Detection and Analysis of

Polycystic Ovary Syndrome

Data Science Boot Camp
Final Presentation

Ethan Bootehsaz, Kaiyuan Ma, Joy Wang

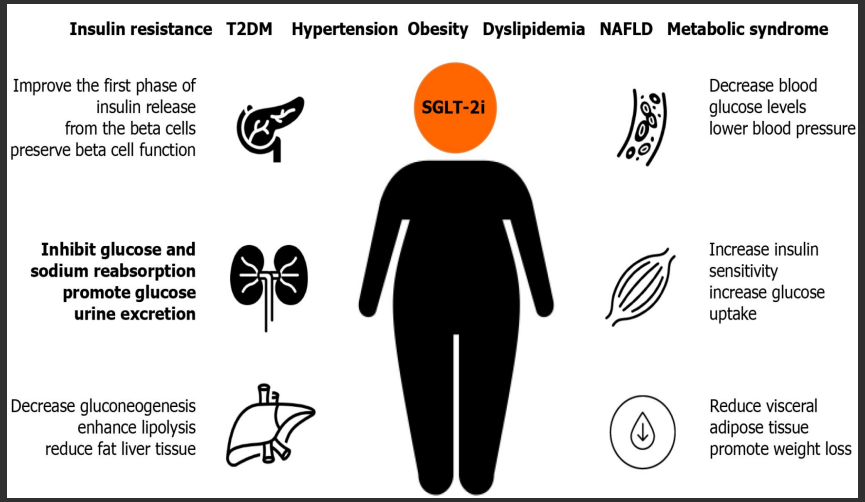
04/30/2024

Agenda

- Goal
- Data Preprocessing Steps
- Exploration of Data
  - More Dataset Features
- Initial Hypothesis
- Model Selection
  - Logistic Regression
  - Deep Learning
  - Deep learning with Cross Validation
- Results & Discussion
  - Model Performance
  - Future Improvements

What is PCOS?

PCOS affects 7–10% of women of and is the most common cause of infertility. In the United States, an estimated 5 to 6 million women have PCOS.



- Condition where ovaries produce an excessive amount of male hormones (androgens)
- May or may not develop cysts due to problems in menstrual cycle (anovulation)

Goal

Our goal was to develop a predictive model for the detection of Polycystic ovary syndrome (PCOS) by utilizing relevant physical and clinical parameters.

- We analyzed the dataset to identify patterns and trends.
- We developed a model that can predict the likelihood of PCOS.
- We evaluated the model's performance using various metrics.
- We created visualizations to support our findings.

Data Preprocessing Steps

- **Initial Data Cleaning**
  - Removed irrelevant columns: `Unnamed: 44`, `Sl. No`, and `Patient File No.` to focus on analytically useful data.
  - Renamed mislabeled columns to ensure consistency: Changed `Marraige Status (Yrs)` to `Marriage Status (Yrs).`

- **Data Formatting**
  - Standardized column names by stripping leading/trailing whitespaces.
  - Cleaned data entries to remove trailing commas and periods for accuracy in analysis.

- **Handling Data Quality Issues**
  - Identified and removed rows where data had non-numeric characters in numeric fields to maintain data integrity.

- **Dataset Post-Processing**
  - Final dataset comprised of 539 rows ready for further analysis and modeling.

```
# Check for missing values in the dataset
missing_data = pcos_data.isnull().sum()
missing_data[missing_data > 0]
```

```
Marraige Status (Yrs)      1
Fast food (Y/N)            1
Unnamed: 44              539
dtype: int64
```

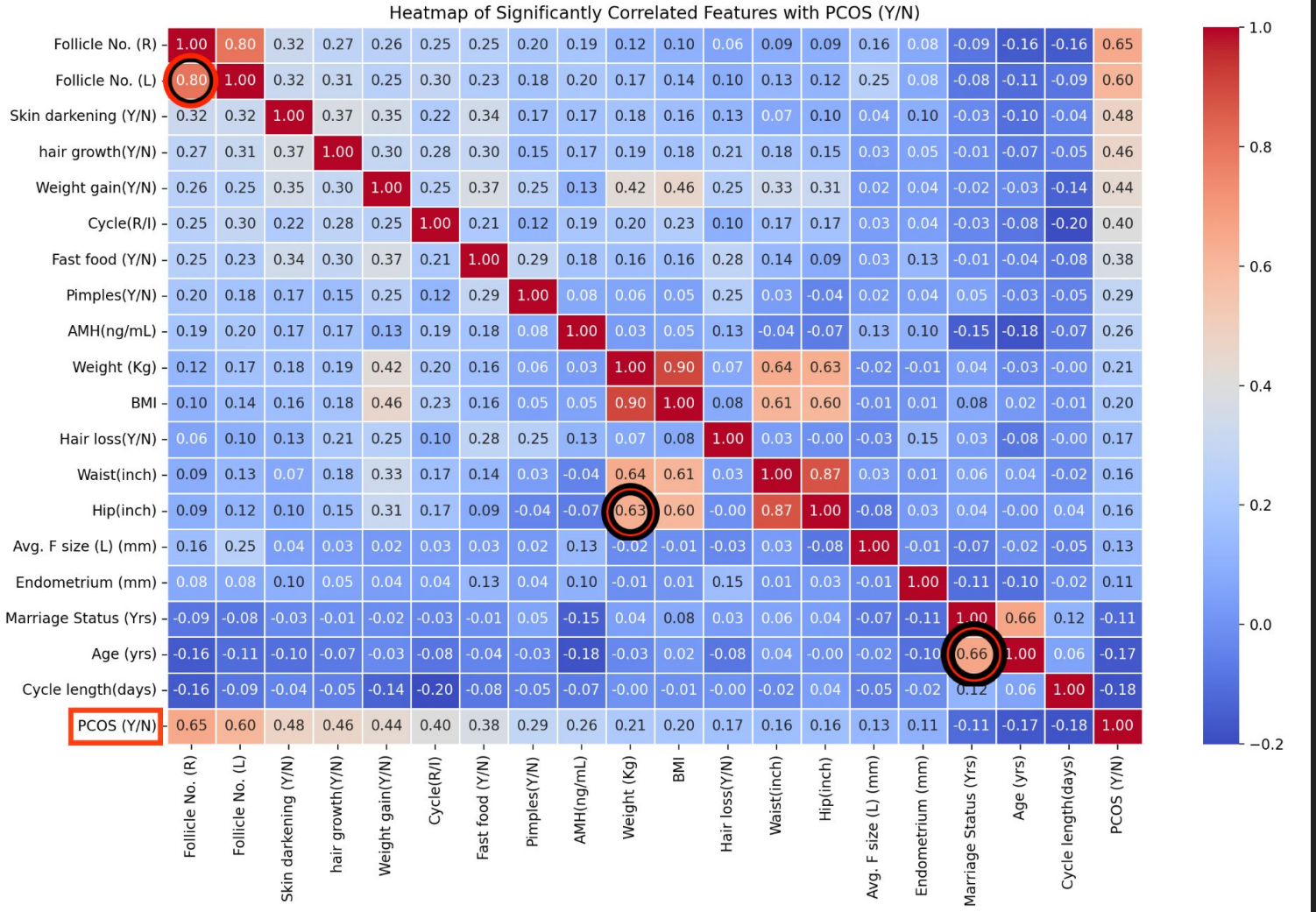
```
Missing values in ../data/PCOS_processed.csv:
PCOS (Y/N)                 0
Age (yrs)                  0
Weight (Kg)                0
Height(Cm)                 0
BMI                        0
Blood Group                0
Pulse rate(bpm)            0
RR (breaths/min)           0
Hb(g/dl)                   0
Cycle(R/I)                 0
Cycle length(days)         0
Marriage Status (Yrs)      0
Pregnant(Y/N)              0
No. of aborptions          0
I beta-HCG(mIU/mL)         0
II beta-HCG(mIU/mL)        0
FSH(mIU/mL)                0
LH(mIU/mL)                 0
FSH/LH                     0
Hip(inch)                  0
Waist(inch)                0
Waist:Hip Ratio            0
TSH (mIU/L)                0
AMH(ng/mL)                 0
...
```

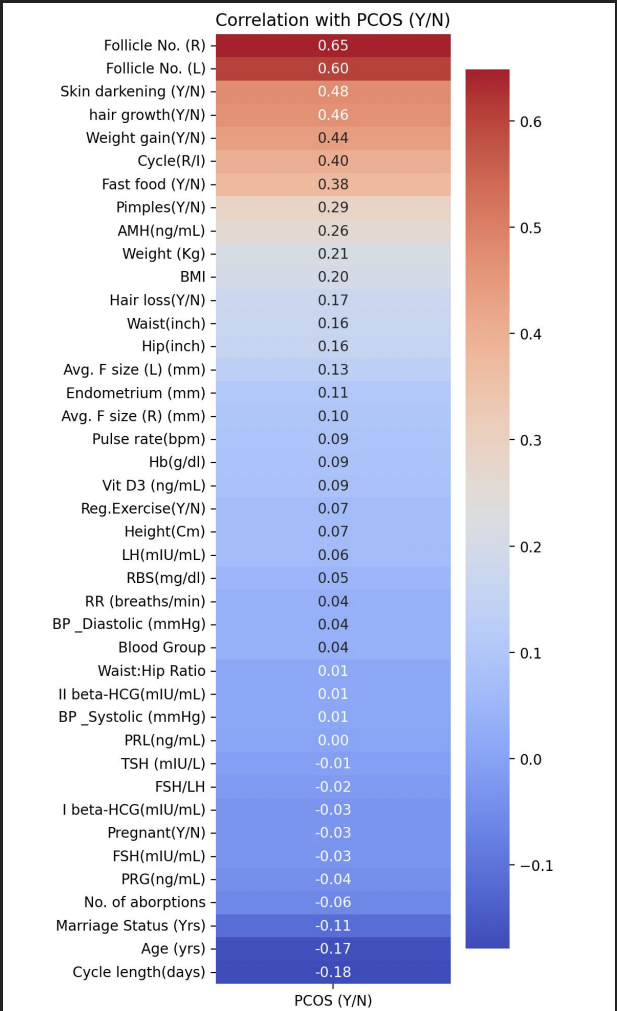
```
#Dropping repeated/unnecessary columns
df = df.drop(['Unnamed: 44', 'Sl. No', 'Patient File No.'], axis=1, errors='ignore')

#Renaming column due to misspelling in original df
df.rename(columns={'Marraige Status (Yrs)': 'Marriage Status (Yrs)'}, inplace=True, errors='ignore')

# Fix column names - optional
df.columns = df.columns.str.strip() # .str.replace(' ', '_').str.lower()
df.columns = [re.sub(r'\s+', ' ', col).strip() for col in df.columns]

# Remove leading/trailing whitespaces from the data, remove trailing commas, periods
df = df.applymap(lambda x: x.strip() if isinstance(x, str) else x)
df = df.applymap(lambda x: x.rstrip('.') if isinstance(x, str) else x)
```

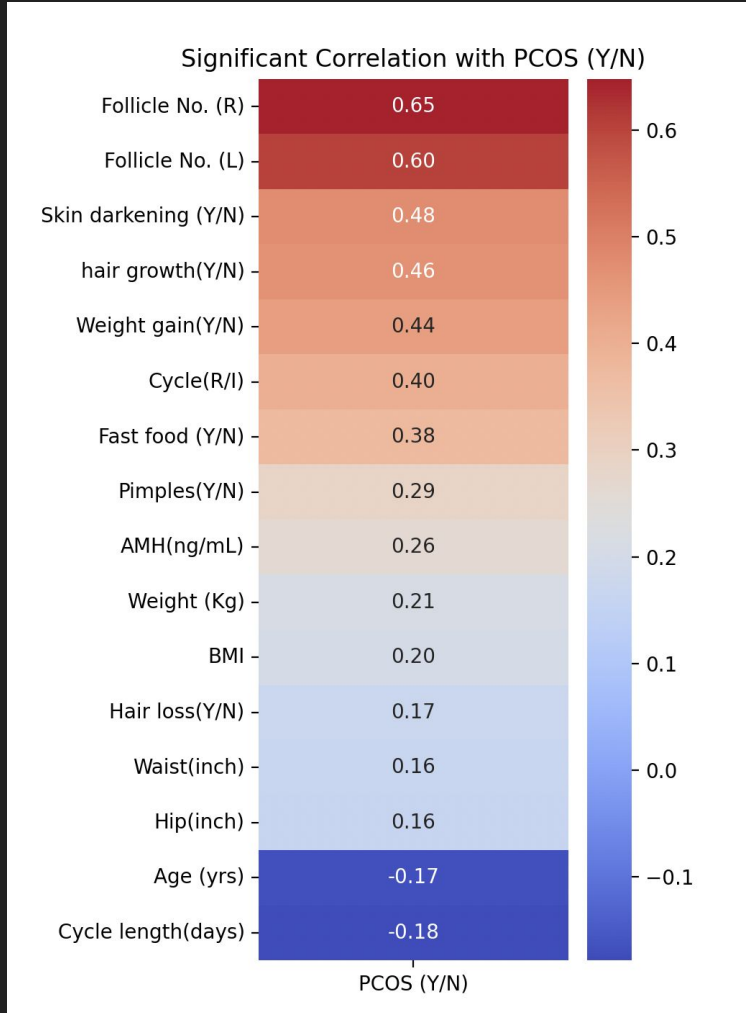
Exploration of Data

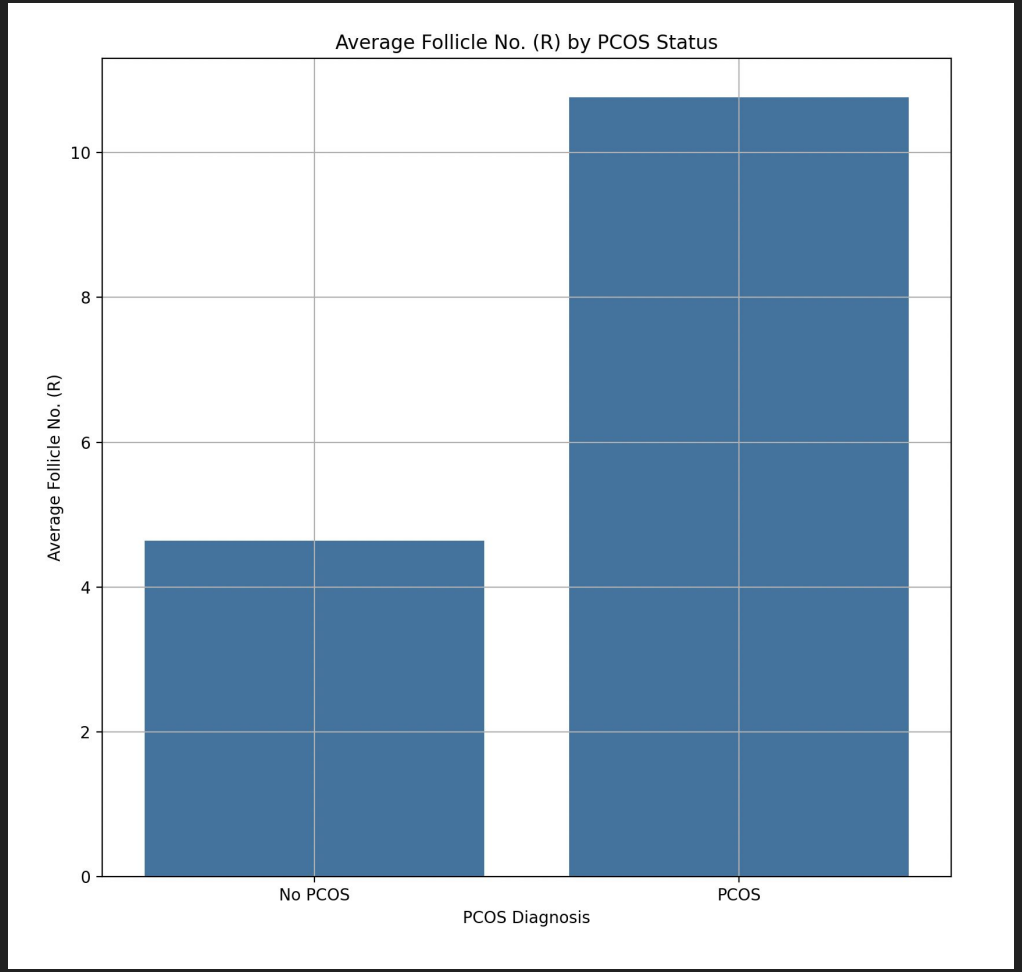
- Understanding Correlation Calculation in Data Analysis
  - df.corr()['PCOS (Y/N)']

- Correlation coefficients range from -1 to 1. A coefficient close to 1 indicates a strong positive correlation, meaning as one variable increases, so does the other. A coefficient close to -1 indicates a strong negative correlation, where one variable increases as the other decreases. A coefficient near 0 suggests no linear relationship.


Heatmap of Significantly Correlated Features with PCOS (Y/N)
Follicle No. (R) 1.00 0.80 0.32 0.27 0.26 0.25 0.25 0.20 0.19 0.12 0.10 0.06 0.09 0.09 0.16 0.08 -0.09 -0.16 -0.16 0.65
Follicle No. (L) 0.80 1.00 0.32 0.31 0.25 0.30 0.23 0.18 0.20 0.17 0.14 0.10 0.13 0.12 0.25 0.08 -0.08 -0.11 -0.09 0.60
Skin darkening (Y/N) 0.32 0.32 1.00 0.37 0.35 0.22 0.34 0.17 0.17 0.18 0.16 0.13 0.07 0.10 0.04 0.10 -0.03 -0.10 -0.04 0.48
hair growth(Y/N) 0.27 0.31 0.37 1.00 0.30 0.28 0.30 0.15 0.17 0.19 0.18 0.21 0.18 0.15 0.03 0.05 -0.01 -0.07 -0.05 0.46
Weight gain(Y/N) 0.26 0.25 0.35 0.30 1.00 0.25 0.37 0.25 0.13 0.42 0.46 0.25 0.33 0.31 0.02 0.04 -0.02 -0.03 -0.14 0.44
Cycle(R/I) 0.25 0.30 0.22 0.28 0.25 1.00 0.21 0.12 0.19 0.20 0.23 0.10 0.17 0.17 0.03 0.04 -0.03 -0.08 -0.20 0.40
Fast food (Y/N) 0.25 0.23 0.34 0.30 0.37 0.21 1.00 0.29 0.18 0.16 0.16 0.28 0.14 0.09 0.03 0.13 -0.01 -0.04 -0.08 0.38
Pimples(Y/N) 0.20 0.18 0.17 0.15 0.25 0.12 0.29 1.00 0.08 0.06 0.05 0.25 0.03 -0.04 0.02 0.04 0.05 -0.03 -0.05 0.29
AMH(ng/mL) 0.19 0.20 0.17 0.17 0.13 0.19 0.18 0.08 1.00 0.03 0.05 0.13 -0.04 -0.07 0.13 0.10 -0.15 -0.18 -0.07 0.26
Weight (Kg) 0.12 0.17 0.18 0.19 0.42 0.20 0.16 0.06 0.03 1.00 0.90 0.07 0.64 0.63 -0.02 -0.01 0.04 -0.03 -0.00 0.21
BMI 0.10 0.14 0.16 0.18 0.46 0.23 0.16 0.05 0.05 0.90 1.00 0.08 0.61 0.60 -0.01 0.01 0.08 0.02 -0.01 0.20
Hair loss(Y/N) 0.06 0.10 0.13 0.21 0.25 0.10 0.28 0.25 0.13 0.07 0.08 1.00 0.03 -0.00 -0.03 0.15 0.03 -0.08 -0.00 0.17
Waist(inch) 0.09 0.13 0.07 0.18 0.33 0.17 0.14 0.03 -0.04 0.64 0.61 0.03 1.00 0.87 0.03 0.01 0.06 0.04 -0.02 0.16
Hip(inch) 0.09 0.12 0.10 0.15 0.31 0.17 0.09 -0.04 -0.07 0.63 0.60 -0.00 0.87 1.00 -0.08 0.03 0.04 -0.00 0.04 0.16
Avg. F size (L) (mm) 0.16 0.25 0.04 0.03 0.02 0.03 0.03 0.02 0.13 -0.02 -0.01 -0.03 0.03 -0.08 1.00 -0.01 -0.07 -0.02 -0.05 0.13
Endometrium (mm) 0.08 0.08 0.10 0.05 0.04 0.04 0.13 0.04 0.10 -0.01 0.01 0.15 0.01 0.03 -0.01 1.00 -0.11 -0.10 -0.02 0.11
Marriage Status (Yrs) -0.09 -0.08 -0.03 -0.01 -0.02 -0.03 -0.01 0.05 -0.15 0.04 0.08 0.03 0.06 0.04 -0.07 -0.11 1.00 0.66 0.12 -0.11
Age (yrs) -0.16 -0.11 -0.10 -0.07 -0.03 -0.08 -0.04 -0.03 -0.18 -0.03 0.02 -0.08 0.04 -0.00 -0.02 -0.10 0.66 1.00 0.06 -0.17
Cycle length(days) -0.16 -0.09 -0.04 -0.05 -0.14 -0.20 -0.08 -0.05 -0.07 -0.00 -0.01 -0.00 -0.02 0.04 -0.05 -0.02 0.12 0.06 1.00 -0.18
PCOS (Y/N) 0.65 0.60 0.48 0.46 0.44 0.40 0.38 0.29 0.26 0.21 0.20 0.17 0.16 0.16 0.13 0.11 -0.11 -0.17 -0.18 1.00
Follicle No. (R)
Follicle No. (L)
Skin darkening (Y/N)
hair growth(Y/N)
Weight gain(Y/N)
Cycle(R/I)
Fast food (Y/N)
Pimples(Y/N)
AMH(ng/mL)
Weight (Kg)
BMI
Hair loss(Y/N)
Waist(inch)
Hip(inch)
Avg. F size (L) (mm)
Endometrium (mm)
Marriage Status (Yrs)
Age (yrs)
Cycle length(days)
PCOS (Y/N)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

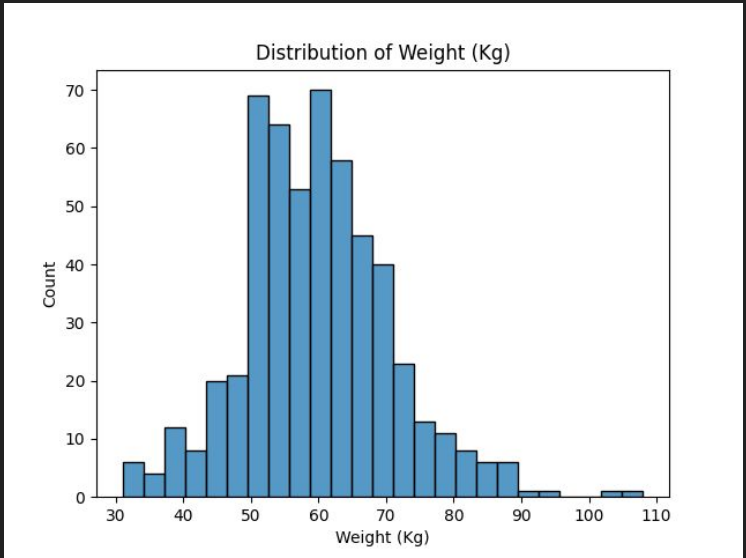

Correlation with PCOS (Y/N)
Follicle No. (R) 0.65
Follicle No. (L) 0.60
Skin darkening (Y/N) 0.48
hair growth(Y/N) 0.46
Weight gain(Y/N) 0.44
Cycle(R/I) 0.40
Fast food (Y/N) 0.38
Pimples(Y/N) 0.29
AMH(ng/mL) 0.26
Weight (Kg) 0.21
BMI 0.20
Hair loss(Y/N) 0.17
Waist(inch) 0.16
Hip(inch) 0.16
Avg. F size (L) (mm) 0.13
Endometrium (mm) 0.11
Avg. F size (R) (mm) 0.10
Pulse rate(bpm) 0.09
Hb(g/dl) 0.09
Vit D3 (ng/mL) 0.09
Reg.Exercise(Y/N) 0.07
Height(Cm) 0.07
LH(mIU/mL) 0.06
RBS(mg/dl) 0.05
RR (breaths/min) 0.04
BP _Diastolic (mmHg) 0.04
Blood Group 0.04
Waist:Hip Ratio 0.01
II beta-HCG(mIU/mL) 0.01
BP _Systolic (mmHg) 0.01
PRL(ng/mL) 0.00
TSH (mIU/L) -0.01
FSH/LH -0.02
I beta-HCG(mIU/mL) -0.03
Pregnant(Y/N) -0.03
FSH(mIU/mL) -0.03
PRG(ng/mL) -0.04
No. of aborptions -0.06
Marriage Status (Yrs) -0.11
Age (yrs) -0.17
Cycle length(days) -0.18
PCOS (Y/N)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1

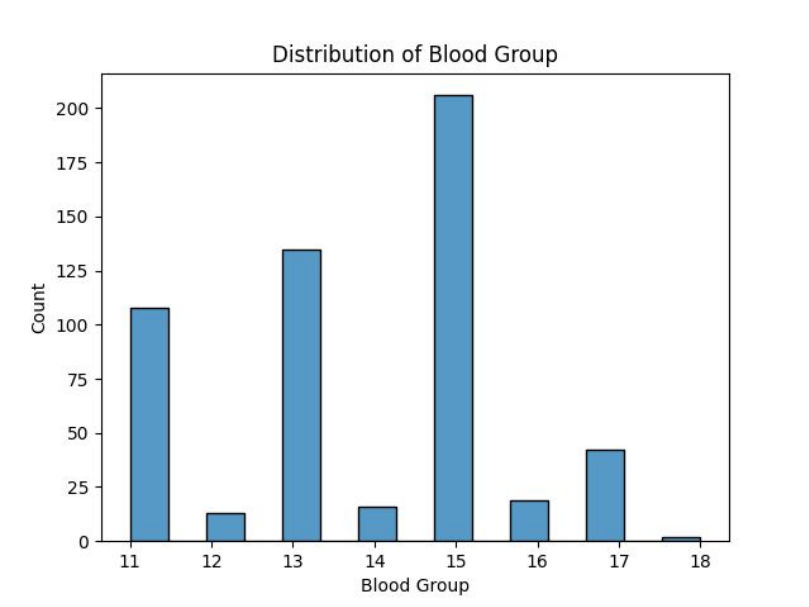

Significant Correlation with PCOS (Y/N)
Follicle No. (R) 0.65
Follicle No. (L) 0.60
Skin darkening (Y/N) 0.48
hair growth(Y/N) 0.46
Weight gain(Y/N) 0.44
Cycle(R/I) 0.40
Fast food (Y/N) 0.38
Pimples(Y/N) 0.29
AMH(ng/mL) 0.26
Weight (Kg) 0.21
BMI 0.20
Hair loss(Y/N) 0.17
Waist(inch) 0.16
Hip(inch) 0.16
Age (yrs) -0.17
Cycle length(days) -0.18
PCOS (Y/N)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1


Average Follicle No. (R) by PCOS Status
Average Follicle No. (R)
0
2
4
6
8
10
No PCOS
PCOS
PCOS Diagnosis

More Dataset Features

Distribution of numerical variables

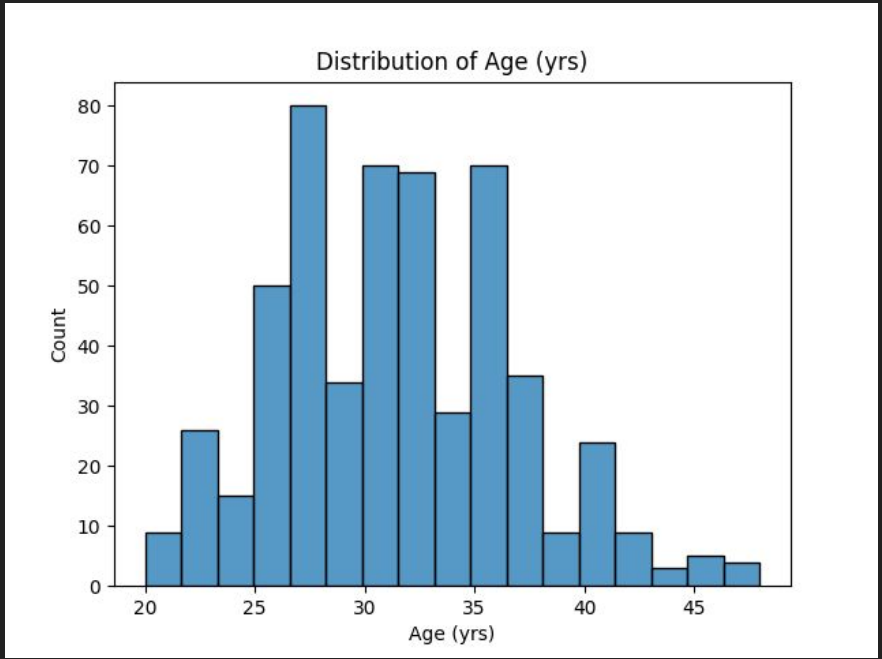


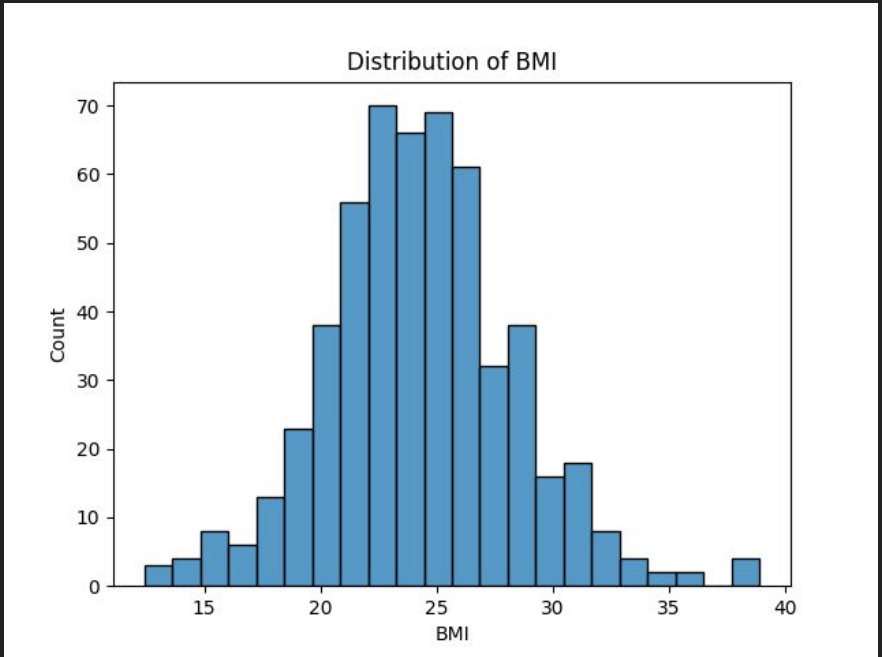


mean	std	min	max
156.484835	6.033545	137.000000	108.000000

Blood Group of the patient A+ = 11, A- = 12, B+ = 13, B- = 14, O+ =15, O- = 16, AB+ =17, AB- = 18

Distribution of numerical variables cont.





Count	mean	std	min	max
541.0	31.430684	5.411006	20.000000	48.000000

Count	mean	std	min	max
541.0	24.311285	4.056399	12.417882	38.900000

Contextualize Dataset Info

The average patient is/has:

- 156.48 lbs
- 31.43 years old
- BMI of 24.311
- Been married 5 years
- More likely to be blood type (O+)

Cycle (R/I) + Cycle length (days)

Cycle R/I

There's no documentation on what the 2, 4, 5 mean >:(((

My guess is that 2 is regular, 4 is irregular, 5 is none but this is unusable

Cycle Length

Varies, but the average is 28 days

beta-HCG II (mlU/ML)

Used in fertility / pregnancy testing. Can also determine the age of the fetus based on gestation period.

FSH → Follicle-stimulating hormone

Levels vary based on puberty, menstruation, menopause

Waist/Hip ratio

Used with BMI to determine health risks

Health risk: obesity, type II diabetes, heart disease

- Low = 0.80 or below
- Moderate = 0.81 - 0.85
- High = 0.86 or higher

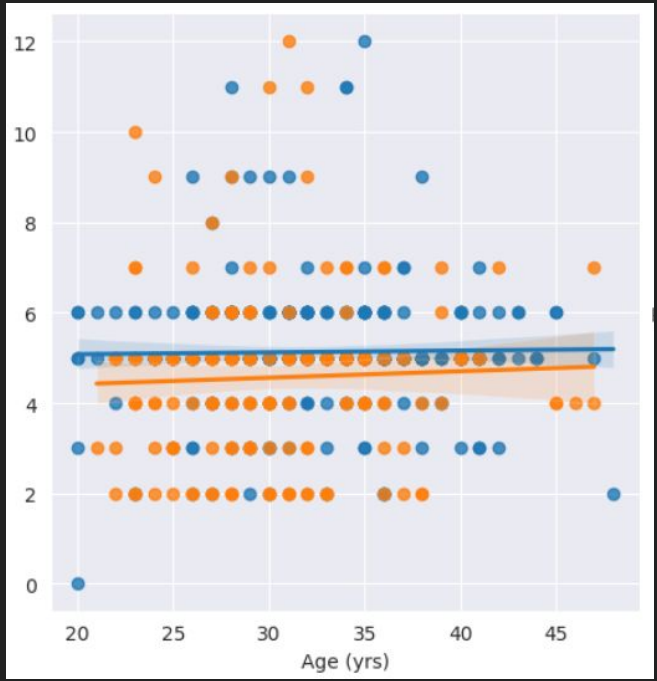
Proportion of diagnoses

	Age Group	PCOS (Y/N)	Proportion
0	(20, 25]	1	0.564516
1	(20, 25]	0	0.435484
2	(25, 30]	0	0.633880
3	(25, 30]	1	0.366120
4	(30, 35]	0	0.702381
5	(30, 35]	1	0.297619
6	(35, 40]	0	0.804348
7	(35, 40]	1	0.195652
8	(40, 45]	0	0.840000
9	(40, 45]	1	0.160000
10	(45, 50]	1	0.600000
11	(45, 50]	0	0.400000

```
# make a linear model plot for the menstrual length, PCOS vs NORMAL
ax = sns.lmplot(data = df, x = "Age (yrs)", y = "Cycle length(days)", hue = "PCOS (Y/N)")
plt.show(ax)
```

Y axis:

Cycle length
→



Higher incidence rate in younger & older pop.

Initial Hypothesis

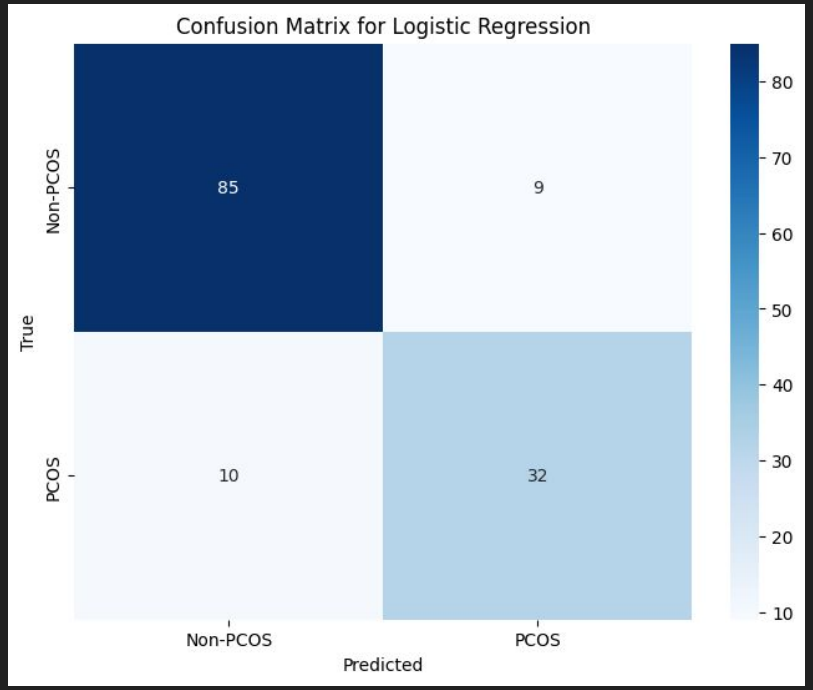
- Age (yrs): Age can influence hormonal balance, which is crucial in PCOS.
- Weight (Kg) and BMI: There's a known correlation between body weight, fat distribution, and PCOS.
- Pulse rate (bpm): Although not a direct indicator, it might provide insights into metabolic rate and overall health.
- Follicle Number (L/R) and Average Follicle Size: Directly related to ovarian function, which is central to PCOS diagnosis.
- Endometrium (mm): Endometrial thickness can be affected by PCOS due to hormonal imbalances.
- Fast food (Y/N): Diet and lifestyle choices are relevant to symptoms and management of PCOS.

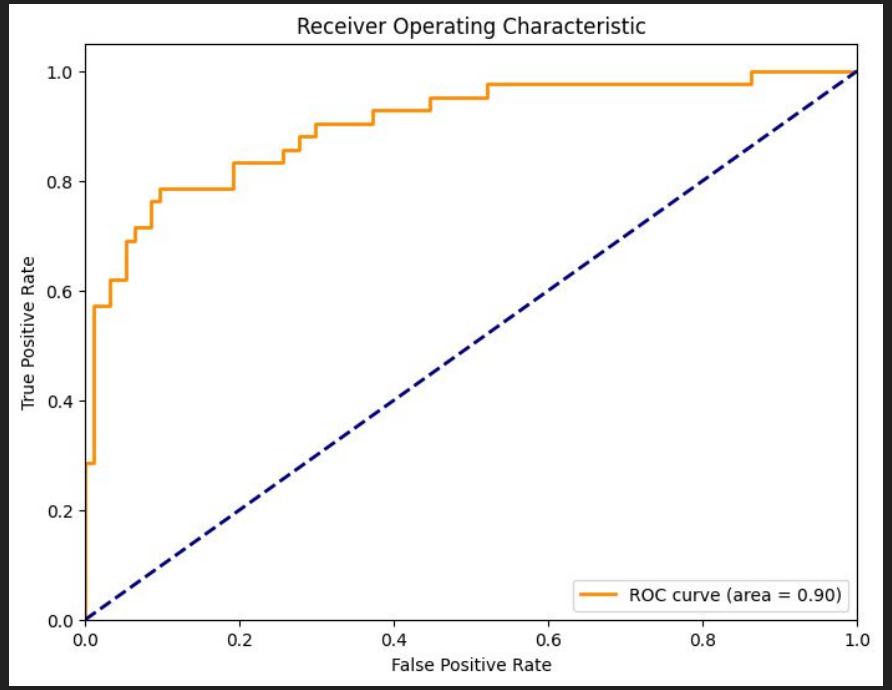
Model Selection

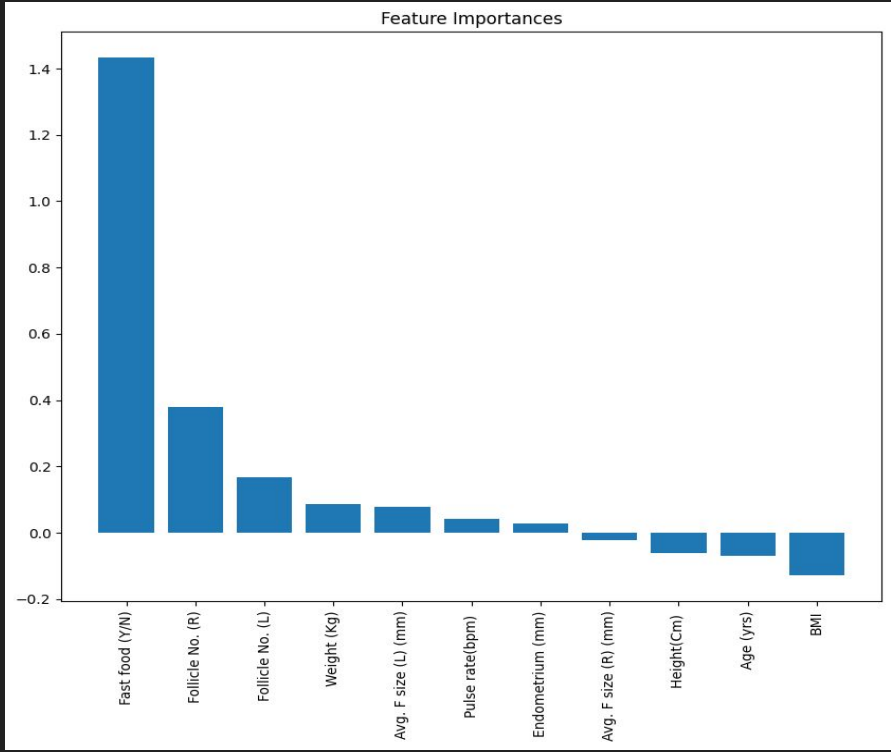
Logistic Regression

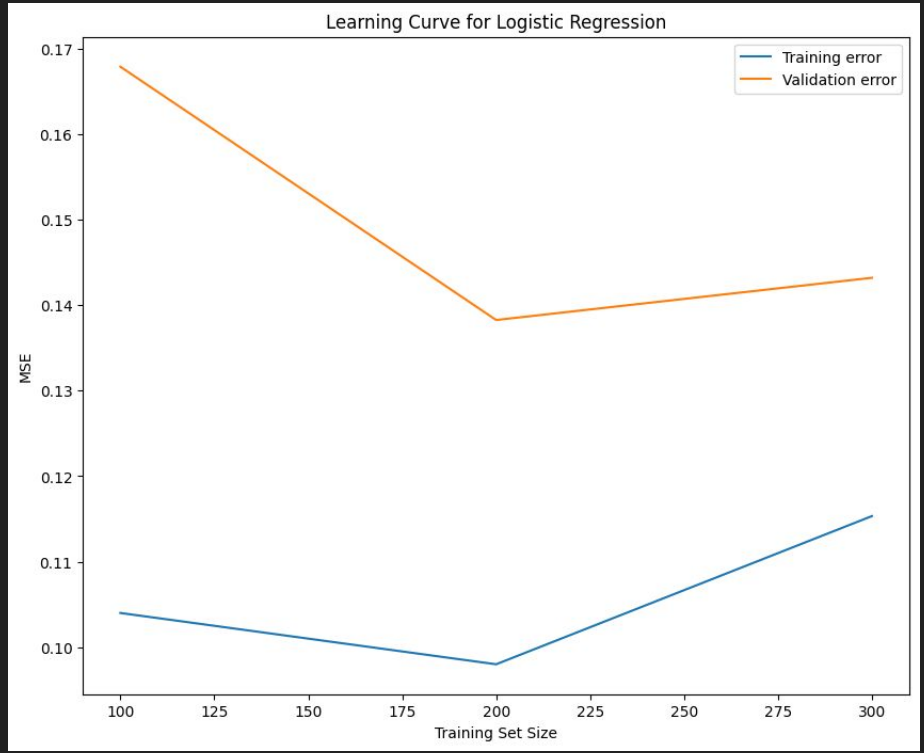
The logistic regression model performed quite well in predicting PCOS based on the selected features:

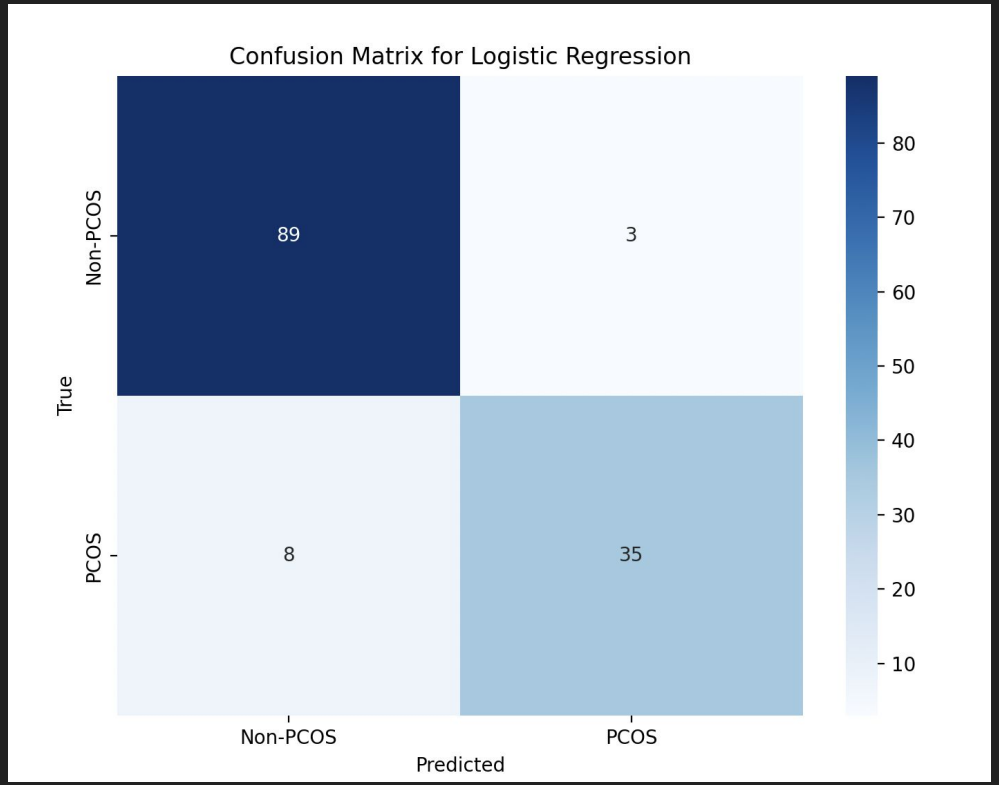
Accuracy: 86.03%
Precision for Class 1 (PCOS): 78%
Recall for Class 1 (PCOS): 76%
F1-Score for Class 1 (PCOS): 77%
These results suggest that the model is quite effective, with good balance between precision and recall for detecting PCOS cases.


Confusion Matrix for Logistic Regression
Non-PCOS
PCOS
True
85
9
10
32
Non-PCOS
PCOS
Predicted
80
70
60
50
40
30
20
10


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.90)

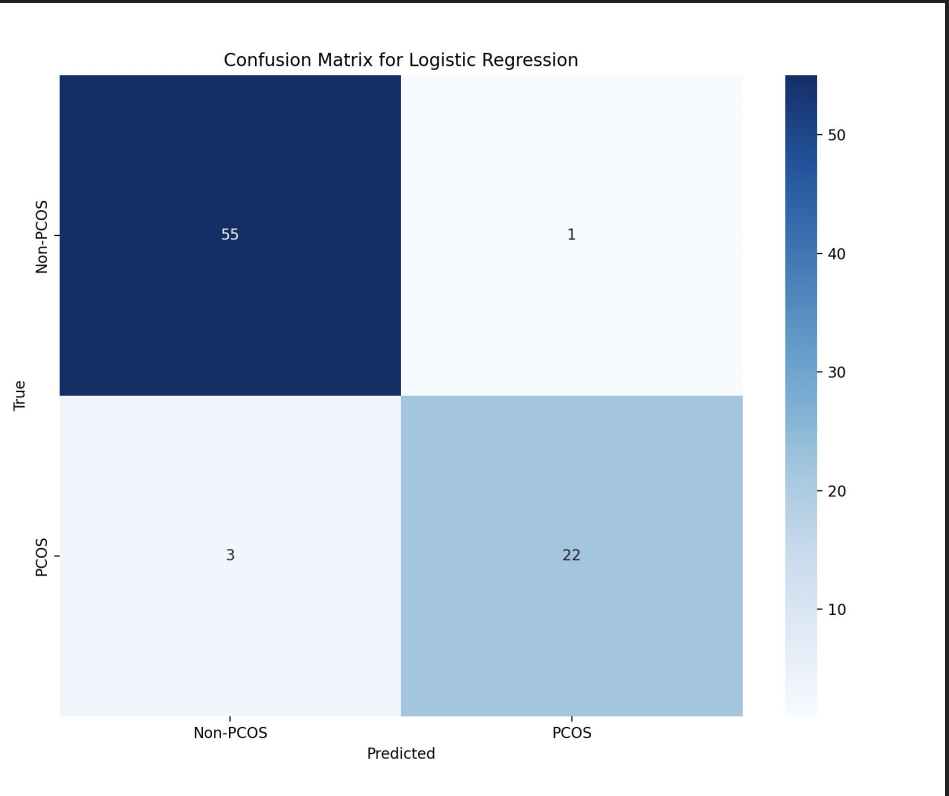

Feature Importances
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Fast food (Y/N)
Follicle No. (R)
Follicle No. (L)
Weight (Kg)
Avg. F size (L) (mm)
Pulse rate(bpm)
Endometrium (mm)
Avg. F size (R) (mm)
Height(Cm)
Age (yrs)
BMI


Learning Curve for Logistic Regression
Training error
Validation error
MSE
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
100
125
150
175
200
225
250
275
300
Training Set Size



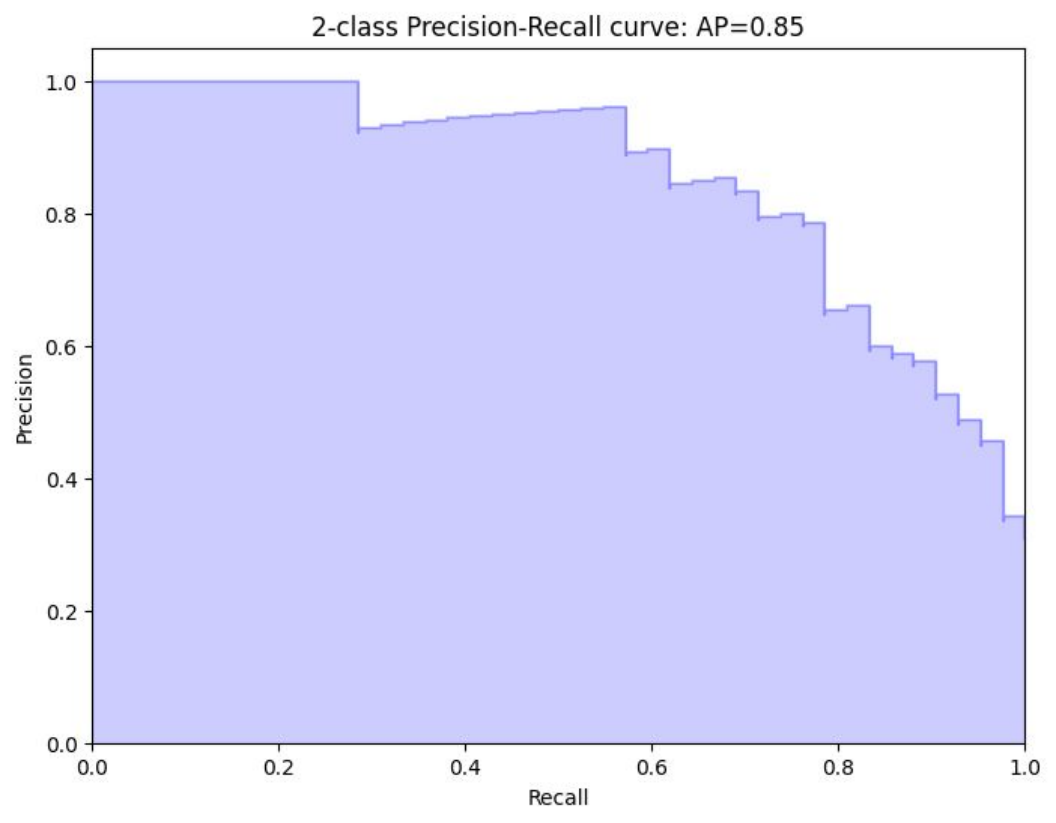

```
Logistic Regression Results:
Accuracy:
0.9185

Classification_report:
                    precision    recall  f1-score   support

                0       0.92      0.97      0.94        92
                1       0.92      0.81      0.86        43

         accuracy                           0.92       135
        macro avg       0.92      0.89      0.90       135
     weighted avg       0.92      0.92      0.92       135


Confusion_matrix:
```

```
Neural Network Results:
Accuracy:
0.9037
```




```
Logistic Regression Results:
Accuracy:
0.9506

Classification_report:
                    precision    recall  f1-score   support

                0       0.95      0.98      0.96        56
                1       0.96      0.88      0.92        25

         accuracy                           0.95        81
        macro avg       0.95      0.93      0.94        81
     weighted avg       0.95      0.95      0.95        81


Confusion_matrix:
```

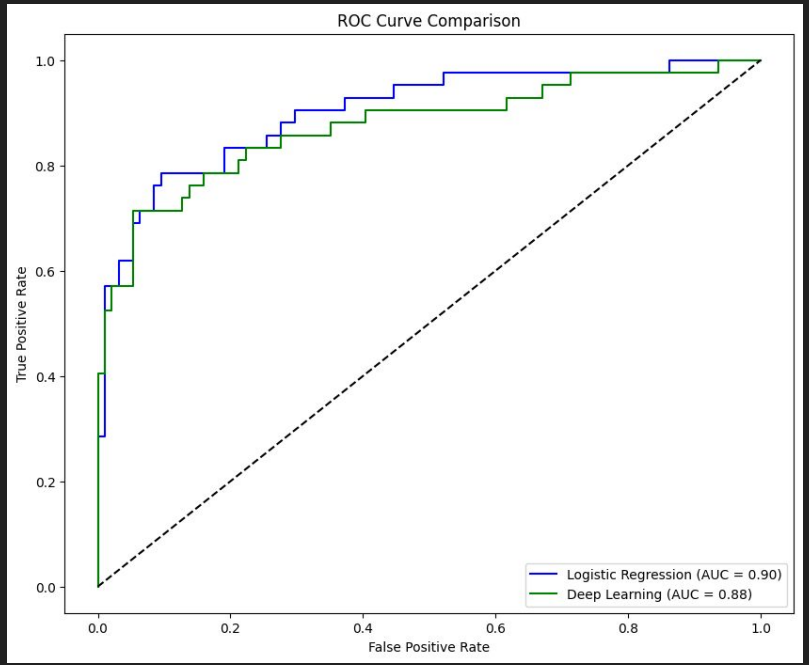
```
Neural Network Results:
Accuracy:
0.9383
```

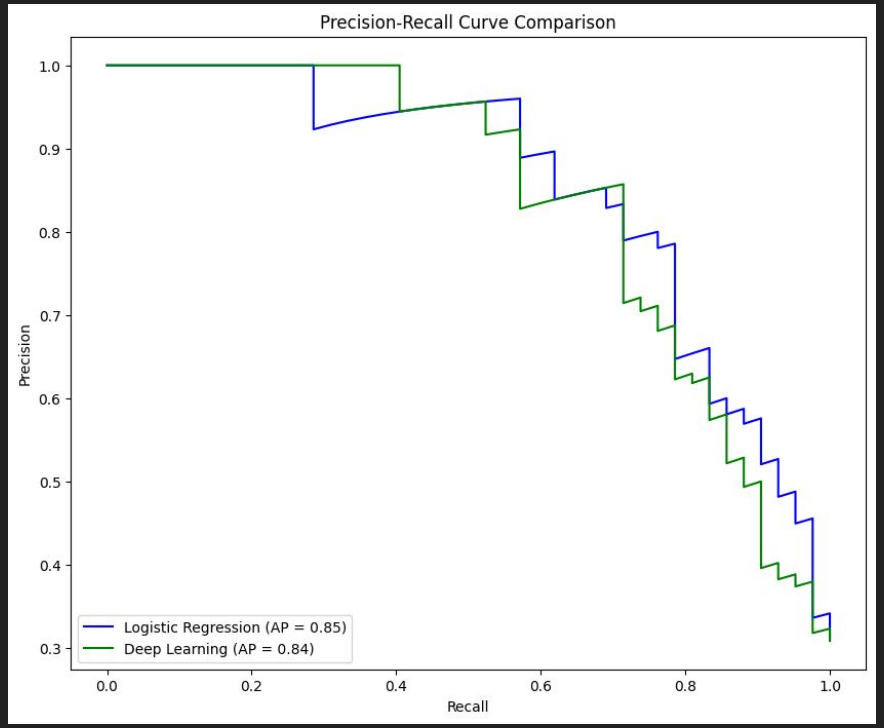

2-class Precision-Recall curve: AP=0.85
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

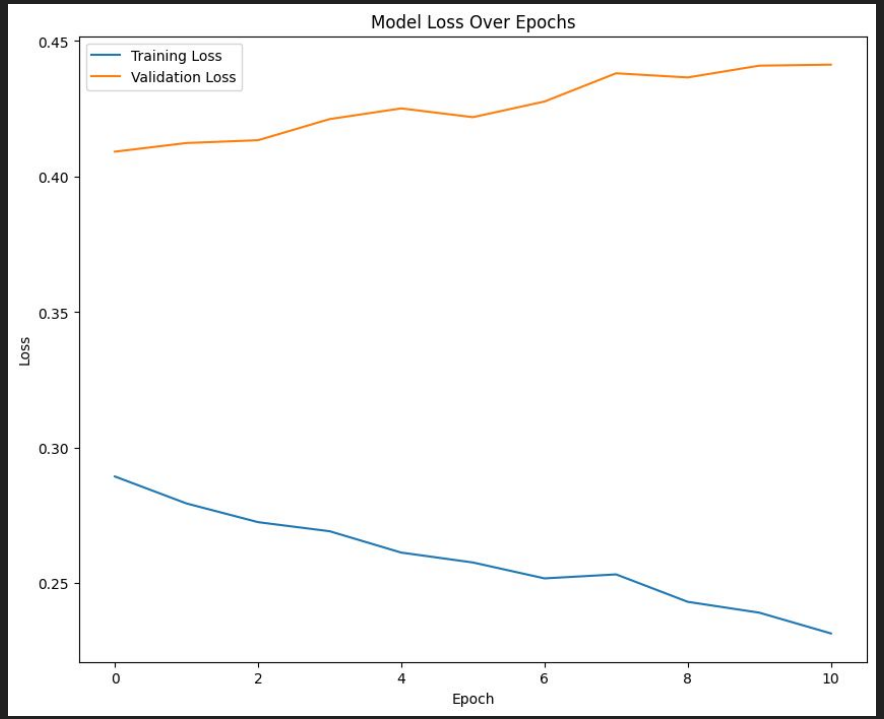
Deep Learning

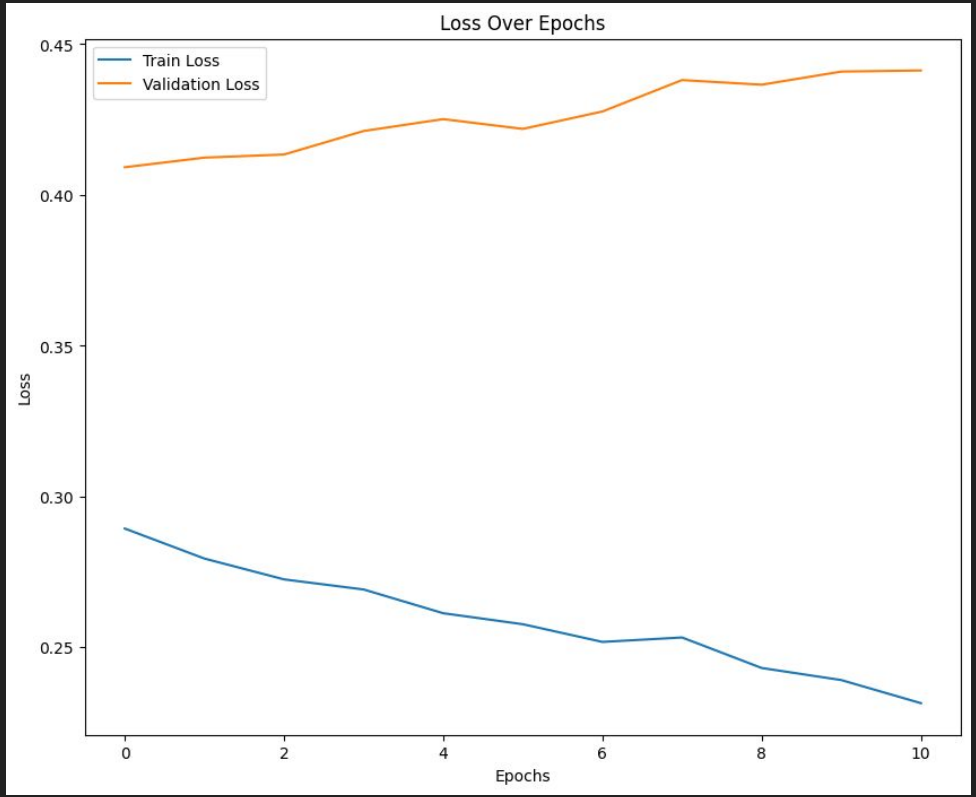
Deep Learning Test Accuracy: 84.56%

```python
# Deep Learning Model
model = Sequential([
    Dense(64, activation='relu', input_shape=(X_train_scaled.shape[1],)),
    Dense(32, activation='relu'),
    Dense(1, activation='sigmoid')
])

model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
early_stopping = EarlyStopping(monitor='val_loss', patience=10, restore_best_weights=True)
model.fit(X_train_scaled, y_train, epochs=50, batch_size=10, validation_split=0.2, callbacks=[early_stopping])

# Evaluate the Deep Learning Model
loss, accuracy = model.evaluate(X_test_scaled, y_test)
print('Deep Learning Test Accuracy:', accuracy)
```




Precision-Recall Curve Comparison
Precision
Recall
Logistic Regression (AP = 0.85)
Deep Learning (AP = 0.84)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

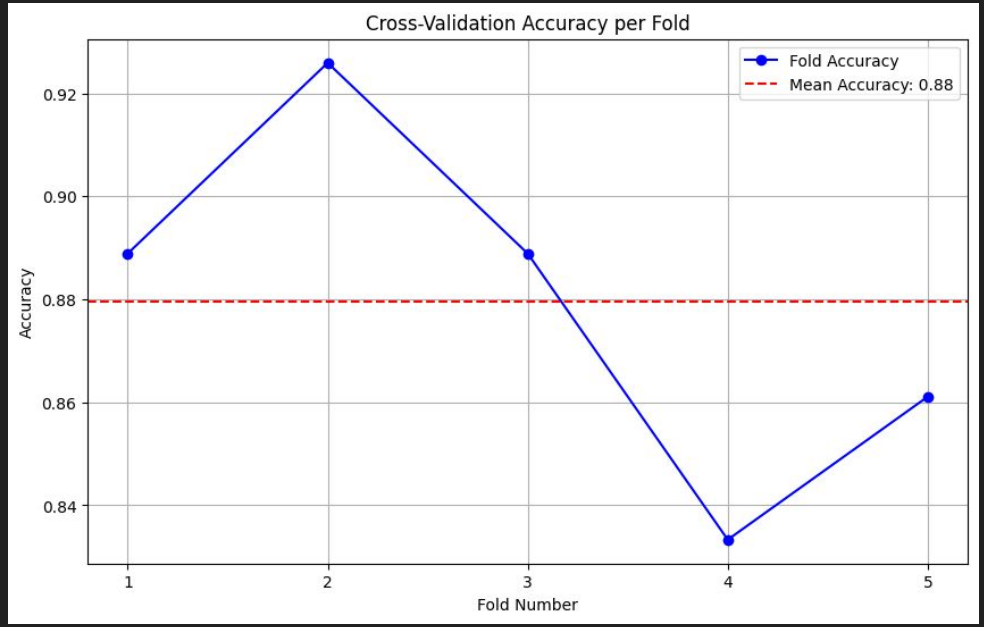

Model Loss Over Epochs
Loss
Epoch
Training Loss
Validation Loss
0.45
0.40
0.35
0.30
0.25
0
2
4
6
8
10

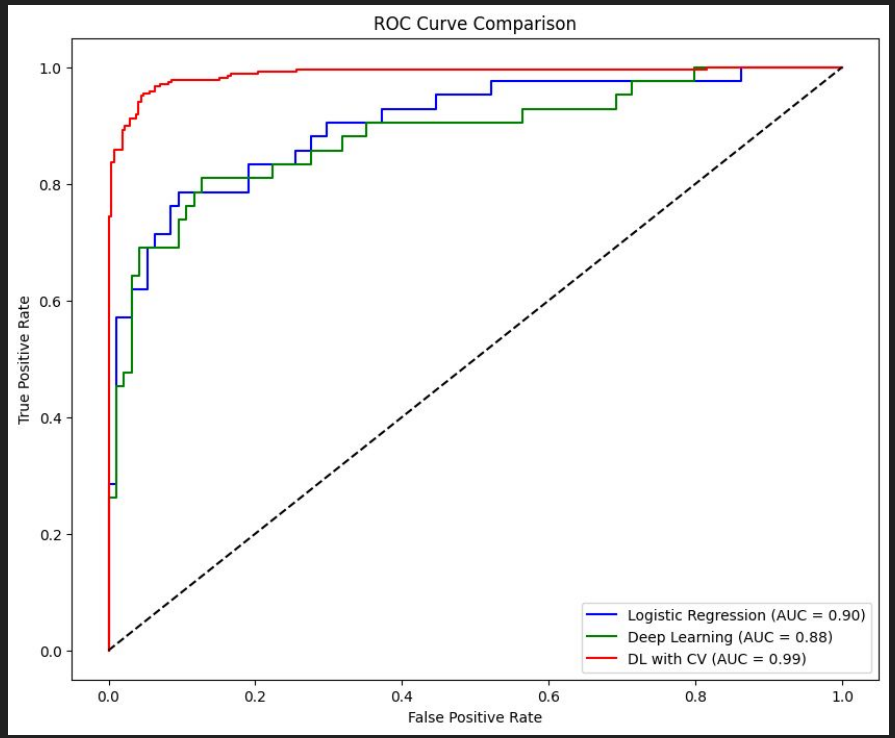

Loss Over Epochs
Train Loss
Validation Loss
0.45
0.40
0.35
0.30
0.25
Loss
0
2
4
6
8
10
Epochs

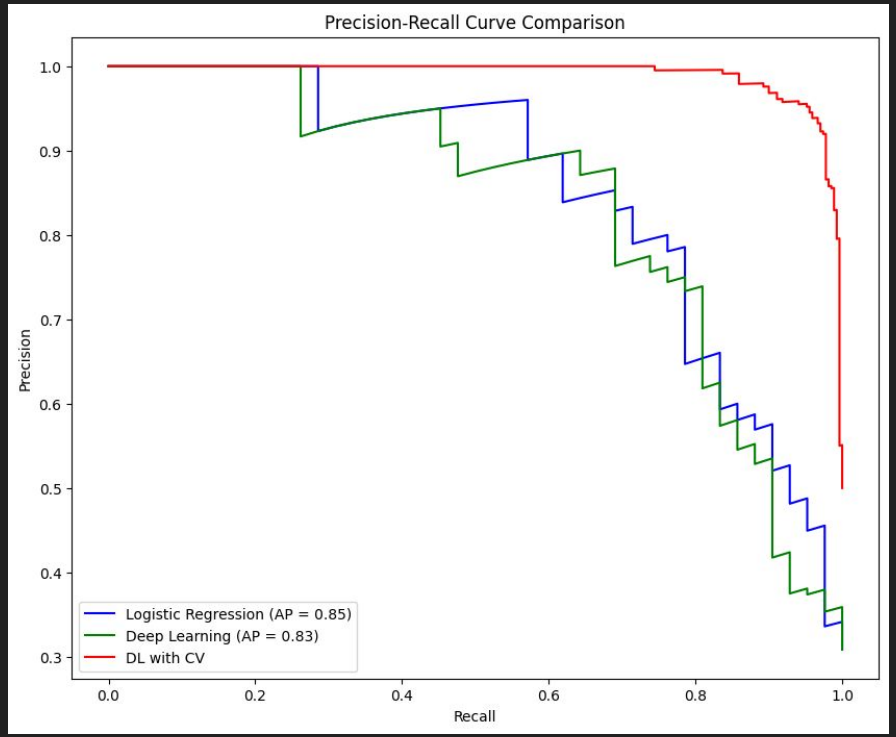
Deep learning with Cross Validation

Mean Accuracy over 5 folds:
0.879629623889923

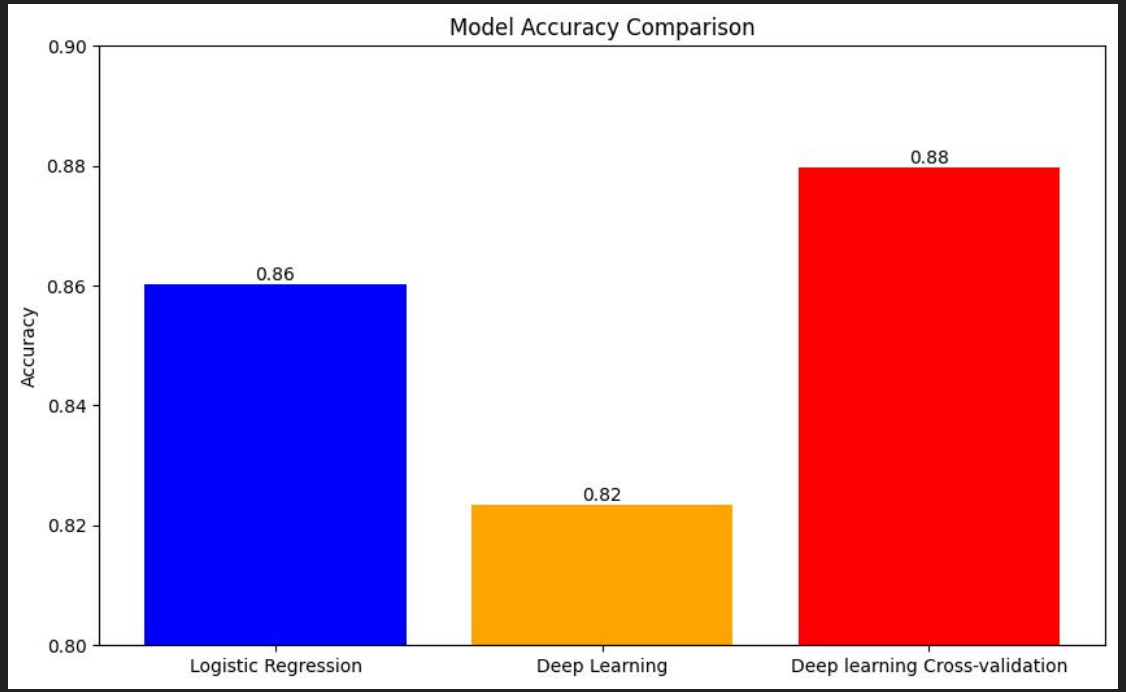
```python
# K-fold cross-validation
kfold = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
fold_no = 1
scores = []

for train, test in kfold.split(X_resampled, y_resampled):
    # Neural network model
    model = Sequential([
        Dense(128, activation='relu', input_shape=(X_train_scaled.shape[1],)),
        BatchNormalization(),
        Dropout(0.5),
        Dense(64, activation='relu'),
        BatchNormalization(),
        Dropout(0.3),
        Dense(32, activation='relu'),
        Dense(1, activation='sigmoid')
    ])

    # Compile the model
    model.compile(optimizer=Adam(learning_rate=0.001), loss='binary_crossentropy', metrics=['accuracy'])

    # Fit the model
    print(f'Training for fold {fold_no} ...')
    model.fit(X_resampled[train], y_resampled[train], epochs=50, batch_size=32, verbose=0)

    # Evaluate the model
    scores.append(model.evaluate(X_resampled[test], y_resampled[test], verbose=0)[1])
    fold_no += 1

print(f'Mean Accuracy over {kfold.n_splits} folds: {np.mean(scores)}')
```




ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.90)
Deep Learning (AUC = 0.88)
DL with CV (AUC = 0.99)


Precision-Recall Curve Comparison
Precision
Recall
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Logistic Regression (AP = 0.85)
Deep Learning (AP = 0.83)
DL with CV

Model Comparisons



Results & Discussion

- The logistic regression model achieved an accuracy of 94%, with precision and recall at 96% and 93%, respectively. Feature importance analysis revealed that ‘Fast Food’, 'Follicle No. (R)', ‘Weight’ and 'Avg. F size (L) (mm)' were pivotal in predicting PCOS.

- Our deep learning model, particularly when enhanced with cross-validation, achieved an impressive accuracy of 88%, indicating robust generalizability.

Model Performance

High Accuracy Factors:

- Advanced Capability: Deep learning ability to capture non-linear relationships in data likely enhanced its performance.
- Cross-Validation: Rigorous cross-validation contributed to robustness, ensuring the model performs well on unseen data.
- Preprocessing and Feature Selection: Careful preprocessing and strategic feature selection based on domain expertise helped focus the model on the most relevant predictors, reducing noise.

Future Improvements

Considerations and Cautions:

- Risk of Overfitting: Deep learning models are prone to memorizing training data rather than generalizing from it, raising concerns about overfitting.
- Future Improvements:
  - Dropout Techniques: Incorporating dropout techniques and regularization could help mitigate overfitting risks.
  - External Validation: For clinical deployment, validating models on an independent external dataset is crucial to confirm their practical utility and reliability.

Exploration of data

relevant physical and clinical parameters for PCOS

- 1: Project presentation lacked any sort of depth and clarity, student(s) presenting did not clearly address issues from prompt
- 2: Project presentation included little detail and inconclusive results, difficult to understand
- 3: Project presentation addresses challenges stated in prompt clearly and concisely, easy for audience to follow and understand

PROJECT DESIGN

- 1: Project not designed well; yields no tangible results or applicability
- 2: Project designed to move forward, but disjointed approach
- 3: Project designed to move forward in systematic way

This is what the syllabus says:

https://docs.google.com/document/d/1mIJUfQBfq0tC8_9HiACewh-5BFG6YwUM-XBBFJKJt0k/edit

Coordination

Exploration of data

ADDRESSING STATED PROBLEMS

- 1: Results did not address stated problem at all
- 2: Results include missing details and inconclusive results
- 3: Results solves stated problem at face value

Fixing outliers
<decision point, challenges>

Choosing the models