**GOMINDZ ACADEMY**

**INTERMEDIATE DATA SCIENCE COURSE**

**OMAR JAWARA- CAPSTONE PROJECT**

**Question 2:** In a healthcare dataset, you have information about patients' age, blood pressure, and cholesterol levels, and you want to predict whether they have a heart disease (yes/no). Which regression approach (linear or logistic) would you employ, and what considerations led you to this decision?

Due to its unsuitability for binary categorical variables, linear regression would not be the appropriate choice for predicting the presence or absence of heart disease in this healthcare dataset. The target variable, 'Heart Disease,' is a binary categorical variable (Presence/Absence), and logistic regression is specifically designed to model the probability of such binary outcomes.

Logistic regression offers significant advantages by modeling the probability of a positive outcome (heart disease) based on independent variables like age, blood pressure, and cholesterol, outputting a value between 0 and 1 that can be interpreted as the likelihood of heart disease. Logistic regression does not require the independent variables to be normally distributed, a common characteristic of real-world healthcare data. However, it's crucial to remember that logistic regression has its own assumptions, such as linearity between the log odds of the outcome and the independent variables and no perfect multicollinearity. Checking these assumptions is a formality and a key part of the modeling process that ensures the model's validity and your responsibility as a data analyst.

Additionally, logistic regression provides interpretable coefficients that indicate the direction and magnitude of the relationship between each independent variable and the likelihood of heart disease. This ensures that you are well-informed about the model's output, giving you a sense of control and empowerment over the predictive process.

**Question 3**: You have a dataset containing historical stock price data (target variable: continuous stock price) and various market indicators such as trading volume, price-to-earnings ratio, and economic indicators. You aim to build a model for predicting future stock prices. Would you opt for linear or logistic regression, and what factors influence your decision?

For predicting future stock prices, linear regression would be more appropriate than logistic regression. Linear regression is designed to model the relationship between a continuous target variable, such as stock price, and one or more predictor variables, such as market indicators and economic factors. The goal is to find a linear equation that best predicts the stock price based on the values of these predictor variables.

In contrast, logistic regression is suitable for predicting categorical target variables, like whether a stock will go up or down, which involves binary classification and uses a sigmoid function to estimate the probability of the target variable belonging to a particular category. Several factors influence this decision. Firstly, the nature of the target variable is crucial; stock prices are continuous variables that fluctuate over a range of values, and linear regression excels at handling such continuous data and predicting future values within that range. This understanding of the target variable's nature is key to making an informed decision. While stock market relationships might be complex, the linearity assumption of linear regression can still provide useful insights and predictions, especially when combined with feature engineering. Additionally, linear regression models are generally more interpretable than logistic regression models, making understanding the impact of different market indicators on stock prices easier.

When considering data availability, if you have a large dataset with historical stock prices and various market indicators, linear regression can be effectively trained to capture patterns and relationships. However, it is important to consider the potential non-linearity of the stock market. In cases where non-linear relationships are present, linear regression may not be the most accurate method, and techniques like polynomial regression or other machine learning algorithms might be more suitable. The time series nature of stock prices should also be accounted for, as they are influenced by past values, necessitating the incorporation of time series analysis techniques into the model.