Assignment No 4

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

df = pd.read_csv("breast_cancer.csv")
df = df.drop(columns=[col for col in df.columns if 'Unnamed' in col], errors='ignore')
df['Status'] = df['Status'].map({'Alive': 1, 'Dead': 0})
df = df.dropna()

X = df.drop(columns=['Status'])
y = df['Status']
X = pd.get_dummies(X, drop_first=True)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy is {accuracy:.2f}")

def recommend(features):
    prediction = model.predict([features])
    return "Alive" if prediction == 1 else "Dead"

test_sample = X_test.iloc[0].values
ans = recommend(test_sample)
print("Prediction for sample:", ans)
```

Viva-Ready Questions & Answers

Q: What is the main goal of this system?
A: To predict whether a breast cancer patient is likely to survive based on their medical data.

Q: Why is Random Forest used?
A: It's robust, handles missing data, works with mixed features, and reduces overfitting by combining multiple decision trees.

Q: What type of problem is this — classification or regression?
A: Classification problem (binary classification — Alive or Dead).

Q: What is the target variable?
A: `Status` — encoded as 1 for Alive, 0 for Dead.

Q: What is One-Hot Encoding used for?
A: To convert categorical text variables into numeric binary columns suitable for machine learning models.

Q: What is accuracy and how is it calculated?
A: Accuracy = (Number of Correct Predictions / Total Predictions). It measures how well the model predicts the correct outcome.

Q: How is data split and why?
A: 80% training and 20% testing to evaluate the model on unseen data and avoid overfitting.

Q: What is the advantage of Random Forest over Decision Tree?
A: Random Forest averages results from many trees, improving accuracy and reducing overfitting.

Q: What is the significance of using `random_state=42`?
A: Ensures reproducibility — same random data split and model behavior every time.

Q: What does the `recommend()` function do?
A: Predicts the survival status ("Alive" or "Dead") for a new patient's features.

Q: What are possible improvements to this model?
A:

- **Perform hyperparameter tuning (n_estimators, max_depth).**
- **Use scaling or feature selection.**
- **Try other models (Logistic Regression, SVM, XGBoost).**
- **Add clinical data like stage, hormone receptor status, etc.**

Q: What kind of ML task does this project represent in healthcare?
A: Predictive analytics — helps forecast patient outcomes and assist doctors in making treatment decisions.

Q: What are the limitations of this system?
A:

- **Accuracy depends on dataset quality.**
- **Limited interpretability for medical explanations.**
- **May need retraining for new hospital data.**