Logistic regression

why not use linear regression?

- it gives continuous, numerical outputs
- it has no boundaries (not applicable for probabilities) → it can predict from -ve infinity uo to infinity while probability lies between o and 1
- not applicable for non-linear relationship between features

purpose of logistic regression

it's a supervised learning algorithm that's used for binary classification, estimating probabilities ,handling non linear boundaries,and providing insight for important features

when to use ?

- on having binary target vars
- when interpretability of feature importance is high
- when you need a simple, fast, and effective model classification

Sigmoid function

a mathematical function takes any real number and convert it to a value between 0 and 1, "has an S shaped curve"

1. used for non-linear transformation
2. used for threshold based classification

why squash the line?

- handling outliers

- gradient-based optimization (used in cost functions by logistic regression)

confusion matrix:

is a simple table that shows how well a classification model is performing by comparing its predictions to the actual results.

it breaks the results into:

correct predictions for both classes (**true positives** and **true negatives**) and incorrect predictions (**false positives** and **false negatives**).

- **True Positive (TP):** The model correctly predicted a positive outcome (the actual outcome was positive).
- **True Negative (TN):** The model correctly predicted a negative outcome (the actual outcome was negative).
- **False Positive (FP):** The model incorrectly predicted a positive outcome (the actual outcome was negative).
- **False Negative (FN):** The model incorrectly predicted a negative outcome (the actual outcome was positive).

Accuracy, Precision, Recall and F1 Score

Accuracy:

"Out of all the predictions we made, how many were true?"

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ negatives + false\ positives}$$

Precision:

"Out of all the positive predictions we made, how many were true?"

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Recall:

*"**Out of all the data points that should be predicted as true, how many did we correctly predict as true?**"*

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

F1 Score:

"a measure that combines recall and precision"

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

AUC and the ROC:

***ROC** curve gives a visual representation of the trade-offs between the true positive rate (TPR) and false positive rate (FPR)*

AUC, or Area Under the Curve, is a single scalar value ranging from 0 to 1, that gives a performance snapshot of the model.

-You only calculate AUC after generating the ROC curve because the AUC represents the area beneath the curve.-