Data-Driven Clinical Decision Support for PCOS: Applying Regression Models to Improve Care

Group 6

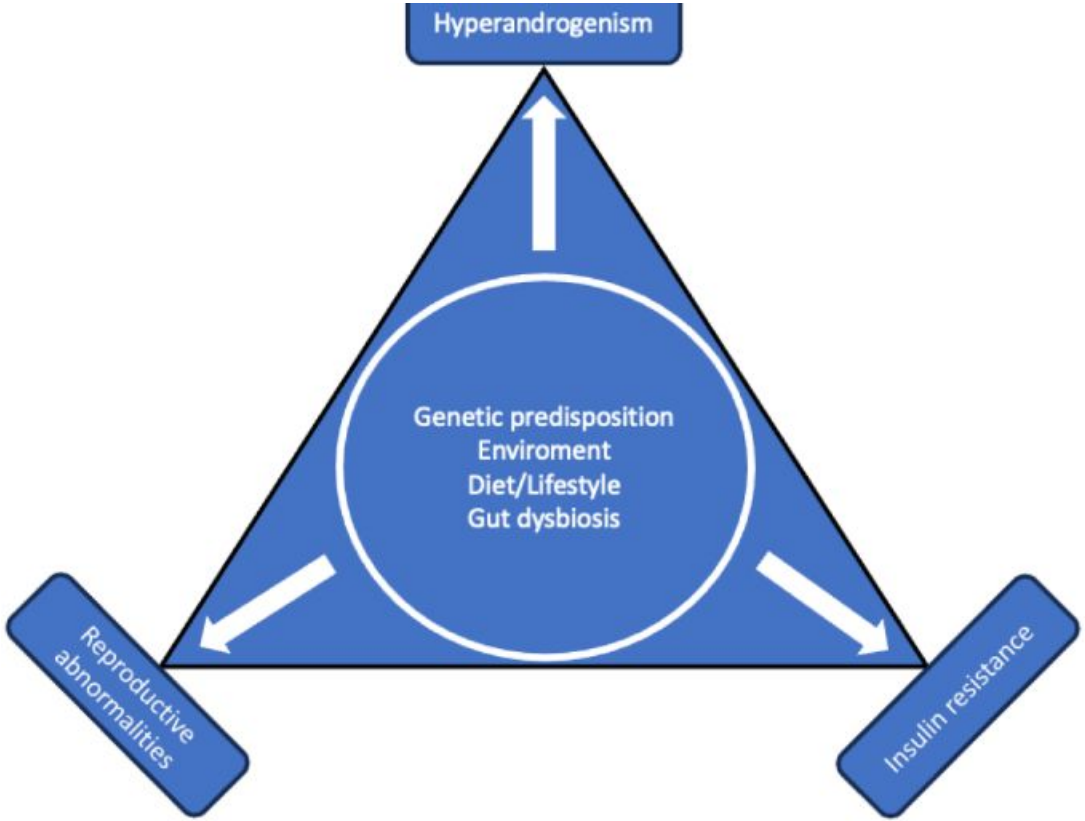
Saketha Kusu, Chrishey Holbrook

Introduction to the Project

- Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine disorder impacting millions of women globally. It carries long-term risks, including infertility, diabetes, and cardiovascular disease.

- This project leverages logistic regression to predict the risks of cardiovascular and metabolic complications in PCOS patients, supporting early clinical interventions to improve care outcomes.

This diagram highlights the multifactorial nature of PCOS, driven by genetic, environmental, and lifestyle factors. Key outcomes include hyperandrogenism, insulin resistance, and reproductive abnormalities, showcasing their interconnections.



Research Objectives

Objective:

- Develop a Clinical Decision Support System (CDSS) using logistic regression to enhance predictive insights in PCOS patient care.

Key Goals:

- Identify key predictors (e.g., insulin resistance, obesity) impacting PCOS complications.
- Enhance early diagnosis and personalized care.

Literature Review Findings

Key Insights:

- Predictive models can identify high-risk PCOS patients effectively, especially through logistic regression for binary outcomes.
- Data-driven insights, including integration of clinical factors, are essential, though current models often lack genetic and lifestyle data, limiting accuracy.
- Limited models targeting PCOS-specific risks like cardiovascular diseases.

Methodology Overview

Model Selection:

- Logistic regression is chosen for its suitability in binary classification tasks, ideal for predicting the likelihood of cardiovascular complications in PCOS patients.

Steps:

1. Data Collection and Cleaning
2. Model Training and Validation
3. Feature Selection and Expansion

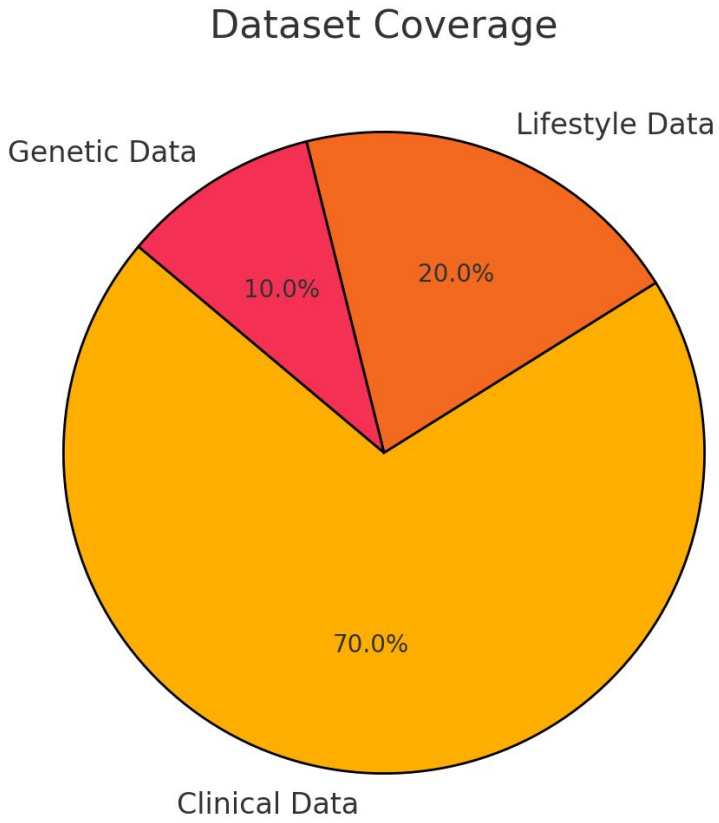
Justification of Population Chosen

- PCOS affects **1 in 10 women** globally.
- Leads to serious complications like cardiovascular disease and diabetes.
- A CDSS provides critical tools for early detection and intervention.

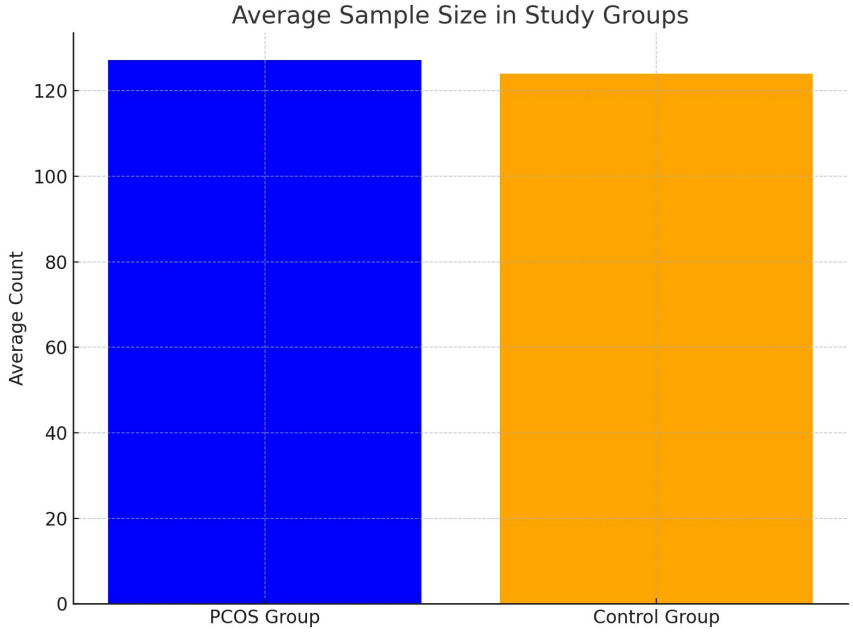
Data Collection and Sources

Dataset Details:

- Retrospective and prospective studies with sample sizes ranging from 80 to 174,000 participants.
- Key data points include BMI, insulin resistance, menstrual irregularities, arterial stiffness, and more.
- Source: Peer-reviewed data from the Journal of Clinical Medicine.



This pie chart illustrates the dataset composition: 70% clinical data, 20% lifestyle data, and 10% genetic data. The focus on clinical data highlights its importance in predicting PCOS-related complications, while the smaller proportions of lifestyle and genetic data indicate areas for future dataset enhancement to improve predictive accuracy.



The bar chart compares the average sample sizes of PCOS and Control groups.

The PCOS group has a slightly larger average size, highlighting the need for balanced data in model training.

Data Preprocessing

We conducted extensive data preprocessing including:

- Data cleaning to handle missing values
- Encoding categorical variables to ensure consistency across data.
- This process is crucial given the complexity and multifaceted nature of clinical data.

```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score
from sklearn.preprocessing import LabelEncoder

# Load the dataset
data = pd.read_csv("C:\\Users\\chris\\Downloads\\PCOS_cardiovascular_data.csv")

# Handle missing values (if any)
data.fillna(data.median(), inplace=True)

# Define predictors and target
predictors = ['Mean Age', 'PCOS cIMT (mm)', 'Control cIMT (mm)', 'PCOS n', 'Control n
target = 'Results'  # Assuming 'Results' is the target variable

# If 'Results' is categorical, we need to encode it
label_encoder = LabelEncoder()
data[target] = label_encoder.fit_transform(data[target])
```

Correlation Matrix - Heat Map

We used the correlation matrix to show how well our model is performing by comparing its predictions to the actual outcomes. The confusion matrix helps us see where the model is making mistakes.

The matrix has four key parts:

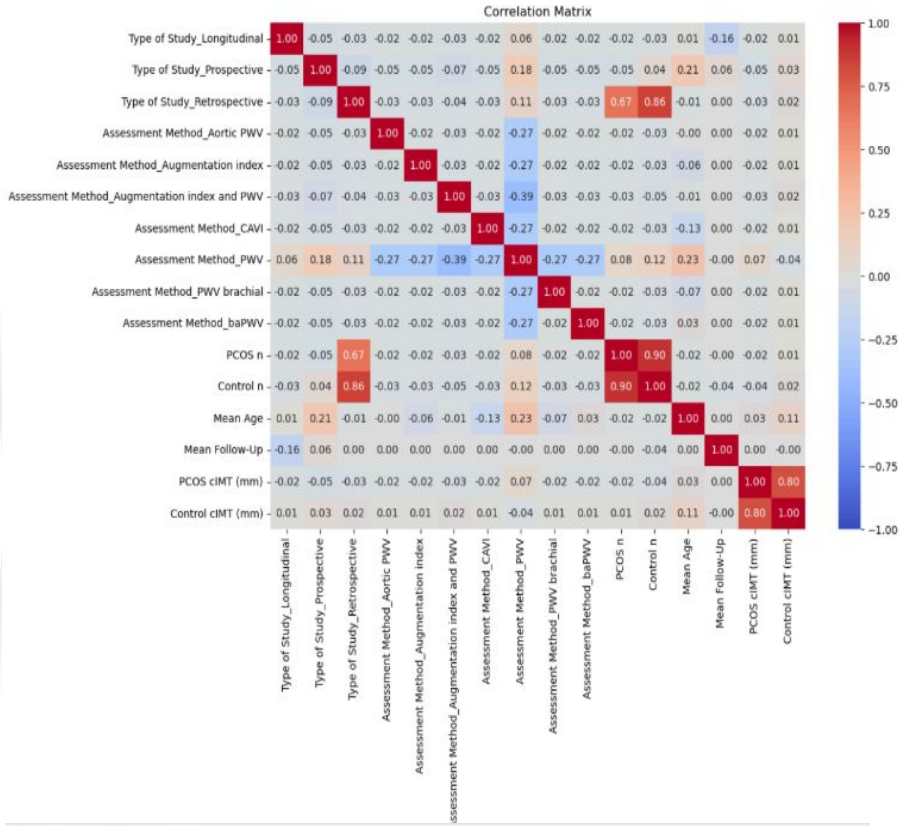
True Positives (TP): The number of times the model correctly predicted the positive class (e.g., predicting a patient has a disease when they actually do).

True Negatives (TN): The number of times the model correctly predicted the negative class (e.g., predicting a patient does not have a disease when they actually don't).

False Positives (FP): The number of times the model incorrectly predicted the positive class (e.g., predicting a patient has a disease when they actually don't).

False Negatives (FN): The number of times the model incorrectly predicted the negative class (e.g., predicting a patient does not have a disease when they actually do).

•The matrix is often color-coded to show how many times each of these happened. Darker colors might mean higher numbers, and lighter colors mean lower numbers. This chart helps us see not just how many predictions were correct, but where the model made mistakes. For example, if the model misses a lot of positives (False Negatives), we know it needs improvement in detecting the positive class.




```
# Compute the correlation matrix
correlation_matrix = df.corr()

# Display the correlation matrix
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f', vmin=-1, vmax
plt.title('Correlation Matrix')
plt.show()
```

Preliminary Findings

Initial findings reveal insulin resistance and BMI as significant predictors of cardiovascular complications in PCOS patients.

These preliminary insights underscore the impact of metabolic factors on risk prediction and lay the foundation for refining our model.

Model Training, Testing, and Validation

Data split into 80% training and 20% testing.

Ensures sufficient data for training while maintaining evaluation integrity.

Logistic Regression (max_iter = 1000)

Random Forest (n_estimators = 100)

Our logistic regression model undergoes cross-validation to enhance reliability.

We track metrics such as accuracy, precision, recall to assess model performance and ensure it aligns with clinical standards.

```
# Logistic Regression Model
model = LogisticRegression(max_iter=1000)
# Fitting the model
model.fit(X_train, y_train)

# Predicting on test data
y_pred = model.predict(X_test)
# Evaluating the model
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy * 100:.2f}%')
```

Model Performance and Accuracy

Accuracy:

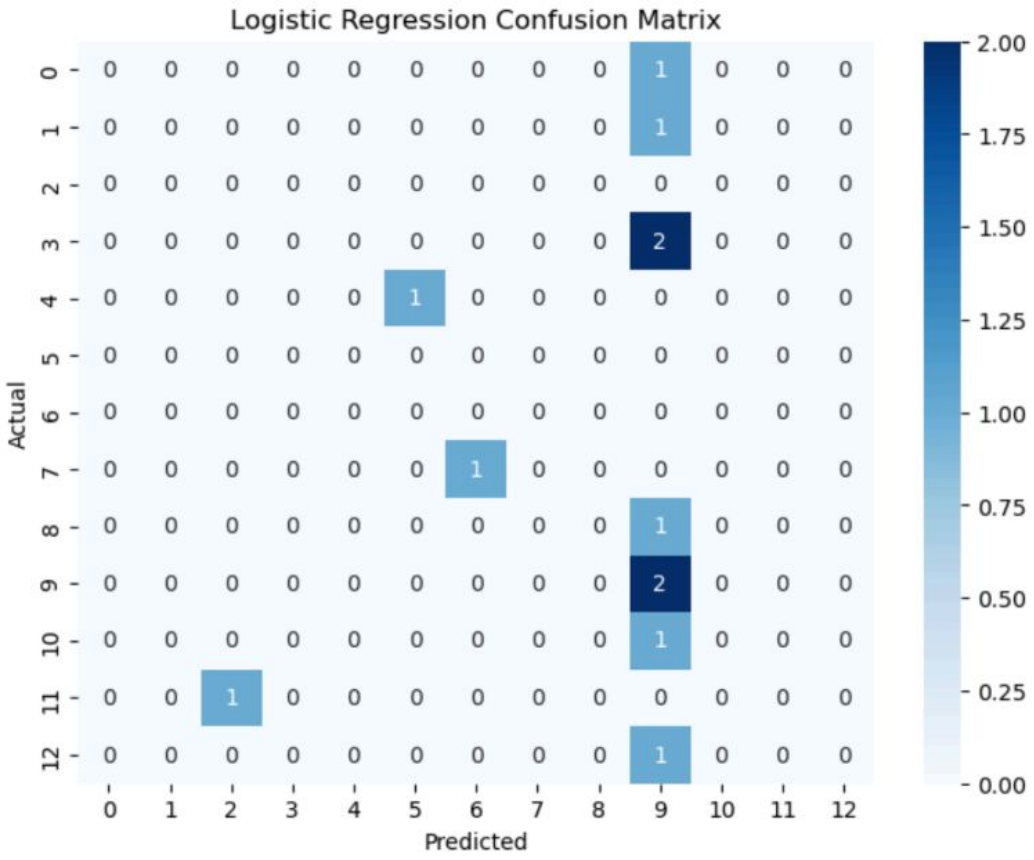
- Logistic Regression: 91.67%
- Random Forest: 25%

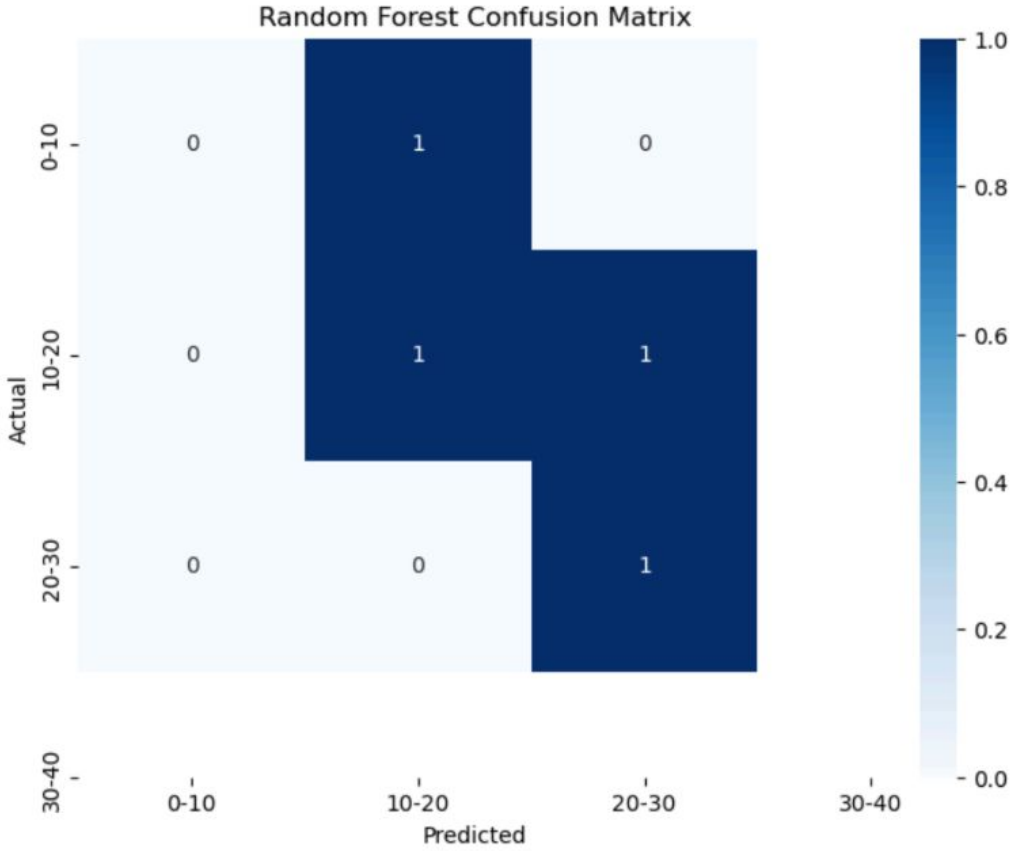
Feature Importance (Random Forest):

- Control n: 36.25%
- Mean Age: 34.61%
- PCOS n: 23.04%

Observations:

- Logistic Regression outperformed Random Forest.
- Limited availability of genetic and lifestyle data reduces model accuracy.
- Simplified assumptions in logistic regression restrict capturing complex interactions.
- AUC-ROC curve analysis not performed, limiting a full comparison of model discrimination.


Logistic Regression Confusion Matrix
Actual
Predicted
0 1 2 3 4 5 6 7 8 9 10 11 12
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

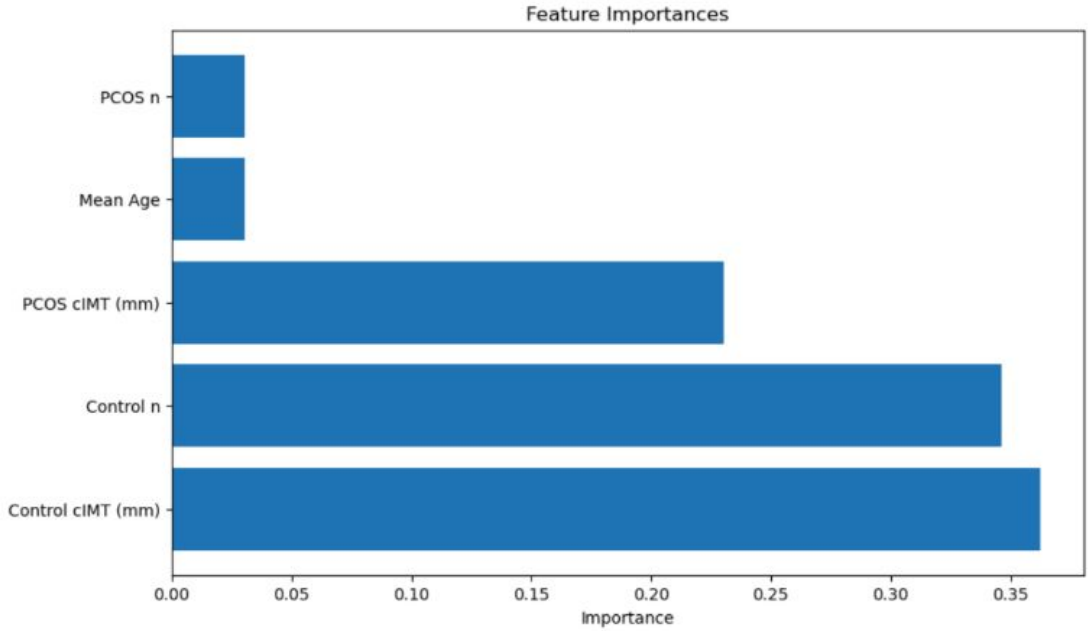

Random Forest Confusion Matrix
Actual
Predicted
0-10
10-20
20-30
30-40
0
1
0
0
1
1
0
0
1
1.0
0.8
0.6
0.4
0.2
0.0

Feature Importances

The feature importance plot shows the relative importance of each feature in the model. From the graph, we can see that PCOS cIMT and Control cIMT have the highest importance, followed by Control n and Mean Age.

The feature PCOS n has the least importance in the model's predictions. This helps us understand which variables are driving the model's decision-making process.

```python
# Plot the feature importances
plt.figure(figsize=(10, 6))
plt.barh(range(len(feature_names)), importances[indices], align='center')
plt.yticks(range(len(feature_names)), np.array(feature_names)[indices])
plt.xlabel('Importance')
plt.title('Feature Importances')
plt.show()
```



Limitations

1. Data Limitations:

- Insufficient genetic and lifestyle data reduces analysis scope.

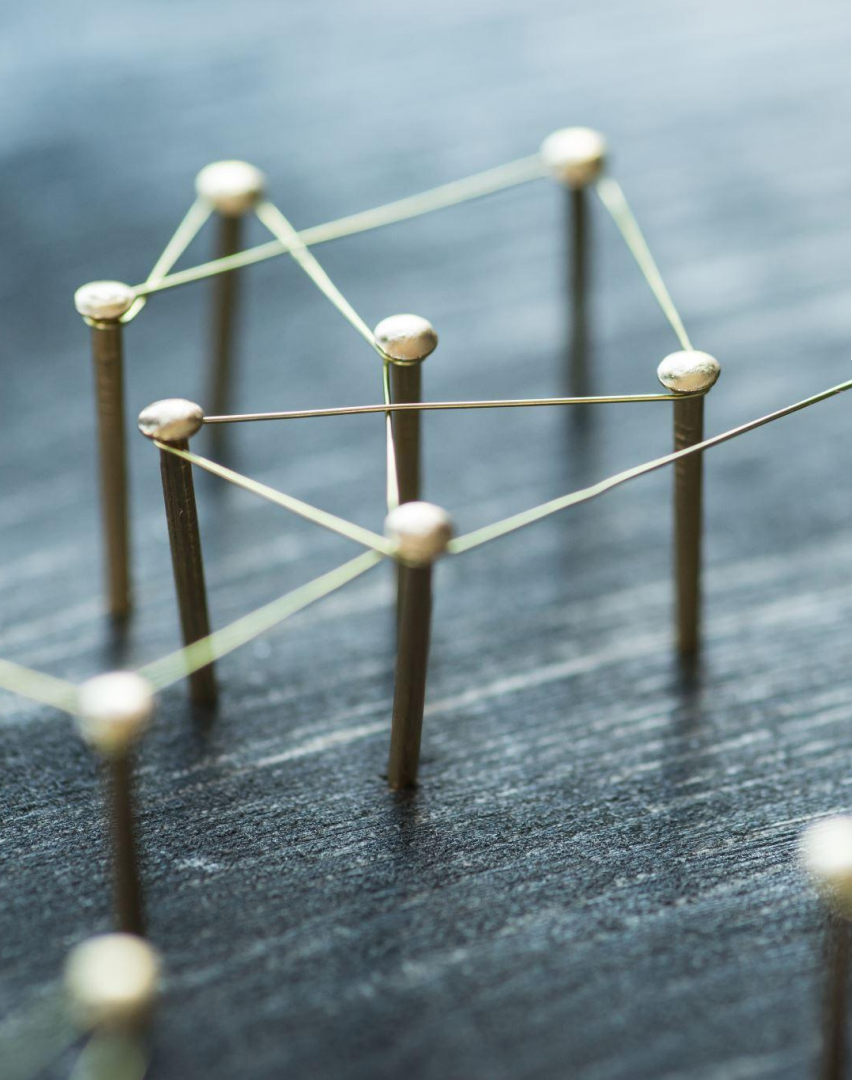
- Sample diversity issues may affect model generalizability.

2. Model Constraints:

- Logistic regression assumptions oversimplify complex interactions.

- Challenges in capturing relationships like insulin resistance and BMI.

Performance metrics like AUC-ROC were not computed, which limits the ability to assess model discrimination comprehensively.



Adjustments for Limitations

1. Enhancing Data:

Expanding features to incorporate genetic and lifestyle data

2. Improving Models:

Exploring advanced models like Random Forest

3. Mitigating Performance Issues:

Performing feature engineering and hyperparameter tuning to refine models.

Conclusion

To Conclude our project demonstrated the potential of using logistic regression for predictive insights in PCOS care.

We identified critical predictors like BMI and cIMT, which provide actionable insights for clinicians.

This work serves as a foundation for building more comprehensive predictive tools to improve health outcomes for PCOS patients.

References

Profili, N. I., Castelli, R., Gidaro, A., Manetti, R., Maioli, M., Petrillo, M., Capobianco, G., & Delitala, A. P. (2024). Possible effect of polycystic ovary syndrome (PCOS) on cardiovascular disease (CVD): An update. *Journal of Clinical Medicine, 13*(3), 698. https://doi.org/10.3390/jcm13030698

Joham, A. E., Kakoly, N. S., Teede, H. J., & Earnest, A. (2021). Incidence and predictors of hypertension in a cohort of Australian women with and without polycystic ovary syndrome. *Journal of Clinical Endocrinology & Metabolism*, 106(5), 1585–1593. https://doi.org/10.1210/clinem/dgab092

Schmidt, J., Landin-Wilhelmsen, K., Brannstrom, M., & Dahlgren, E. (2011). Cardiovascular disease and risk factors in PCOS women of postmenopausal age: A 21-year controlled follow-up study. *Journal of Clinical Endocrinology & Metabolism*, 96(12), 3794–3803. https://doi.org/10.1210/jc.2011-1750

Amiri, M., Ramezani Tehrani, F., Behboudi-Gandevani, S., Bidhendi-Yarandi, R., & Carmina, E. (2020). Risk of hypertension in women with polycystic ovary syndrome: A systematic review, meta-analysis, and meta-regression. Reproductive Biology and Endocrinology, 18(1), 23. https://doi.org/10.1186/s12958-020-00583-4

Ozkan, S., Yilmaz, O. C., & Yavuz, B. (2020). Increased masked hypertension prevalence in patients with polycystic ovary syndrome (PCOS). *Clinical and Experimental Hypertension*, 42(7), 681–684. https://doi.org/10.1080/10641963.2020.1762591

Thank you