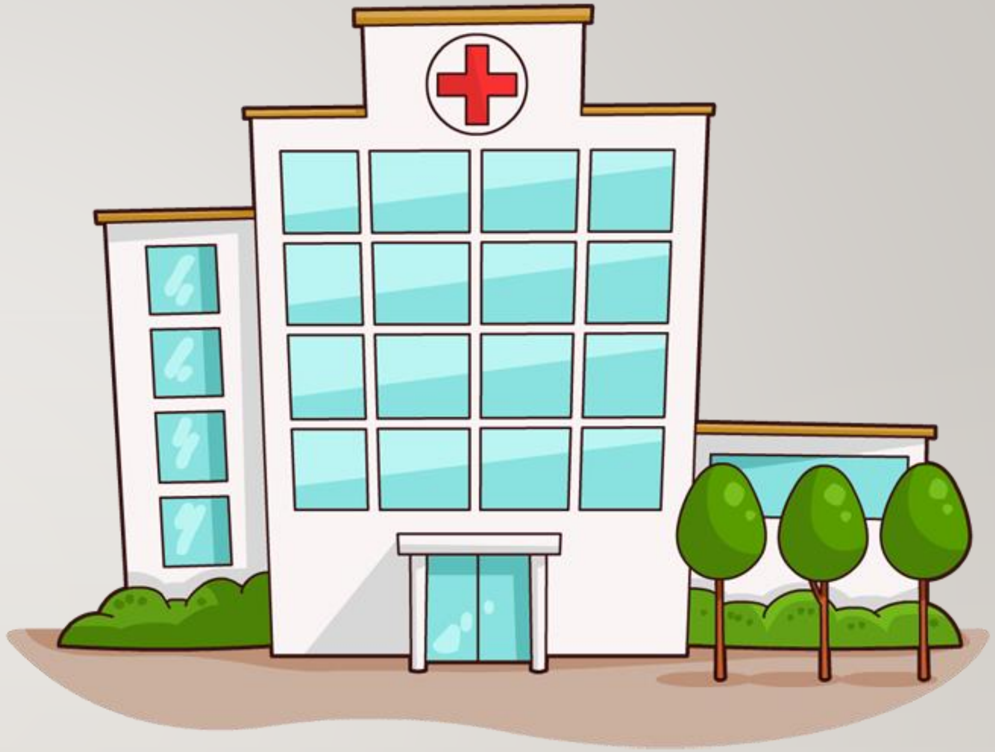
OVARIAN CANCER DETECTION USING MACHINE LEARNING

EARLY DETECTION FOR BETTER PATIENT OUTCOMES



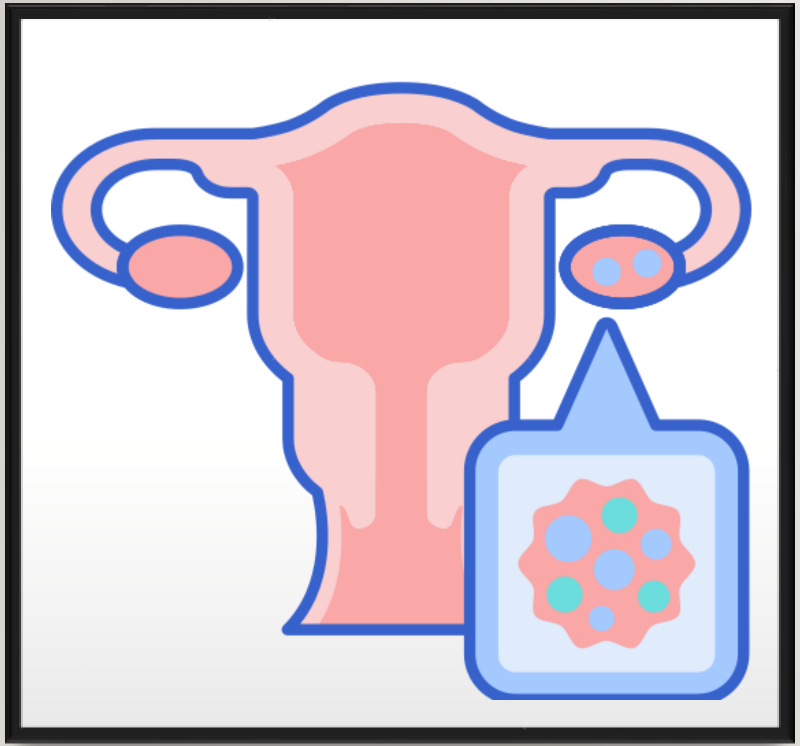
Presented by:

Mayank Maletha

Deepanshu

Nityam Tripathi

INTRODUCTION



- **What is Ovarian Cancer?**
  Ovarian cancer is one of the most challenging cancers to detect early due to subtle symptoms and lack of effective early diagnostic tools.
- **Importance of Early Detection:**
  Early detection significantly improves treatment success and survival rates. Machine learning offers a novel approach to detecting ovarian cancer at an early stage.
- **Objective:**
  Develop an effective, accessible machine learning system for predicting ovarian cancer using clinical and biochemical data.

DATASET OVERVIEW

Dataset Details:

- Total Records: 349 records with 51 attributes.
- Data Types: Clinical markers (e.g., CA125, HE4), biochemical markers (e.g., ALB, ALT), and demographics (e.g., age, menopause).
- Target Variable: Binary classification (0 = No cancer, 1 = Cancer).

Challenges:

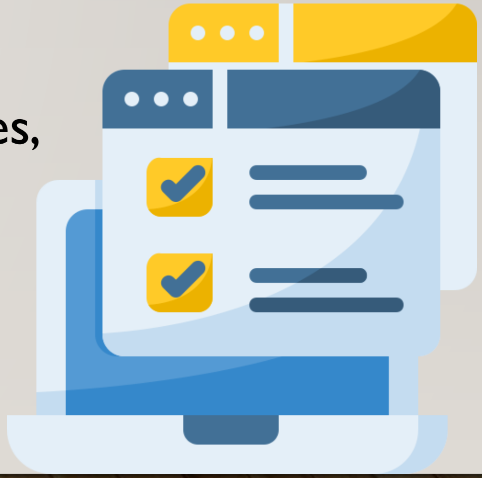
- Missing values in key features.
- Handling imbalanced class distribution.
- Selecting relevant features to improve model accuracy.

DATA PREPROCESSING

- Removed whitespace and non-numeric entries.
- Missing values imputed using column means.
- Used SelectKBest with ANOVA F-test to select top 10 features: **Age, ALB, CA125, HE4, LYM#, LYM%, Menopause, NEU, PCT, PLT**.
- Split dataset into 80% training and 20% testing data for robust evaluation.

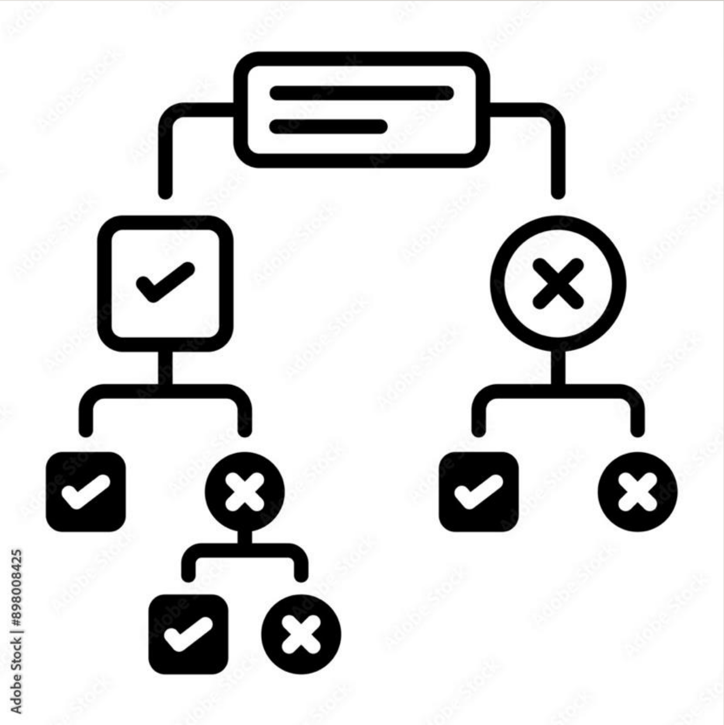
FEATURE SELECTION

- **Feature Selection Method:**
- Applied ANOVA F-test using SelectKBest for statistical relevance.
- Top features like **CA125** and **HE4** were identified as having the highest predictive power.
- **Significance:**
  These features are critical for distinguishing between positive and negative cases, reflecting real-world clinical relevance.



ALGORITHMS USED

- **Logistic Regression:** Baseline model for binary classification.
- **Random Forest:** Most accurate (89%) with robust feature handling.
- **Decision Tree:** Provides interpretable rules for predictions.
- **SVM:** Finds optimal boundaries between classes using kernel functions.
- **KNN:** Classifies based on proximity to similar cases.



MODEL EVALUATION

- **Metrics Used:** Accuracy, precision, recall, F1-score, and confusion matrix.
- **Best Model:** Random Forest achieved 89% accuracy.
- Key features like **CA125** and **HE4** had the most influence.

RESULTS AND ACCURACY

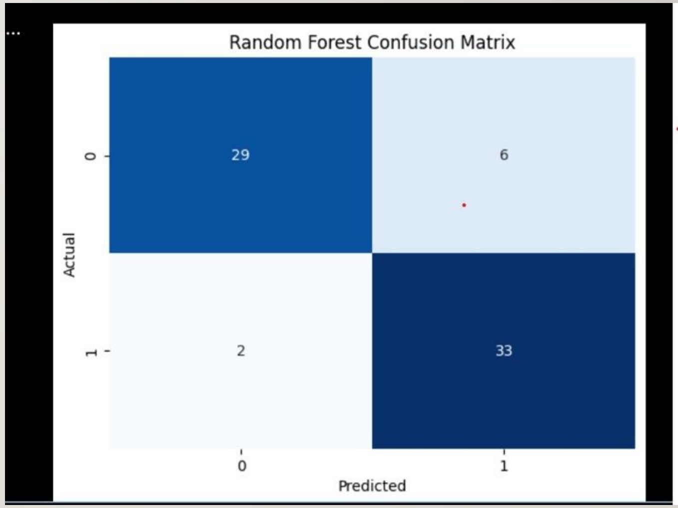
- **Model Comparison:**
  - Logistic Regression: 87.14
  - Random Forest: 89% (Best performer)
  - Decision Tree: 85.71%
  - SVM: 77.14%
  - KNN: 85.71%
- **Key Observations:**
  - Features **CA125** and **HE4** consistently contributed to accurate predictions.
  - Random Forest balanced interpretability with performance.

RANDOM FOREST

```
Optimized Model Accuracy: 88.57142857142857
Classification Report:
              precision    recall  f1-score   support

         0.0       0.94      0.83      0.88        35
         1.0       0.85      0.94      0.89        35

    accuracy                           0.89        70
   macro avg       0.89      0.89      0.89        70
weighted avg       0.89      0.89      0.89        70
```



PREDICTION PIPELINE

Workflow:

1. Input new patient data via web interface.
2. Preprocess data to handle missing values and scale features.
3. Apply feature selection to retain top predictors.
4. Use the trained Random Forest model for prediction.
5. Display results and confidence scores in real-time.

DEPLOYMENT

Technologies Used:

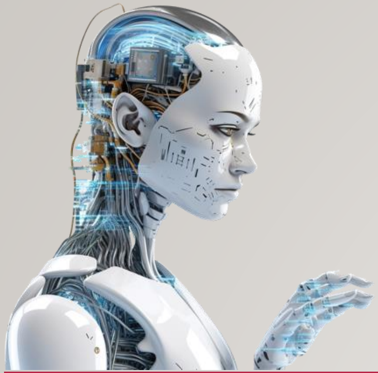
- Flask framework for web app development.
- Model serialized using Pickle for deployment.

Web Application Features:

- User-friendly interface for entering patient data.
- Real-time predictions displayed with confidence scores.

Example Workflow:

Input values for CA125, HE4, and demographics → Receive prediction: "High possibility of cancer" or "Safe".



CHALLENGES

Challenges Faced:

- Handling missing data entries without introducing bias.
- Ensuring the model is interpretable for non-technical medical professionals.
- Addressing imbalanced datasets for reliable predictions.

Solutions:

- Used imputation techniques for missing data.
- Prioritized Random Forest for its balance of accuracy and interpretability.

FUTURE SCOPE

- Expand the dataset to include more samples for better generalization.
- Explore advanced algorithms like XGBoost and deep learning to improve predictions further.
- Add visualizations to explain model decisions to healthcare practitioners.
- Integrate real-time data from wearable or hospital monitoring systems.

CONCLUSION

- The project demonstrates the potential of machine learning in early ovarian cancer detection.
- By combining high-performing models with a user-friendly web interface, this system can assist healthcare professionals in improving patient outcomes.
- Future work will focus on scaling the model, improving interpretability, and real-time integration.