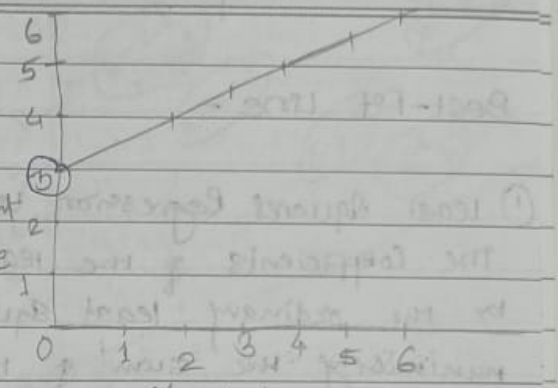
Regression



① Intercept of the straight line -
What is the intercept of the given line? use the graph given above to answer this question

- 3, the value of y when x=0 in the given straight line is 3, so 3 would be the intercept in this case

② Slope of a straight line -
What is the slope of the given line? use the graph given above to answer this question.

- 1/2, the slope of any straight line can be calculated by $(y_2-y_1)/(x_2-x_1)$ where (x_1, y_1) & (x_2, y_2) are any two points through which the given line passes. This line passes (0, 3) & (2, 4), so the slope of this line would be

$$((4-3)/(2-0)) = 1/2$$

$$\frac{(y_2-y_1)}{(x_2-x_1)} \quad \because \underset{0,\ 3}{(x_1, y_1)} \quad \underset{(2,\ 4)}{(x_2, y_2)} \quad \therefore \frac{y_2-y_1}{x_2-x_1}$$

③ What - Equation of a straight line -
What would be the equation of the given line.

$-y = \frac{x}{2} + 3$ ∵ the std eqn of a straight line $y = mx + c$

Where m is the slope & c is the intercept. In this case m = 1/2 c = 3 ∴ $Y = \frac{X}{2} + 3$

Best-Fit Line -

① Least Squares Regression Line -
The coefficients of the least squares regression line are determined by the ordinary least squares method - which basically means minimising the sum of the squares of the

- y-coordinates of actual data - x-coordinates of predicated data

The ordinary Least Squares method has the criterion of the minimisation of the sum of squares of residuals. Residuals are defined as the differences between the y-coordinates of actual data & the y-coordinates of predicted data.

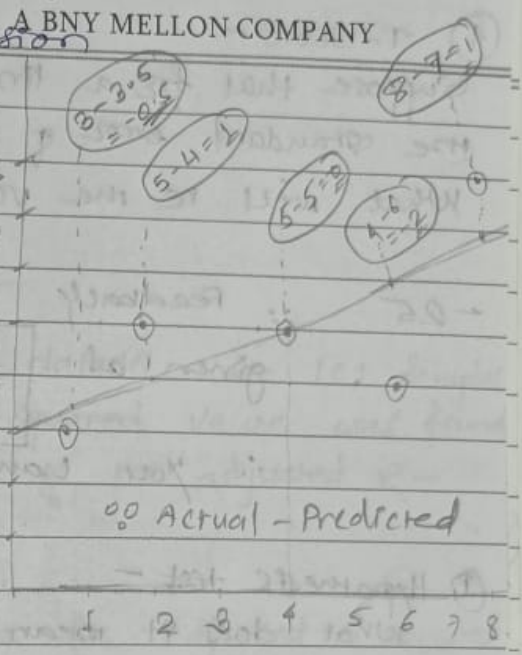
② Best Fit Regression Line -
What is the main criterion used to determine the best fitting regression line?

- The line that minimises the sum of squares of distances of points from the regression line -
The criterion is given by the Ordinary least Squares (OLS) method, which states that the sum of the squares of residuals should be minimum. This is explained by this option.

~~Straight~~ Strength of Simple Linear Regression

(1) Residual Sum of Squares (RSS)
Find the value of RSS for this Regression line

- 6.25 ∵ the residual for all 5 points are -0.5, 1, 0, -2, 1. The sum of squares of all 5 residuals would be
0.25 + 1 + 4 + 1 = 6.25



(2) Total Sum of Squares (TSS)
Find the value of TSS for this regression line

- 14 ∵ the average of y-value for all data points (3+5+5+4+8)/5 = 25/5 = 5. So $y - \bar{y}$ term for each data pot would be -2, 0, 0, 1, 3. So the squared sum of these terms would be 4 + 1 + 9 = 14

(3) R^2 - The RSS for this example comes out to be 6.25 & TSS comes out to be 14.
What would be the R^2 for this regression line

- 1 - (6.25 / 14) ∵ R^2 value is given by $1 - \frac{RSS}{TSS}$

So in this case $R^2 = 1 - \left(\frac{6.25}{14}\right) = 1 - 0.4464 = 0.5536$

Hypothesis Testing in Linear Regression -

④ T-score -

Suppose that for a linear model, you got β_1 as 0.5. Also the standard error of β_1 was found out to be 0.02. What will be the value of t-score for β_1

-25 ∴ Feedback Recall that the t-score for β_1 is given as $\frac{\beta_1}{SE(\beta_1)}$

∴ You have: $t\text{-score} = \frac{0.5}{0.02} = 25$

① Hypothesis test -

What does it mean if you fail to reject the Null hypothesis in the case of simple linear Regression.

Given - Null Hypothesis $(H_0): \beta_1 = 0$

Alternate Hypothesis $(H_A): \beta_1 \neq 0$

—

β_1 of Null, the independent variable it is associated with is insignificant in the prediction of the dependent variable.

Feedback - Correct! the Null Hypothesis in simple linear Regression is: $\beta_1 = 0$

Thus, if we fail to reject the Null Hypothesis, it means that β_1 is indeed zero, & thus insignificant for the prediction of the independent variable

Model Assessment & Comparison -

$$\text{adjusted } R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

(1) Calculating Adjusted R-Squared -

When a model was built from a dataset with 101 samples and 10 predictor variables, the R-squared value was found to be 0.7. What will the value of the adjusted R-squared be for the same model?

- 0.67 : Feedback : The formula for adjusted R-squared is given as :

$$1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

So, substituting the given values at the appropriate places gives us

$$1 - \frac{(1-0.7)(101-1)}{101-10-1} \simeq 0.67$$

(2) Model Assessment

- After performing inferences on a linear model built with several variables, you concluded that the variable 'r' was insignificant. This meant that the variable 'r'

- Had a high p-value

A high p-value means that the variable is not significant, and hence, doesn't help much in prediction.

(1) R-Squared values -

Suppose you built a model with some features. Now you add & add another variable to the model. Which of the following statements would be true.

- R-squared value will either increase or remain the same.
- The Adjusted R-Squared value may increase or decrease

(2) Overfitting -

- Number of data points are less

∴ Correct! Overfitting is the condition where in the model is so complex that it ends up memorizing almost all the data points on the train set. Hence, this condition is more probable if the number of data points is less since the model passing through almost every point becomes easier.

(3) VIF -

VIF is a measure of:

- How well a predictor variable is correlated with all the other variables, excluding the target variable

(4) Calculating VIF -

Suppose you were predicting Sales of a company using two variables 'Social Media Marketing' and 'TV Marketing'. You found out that the correlation between 'Social Media Marketing' & 'TV Marketing' is 0.9. What will be the approximate value of VIF for either of them?

- 5.26

Correct! The formula for VIF is given by:

$$VIF = \frac{1}{1 - R_i^2}$$

$\therefore$ VIF = $\frac{1}{1 - 0.9^2}$

≈ 5.26

$\therefore$ Here, the R-squared variable will be simply the correlation coefficient squared since we have only 2 variables. Hence, you have:

$= \frac{1}{1 - 0.81}$

$= 1/0.19$

≈ 5.26

(1) Dummy Variables –

Suppose you have 'n' categorical variables, each with 'm' levels. How many dummy variables would you need to represent all the levels of all the categorical variables?

$\rightarrow$ $(m-1) * n$ $\therefore$ Each of the dummy variable has 'm' levels. So to represent one categorical variable, you would require (m-1) levels. Hence, to represent 'n' categorical variables, you would need $(m-1) * n$ dummy variables.

(2) Automated feature selection –

Which of the following is/are an example of an automated approach for linear regression

- Recursive feature Elimination
- Stepwise selection using AIC
- Regularisation

⑤ Redundant Variables —

After performing inferences on a linear model built with several variables, you concluded that the variable 'r' was almost being described by other feature variable. This meant that the variable 'r':

— Had a high VIF ∵ correct! If the variable is being described well by the rest of the feature-variable, it means it has a high VIF meaning it is redundant in the presence of the other variables.

① Comprehension questions —

If $\beta_1 = \beta_2 = 0$ holds & $\beta_3 = 0$ fails to hold, then what can you conclude?

— there is a linear relationship between the outcome variable (Y) & x_3

Feedback — since $\beta_3 = 0$ fails to hold, this means that x_3 is a significant variable in this linear regression model thus, we can say that there is a linear relationship between the outcome variable (Y) & x_3

② comprehension questions —

If $\beta_1 = \beta_2 = \beta_3 = 0$ holds true, then what can you conclude

— there is no linear relationship between Y and any of the 3 independent variables

Feedback — If all the coefficients are found insignificant, then there cannot be a linear relationship between Y & any of the variables

Python Coding Problem
Mapping Variables

'Yes' → 1 'No' → 0 'Maybe' → 0.5
'yes' → 1 'no' → 0 'maybe' → 0.5
'YES' → 1 'NO' → 0 'MAYBE' → 0.5

```
import pandas as pd
df = pd.DataFrame({'Name': name, 'Response': response})
df['Response'] = df['Response'].str.lower()
df['Response'] = df['Response'].map({'yes': 1.0, 'maybe': 0.5
                                     ,'no': 0})
print(df)
```


(2) Residuals-
In regression analysis, which of the statement is true.

- The mean of residuals is always equal to zero
Feedback - When a model gives you a "best fit" line, by design it is made such that the mean of all residuals is always zero

- The Sum of residuals is always equal to zero
Feedback - When a model gives you a "best fit" line, by design it is made such that the Sum of all residuals is always zero.

Sigmoid Curve –

This is the sigmoid curve equation:

$$Y = P(\text{Diabetes}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

. Here let's say you take $\beta_0 = -15$ & $\beta_1 = 0.065$. Now, what will be the probability of diabetes for a patient with sugar level 220?

– 0.33 ∵ $P(\text{Diabetes}) = \frac{1}{1+e^{-(-15+0.065*220)}}$

$= \frac{1}{1+e^{(0.7)}}$

$= \frac{1}{1+2.01}$

$= \frac{1}{3.01}$

$= 0.33$

(2) Sigmoid curve –

For the sigmoid curve ($\beta_0 = -15$ & $\beta_1 = 0.065$), what will be the probability of diabetes for a patient with sugal level 240?

– 0.645 ∵ $Y = P(\text{Diabetes}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

$= \frac{1}{1+e^{-(-15+0.065\times 240)}}$

$= \frac{1}{1+e^{-(-15+15.6)}}$

$= \frac{1}{1+e^{-0.6}}$

$= \frac{1}{1+0.5488}$

$= \frac{1}{1.5488}$

$= 0.645$

① Likelihood

Now, let's say that for the ten points in our example, the labels are as follows:

Points nos:	1	2	3	4	5	6	7	8	9	10
Diabetes:	no	no	no	Yes	no	Yes	Yes	Yes	Yes	Yes

In this case, the likelihood would be equal to:

$$(1-P_1)(1-P_2)(1-P_3)(1-P_5)(P_4)(P_6)(P_7)(P_8)(P_9)(P_{10})$$

∵ Feedback –
Recall that likelihood is the product of $(1-P_i)$ for all non-diabetic patients & (P_i) for all diabetic patients. Hence, the likelihood is given by $(1-P_1)(1-P_2)(1-P_3)(1-P_5)$, (all non-diabetic patients) multiplied by $(P_4)(P_6)(P_7)(P_8)(P_9)(P_{10})$ (all diabetic patients)

① Log Odds –

So, let's say that the equation for the log odds is –

$$\ln\left(\frac{P}{1-P}\right) = -13.5 + 0.06x$$

$\because \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$

For $x = 220$, the log odds are equal to

$-13.5 + (0.06 * 220) = -0.3$

For $x = 231.5$, log odd are equal to:

-0.39 $\because -13.5 + (0.06 \times 231.5)$

$= 0.39$

$\frac{P}{1-P} = \text{odds}$

$\ln\left(\frac{P}{1-P}\right) = \text{Log Odds}$

① Standardising Variables –

In a dataset with mean 50 & standard deviation 12, what will be the value of a variable with an initial value of 20 after you standardise it?

-2.5

∵ The formula for standardising a value in a dataset is given by $\frac{(X - \mu)}{\sigma}$

$\therefore \frac{(20 - 50)}{12} = \frac{-30}{12} = -2.5$

② Correlation Table:

Command can be used to view the correlation table for the dataframe telecom?

⇒ telecom.corr() ; ∵ telecom.corr() will give you correlation table for the dataframe telecom.

③ p-values –

After learning the coefficients of each variable, the model also produces a 'p-value' of each coefficient. Fill in the blanks so that the statement is correct:

"The null hypothesis is that the coefficient is _____. If the p-value is small, you can say that the coefficient is significant & hence the null hypothesis _____"

– zero, can be rejected

∵ Feedback: Yes! Recall that the null hypothesis for any beta was $\beta_i = 0$

And if the p-value is small, you can say that the coefficient is significant, & hence you can reject the null hypothesis that $\beta_i = 0$

(3) Accuracy Calculation -

From the table you used for the last 2 questions, what will be the accuracy of the model?

Actual / Predicted	NO	Yes
No	400	100
Yes	50	150

- 78.5% ∴ The accuracy of a model is given by:

$$\text{Accuracy} = \frac{\text{Correctly Predicted Labels}}{\text{Total Numbers of Labels}}$$

or

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

∴ The number of correctly Predicted labels as you found out from the last question is equal to 550.

The total numbers of Label is $(400 + 100 + 50 + 150) = 700$.

Hence the accuracy becomes:

$$\text{Accuracy} = \frac{550}{700} = 0.785 \approx 78.5\%$$

(1) Logistic Regression in Python -

Which of these methods is used for fitting a logistic regression model using statsmodels

- GLM() ∴ feedback : Correct! the GLM() method is used to fit a logistic regression model using statsmodels

Log Odds :

Suppose you are working for a media service company like Netflix. They're launching a new show called 'Sacred Games' & you are building a logistic Regression model which will predict whether a person will like it or not based on whether consumers have liked / disliked some previous shows. You have the data of five of the previous shows & you're just using the dummy variables for these five shows to build the model. If the variable is 1, it means that the consumer liked the show & if the variable is zero it means that the consumer didn't like the show. The following table shows the value of coefficients for these five shows that you got after building the logistic regression model.

Variable Name	Coefficient Value
TrueDetective_Liked	0.47
ModernFamily_Liked	-0.45
Mindhunter_Liked	0.39
Friends_Liked	-0.23
Narcos_Liked	0.55

Now you have the data of three consumers Reetesh, Kshitij & Shruti for these 5 shows indicating whether or not they liked these shows. This is shown in the table below :

Consumer	True Detective-Liked	Modern Family-Liked	Mindhunter-Liked	Friends-Liked	Narcos-Liked
Reetesh	1	0	0	0	1
Kshitij	1	1	1	0	1
Shruti	0	1	0	1	1

Based on this data, which one of these three consumers is most likely to like to new show 'Sacred Games'?

- Reetesh Feedback: Correct!

To find the person who is most likely to like the show, you can use log odds. Recall the log odds is given by-

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_n x_n$$

Here, there are five variable for which the coefficients are given. Hence, the log odds become:

$$\ln\left(\frac{p}{1-p}\right) = 0.47x_1 - 0.45x_2 + 0.39x_3 - 0.23x_4 + 0.55x_5$$

As you can see, we have ignored the β_0 since it will be the same for all the three consumers. Now using the values of the 5 variables given, you get-

$$(\text{Log Odds})_{\text{Reetesh}} = (0.47\times1) - (0.45\times0) + (0.39\times0) - (0.23\times0) + (0.55\times1)$$
$$= 1.02$$

$$(\text{Log Odds})_{\text{Kshitij}} = (0.47\times1) - (0.45\times1) + (0.39\times1) - (0.23\times0) + (0.55\times1)$$
$$= 0.96$$

$$(\text{Log Odds})_{\text{Shruti}} = (0.47\times0) - (0.45\times1) + (0.39\times0) - (0.23\times1) + (0.55\times1)$$
$$= -0.13$$

You can clearly see, the log Odds of Recresh is the highest, hence, the odds of Recresh liking the show of the guest & hence, he is most likely to like the new show, Sacred Games

Sensitivity & Specificity –

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Actual / Predicted	Not Churn	Churn
Not Churn	TN	FP
Churn	FN	TP

$$\text{False Positive Rate} = \frac{FP}{TN + FP}$$ ← False Positive Rate (FPR)

$$\text{Positive Predictive Value} = \frac{TP}{TP + FP}$$ ← ~~True Positive Rate (TPR)~~

$$\text{Negative Predictive Value} = \frac{TN}{TN + FN}$$ ←

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN}$$

$$\therefore TPR = \text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\therefore FPR = 1 - \text{Specificity} = 1 - \frac{TN}{TN + FP}$$

$$= \frac{FP}{TN + FP}$$

True Positive Rate -

You have the following table showcasing the actual 'churn' labels & the predicted probabilities for 5 customers.

Customer	Churn	Predicted Churn Probability
Thulasi	1	0.52
Aditi	0	0.56
Jaideep	1	0.78
Ashok	0	0.45
Amulya	0	0.22

Calculate the True Positive Rate & False Positive Rate for the cutoffs of 0.4 & 0.5. Which of these cutoffs, will give you a better model?

Note - The good model is the one in which TPR is high & FPR is low.

- Cutoff of 0.5

Feedback:

Now, at the cutoff of 0.4, you get the following values of the predicted probabilities -

Customer	Churn	Pred. Chur Prob.	Predi. Churn Label
Thulasi	1	0.52	1
Aditi	0	0.56	1
Jaideep	1	0.78	1
Ashok	0	0.45	1
Amulya	0	0.22	0

From the above table, you can easily calculate:

True Positives = 2
False Positives = 2

Also, from the original table, you have:

Actual Positives = 2
Actual Negatives = 3

Hence you get:

$$TPR = \frac{\text{True Positives}}{\text{Total Actual Positives}} = \frac{2}{2} = 100\%$$

$$FPR = \frac{\text{False Positives}}{\text{Total Actual Negatives}} = \frac{2}{3} \approx 67\%$$

Performing similar steps for a cutoff of 0.5 will give you—

Customer	Churn	Pred. Churn Prob.	Pred. Churn Label
Thulasi	1	0.52	1
Aditi	0	0.56	1
Jaideep	1	0.78	1
Ashok	0	0.45	0
Amulya	0	0.22	0

From the above table, you can easily calculate:

True Positive = 2
False Positive = 1

Also, from the original table, you have

Actual Positives = 2
Actual Negatives = 3

Hence you get -

$$TPR = \frac{\text{True Positives}}{\text{Total Actual Positives}} = \frac{2}{2} = 100\%$$

$$FPR = \frac{\text{False Positives}}{\text{Total Actual Negatives}} = \frac{1}{3} = 33\%$$

As you can see, with both the cutoffs, the TPR is 100%. but for the cutoff of 0.5 you have a lower value of FPR. So clearly, a cutoff of 0.5 gives you a better model.

Please note that 0.5 just gives the better model among 0.4 & 0.5. It might be possible that there is a cutoff point which gives an even better model.

(2) TPR and FPR -

Fill in the blanks:

When the value of TPR increases, the value of FPR _____

- Increases. : Feedback - Correct! This can be clearly seen from the ROC curve as well, when the value of TPR (on the Y-axis) is increasing, the value of FPR (on the X-axis) also increases

(3) Area Under the curve -

You have the following five AUC (Area Under Curve) for ROC's plotted for five different models, which of these model is the best?

Model -	A	B	C	D	E
AUC -	0.54	0.82	0.79	0.66	0.56

Ans - B (0.82)

1) Choosing the optimal Cut-off -

Suppose you created a dataframe to find out the optimal cut-off point for a model you built. The dataframe looks like the following:

Threshold	Probability	Accuracy	Sensitivity	Specificity
0.0	0.0	0.21	1.00	0.00
0.1	0.1	0.39	0.96	0.22
0.2	0.2	0.56	0.88	0.49
0.3	0.3	0.59	0.81	0.53
0.4	0.4	0.62	0.78	0.63
0.5	0.5	0.74	0.73	0.74
0.6	0.6	0.81	0.64	0.79
0.7	0.7	0.78	0.42	0.83
0.8	0.8	0.63	0.21	0.92
0.9	0.9	0.56	0.03	0.98

Based on the table above, what will the approximate value of the optimal cut-off be?

- 0.5

∴ Feedback -

Correct! The optimal cut-off point exists where the values of accuracy, sensitivity, and specificity are fairly decent & almost equal. At the cut-off of 0.5, the metric value are 0.74, 0.73 & 0.74 respectively. This is the optimal value of threshold that you can have.

Precision & Recall –

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

(Recall = Sensitivity = TPR)

① Calculating Precision –

Calculate the precision value for the following model:

Actual / Predicted	Not Churn		Churn	
Not Churn	400	(TN)	100	(FP)
Churn	50	(FN)	150	(TP)

– 60% ∵ Feedback : Correct!

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{150}{100+150} = \frac{150}{250} = \frac{3}{5}$$

$$= 0.6 \approx 60\%$$

② F1 – Score –

There is a measure known as F1-score which essentially combines both precision & Recall. It is the basically the harmonic mean of precision & Recall & this formula is given by:

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{precision} + \text{Recall}}$$

The F1-Score is useful when you want to look at the performance of precision & recall together. Calculate the F1-score for the model below:

Actual / Predicted	Not Churn	Churn
Not Churn	400	100
Churn	50	150

67 % ∵ Feedback : Correct!

$$\text{F1-Score} = 2 \times \frac{\text{precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$= 2 \times \frac{(150/100+150) \times (150/150+50)}{(150/100+150) + (150/150+50)}$$

$$= 2 \times \frac{(150/250) \times (150/200)}{(150/250) + (150/200)}$$

$$= 2 \times \frac{(3/5) \times (3/4)}{(3/5) + (3/4)}$$

$$= 2 \times \frac{(0.6 \times 0.75)}{0.6 + 0.75}$$

$$= 2 \times \frac{0.45}{1.35}$$

$$= 2 \times 0.333$$

$$= 0.6666 \approx 66.67\%$$

$$\approx 67\%$$

(5) Evaluation Metrics :

Consider the same model given in the last question.

Patient ID	Heart Disease	Pred. Prob for Heart D	Pred. Lab
1001	0	0.34	0
1002	1	0.58	1
1003	1	0.79	1
1004	0	0.68	1
1005	0	0.21	0
1006	0	0.04	0
1007	1	0.48	0
1008	1	0.64	1
1009	0	0.61	1
1010	1	0.86	1

Calculate the values of Accuracy, Sensitivity, Specificity & Precision. Which of these four metrics is the highest for the model.

- Sensitivity

∵ Feedback: Correct!

TP = 4

FN = 1

FP = 2

TN = 3

Actual → Predicted

∵ 1 ⟷ 1 ⟷ TP

∵ 1 ⟷ 0 ⟷ FN

∵ 0 ⟷ 1 ⟷ FP

∵ 0 ⟷ 0 ⟷ TN

TN 3	FP 2
FN 1	TP 4

Precision = $\frac{4}{6}^{2}$ = 0.66 ≈ 67%

Sensitivity = $\frac{4}{5}$ = 0.8 ≈ 80%

Accuracy = $\frac{7}{10}$ = 0.7 ≈ 70%

Specificity = $\frac{3}{5}$ = 0.6 ≈ 60%

Naive Bayes

Bayes' Theorem & It's Building Blocks -

Graded Questions -

Courses	DS	ML	DL	BD	AI	Total
Male	80	60	40	50	30	260
Female	70	40	50	70	10	240
Total	150	100	90	120	40	500

(1) Given this contingency table, what is the probability that a randomly selected person joined Data Science DS?

- $\frac{150}{500}$ ∵ P(Person who Joined DS) = 150/500 = .3 ≈ 30%.

(2) Given this Contingency table, what is the probability that randomly selected female joined DS? In other words, what is the probability of a person joining DS. Given that she is female

- $\frac{70}{240}$ ∵ P(DS | female) = 70/240 = .29166 ≈ 29.16%.

(3) Consider a set containing all DL students OR all male students what is the probability that a randomly selected person will belong to this set?

Hint: Use the formula : (A or B) = P(A) + P(B) - P(A & B)

- $\frac{310}{500}$

:: Feedback :

This question deals with a probability concept called 'OR'. There is a formula for OR, which is —

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$$

In this example, you're looking at two things : DL & Male
So the question asked is — P(DL OR Male) =

$$P(DL) + P(Male) - P(DL \text{ AND } Male)$$

Using table in the question description, you see that —

$P(DL) = 90/500$
$P(Male) = 260/500$
$P(DL \text{ and } Male) = 40/500$

Therefore $P(DL \text{ OR } Male) = P(DL) + P(Male) - P(DL \text{ and } Male)$

$= 90/500 + 260/500 - 40/500$
$= (90 + 260 - 40)/500$
$= 310/500$
$= .62$
$\approx 62\%$

.. Now, why do you substract the probability of (Male & DL) ? The Answer is that when you count all the males and then count all the people who joined DL, there is an overlap because some males joined DL. This means you counted them twice, and so you have to substract the extra count.

$P(A \text{ AND } B) = P(A) * P(B)$

Naive Bayes – With One Feature

$$P(C = \text{Edible} / x = \text{Convex}) = \frac{P(x = \text{Convex} / C = \text{Edible}) \times P(C = \text{Edible})}{P(x)}$$

$$P(C_i|x) = \frac{P(x|C_i)\,P(C_i)}{P(x)}$$

Probability Calculation –

The probability of a convex mushroom being edible, $P(C = \text{edible} | X = \text{CONVEX})$ is given by:

$$P(X = \text{CONVEX} | C = \text{edible}) \cdot P(C = \text{edible}) / P(X = \text{CONVEX})$$

Mushroom Data Set

Sr.No.	Type of Mushroom	Cap. Shape	Cap. Surface
1	Poisonous	Convex	Scaly
2.	Edible	Convex	Scaly
3.	Poisonous	Convex	Smooth
4.	Edible	Convex	Smooth
5.	Edible	Convex	Fibrous
6.	Poisonous	Convex	Scaly
7.	Edible	Bell	Scaly
8.	Edible	Bell	Scaly
9.	Edible	Convex	Scaly
10.	Poisonous	Convex	Scaly
11.	Edible	Flat	Scaly
12.	Edible	Bell	Smooth

① Probability Calculation –
Now let's say you picked a new mushroom whose cap-shape is CONVEX. What are the chances of this happening, i.e. What is the value of P(X = CONVEX)

– 8/12 ∵ P(X = CONVEX) = 8/12

② P(x = CONVEX | C = edible) ?

– 4/8 ∵ P(X = CONVEX | C = edible) means out of all the edible mushrooms, how many are CONVEX. Out of total of 8 edible mushrooms, 4 are CONVEX. Thus, it is 4/8

③ Probability Calculation –
In the previous questions, you have calculated that P(C = edible is 8/12, P(X = convex) is 8/12 & P(X = CONVEX | C = edible) is 4/8.
What is the probability that the convex mushroom is edible, P(C = edible | X = CONVEX)?

– 4/8 ∵ Feedback –

$$P(X = CONVEX \mid C = edible) \cdot P(C = edible) / P(X = CONVEX)$$

$$= \frac{\left(\frac{4}{8}\right) \times \left(\frac{8}{12}\right) \times}{\frac{8}{12}}$$

$$= \frac{4}{12} \times \frac{12}{6}$$

$$= 4/8$$

④ Probability Calculation –
In the previous question, you find the probability of the convex mushroom being edible. what is the probability of the CONVEX mushroom being poisonous, P(C = poisonous | X = CONVEX)?

– 4/8 ∴ Feedback : Since a mushroom can either be edible or poisonous,
P(C = poisonous | X = convex) is 1 – P(C = edible | X = CONVEX)

= 1 – 4/8 = 4/8

⑤ P(C = poisonous) ⇒ 4/12

⑥ Probability calculation –
what are the chances of a mushroom being CONVEX given it is poisonous. i.e. (P X = CONVEX | C = poisonous)?

= 1 Feedback = P(X = CONVEX | C = poisonous) means out of all the poisonous mushroom, how many are CONVEX
Out of the total 4 poisonous mushroom, all the 4 are CONVEX. Thus it is 4/4 = 1

P(C = edible | X = CONVEX) =
P(X = CONVEX | C = edible)
* P(C = edible) / P(X = CONVEX)

Refer Mushroom Dataset :—

(1) Calculating conditional probability —
Say you take a new mushroom which is (CONVEX, SMOOTH). What is the numerator of P(C = edible | X = CONVEX, SMOOTH)

→ P(edible) × P(CONVEX | edible) × P(SMOOTH | edible)

(2) what is P(CONVEX | edible) ⇒ 4/8

(3) what is P(SMOOTH | edible) ⇒ 2/8

(4) What is P(CONVEX | Poisonous) ⇒ 1 (∵ 4/4 = 1)

(5) What is P(SMOOTH | poisonous) ⇒ 1/4

(6) P(CONVEX | edible) = 4/8
P(SMOOTH | edible) = 2/8
P(CONVEX | poisonous) = 1
P(SMOOTH | poisonous) = 1/4

∴ If any mushroom above 50% probability of being edible are classified as edible, is the CONVEX, SMOOTH mushroom edible.

⇒ Cannot be decided, it is a tie

$$P(\text{Edible} \mid \text{CONVEX}, \text{SMOOTH}) = \frac{P(\text{edible}) \cdot P(\text{CONVEX} \mid \text{edible}) \cdot P(\text{SMOOTH} \mid \text{edible})}{\text{denominator}}$$

$$= (8/12)(4/8)(2/8)/d = \underline{1/12d}$$

$$P(\text{poisonous} \mid \text{CONVEX}, \text{SMOOTH}) = \frac{P(\text{poisonous}) \cdot P(\text{CONVEX} \mid \text{poisonous}) \cdot P(\text{SMOOTH} \mid \text{poisonous})}{\text{denominator}}$$

$= (1/12)(1)(1/4)/d = \underline{1/12d}$

Since both numerators are equal to $1/12d$, this mushroom cannot be classified with a 50% threshold. Although if you would take a higher threshold, like 60%. (which is reasonable since you don't want to take responsibility of people eating poisonous mushrooms), then it will be classified as poisonous. why? Because, when you set the threshold as 60%, you want the probability of edible | CONVEX, SMOOTH to atleast 60%

Prior, Posterior and Likelihood —

- Prior Probability — P(class = edible) or P(class = poisonous)
- Likelihood — P(X | class)
- P(class = edible | X) — Posterior Probability

Refer Mushroom Dataset —

② Likelihood caculation —

Say you consider a (CONVEX, SCALY) mushroom. The likelihood is higher for it being —

— Poisonous ∵ Feedback —

likelihood ∝ P(X = (CONVEX, SCALY) | class) . class = edible

∴ P(CONVEX | Edible) . P(SCALY | Edible) = $(4/8)(5/8) = \frac{20}{64} = \underline{31.25\%}$

class = poisonous : P(CONVEX | poisonous) . P(SCALY | poisonous)

$= (4/4)(3/4) = \underline{75\%}$

③ Likelihood Calculation –

The value of $P(X|Class) \cdot P(Class)$ where $X = (CONVEX, SCALY)$ for both classes (edible and poisonous) are respectively:

– Edible = 20.8% ; Poisonous = 25.0%

∵ Feedback – $P(CONVEX|Edible) \cdot P(SCALY|EDIBLE) \cdot P(Edible)$
$= (4/8)(5/8)(8/12) = 20.8\%$;

Poisonous: $P(CONVEX|poisonous) \cdot P(SCALY|poisonous) \cdot P(poisonous)$
$= (4/4)(8/4)(4/12) = 25\%$

④ For the (CONVEX, SCALY) mushroom:

– The prior is in favour of edible, posterior in favor of poisonous.

∵ prior = 8/12 & 4/12 for edible & poisonous respectively;
posterior is 20.8% & 25%.

① Prior Probability – (Repeat Dataset on next page)
What is the prior probability of a mail being Spam, P(class = spam)

– 7/15 – ∵ Feedback: There are 7 spam mails in the data set.

② Naive Bayes Assumption –
What does Naive Bayes assume while classifying Spam or Ham mails?

— That frequency of keywords like lucky, free, offer etc are conditionally independent of each other

③ Likelihood Calculation —
Consider an email with the vector of feature.
X = (free, data, weekend, click). What is the likelihood, P(X | spam) ?

— $\frac{4}{2401}$ ∵ Feedback —

$P(X|spam) = P(free|spam) \cdot P(data|spam) \cdot P(weekend|spam) \cdot P(click|spam)$

$= (2/7)(1/7)(1/7)(2/7)$

$= \frac{4}{2401}$

④ Likelihood Calculation —
features X = (free, data, weekend, click). What is the likelihood, P(X | ham) ?

— $\frac{2}{4096}$ ∵ Feedback

$P(X|ham) = P(free|ham) \cdot P(data|ham) \cdot P(weekend|ham) \cdot P(click|ham)$

$= (1/8)(2/8)(1/8)(1/8)$

$= \frac{2}{4096}$

Dataset -

Sr.No.	Class	Freq1	Freq2	Freq3	Freq4
1	Spam	free	buy	limited	hurry
2	Ham	reply	data	report	presentation
3	Ham	report	present	file	end of day
4	Spam	limited	file	buy	click
5	Ham	meeting	timelines	limited	documents
6	Spam	hurry	data	buy	Stock
7	Spam	limited	sex	click	viagra
8	Ham	present	end of day	data	report
9	Ham	reply	data	present	click
10	Spam	free	reply	weekend	click
11	Spam	limited	click	free	hurry
12	Ham	meeting	end of day	weekend	data
13	Spam	hurry	weekend	Stock	offer
14	Ham	report	present	file	end of day
15	Ham	free	timelines	reply	offer

(5) Calculate Conditional probability
The value of $P(X|class) \cdot P(class)$ for class = spam for
X = (free, data, weekend, click)?

$= (4/2401)(7/15)$

∵ feedback - $P(class = spam|X) = P(class = spam) \cdot P(X|class = spam) = (7/15)(4/2401)$

(6) Posterior Probability -
What is the posterior for class = ham (i.e. without division by denominator) for the feature vector X = (free, data, weekend, click)?
-

$= (2/4096)(8/15)$ ∴ feedback:

$P(\text{class} = \text{ham} | x) = P(\text{class} = \text{ham}) \cdot P(x | \text{class} = \text{ham})$

$= (8/15)(2/4096)$

Learning from Model Selection:

* Central issue in Machine learning can be said to be the study of — How to extrapolate learnings from a finite amount of data to explain or predict all possible inputs of the same type

* Occam's Razor — A model should be simple as possible but robust

* Regression — $Y = X + 3W$.

* Model & Learning Algorithm — The learning algorithm is instructed what needs to be done; it figures out how it needs to be done & returns a model.

Finally, you learnt about the following four unique points regarding the usage of a simpler model wherever possible:-

① A simple model is usually more generic than a complex model. This becomes important because generic models are bound to perform better on unseen data sets.

② A simpler model requires fewer training data points. This becomes extremely important because in many cases, one has to work with limited data points.

③ A simple model is more robust and does not change significantly if the training data points undergo small changes.

(40) A simple model makes more errors on the training phase but it outer may perform complex models when it views new data. This happens because of overfitting.

* Simplicity and Complexity —
Why are simpler models considered to be better than complex models? (Note: More than one option may be correct.)

⇒ Simpler models are generic. i.e., they apply to a wider range of data
⇒ Complex models makes assumptions about the data, which are likely to be wrong
⇒ Simpler models requires less training data compared with Complex
⇒ Simpler models are more robust

* Overfitting — Is an Extreme case of overfitting?
⇒ The first person has mugged up all the possible questions from numerous textbooks & preparation material.

* Disadvantage of Complexity —
⇒ They will need more training data to 'learn'
⇒ Despite the training, they may not learn and perform poorly in the real world

* Overfitting — Possibility of overfitting exist primarily?
⇒ Models are trained on a set of training data, but their efficacy is determined by their ability to perform well on unseen (test) data.

* Overfitting in Linear Regression — Clear sign of overfitting in Linear Regression
The R-squared value is 0.90 and 0.30 on train and test data, respectively.

Bias & Variance

High Variance — Unstable & sensitive to changes in the training data

High Bias — Quantifies how accurate the model is likely to be on future (test) data. Extremely simple models are likely to fail in predicting complex real-world phenomena. Simplicity has its own disadvantage

* In practise, however, we often cannot have a model with a low bias and a low variance. As the model complexity increases, the bias reduces, whereas the variance increases & hence, the trade-off.

① Bais & variance

⇒ Student 1 — high variance, Student 2 high bias
⇒ student 1 — low bias, Student 2 — low variance

② Model variance — Regression
⇒ Model2 — ∵ since it can change it's 11 coefficients (& the constant term) to fit the new training data

③ ⇒ Straight line > Degree-15 > Polynomial
⇒ ∵ feedback — The bias is high when the model is highly simple.

④ Variance — Why is the variance in the higher degree polynomial said to be higher than the other two models.?

⇒ The model will change drastically from its current state if the current training data is altered.

Feedback : variance refers to changes in the model as a whole when trained on a different data set. Since the polynomial is trying to overfit the data, it will change drastically with respect to it.

⑤ Regularization — When is Regularization typically performed?

⇒ While the learning algorithm uses the training data to produce a model.

⑥ Cross Validation — Why is it often not possible to use the validation set approach?

⇒ the data available for training & testing is limited.

* In K-Fold Cross Validation, you divide the training data into K-Groups of samples. If K=4 (say), you use K-1 folds to build the model and test the model on the Kth fold.

* Hyperparameters are used to 'fine-tune' or regularize the model to keep it optimally complex.

* the learning algorithm is given the hyperparameters as the input, & it returns the model parameters as the output.

* Hyperparameters are not part of the final model output.

(1) Low Variance —
— Weak learner ∵ weak learner creates simpler models that have a lower variance. They are not able to model complex relationships and, hence, create a more generic model.

(2) Bias - Variance —
— Linear regression will have a high bias and a low variance.
— A polynomial equation of degree 4 will have a low bias & a low variance

(1) Model variance — Measure variance of a model.
— By measuring how much does the estimates of the model change on the test data on changing the training data

(2) Regularization —
— It is a technique that is used to strike a balance between model complexity and model accuracy on training data
∵ Reg^n doesn't improve accuracy; it improves the balance betn accuracy and complexity.

(3) Simple Models —
— Simpler Models will always have fewer test errors than a complex model

(4) simplicity of a Models —
— By the no. of features used in the model
— By the no. of nodes & depth of trees in case of a tree model.

(5) K-FOID Cross Validation -

- As K increases, the training time for k-fold cross validation increases
- With higher number of folds, the estimated error, on an average, is usually lower.

- You repeat the cross validation process 'K' times
- Each 'Kth' fold is used as the validation data once.
- A model trained with k-fold cross validation will overfit.