4. What is supervised learning?

Answer:

Supervised learning is when a model is trained on labeled data (i.e., input + known output). In our project, we trained the model using tumor measurements and their known diagnoses (Benign/Malignant).

5. Which algorithm did you use in your project and why?

Answer:

I used the Support Vector Machine (SVM) with an RBF kernel. It's good for classification problems and works well with smaller datasets like the breast cancer dataset.

6. What is the purpose of the training and testing split?

Answer:

Training is used to teach the model, and testing is used to check how well the model performs on new, unseen data. This helps avoid overfitting.

7. What is a confusion matrix?

Answer:

It’s a table that shows how many correct and incorrect predictions a model made:

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

8. What is ROC curve and AUC?

Answer:

ROC curve shows the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate).

AUC (Area Under Curve) tells how well the model can distinguish between classes. Higher AUC means better performance.

☑ C. Explainability (XAI)

9. What is Explainable AI (XAI)?

Answer:

Explainable AI refers to techniques that help humans understand how an AI model makes decisions.

10. What are SHAP and LIME?

Answer:

SHAP: Shows how much each feature contributed to a specific prediction.

LIME: Explains individual predictions locally by approximating the model near that prediction.

☑ D. Use Case Awareness

11. Why is your project important in real life?

Answer:

It helps in early detection of breast cancer, which can save lives. It also provides doctors with quick, explainable predictions to support clinical decisions.

12. Can this model be deployed in real hospitals?

Answer:

Yes, especially in screening or decision-support systems. However, it would need to be integrated into a certified medical platform and validated clinically.

☑ Bonus: Personal Questions

13. What challenges did you face in this project?

Sample Answer:

Understanding and tuning the SVM model was tricky. Also, integrating ML with Shiny UI and interpreting the results with SHAP/LIME required extra effort and learning.

14. What did you learn from this capstone?

Sample Answer:

I learned how to apply machine learning to real-world healthcare data, build an interactive UI using Shiny, and make my model interpretable with Explainable AI techniques.

Why SVM with RBF Kernel Was Used

🔍 Problem:

You’re trying to classify breast tumors into Benign and Malignant.

The data is not perfectly linearly separable — the two classes overlap based on complex patterns of features (like radius, concavity, texture, etc.).

☑ Why SVM?

SVM (Support Vector Machine) is a powerful supervised learning algorithm used for classification tasks.

It works by finding the optimal boundary (hyperplane) that best separates the classes.

It’s especially good when you have a smaller but clean dataset (like the WBCD dataset).

☑ Why Non-linear RBF Kernel?

RBF stands for Radial Basis Function. Here's why you used it:

◇ Reason 🔍 Explanation

1. Non-linearity Real-world medical data is rarely perfectly linear. RBF kernel allows the model to map input features into a higher-dimensional space where it can separate the two classes better.

2. Captures Complex Patterns Tumor characteristics like texture, radius, and concavity are not linearly related to diagnosis. The RBF kernel handles such complexity effectively.

3. High Accuracy In most medical datasets (like WBCD), RBF kernel SVM has shown higher classification accuracy than linear models.

4. Fewer Parameters Compared to neural networks or decision trees, SVM with RBF needs only a few hyperparameters (like C and gamma), making it easier to optimize.

5. Works well on small to medium datasets The Wisconsin dataset has fewer than 600 records — perfect for SVM to shine.

🧠 Viva-Ready Answer (Say this):

"I used SVM with an RBF kernel because the relationship between tumor features and diagnosis is non-linear. The RBF kernel helps project the data into a higher-dimensional space where the classes become more separable. It works well with small datasets like ours and gives high accuracy in classification. Also, SVM is less prone to overfitting and provides a solid decision boundary between benign and malignant tumors."

What is Overfitting?

🧠 Definition:

Overfitting happens when a machine learning model memorizes the training data too well, including the noise and outliers, instead of learning the general patterns.

This makes the model perform very well on training data, but poorly on new, unseen data.

📊 Example (Viva Style):

“Imagine a student who memorizes all the questions from last year’s paper. In the test, if the same questions appear, they score 100%. But if a new question is asked, they fail to answer — that’s like overfitting in ML.”

✅ How to Detect Overfitting?

High training accuracy, but low test accuracy.

Large gap between training loss and validation loss.

✅ How to Prevent Overfitting?

Use cross-validation.

Use simpler models.

Apply regularization (e.g., L1, L2).

Reduce number of features (feature selection).

Use more data (if available).

✅ Commonly Used Terms in Machine Learning (with simple meanings):

◇ Term	🧠 Meaning
Feature	A variable or column used to make predictions (e.g., radius_mean)
Label / Target	The output you're trying to predict (e.g., diagnosis)
Training Set	Data used to train the model
Testing Set	Data used to evaluate the model
Accuracy	Percentage of correct predictions made by the model
Precision	How many predicted positives are actually positive
Recall	How many actual positives are correctly predicted
F1 Score	A balance between precision and recall
ROC Curve	Graph showing model performance at all classification thresholds
AUC (Area Under Curve)	Measures how well the model separates classes (1.0 = best)
Confusion Matrix	Table showing TP, TN, FP, FN to evaluate classification performance
Bias	Error due to overly simple assumptions (underfitting)
Variance	Error due to model sensitivity to training data (overfitting)

Hyperparameters	Settings like C, gamma, learning rate, chosen before training
Cross-Validation	Splitting data into folds to ensure model generalizes well
Model Evaluation	Checking how good a model is using metrics like accuracy, F1, etc.

🗣 Viva Tip:

When asked a term like "Recall", say:

"Recall is the model’s ability to find all actual positive cases. In our breast cancer project, it means catching as many malignant tumors as possible."