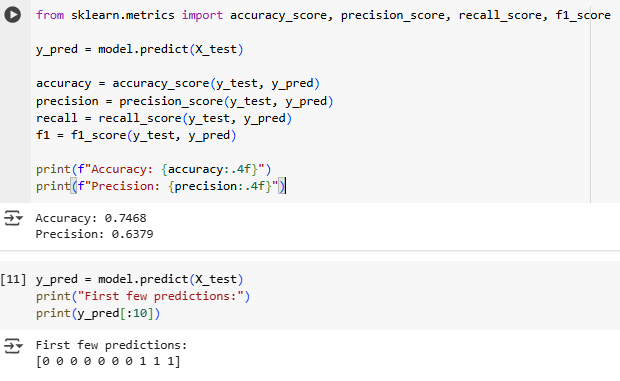


Build a **Logistic Regression model** to predict the Outcome (whether a person has diabetes or not).

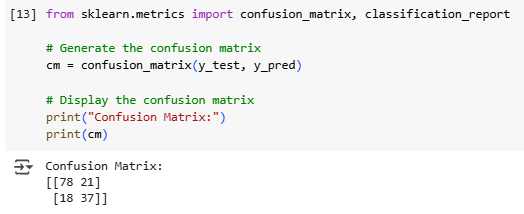


Evaluate the performance of your model using appropriate metrics such as:

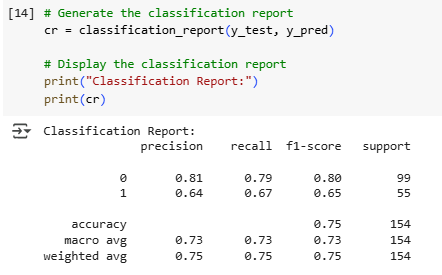
Accuracy



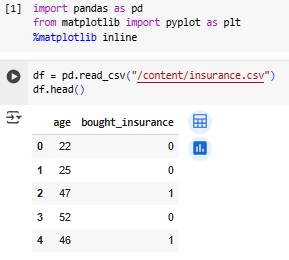
Confusion Matrix

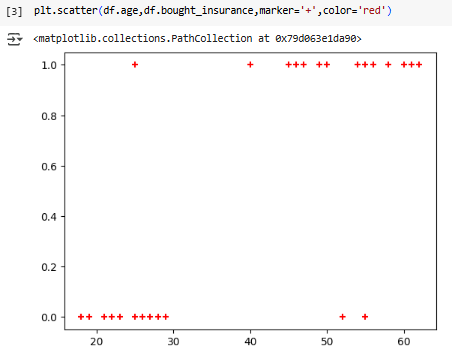


Classification Report

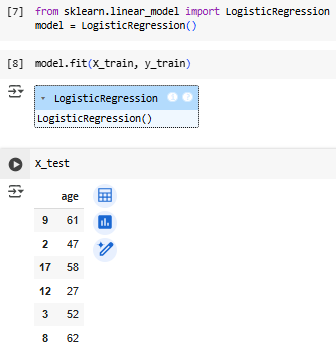


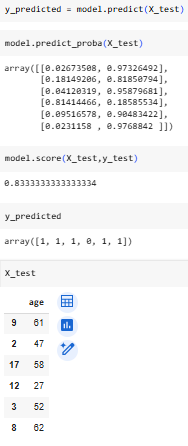
1. **Insurance**

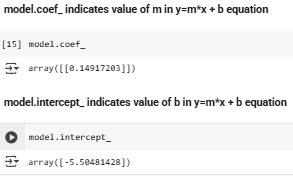


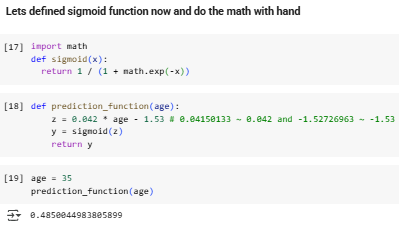


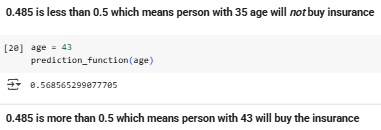












**Conclusion:** Hence, through this experiment, we learned how to apply Logistic Regression for binary classification problems, train the model using given datasets, and evaluate its performance using appropriate metrics such as accuracy, precision, recall, and F1-score. This process is essential for building reliable classification models and making accurate predictions in real-world applications.