CMP5101 Data Mining Midterm Project

Student ID	Student Name
1900812	Umniyah Sameer Hiatham Abood

Feature Correlation in the Diagnosis of Endometriosis and PCOS in Female Infertility

In this project, I am diving into a topic that is important to me: **female infertility**, with a special focus on **endometriosis** as my main subject, and also exploring another condition called **Polycystic Ovary Syndrome (PCOS)**. Both are major causes of infertility in women, yet are often misunderstood or misdiagnosed.

Endometriosis is a chronic inflammatory condition where tissue similar to the lining of the uterus grows outside the uterus, causing pain, irregular bleeding, and sometimes infertility. According to the European Society of Human Reproduction and Embryology (ESHRE), it affects about **10% of reproductive-age women globally** [Johnson et al., 2022].

PCOS, on the other hand, is a hormonal disorder characterized by irregular periods, excess androgen levels, and polycystic ovaries. It is the most common endocrine disorder in women of reproductive age, affecting approximately **8–13% of women worldwide** [Teede et al., 2018; Lim et al., 2019].

When it comes to symptoms, **chronic pelvic pain, painful periods (dysmenorrhea), and infertility** are the most correlated with **endometriosis**. For **PCOS**, the most common symptoms include **irregular menstruation, acne, hirsutism (excess hair growth), and difficulty getting pregnant**.

Research in recent years has tried to uncover how these conditions are linked to infertility. Endometriosis may cause infertility by creating inflammation and changing the shape of organs in the pelvis [Dunselman et al., 2014]. PCOS mainly affects ovulation due to hormone problems like high androgens or insulin resistance [Teede et al., 2018]. Although both conditions reduce fertility, they do it in different ways—one by physical changes and inflammation, the other by stopping regular ovulation.

By combining medical data on endometriosis, PCOS, and female infertility, my goal is to identify the most relevant and correlated features for diagnosing endometriosis, especially since it is often underdiagnosed.

1. Endometriosis Dataset

1.1 Data Exploration

The dataset has 5 features and one target feature, which is the **Diagnosis**. Features like **Menstrual_Irregularity**, **Hormone_Level_Abnormality**, and **Infertility** are categorical and already encoded as 0 and 1. **Age** ranges from 18 to 49, **Chronic_Pain_Level** is on a scale from 0 to 10, and **BMI** ranges between 15 and 37. You can check the notebook named *Data Mining Endometriosis* for full data exploration. There are no null values or duplicates. However, the target label is imbalanced, so I will apply undersampling—you can see this in **Image 4**.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Age                        10000 non-null  int64  
 1   Menstrual_Irregularity     10000 non-null  int64  
 2   Chronic_Pain_Level         10000 non-null  float64
 3   Hormone_Level_Abnormality  10000 non-null  int64  
 4   Infertility                10000 non-null  int64  
 5   BMI                        10000 non-null  float64
 6   Diagnosis                  10000 non-null  int64  
dtypes: float64(2), int64(5)
memory usage: 547.0 KB
None
```

Figure 1 Data exploration for Endometriosis Dataset

Descriptive Statistics

	Age	Menstrual_Irregularity	Chronic_Pain_Level	Hormone_Level_Abnormality	Infertility	BMI	Diagnosis
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	33.692300	0.697500	5.030619	0.591100	0.298300	23.052865	0.407900
std	9.205308	0.459364	1.983955	0.491655	0.457535	3.891615	0.491469
min	18.000000	0.000000	0.000000	0.000000	0.000000	15.000000	0.000000
25%	26.000000	0.000000	3.671697	0.000000	0.000000	20.329327	0.000000
50%	34.000000	1.000000	5.035825	1.000000	0.000000	23.036315	0.000000
75%	42.000000	1.000000	6.396854	1.000000	1.000000	25.712923	1.000000
max	49.000000	1.000000	10.000000	1.000000	1.000000	37.146127	1.000000

Figure 2 Descriptive Statistics for Endometriosis Dataset

```
Unique Values Count:
 Age                          32
Menstrual_Irregularity         2
Chronic_Pain_Level          9875
Hormone_Level_Abnormality      2
Infertility                    2
BMI                         9776
Diagnosis                      2
dtype: int64
```

Figure 3 Unique Values Coun for Endometriosis Dataset

```
Target Variable Distribution:
0    0.5921
1    0.4079
Name: Diagnosis, dtype: float64
```

Figure 4 Target Variable Distribution for Endometriosis Dataset

1.2 Data Visualization

In the correlation matrix below in image 5 and image 6, none of the features shows a strong linear relationship with the **Diagnosis**. The highest correlation is with **Hormone Level Abnormality** (0.187), followed by **Chronic Pain Level** (0.117), **Infertility** (0.096), and **Menstrual Irregularity** (0.095)—all of which are weak but clinically relevant. **BMI** shows a very weak correlation (0.08), and **Age** has no correlation (-0.012).

As a result, there is little to no correlation among the symptoms themselves, suggesting they may contribute independently to the condition or are measured in a way that does not reflect symptom clustering.

```
Correlation with Diagnosis:
 Diagnosis                   1.000000
Diagnosis_numeric            1.000000
Hormone_Level_Abnormality    0.187039
Chronic_Pain_Level           0.116996
Infertility                  0.096172
Menstrual_Irregularity       0.095197
BMI                          0.080310
Age                         -0.011559
Name: Diagnosis, dtype: float64
```

Figure 5 Correlation with Diagnosis for Endometriosis Dataset

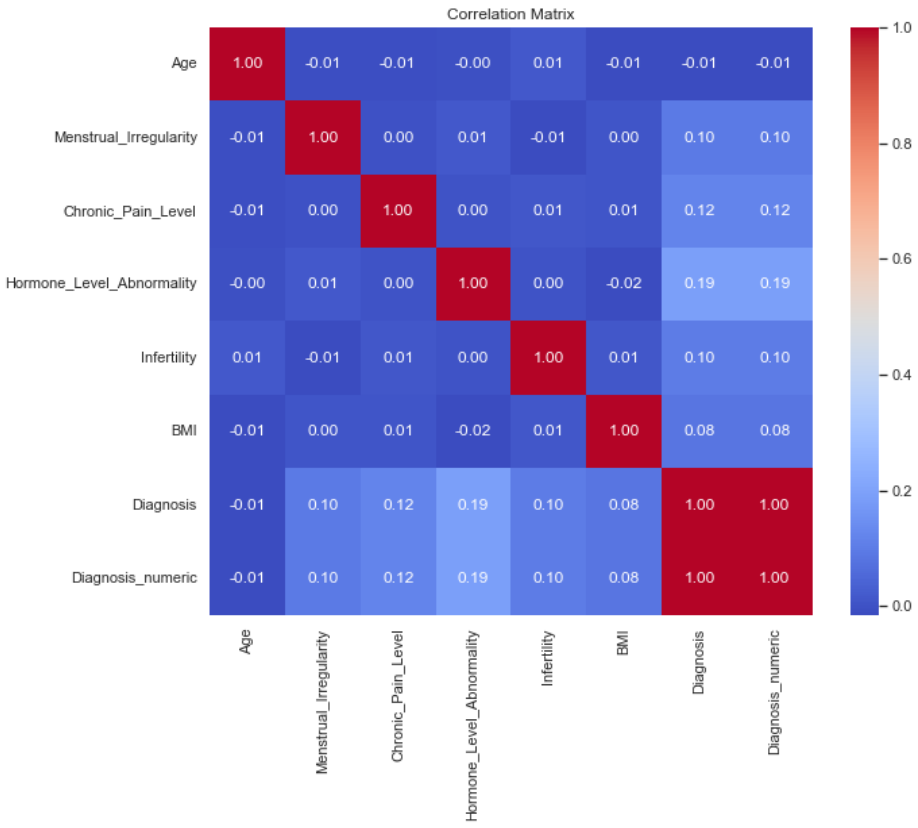

Figure 6 Correlation Matrix for Endometriosis Dataset

We can observe in Figure 7 that **chronic pain levels**, rated on a scale from 0 to 10, are most commonly reported between **5 and 6**, indicating moderate to severe pain among many participants. Figure 8 shows that as well. Regarding **BMI**, the majority of values fall within the **healthy (18.5–23)** and **overweight (23–27.5)** ranges. According to the standard BMI categories:

- Underweight: <18.5
- Healthy: 18.5–23
- Overweight: 23–27.5
- Obese: >27.5

This distribution is not surprising, as women with **endometriosis** often experience **insulin resistance**, which can make it more difficult to lose weight. Several studies have found a link between endometriosis and metabolic disturbances. Same thing can be seen in Figure 9.

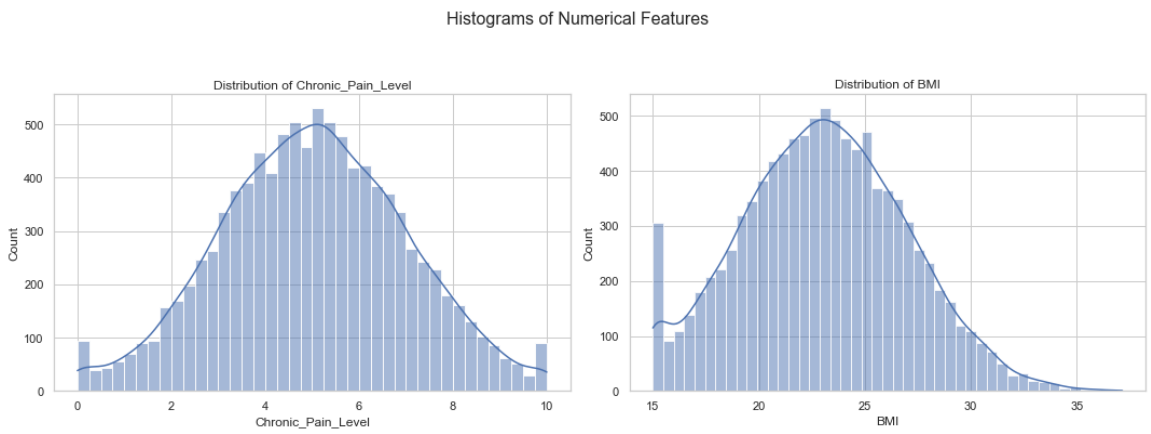

Figure 7 KDE of Chronic_Pain_Level and BMI

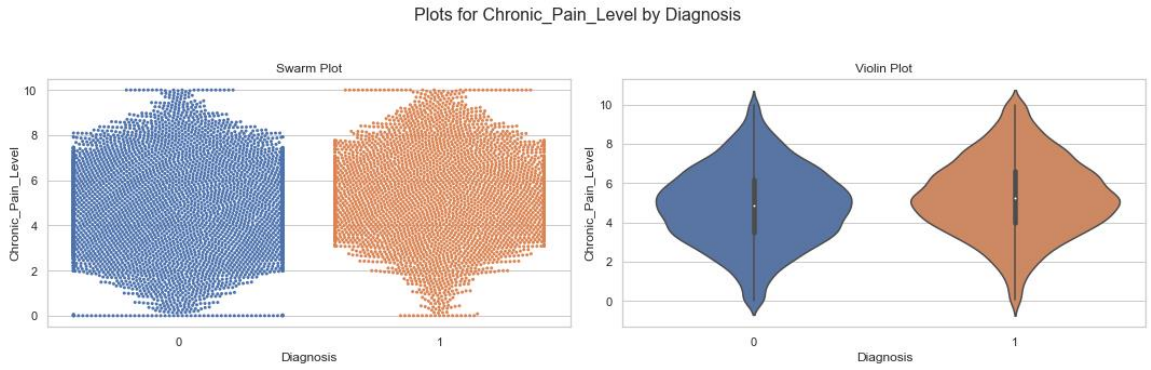

Figure 8 Plots for Chronic_Pain_Level by Diagnosis

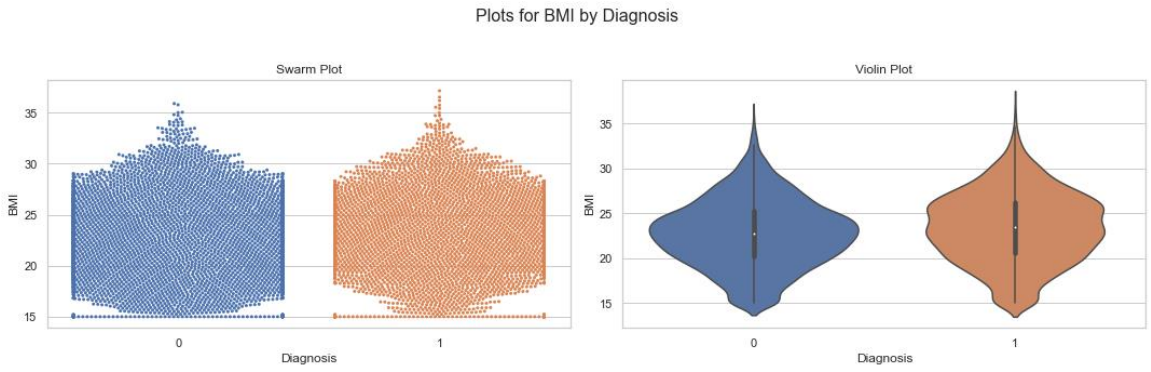

Figure 9 Plots for BMI by Diagnosis

We can see in Figure 10, the age distribution is fairly even, with noticeable peaks between ages 18–20 and 45–48.

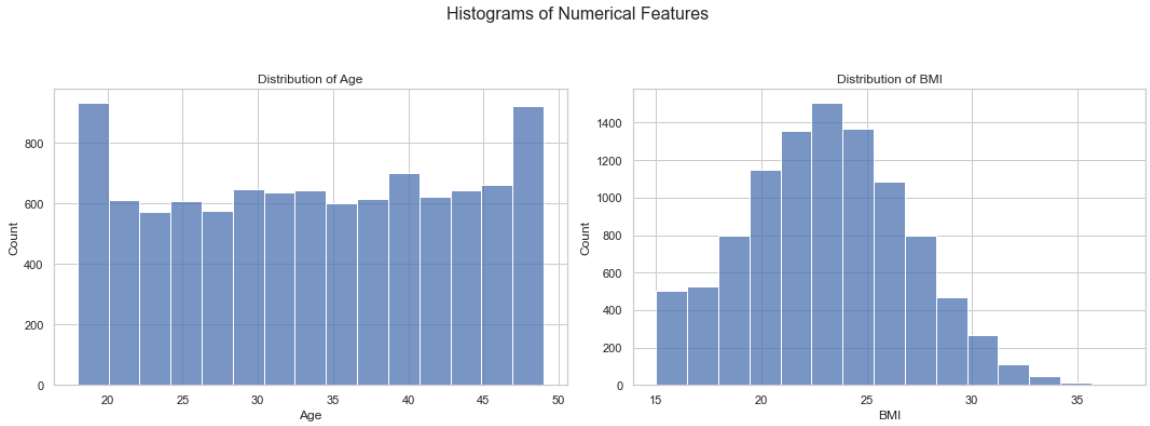

Figure 10 Histogram of Age and BMI

The box plots in Figure 11 below show how **Age**, **Chronic Pain Level**, and **BMI** are distributed across the two diagnosis groups (0 = no diagnosis, 1 = diagnosed).

- **Age**: The median age is very close in both groups, and the spread is similar. This supports the earlier finding that age does not have a strong effect on diagnosis in this dataset.
- **Chronic Pain Level**: The median pain level is slightly higher in the diagnosed group, and the overall range is a bit wider. This aligns with the clinical expectation that chronic pain is more common in those with endometriosis.
- **BMI**: The diagnosed group shows a slightly higher median BMI and more upper outliers. Most values still fall in the healthy-to-overweight range, which fits the earlier observation that women with endometriosis may struggle with weight, possibly due to insulin resistance.

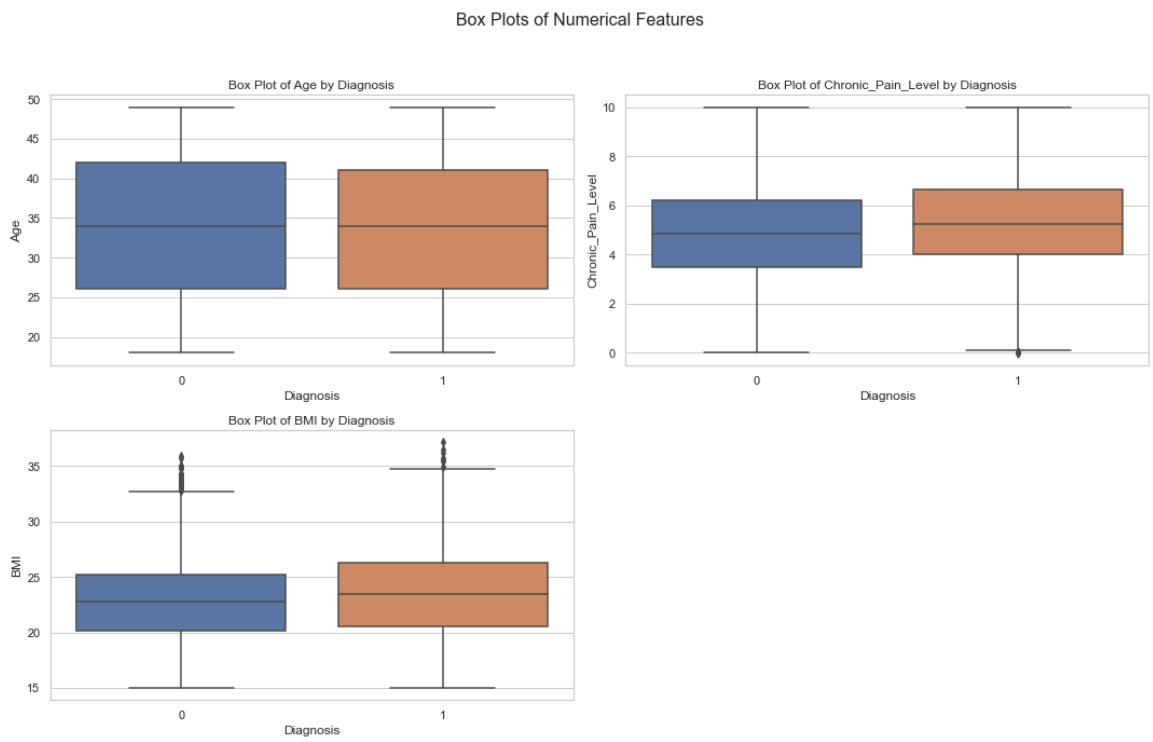

Figure 11 Box Plots of Numerical Features for Endometriosis Dataset

The bar plots in Figure 12 shows how **Menstrual Irregularity**, **Hormone Level Abnormality**, and **Infertility** are distributed between the diagnosed and non-diagnosed groups:

Menstrual Irregularity

- **No Irregularity (0):** Most people without menstrual irregularity do **not** have a diagnosis.
- **With Irregularity (1):** The majority of those who have menstrual irregularities are in the **diagnosed** group. This suggests that irregular cycles are more common among those diagnosed, but they also appear in many without a diagnosis.

Hormone Level Abnormality

- **Normal Hormone Levels (0):** People without hormone abnormalities are mostly **not diagnosed**, with fewer diagnosed cases.
- **Abnormal Hormone Levels (1):** A large number of both diagnosed and undiagnosed individuals show hormone abnormalities, but **more diagnosed people have abnormal hormone levels** than those with normal levels, supporting its relevance as a symptom.

Infertility

- **No Infertility (0):** Most people without infertility are **not diagnosed**, indicating that fertility problems are less common in the undiagnosed group.
- **With Infertility (1):** The counts are close between diagnosed and undiagnosed, but there are **slightly more diagnosed cases** with infertility, pointing to a modest association with diagnosis.

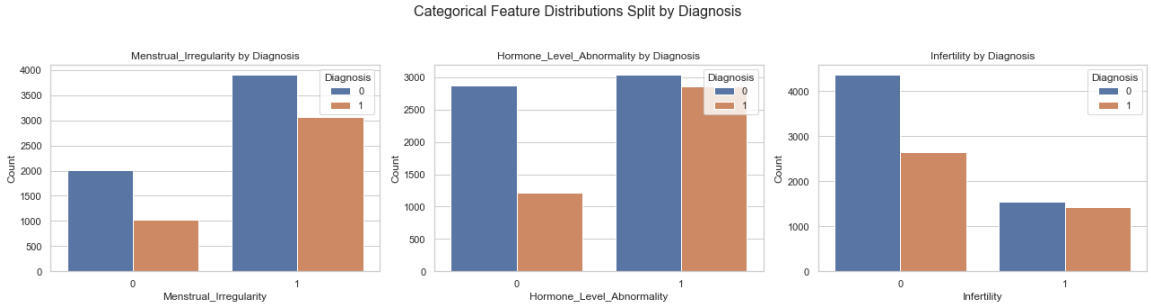

Figure 12 Categorical Feature Distributions Split by Diagnosis for Endometriosis Dataset

1.3 Data Analysis

To better understand which features are related to the diagnosis, I ran both correlation and statistical tests:

```
Pearson Correlation with Diagnosis (assuming Diagnosis is numerical or can be treated as such for correlation):
  Age vs. Diagnosis: correlation = -0.0116, p-value = 0.2478
  Chronic_Pain_Level vs. Diagnosis: correlation = 0.1170, p-value = 0.0000
  Hormone_Level_Abnormality vs. Diagnosis: correlation = 0.1870, p-value = 0.0000
  BMI vs. Diagnosis: correlation = 0.0803, p-value = 0.0000

Statistical Tests:
Chi-squared test for Menstrual_Irregularity vs. Diagnosis:
  p-value: 2.149337936523963e-21
Chi-squared test for Infertility vs. Diagnosis:
  p-value: 8.396255071581354e-22

T-test for Chronic_Pain_Level vs. Diagnosis:
  p-value: 8.09365184751324e-32

T-test for BMI vs. Diagnosis:
  p-value: 8.764850160648458e-16
```

Figure13 Correlation and Statistical Test for Endometriosis Dataset

Pearson Correlation

- **Age** shows almost no correlation with diagnosis (**-0.0116**, $p = 0.24$), meaning age does not have a meaningful linear relationship here.
- **Chronic Pain Level** has a weak positive correlation (**0.117**, $p < 0.001$).

- **Hormone Level Abnormality** shows the highest correlation (**0.187**, $p < 0.001$), suggesting a slightly stronger relationship with diagnosis.
- **BMI** also shows a weak correlation (**0.0803**, $p < 0.001$).

Statistical Tests

- **Menstrual Irregularity** and **Infertility** both have extremely small p-values from the **Chi-squared test** (≈ 2e-21 and 8e-22), confirming that they are **significantly associated with diagnosis**, even if their correlation values are low.
- **T-tests** also show that both **Chronic Pain Level** and **BMI** differ significantly between diagnosed and non-diagnosed groups ($p < 0.001$ for both), which supports their importance in further analysis.

Even though most correlations are weak, the p-values show that features like **hormone abnormality**, **chronic pain**, **infertility**, and **menstrual irregularity** are statistically important and should not be ignored.

1.4 Model Training

For model training, I followed a step-by-step approach for all three datasets. I started by training individual models, including **K-Nearest Neighbors (KNN)**, **Decision Tree**, **Random Forest**, and **Gradient Boosting**, all using **10-fold cross-validation**.

Next, when the data was imbalanced, I retrained the models using **under-sampling or oversampling techniques** to improve performance.

After that, I explored **feature selection** using **SelectKBest**, and finally, I built an **ensemble model using soft voting**, combining the best-performing classifiers.

1.4.1 Model Training - with 10-fold cross-validation only

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
KNeighborsClassifier	0.5912	0.4987	0.4336	0.4638	0.5932
DecisionTreeClassifier	0.5525	0.4521	0.4610	0.4564	0.5380
RandomForestClassifier	0.5937	0.5026	0.4018	0.4462	0.6040
GradientBoostingClassifier	0.6252	0.5661	0.3473	0.4301	0.6505

1.4.2 Model Training - with 10-fold cross-validation only with balanced data

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
KNeighborsClassifier	0.5682	0.4755	0.5735	0.5197	0.5894
DecisionTreeClassifier	0.5320	0.4385	0.5281	0.4789	0.5313
RandomForestClassifier	0.5678	0.4751	0.5712	0.5184	0.5987
GradientBoostingClassifier	0.6042	0.5115	0.6408	0.5687	0.6480

1.4.3 Model Training - with 10-fold cross-validation And SelectKBest K=4

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Voting Ensemble	0.637056	0.628169	0.671848	0.649167	0.687918
Gradient Boosting	0.624893	0.611787	0.683333	0.645502	0.673736
Random Forest	0.624640	0.623831	0.627428	0.625501	0.676288
KNN	0.613157	0.604755	0.653270	0.628011	0.645747
Decision Tree	0.573635	0.572470	0.581993	0.577053	0.573636

1.4.4 Model Training - with 10-fold cross-validation And SelectKBest K=5

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Voting Ensemble	0.645669	0.635566	0.683333	0.658451	0.702891
Random Forest	0.641278	0.638809	0.650396	0.644442	0.697887
KNN	0.624979	0.613353	0.676576	0.643383	0.659452
Gradient Boosting	0.620840	0.609693	0.672189	0.639222	0.675408
Decision Tree	0.594494	0.592876	0.603614	0.598032	0.594493

1.4.5 Model Training - with 10-fold cross-validation And SelectKBest K=6

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Voting Ensemble	0.660191	0.648049	0.700895	0.673318	0.717299
Random Forest	0.652424	0.649130	0.663909	0.656336	0.714710
KNN	0.638911	0.625128	0.694478	0.657908	0.675819
Gradient Boosting	0.627341	0.619874	0.658169	0.638363	0.678630
Decision Tree	0.602348	0.599739	0.615604	0.607489	0.602346

In **Figure 14**, I trained the model using **10-fold cross-validation** with **SelectKBest applied to all features**. The results show that **Hormone Level Abnormality** is the most important feature, followed by **BMI** and **Menstrual Irregularity**. These are all clinically relevant and make sense as key indicators for diagnosing endometriosis.

	Feature	Score
3	Hormone_Level_Abnormality	0.024976
5	BMI	0.008848
1	Menstrual_Irregularity	0.004289
2	Chronic_Pain_Level	0.003605
0	Age	0.003223
4	Infertility	0.002277

Figure 14 Feature Importance for endometriosis dataset

1.5 Model Evaluation

As seen in **Figures 15 to 19**, the results can be interpreted as follows:

Overall Observations First:

1. **Impact of Balancing (results_10_Fold vs. results_10_Fold_Balanced):**
    - For Gradient Boosting, balancing significantly improved its Recall (0.3473 to 0.6408) and F1 Score (0.4301 to 0.5687), making it a much more viable model. Other models also saw shifts, generally towards better Recall.
2. **Impact of Feature Selection (results_k4, results_k5, results_k6):**

- Generally, selecting a subset of features (k=4, 5, or 6) led to better performance across all metrics compared to the 10-Fold cross-validation on all features (even the balanced version).
- The **Voting Ensemble** model appears in these feature-selected results and often performs very well.

Interpreting the "Best" Model based on Different Criteria:

1. results_k4 (4 features selected):

- **Voting Ensemble:** Best Accuracy (0.6370), Precision (0.6281), F1 Score (0.6491), ROC AUC (0.6879).
- **Gradient Boosting:** Best Recall (0.6833). Very competitive F1 (0.6455) and ROC AUC (0.6737).
- **Conclusion** for k=4: Voting Ensemble offers the best overall balance, but for maximizing recall, Gradient Boosting is better.

2. results_k5 (5 features selected):

- **Voting Ensemble:** Best Accuracy (0.6456), Recall (0.6833 - tied with GB in k4, better here), F1 Score (0.6584), ROC AUC (0.7028).
- **Random Forest:** Best Precision (0.6388). Very competitive Accuracy (0.6412), F1 (0.6444), and ROC AUC (0.6978).
- **Conclusion** for k=5: Voting Ensemble is again very strong, leading in most key metrics. Random Forest is a close competitor, especially on precision.

3.results_k6 (6 features selected):

- **Voting Ensemble:** Best Accuracy (0.6601), Recall (0.7008), F1 Score (0.6733), ROC AUC (0.7172).
- **Random Forest:** Best Precision (0.6491). Very competitive Accuracy (0.6524), F1 (0.6563), and ROC AUC (0.7147).
- **Conclusion** for k=6: This set shows the highest metrics overall. The **Voting Ensemble** is the top performer across Accuracy, Recall, F1-score, and ROC AUC. **Random Forest** is a very strong second, particularly excelling in Precision and being very close on ROC AUC and F1

Conclusion :The **Voting Ensemble using 6 selected features** demonstrates the best overall performance, particularly in Recall, F1-score, and ROC AUC, which are critical for a reliable diagnostic model. Random Forest with 6 features is also a highly competitive alternative, especially for maximizing precision.

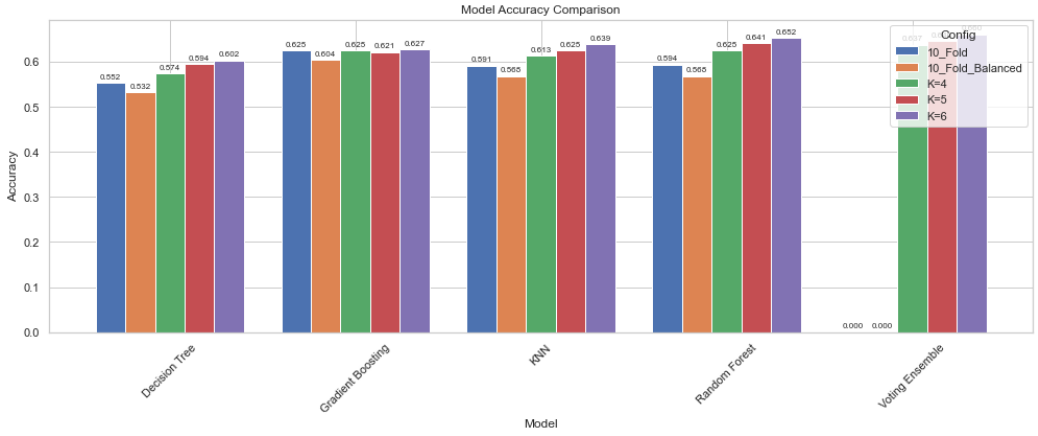

Figure 15 Accuracy Comparison For Endometriosis Dataset

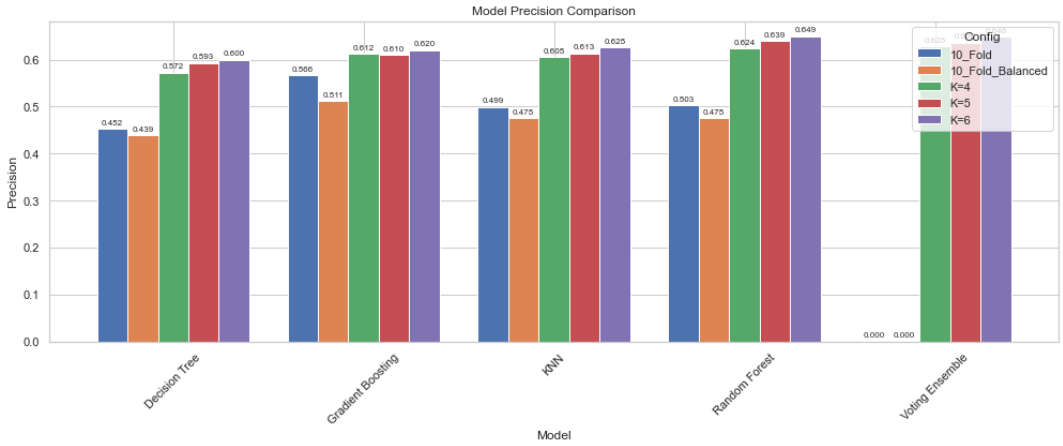

Figure 16 Precision Comparison For Endometriosis Dataset

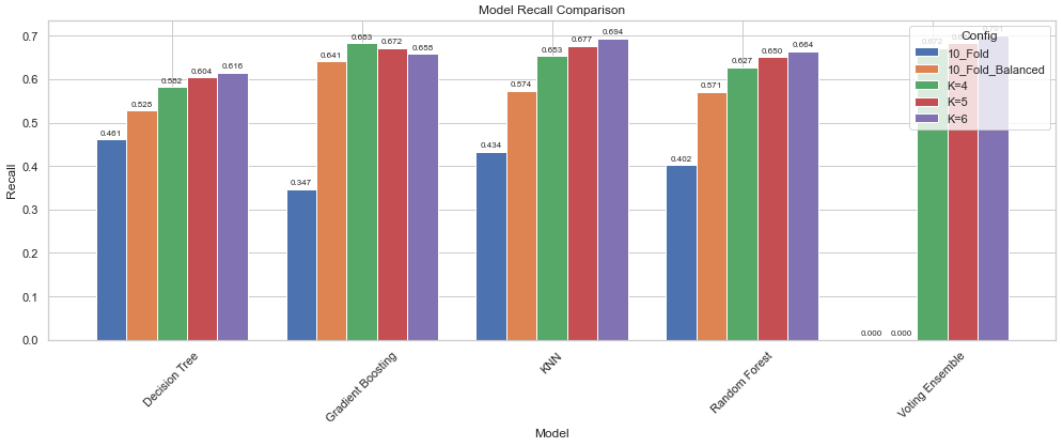

Figure 17 Recall Comparison For Endometriosis Dataset

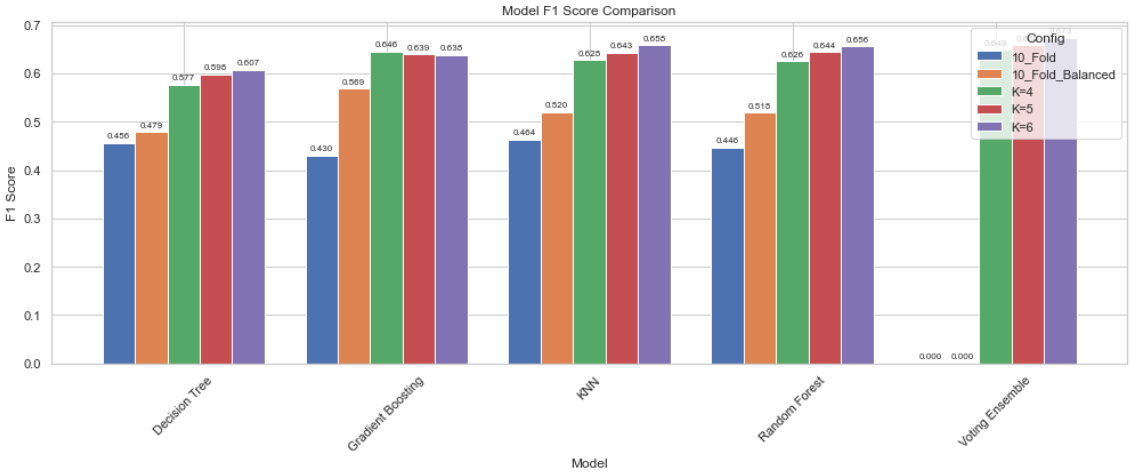

Figure 18 F1Score Comparison For Endometriosis Dataset

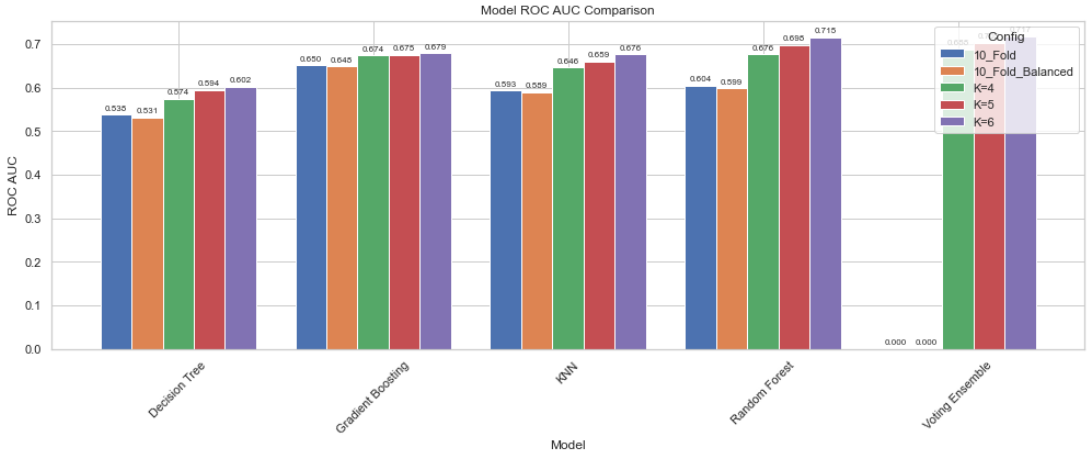

Figure 19 Roc AUC Comparison For Endometriosis Dataset

2. POCS Dataset

2.1 Data Exploration

The dataset contains **4 features and 1 target feature**, which is **PCOS_Diagnosis**. All records are complete, there are **no missing values or duplicates**..
Menstrual_Irregularity and **PCOS_Diagnosis** are **categorical** and already encoded as **0 or 1**. **Age** ranges from **18 to 45**, and **BMI** ranges from **18.1 to 35**.
Testosterone_Level(ng/dL) ranges between **20 and 99.8**, and **Antral_Follicle_Count** spans from **5 to 29**. You can refer to the notebook **titled *Data Mining PCOS*** for full exploratory data analysis. The target feature is unbalanced, as you can see in Figure 23. Therefore, I applied techniques like **oversampling** using SMOTE.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 6 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Age                         1000 non-null   int64
 1   BMI                         1000 non-null   float64
 2   Menstrual_Irregularity      1000 non-null   int64
 3   Testosterone_Level(ng/dL)   1000 non-null   float64
 4   Antral_Follicle_Count       1000 non-null   int64
 5   PCOS_Diagnosis              1000 non-null   int64
dtypes: float64(2), int64(4)
memory usage: 47.0 KB
None
```

Figure 20 Data exploration for POCS Dataset

	Age	BMI	Menstrual_Irregularity	Testosterone_Level(ng/dL)	Antral_Follicle_Count	PCOS_Diagnosis
count	1000.000000	1000.00000	1000.000000	1000.000000	1000.000000	1000.000000
mean	31.771000	26.38700	0.530000	60.159500	17.469000	0.199000
std	8.463462	4.93554	0.499349	23.160204	7.069301	0.399448
min	18.000000	18.10000	0.000000	20.000000	5.000000	0.000000
25%	24.000000	21.90000	0.000000	41.700000	12.000000	0.000000
50%	32.000000	26.40000	1.000000	60.000000	18.000000	0.000000
75%	39.000000	30.50000	1.000000	80.300000	23.250000	0.000000
max	45.000000	35.00000	1.000000	99.800000	29.000000	1.000000

Figure 21 Descriptive Statistics for POCS Dataset

```
Unique Values Count:
 Age                          28
BMI                          170
Menstrual_Irregularity         2
Testosterone_Level(ng/dL)    577
Antral_Follicle_Count         25
PCOS_Diagnosis                 2
dtype: int64
```

Figure 22 Unique Values Count for POCS Dataset

```
Target Variable Distribution:
0    0.801
1    0.199
Name: PCOS_Diagnosis, dtype: float64
```

Figure 23 Target Variable Distribution for POCS Dataset

2.2 Data visualization

Correlation Analysis

The heatmap and correlation below in Figures 24 and 26, none of the features show a strong linear relationship with the PCOS_Diagnosis target. The most important observations are:

- **Menstrual Irregularity** has the strongest positive correlation with diagnosis (**0.47**), which is expected since irregular periods are one of the most common symptoms of PCOS.
- **BMI** also shows a moderate positive correlation (**0.38**), suggesting that weight may play a role in diagnosis, This aligns with known links between PCOS and metabolic issues.
- **Testosterone Level** (**0.20**) and **Antral Follicle Count** (**0.19**) show weaker, but still relevant, positive correlations with diagnosis.
- **Age** has a very weak negative correlation (**-0.06**), meaning it does not affect the diagnosis in this dataset.

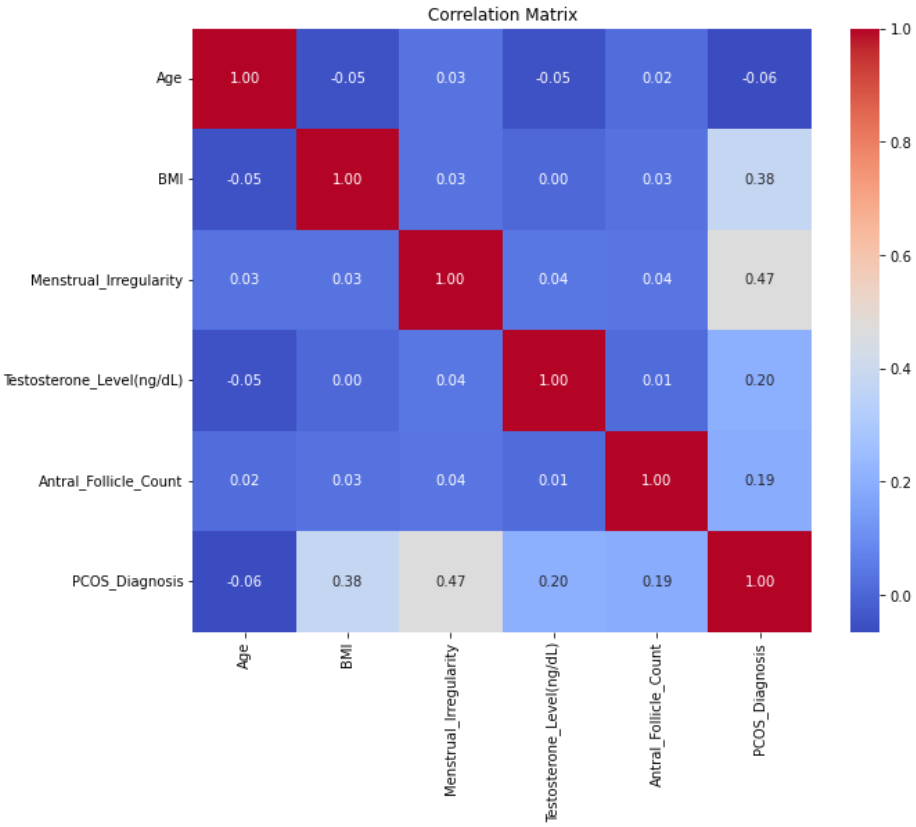

Figure 24 Correlation Matrix for POCS Dataset

```
Correlation with Diagnosis:
 PCOS_Diagnosis               1.000000
Menstrual_Irregularity        0.469376
BMI                           0.377852
Testosterone_Level(ng/dL)     0.200817
Antral_Follicle_Count         0.192014
Age                          -0.064675
Name: PCOS_Diagnosis, dtype: float64
```

Figure 25 Correlation with POCS for Endometriosis Dataset

These KDE plots below in Figure 26 show how each feature is distributed across PCOS Positive and Negative cases:

- **Age**: The distribution is similar for both groups, with no major difference, supporting earlier findings that age has little correlation with diagnosis.
- **BMI**: Women with PCOS tend to have slightly higher BMI values. This supports the moderate correlation found earlier and is consistent with known links between PCOS and weight gain/metabolic issues.

- **Testosterone Level (ng/dL)**: PCOS Positive cases show a shift toward higher testosterone levels, which is expected, as elevated androgen levels are a diagnostic marker for PCOS.
- **Antral Follicle Count**: The PCOS group shows a wider spread and slightly higher follicle counts, which fits clinical criteria used in PCOS diagnosis.

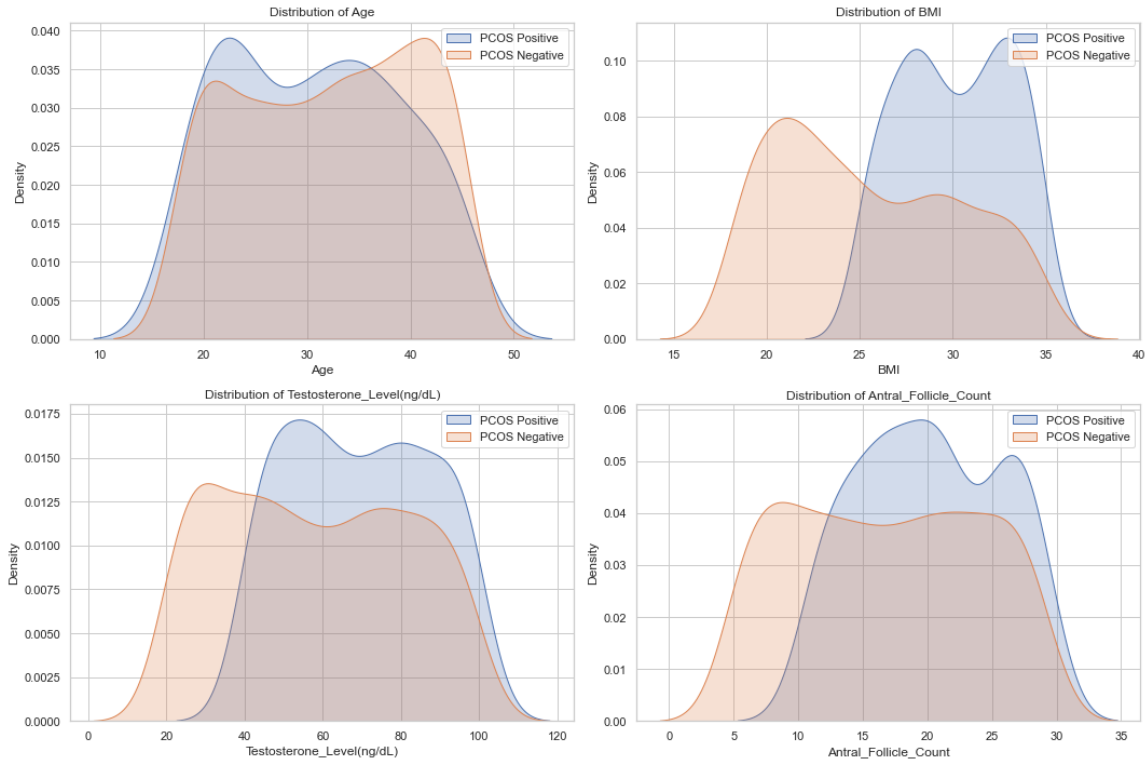

Figure 26 KDE Distribution For POCS Features

The histograms below in Figure 27 show the overall distribution of **Testosterone Level**, **Antral Follicle Count**, and **BMI** across the entire dataset (both PCOS-positive and negative cases combined). It confirmed to finding above and shows that distribution is quite spread out and roughly uniform.

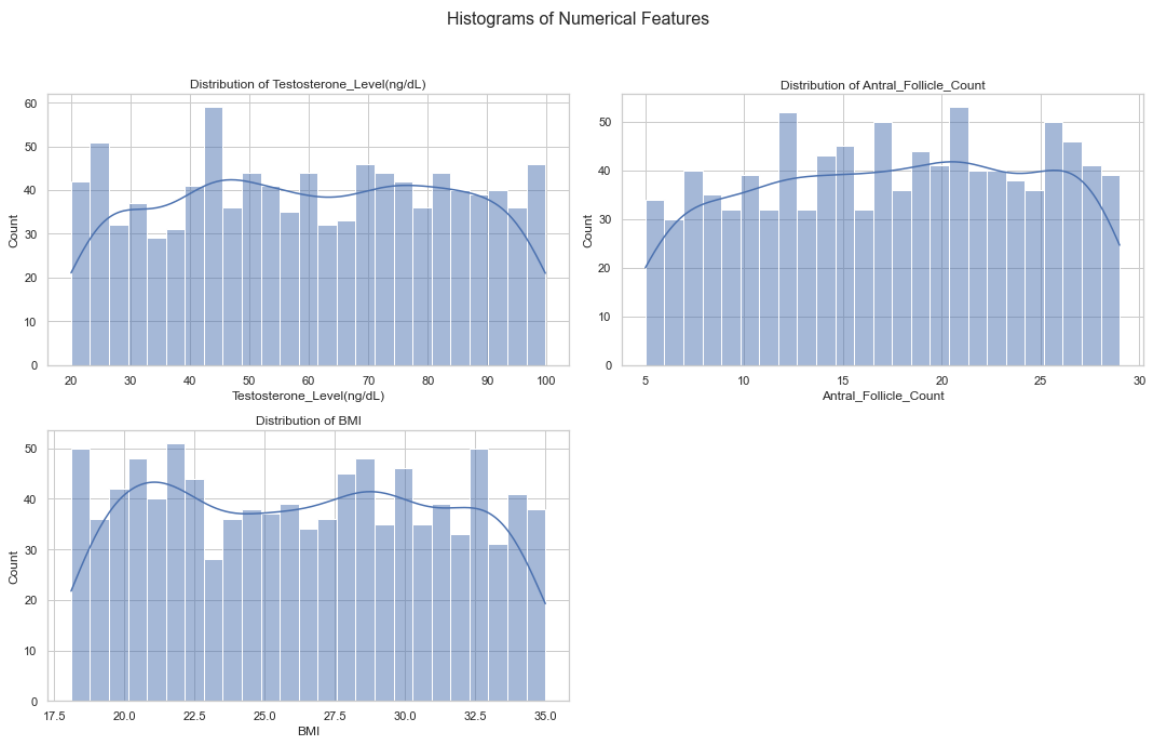

Figure 27 KDE Distribution For POCS Numerical Features

The swarm and violin plots in Figures 28 to 30 provide a more focused comparison between PCOS-positive and PCOS-negative cases by visualizing both individual data points and the full distribution shape for each numerical feature. For **Testosterone Level**, the plots show that PCOS cases are concentrated at higher values (around 70–100), while non-PCOS values are more spread out. **Antral Follicle Count** also shows a denser cluster around 15–25 in the PCOS group, which aligns with clinical expectations. For **BMI**, both plots confirm that PCOS cases tend to fall in the higher range (27–34), reflecting overweight or obese categories. Compared to previous KDE and histogram plots, these visuals give more detailed insight, especially by showing **how concentrated and separated the values are across diagnoses**.

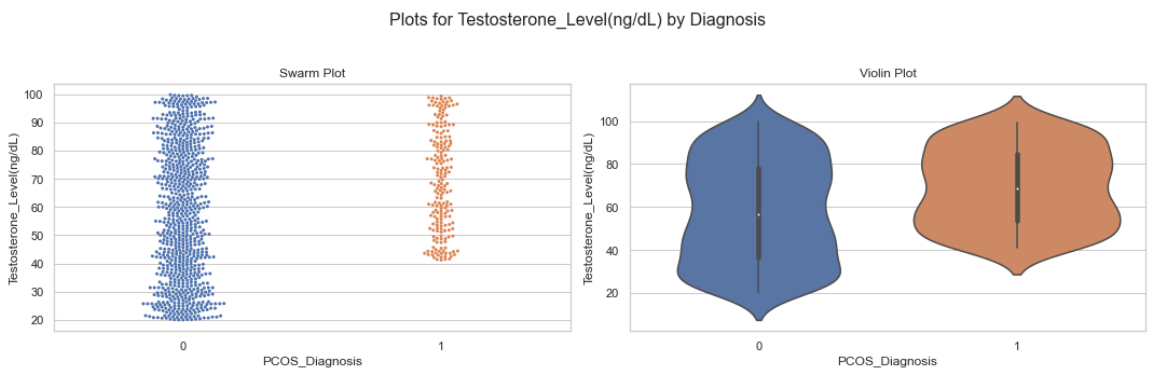

Figure 28 Plots for Testosterone_Level(ng/dL) by Diagnosis

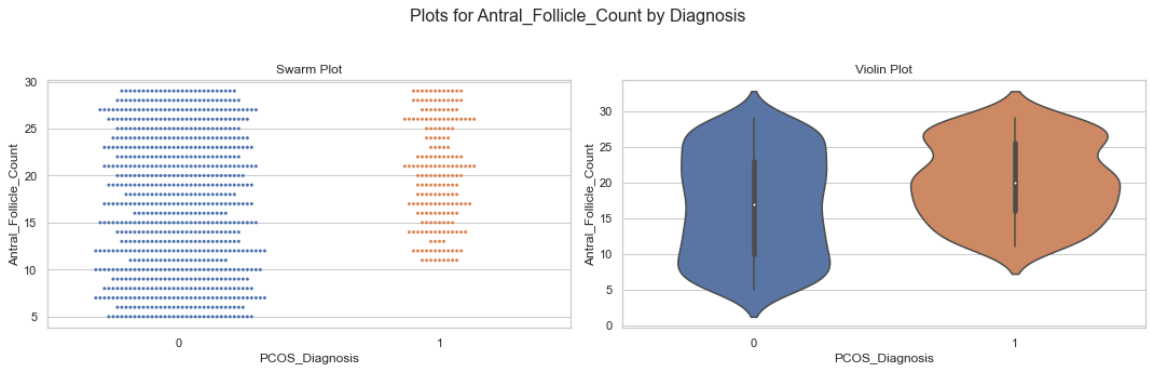

Figure 29 Plots for Antral_Follicle_Count by Diagnosis

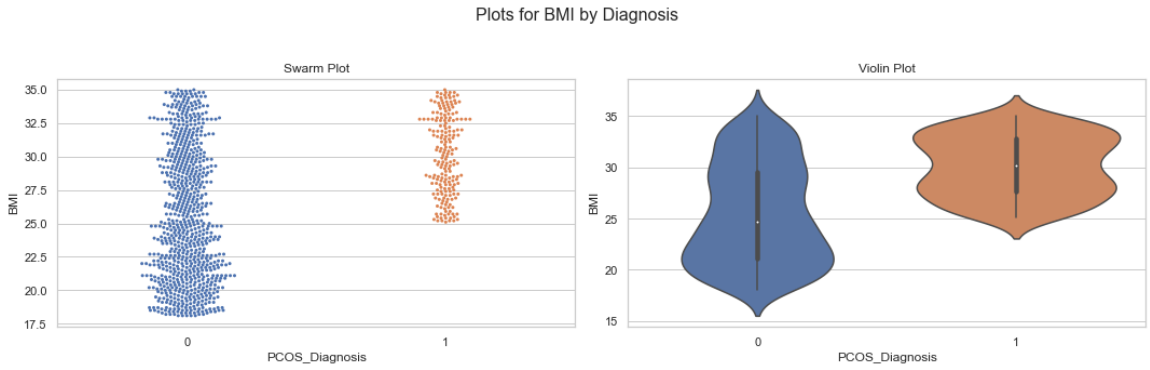

Figure 30 Plots for BMI by Diagnosis

The box plots shown in Figure 31 provide a clear summary of how each numerical feature differs between PCOS-positive and PCOS-negative groups. **Testosterone Level** shows the most noticeable separation; PCOS-positive cases have a higher median of around **68 ng/dL**, compared to **56 ng/dL** for the negative group, with a tighter interquartile range (IQR), reinforcing its diagnostic importance. **BMI** also shows a clear shift: the median BMI for PCOS cases is around **30**, while for non-PCOS cases it is approximately **24**, with the upper quartile in PCOS reaching close to **34**. This aligns with the known association between PCOS and weight issues. **Antral Follicle Count** is

slightly higher in the PCOS group, with a median around **20**, compared to **17** in the negative group, showing more concentration in the diagnostic range. In contrast, **Age** appears similar across both groups, with a slightly lower median in PCOS cases (around **30** vs. **33**), but no strong separation. These plots confirm earlier findings from KDE, violin, and swarm plots, while adding clarity by highlighting **distribution spread, medians, and potential outliers**, all of which help validate feature importance for modeling.

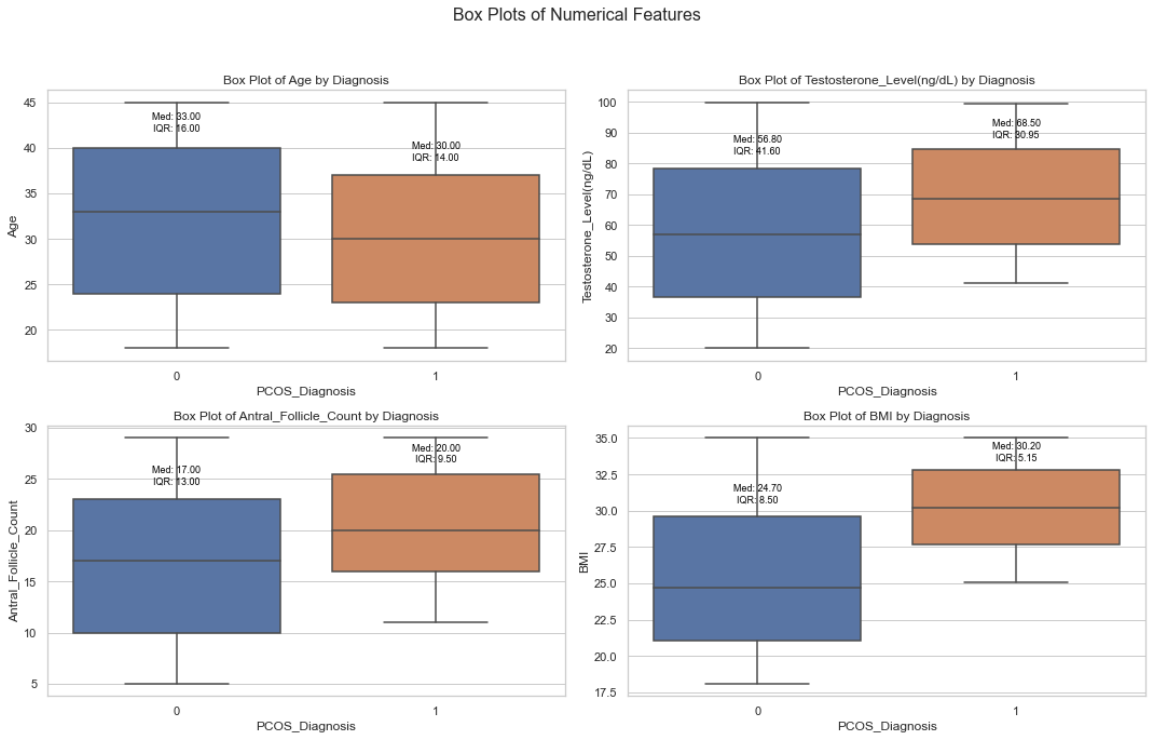

Figure 31 Box Plots of Numerical Features

This bar plot shows how **Menstrual Irregularity** is distributed across PCOS diagnosis groups. Most of the cases without irregularity (**value = 0**) belong to the **non-PCOS group**, with around **460 cases**. On the other hand, among those who do have irregular periods (**value = 1**), the PCOS group makes around **200 cases out of 540**. This supports the clinical understanding that **menstrual irregularity is one of the most common**

symptoms in PCOS and is more frequent among diagnosed cases.

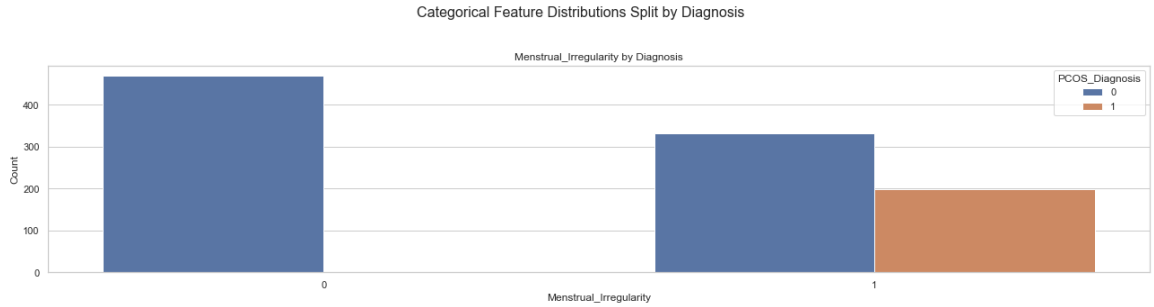

Figure 32 Categorical Feature Distributions Split by Diagnosis

2.3 Data Analysis

This statistical summary below in Figure 33 helps confirm which features are most important for predicting PCOS. The **Pearson correlation** shows that **BMI (0.378)** has the strongest linear relationship with diagnosis, followed by **Testosterone Level (0.200)** and **Antral Follicle Count (0.192),** all positive and statistically significant (**p < 0.001**). **Age** has a weak negative correlation (**-0.065**) and a marginal p-value (**0.0409**), suggesting it is less reliable as a predictive feature, as we already know that before. The **Chi-Squared test** for the categorical feature **Menstrual Irregularity** also shows a **very strong association with PCOS (p = 0.0000).** Finally, **independent t-tests** confirm that the distributions of all numerical features(**Age, BMI, Testosterone Level, and Antral Follicle Count**)are significantly different between PCOS and non-PCOS groups.

```
Pearson Correlation with Diagnosis:
  Age vs. Diagnosis: correlation = -0.0647, p-value = 0.0409
  BMI vs. Diagnosis: correlation = 0.3779, p-value = 0.0000
  Testosterone_Level(ng/dL) vs. Diagnosis: correlation = 0.2008, p-value = 0.0000
  Antral_Follicle_Count vs. Diagnosis: correlation = 0.1920, p-value = 0.0000

Chi-Squared Tests for Categorical Features:
  Menstrual_Irregularity vs. Diagnosis: p-value = 0.0000

 Independent T-Tests for Numerical Features:
  Age vs. Diagnosis: p-value = 0.0409
  BMI vs. Diagnosis: p-value = 0.0000
  Testosterone_Level(ng/dL) vs. Diagnosis: p-value = 0.0000
  Antral_Follicle_Count vs. Diagnosis: p-value = 0.0000
```

Figure 33 Correlation and Statistical Test for POCS Dataset

2.4 Model Training

2.4.1 Model Training - with 10-fold cross-validation only and without balancing the y target

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
KNeighborsClassifier	0.961	0.896357	0.911240	0.902446	0.986023
DecisionTreeClassifier	0.999	1.000000	0.995652	0.997778	0.997826
RandomForestClassifier	0.999	1.000000	0.995652	0.997778	1.000000
GradientBoostingClassifier	0.999	1.000000	0.995652	0.997778	0.998899

2.4.1 Model Training - with 10-fold cross-validation only and with balancing the y target

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Gradient Boosting	0.999379	1.00000	0.99875	0.999371	0.999985
Random Forest	0.998129	1.00000	0.99625	0.998105	1.000000
Decision Tree	0.996879	1.00000	0.99375	0.996807	0.996875
KNN	0.962535	0.93518	0.99500	0.963933	0.991907

2.4.3 Model Training - with 10-fold cross-validation and SelectKBest K=2

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Gradient Boosting	0.884	0.643716	0.955000	0.767652	0.926106
Voting Ensemble	0.880	0.646607	0.905000	0.751786	0.923100
KNN	0.876	0.645751	0.875000	0.739545	0.921495
Random Forest	0.870	0.641264	0.814211	0.715348	0.921286
Decision Tree	0.848	0.617512	0.634211	0.623641	0.879693

2.4.4 Model Training - with 10-fold cross-validation and SelectKBest K=3

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Voting Ensemble	0.957	0.834076	0.995000	0.904817	0.980718
Random Forest	0.954	0.835845	0.975000	0.896837	0.980061
Gradient Boosting	0.954	0.834756	0.975000	0.896699	0.977016
KNN	0.944	0.796557	0.990000	0.878939	0.971992
Decision Tree	0.938	0.840438	0.864737	0.847742	0.910547

2.4.5 Model Training - with 10-fold cross-validation and SelectKBest K=4

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Decision Tree	0.999	1.000000	0.995	0.997436	0.997500
Random Forest	0.999	1.000000	0.995	0.997436	1.000000
Gradient Boosting	0.999	1.000000	0.995	0.997436	1.000000
Voting Ensemble	0.999	1.000000	0.995	0.997436	0.999875
KNN	0.955	0.830921	0.975	0.896327	0.988753

2.4.6 Model Training - with 10-fold cross-validation and SelectKBest K=5

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Decision Tree	0.999	1.000000	0.995	0.997436	0.997500
Random Forest	0.999	1.000000	0.995	0.997436	1.000000
Gradient Boosting	0.999	1.000000	0.995	0.997436	1.000000
Voting Ensemble	0.999	1.000000	0.995	0.997436	1.000000
KNN	0.949	0.811248	0.985	0.887340	0.986681

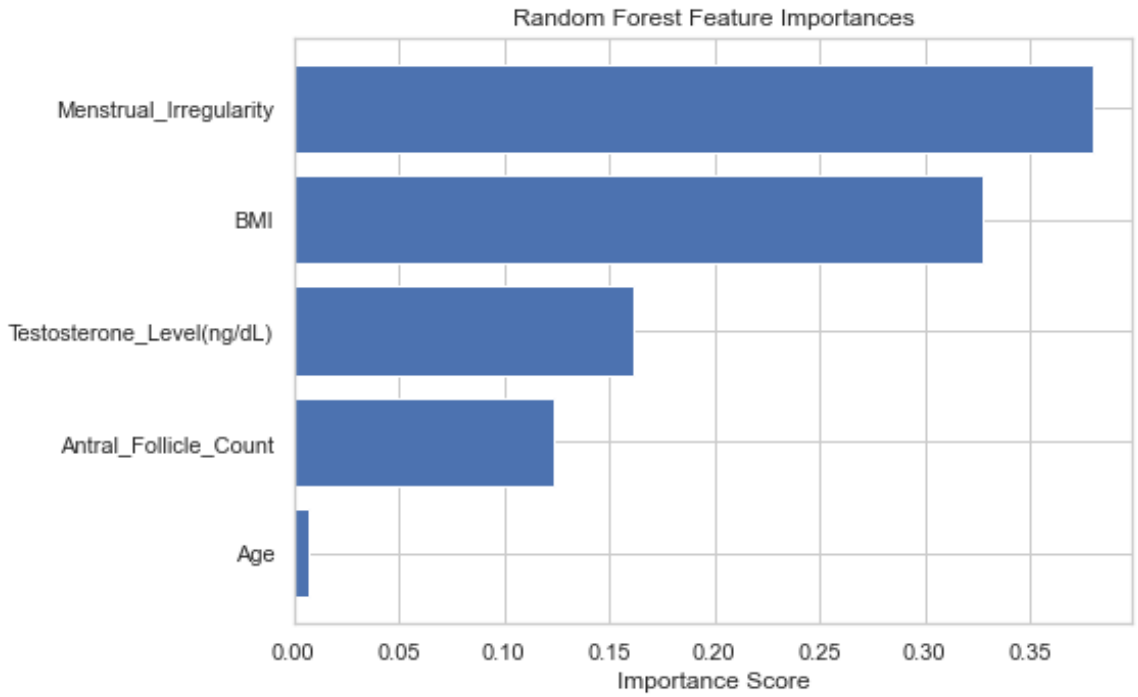

Figure 34 Feature Importance for PCOS Dataset

In **Figure 34**, I trained the model using **10-fold cross-validation** with **SelectKBest applied to all features**. The results show that **Menstrual Irregularity** is the most important feature, which makes sense as it was only shown for the participants who were diagnosed with PCOS, followed by BMI and Testosterone, and Follicle count feature. These are all clinically relevant and make sense as key indicators for diagnosing POCS. And as was proven that before multiple time that Age has zero correlation of effect.

2.5 Model Evaluation

The models achieved **extremely high accuracy**, with most scores above **99%** even without balancing. While these results might seem **too good to be true**, I checked my code for **data leakage,** and it looked fine. I also tried different things, still all implementation gives high accuracy.

Upon checking the dataset, a few characteristics may explain the unusually high performance:

- **PCOS_Diagnosis is highly imbalanced**: Only **199 out of 1000** entries are positive cases (i.e., 20%). Even with me trying to balance out the data with oversampling, as it is a medical dataset, I guess it still did not reflect the real-life complexity.
- The features like **Menstrual Irregularity**, **BMI**, and **Testosterone Level** show **strong signal separation** between PCOS-positive and negative cases (as

confirmed by statistical tests and visualizations). This means the dataset is **inherently easy to classify**, at least based on this sample.

- There are **no missing values or noise**, which makes it ideal for model learning

That said, although the results are statistically valid, they might reflect a **clean, idealized dataset**, rather than a challenging real-world scenario. However, as I did not have any available POCS dataset. I had no other choice but to go with this.

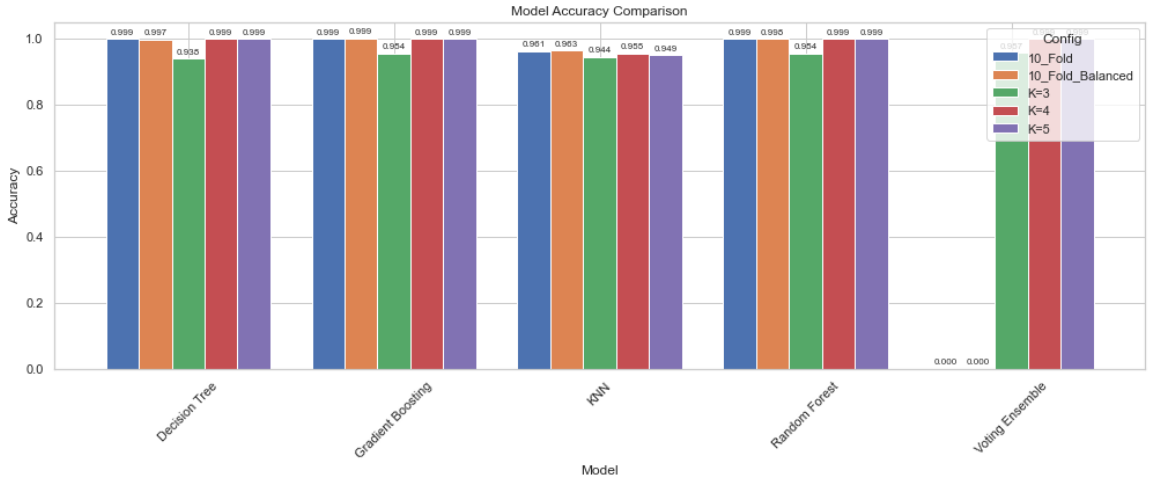

Figure 35: Accuracy Comparison For POCS Dataset

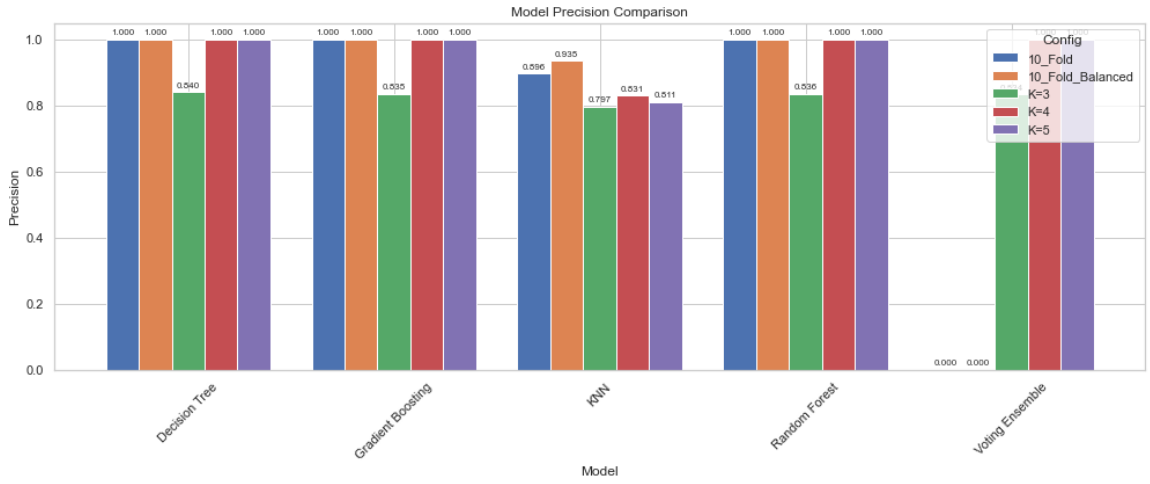

Figure 36 Precision Comparison For POCS Dataset

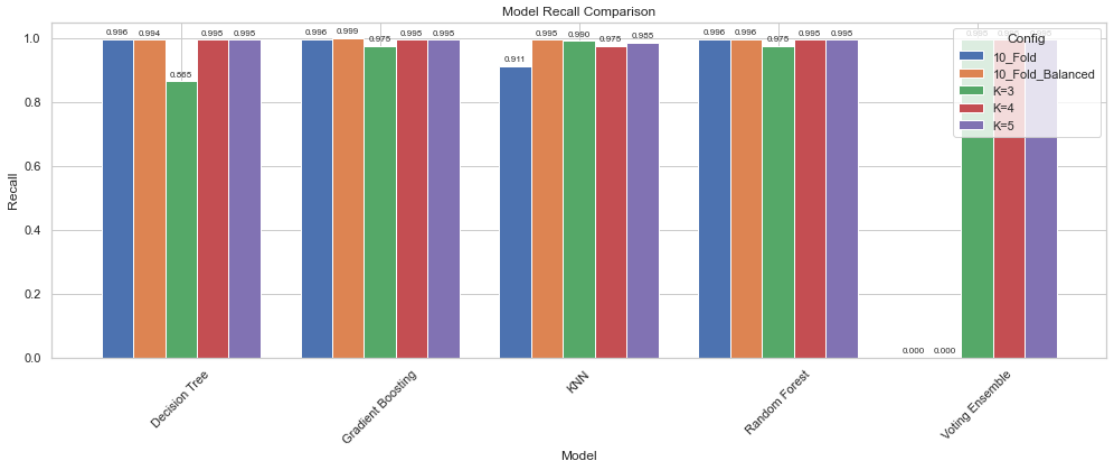

Figure 37 Recall Comparison For POCS Dataset

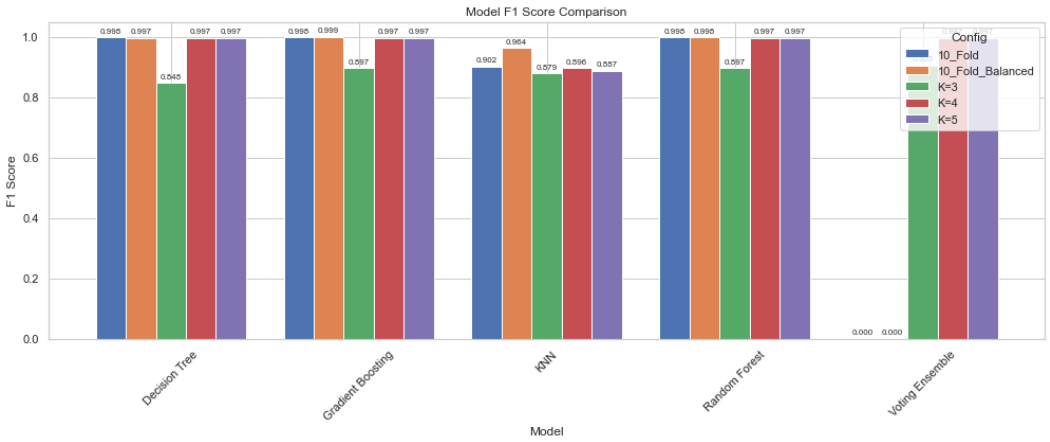

Figure 38 F1Score Comparison For POCS Dataset

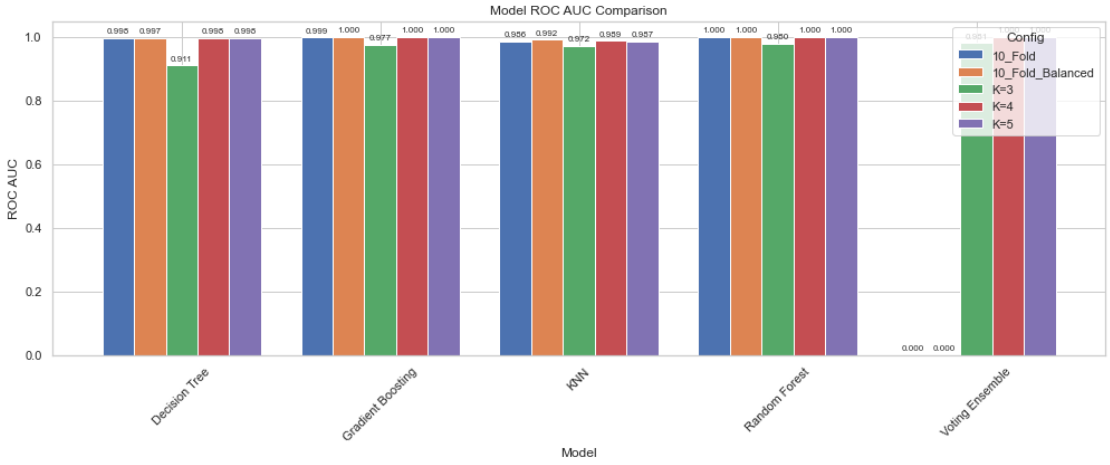

Figure 39 Roc AUC Comparison For POCS Dataset

3. Infertility Dataset

3.1 Data Exploration

The dataset has 11 features and one target feature, which is the **Infertility Prediction,** as you can **see in Figure 40**. All features are binary and already encoded as 0 or 1, including features such as **Ovulation Disorders**, **Blocked Fallopian Tubes**, and **Endometriosis**. The only numerical feature is **Age**, which ranges across 36 unique values, as shown in **Figure 46**. You can check the notebook named **Data Mining Infertility** for full data exploration. There are no null values or duplicates. The dataset contains 705 entries. In training, I dropped Patient_ID column as it does not contribute to the model prediction

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 705 entries, 0 to 704
Data columns (total 13 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Patient ID                      705 non-null    int64
 1   Age                             705 non-null    int64
 2   Ovulation Disorders             705 non-null    int64
 3   Blocked Fallopian Tubes         705 non-null    int64
 4   Endometriosis                   705 non-null    int64
 5   Uterine Abnormalities           705 non-null    int64
 6   Pelvic Inflammatory Disease     705 non-null    int64
 7   Hormonal Imbalances             705 non-null    int64
 8   Premature Ovarian Insufficiency 705 non-null    int64
 9   Autoimmune Disorders            705 non-null    int64
 10  Previous Reproductive Surgeries 705 non-null    int64
 11  Unexplained Infertility         705 non-null    int64
 12  Infertility Prediction          705 non-null    int64
dtypes: int64(13)
memory usage: 71.7 KB
None
```

Figure 40 Data exploration for Infertility Dataset

```
Unique Values Count:
 Patient ID                      705
Age                              36
Ovulation Disorders               2
Blocked Fallopian Tubes           2
Endometriosis                     2
Uterine Abnormalities             2
Pelvic Inflammatory Disease       2
Hormonal Imbalances               2
Premature Ovarian Insufficiency   2
Autoimmune Disorders              2
Previous Reproductive Surgeries   2
Unexplained Infertility           2
Infertility Prediction            2
dtype: int64
```

Figure 46 Unique Values Coun for Infertility Dataset

	Patient ID	Age	Ovulation Disorders	Blocked Fallopian Tubes	Endometriosis	Uterine Abnormalities	Pelvic Inflammatory Disease	Hormonal Imbalances	Premature Ovarian Insufficiency	Autoimmune Disorders	Previous Reproductive Surgeries	U
count	705.000000	705.000000	705.000000	705.000000	705.000000	705.000000	705.000000	705.000000	705.000000	705.000000	705.000000	
mean	353.000000	37.441135	0.648227	0.618440	0.594326	0.595745	0.665248	0.720567	0.641135	0.581560	0.626950	
std	203.660256	9.259041	0.477863	0.486114	0.491371	0.491096	0.472239	0.449039	0.480008	0.493653	0.483958	
min	1.000000	25.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	177.000000	30.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	353.000000	35.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
75%	529.000000	44.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
max	705.000000	60.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

Figure 41 Descriptive Statistics for Infertility Dataset

	Value	Frequency	Feature
0	1	457	Ovulation Disorders
1	0	248	Ovulation Disorders
2	1	436	Blocked Fallopian Tubes
3	0	269	Blocked Fallopian Tubes
4	1	419	Endometriosis
5	0	286	Endometriosis
6	1	420	Uterine Abnormalities
7	0	285	Uterine Abnormalities
8	1	469	Pelvic Inflammatory Disease
9	0	236	Pelvic Inflammatory Disease
10	1	508	Hormonal Imbalances
11	0	197	Hormonal Imbalances
12	1	452	Premature Ovarian Insufficiency
13	0	253	Premature Ovarian Insufficiency
14	1	410	Autoimmune Disorders
15	0	295	Autoimmune Disorders
16	1	442	Previous Reproductive Surgeries
17	0	263	Previous Reproductive Surgeries
18	1	368	Unexplained Infertility
19	0	337	Unexplained Infertility
20	1	579	Infertility Prediction
21	0	126	Infertility Prediction

Figure 42 Variable Distribution for Infertility Dataset

3.2 Data Visualization

In the correlation matrix shown in **Figure 43** and **Figure 44**, none of the features demonstrates a strong linear relationship with the target variable **Infertility Prediction**. The highest correlation is observed with:

- **Unexplained Infertility (0.41)**
- **Previous Reproductive Surgeries (0.38)**
- **Ovulation Disorders (0.21)**
- **Premature Ovarian Insufficiency (0.15)**

All these correlations are weak to moderate in strength but **clinically relevant**, especially considering their known roles in reproductive health.

Features like **Hormonal Imbalances (0.07)**, **Endometriosis (0.04)**, and **Pelvic Inflammatory Disease (0.15)** show only **very weak correlations**, and **Uterine Abnormalities** and **Blocked Fallopian Tubes** even show **slightly negative or negligible correlations**. Furthermore, **Age** (0.07) and **Patient ID** (0.11) have low or irrelevant statistical correlation, as expected.

As a result, the dataset exhibits **little to no multicollinearity** among the features, implying that these conditions contribute **independently to infertility** or are possibly recorded in a way that does not reflect underlying clinical interactions. This may justify using ensemble or tree-based models that handle weakly correlated and potentially nonlinear relationships better.

To gain a broader understanding of endometriosis and PCOS within the context of female infertility, I analyzed how their related features behave in the infertility dataset, as shown in **Figure 46.**

Endometriosis shows very weak correlations with most variables. The highest are with **Pelvic Inflammatory Disease (0.20) and Autoimmune Disorders (0.07),** while features **like Premature Ovarian Insufficiency (–0.27) and Uterine Abnormalities (–0.16)** are negatively correlated. This suggests that symptoms of endometriosis tend to appear independently in this dataset, reflecting its often isolated and underdiagnosed nature.

Ovulation Disorders, used here as a proxy for PCOS, shows slightly more connected patterns. It correlates moderately with **Unexplained Infertility (0.27) and Infertility Prediction (0.21)**, with additional weak relationships **to Hormonal Imbalances and Pelvic Inflammatory Disease (both around 0.08)**. These results are consistent with PCOS presenting across multiple weak but clinically relevant features.

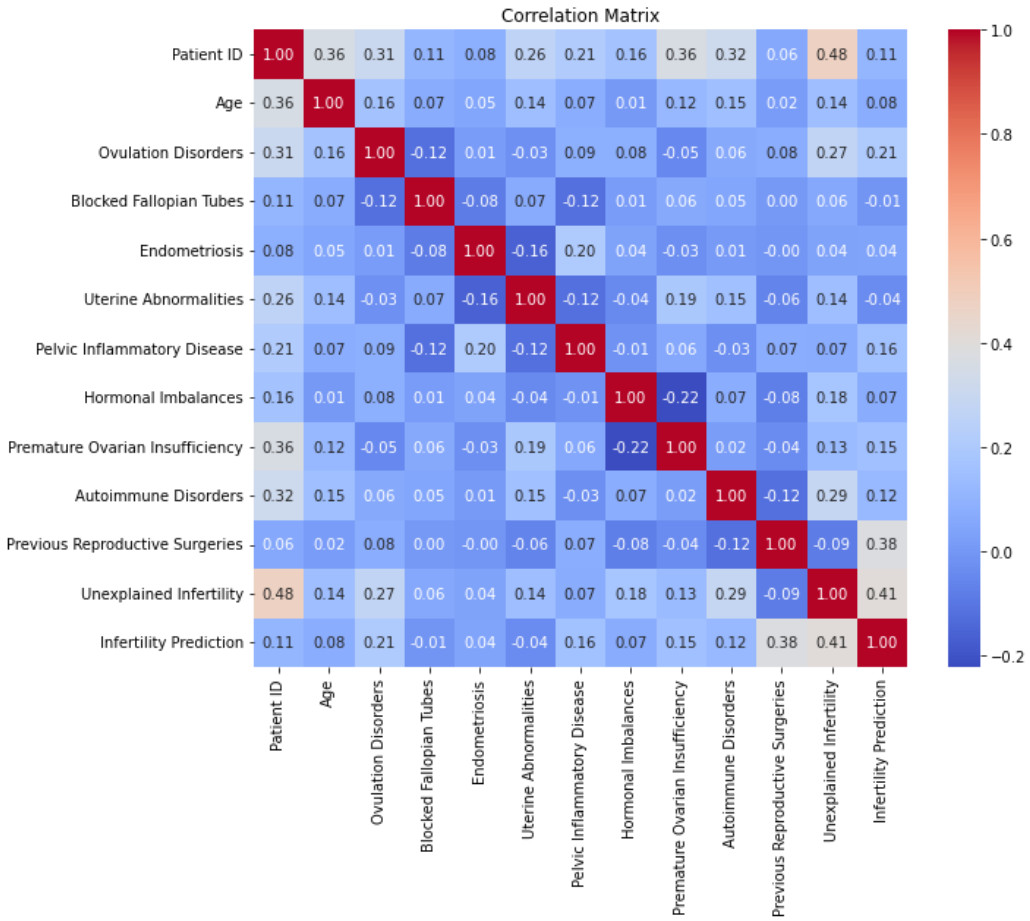

Figure 44 Correlation Matrix for Infertility Dataset

```
Correlation with Diagnosis:
 Infertility Prediction            1.000000
Unexplained Infertility            0.405948
Previous Reproductive Surgeries    0.375089
Ovulation Disorders                0.206829
Pelvic Inflammatory Disease        0.155508
Premature Ovarian Insufficiency    0.152696
Autoimmune Disorders               0.122159
Patient ID                         0.110388
Age                                0.076261
Hormonal Imbalances                0.072537
Endometriosis                      0.044374
Blocked Fallopian Tubes           -0.008205
Uterine Abnormalities             -0.037240
Name: Infertility Prediction, dtype: float64
```

Figure 45 Correlation with Diagnosis for Infertility Dataset

```
Correlation with Diagnosis:
 Endometriosis                      1.000000
Pelvic Inflammatory Disease        0.203607
Patient ID                         0.084058
Age                                0.048132
Infertility Prediction             0.044374
Hormonal Imbalances                0.039156
Unexplained Infertility            0.036364
Ovulation Disorders                0.014476
Autoimmune Disorders               0.007766
Previous Reproductive Surgeries   -0.004135
Premature Ovarian Insufficiency   -0.027917
Blocked Fallopian Tubes           -0.084005
Uterine Abnormalities             -0.162566
Name: Endometriosis, dtype: float64

Correlation with Diagnosis:
 Ovulation Disorders                1.000000
Patient ID                         0.305323
Unexplained Infertility            0.270286
Infertility Prediction             0.206829
Age                                0.162254
Pelvic Inflammatory Disease        0.088007
Hormonal Imbalances                0.084075
Previous Reproductive Surgeries    0.076676
Autoimmune Disorders               0.061581
Endometriosis                      0.014476
Uterine Abnormalities             -0.025757
Premature Ovarian Insufficiency   -0.049532
Blocked Fallopian Tubes           -0.120016
Name: Ovulation Disorders, dtype: float64
```

Figure 47 Correlation with Endometriosis and Ovulation Disorders in Infertility Dataset

The KDE plots in Figures 48 and 49 show the age distribution for **Infertility Prediction** and **Endometriosis**. In both cases, diagnosed individuals (label 1) tend to be slightly older on average than non-diagnosed (label 0). However, the difference is not sharp—both distributions peak between ages **22–30**, and gradually decline with age. This aligns with the clinical reality that **infertility, PCOS, and endometriosis are most commonly diagnosed during the reproductive years**, especially as women begin to seek fertility support. While age does not strongly correlate with diagnosis statistically, these plots confirm that it still plays a **modest role** in both conditions' occurrence patterns.

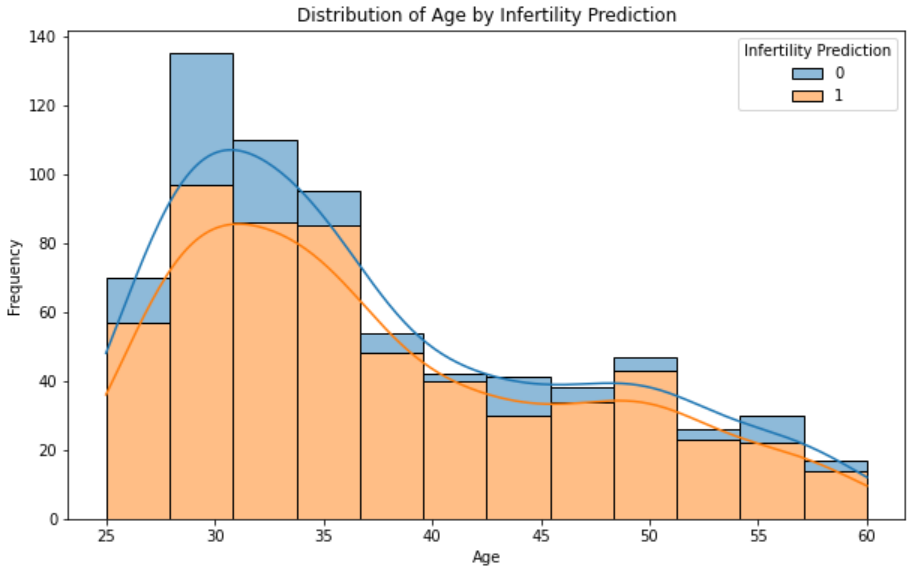

Figure 48 Distribution of Age by Infertility Prediction

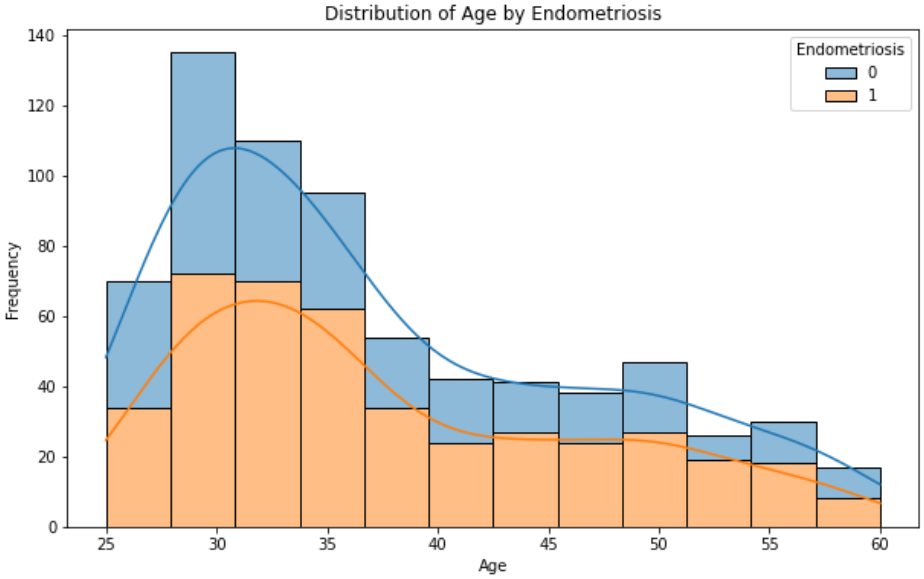

Figure 49 Distribution of Age by Endometriosis

Figure 50 presents the distribution of key binary features across infertility outcomes. **In every case, individuals with the condition (value 1) are more likely to be in the infertile group (label 1**), as shown by the consistently taller orange bars. This includes conditions like ovulation disorders, endometriosis, blocked fallopian tubes, pelvic inflammatory disease, hormonal imbalances, uterine abnormalities, autoimmune disorders, premature ovarian insufficiency, previous reproductive surgeries, and unexplained infertility. **Although individual correlations with infertility are weak, this pattern highlights a**

clear trend: the presence of these conditions is strongly associated with infertility. This reinforces the multifactorial nature of infertility and supports the use of these features in predictive modeling

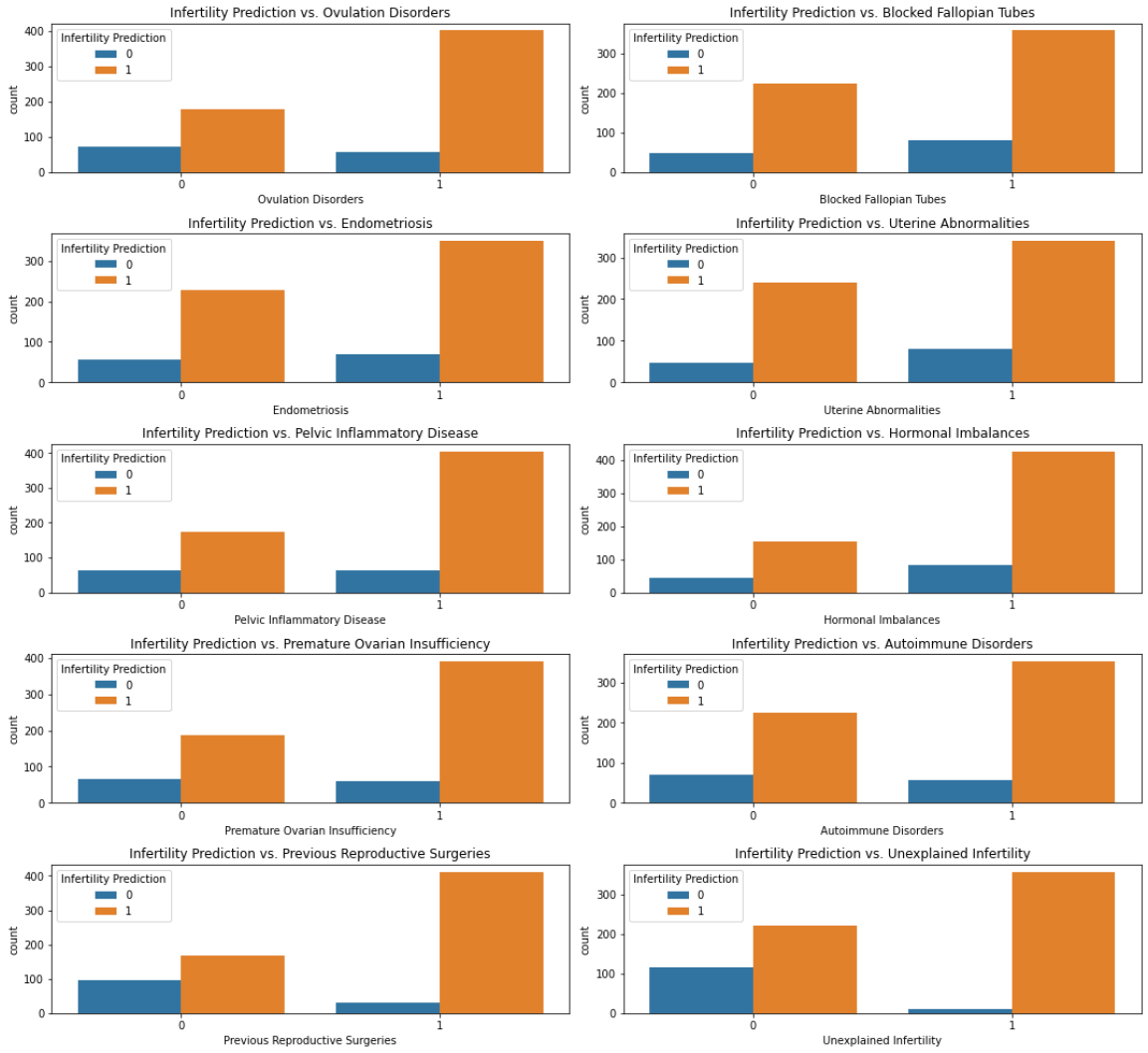

Figure 50 Distribution of Binary Features by Diagnosis

Figure 51 displays the distribution of **key binary features in relation to ovulation disorders**. In all three subplots (Hormonal Imbalances, Premature Ovarian Insufficiency, and Previous Reproductive Surgeries) there is a noticeable increase in the number of individuals with the condition among those who also have ovulation disorders (label 1). This trend is especially seen in hormonal imbalances and premature ovarian insufficiency, where the presence of the condition aligns strongly with ovulatory dysfunction.

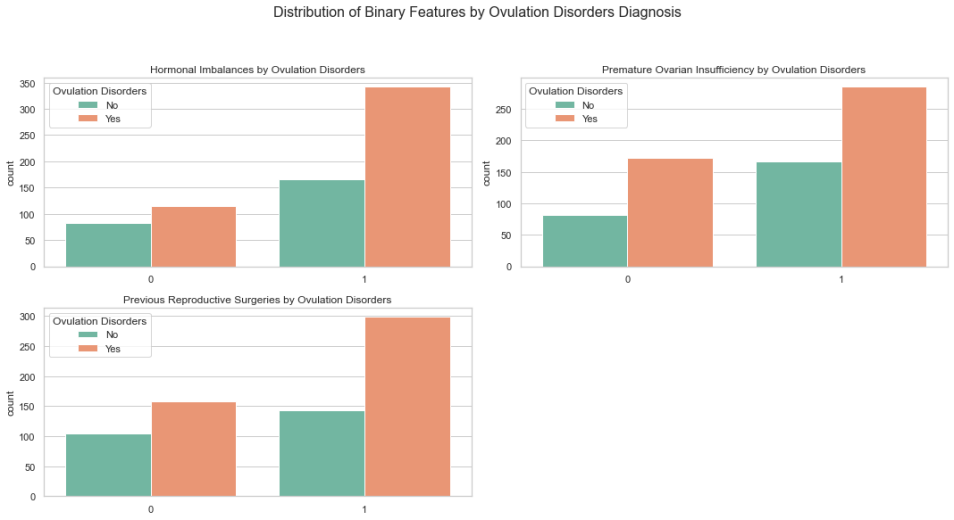

Figure 51 Distribution of Binary Features by ovulation disorders in Infertility Dataset

Figure 52 illustrates the **distribution of several binary clinical features in relation to endometriosis diagnosis**. Across all plots, individuals diagnosed with endometriosis (label 1) tend to show higher counts of associated conditions compared to non-diagnosed individuals. Notably, hormonal imbalances and pelvic inflammatory disease show the clearest increase among diagnosed cases, suggesting a strong clinical connection. Autoimmune disorders, previous reproductive surgeries, and uterine abnormalities also show higher frequencies in the endometriosis group, though the differences are more moderate.

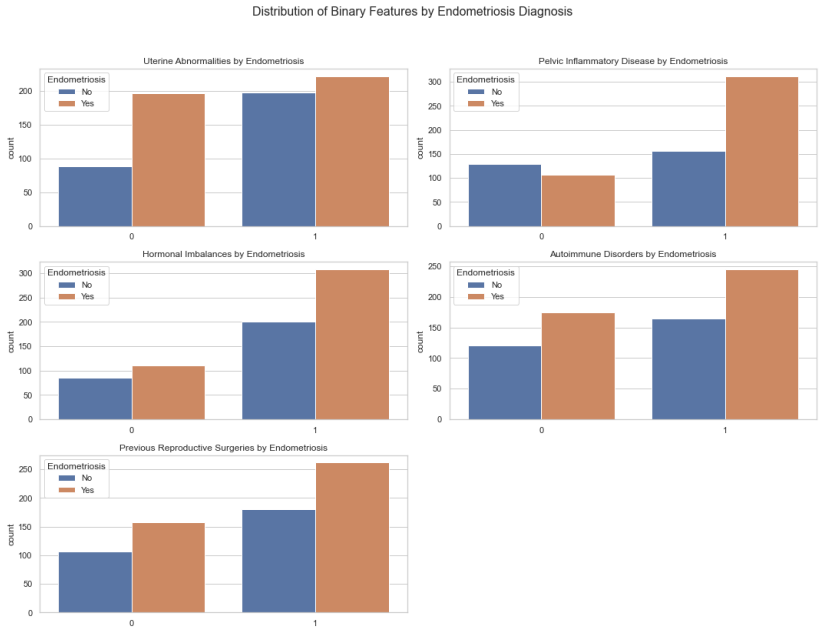

Figure 52 Distribution of Binary Features by Endometriosis Diagnosis in Infertility Dataset

3.3 Data Analysis

The statistical results in Figure 53 highlight several important insights. Based on the p-values, **Ovulation Disorders**, **Pelvic Inflammatory Disease**, **Premature Ovarian Insufficiency**, **Autoimmune Disorders**, **Previous Reproductive Surgeries**, and **Unexplained Infertility** are all significantly associated with infertility prediction ($p < 0.05$). **Age** also shows a statistically significant relationship ($p = 0.009$), while **Endometriosis**, **Hormonal Imbalances**, and **Uterine Abnormalities** do not show significance in this dataset.

The Pearson correlation coefficients confirm that while most features have weak linear relationships with infertility, **Previous Reproductive Surgeries** ($r = 0.375$) and **Unexplained Infertility** ($r = 0.406$) stand out with moderate positive correlations. These are the only features exceeding the threshold for strong correlation ($|r| > 0.3$), making them the most predictive indicators in the analysis. Overall, these results support the multifactorial nature of infertility, where a combination of statistically significant features, despite weak individual correlations, can provide valuable insights when considered together.

```
Summary of Statistical Analysis:

P-values:
Age: 0.009
Ovulation Disorders: 0.000
Blocked Fallopian Tubes: 0.907
Endometriosis: 0.281
Uterine Abnormalities: 0.374
Pelvic Inflammatory Disease: 0.000
Hormonal Imbalances: 0.069
Premature Ovarian Insufficiency: 0.000
Autoimmune Disorders: 0.002
Previous Reproductive Surgeries: 0.000
Unexplained Infertility: 0.000

Pearson Correlation Coefficients:
Age: 0.076
Ovulation Disorders: 0.207
Blocked Fallopian Tubes: -0.008
Endometriosis: 0.044
Uterine Abnormalities: -0.037
Pelvic Inflammatory Disease: 0.156
Hormonal Imbalances: 0.073
Premature Ovarian Insufficiency: 0.153
Autoimmune Disorders: 0.122
Previous Reproductive Surgeries: 0.375
Unexplained Infertility: 0.406

Statistically Significant Relationships (p < 0.05):
Age: p-value = 0.009
Ovulation Disorders: p-value = 0.000
Pelvic Inflammatory Disease: p-value = 0.000
Premature Ovarian Insufficiency: p-value = 0.000
Autoimmune Disorders: p-value = 0.002
Previous Reproductive Surgeries: p-value = 0.000
Unexplained Infertility: p-value = 0.000

Strong Correlations (|r| > 0.3):
Previous Reproductive Surgeries: r = 0.375
Unexplained Infertility: r = 0.406
```

Figure 53 Correlation and Statistical Test for Infertility Dataset

3.4 Model Training

In the infertility dataset, I tried multiple models using cross-validation. I experimented with both balanced and unbalanced versions of the target variable, and varied the number of selected features from k=2 to k=11. However, for this report, I selected only three models to showcase their results with and without balancing. That said, I have results for about 20 models documented in the Jupyter notebook, if you would like to review them**.**

3.4.1 Model Training - with 10-fold cross-validation, only without balancing the Target

Model	accuracy	precision	recall	f1	roc_auc
KNeighbors	0.87662	0.886581	0.974348	0.927893	0.859958
DecisionTree	0.870946	0.914857	0.930067	0.921537	0.77143
RandomForest	0.913501	0.925698	0.972849	0.948364	0.94154
GradientBoosting	0.917746	0.933178	0.969278	0.950669	0.933422

3.4.2 Model Training - with 10-fold cross-validation, only with balancing the Target

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
RandomForest	0.910684	0.932493	0.960663	0.946064	0.943795
VotingClassifier	0.903521	0.941417	0.942388	0.941251	0.941652
GradientBoosting	0.902213	0.940987	0.939979	0.940187	0.937325
KNN	0.878109	0.941327	0.908826	0.924127	0.882090
DecisionTree	0.875231	0.931578	0.916503	0.922938	0.808295

3.4.3 Model Training - with 10-fold cross-validation and SelectKBest K=4 without balancing

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
GradientBoosting	0.909235	0.933907	0.956961	0.945203	0.905691
KNN	0.909235	0.940069	0.950396	0.944930	0.835722

RandomForest	0.906378	0.942960	0.943250	0.942727	0.903462
VotingClassifier	0.906378	0.942960	0.943250	0.942727	0.906523
DecisionTree	0.890865	0.943619	0.920278	0.930623	0.903851

3.4.4 Model Training - with 10-fold cross-validation and SelectKBest K=4 with balancing

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
KNN	0.907787	0.907504	0.988259	0.945920	0.823068
DecisionTree	0.882233	0.937985	0.916317	0.926885	0.889994
RandomForest	0.882233	0.937985	0.916317	0.926885	0.890508
GradientBoosting	0.882233	0.937985	0.916317	0.926885	0.892660
VotingClassifier	0.882233	0.937985	0.916317	0.926885	0.890924

3.4.5 Model Training - with 10-fold cross-validation and SelectKBest K=9 without balancing

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
GradientBoosting	0.900724	0.923475	0.959524	0.940787	0.926596
KNN	0.879416	0.897674	0.962202	0.928509	0.859622
RandomForest	0.878149	0.938620	0.912037	0.924378	0.916407
VotingClassifier	0.859678	0.918305	0.910565	0.913783	0.920008
DecisionTree	0.848410	0.929485	0.882405	0.904547	0.811082

3.4.6 Model Training - with 10-fold cross-validation and SelectKBest K=9 with balancing

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
GradientBoosting	0.907887	0.941343	0.947052	0.943970	0.936889
RandomForest	0.902133	0.930406	0.951988	0.940635	0.926347
VotingClassifier	0.892173	0.933062	0.936680	0.934021	0.940596

DecisionTree	0.880845	0.933987	0.921451	0.926584	0.824528
KNN	0.828471	0.923677	0.861916	0.890600	0.849518

Figure 59 shows the feature importance scores from the Random Forest model in the balanced dataset and with K=all. The most influential feature in predicting infertility is **Unexplained Infertility**, followed closely by **Previous Reproductive Surgeries**. These two features contribute the most to the model's decision-making, confirming earlier statistical findings where both had the highest correlation with infertility.

Age and **Ovulation Disorders** also show moderate importance, indicating their relevance as secondary predictors. All remaining features, including **Hormonal Imbalances**, **Pelvic Inflammatory Disease**, and **Endometriosis**, have low importance in this model. This does not mean they are clinically irrelevant, but rather within this dataset and model, they contribute less to distinguishing between infertile and non-infertile cases.

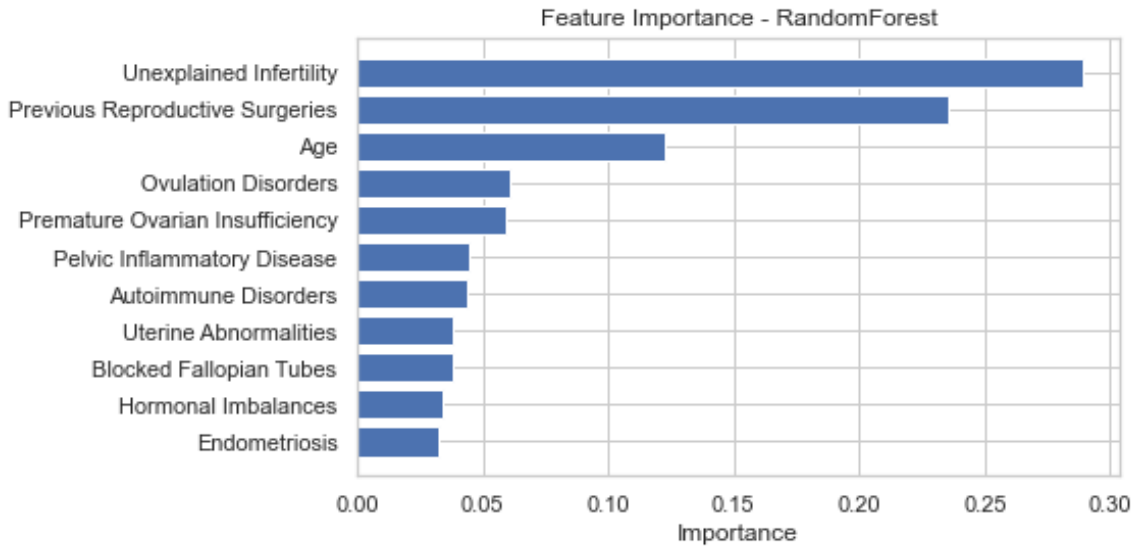

Figure 59 Feature Importance in infertility dataset

3.5 Model evaluation

Overall Observations: Model Performance on Imbalanced vs. Balanced Data

In this dataset, the majority class is infertility cases (label = 1), while the minority class (label = 0) represents non-infertile women. This imbalance can lead models to favor the majority, achieving high scores by mostly predicting infertility. Missing critical "no infertility" cases.

1. Impact of Balancing (results_10_Fold vs. results_10_Fold_Balanced)

- When comparing models trained on **imbalanced data** vs. **balanced data**, we observe a consistent **increase in Recall and F1 Score** after balancing.
    - For example, **KNeighbors' Recall** increases from 0.974 to 0.988 and Gradient Boosting F1 performance (from 0.950 to 0.940), but now performs better across minority class cases as well.
- **ROC AUC also improves**, showing better separation between classes. This confirms that balancing **helps models better detect the underrepresented group (non-infertility)**.
- However, **Accuracy typically decreases slightly after balancing**, because the models are no longer biased toward the dominant class (infertility). This drop is expected and acceptable, as **high accuracy on imbalanced data often reflects poor minority class performance**.
    - For instance, **Random Forest's accuracy** slightly dropped from **0.913 (imbalanced)** to **0.910 (balanced)**, but its **ROC AUC and F1 Score remained strong**, proving its predictions became more meaningful.

2. Impact of Feature Selection (results_k4 vs. results_k9)

- Selecting a reduced number of features (e.g., **K=4 or K=9**) often leads to **more stable and competitive results**, even compared to the full feature set.
- Across different selections, models like **Gradient Boosting, Random Forest, and Voting Classifier** consistently show strong F1 and ROC AUC values, highlighting that reducing input features **does not hurt performance** and may even improve generalization.

Interpreting the "Best" Models Based on SelectKBest + Balancing

results_k4_balanced (4 features + balanced):

- **KNeighbors** gives the **best Recall (0.988)** and F1 Score (0.945).
- **Decision Tree, Random Forest, and Gradient Boosting** follow closely in Precision and overall ROC AUC (~0.89).
- **Conclusion**: This configuration maximizes sensitivity to non-infertile cases without major losses in precision, making it highly suitable in a clinical context where **missing a healthy case matters**.

results_k9_balanced (9 features + balanced):

- **Gradient Boosting** is the strongest overall:
    - **Best Accuracy (0.908)**
    - **F1 Score (0.944)**
    - **ROC AUC (0.937)**
- **Random Forest** and **Voting Classifier** also perform extremely well across all metrics.
- **Conclusion**: With more features, Gradient Boosting remains highly accurate and robust. Random Forest is a close second and especially strong in Precision (0.930).

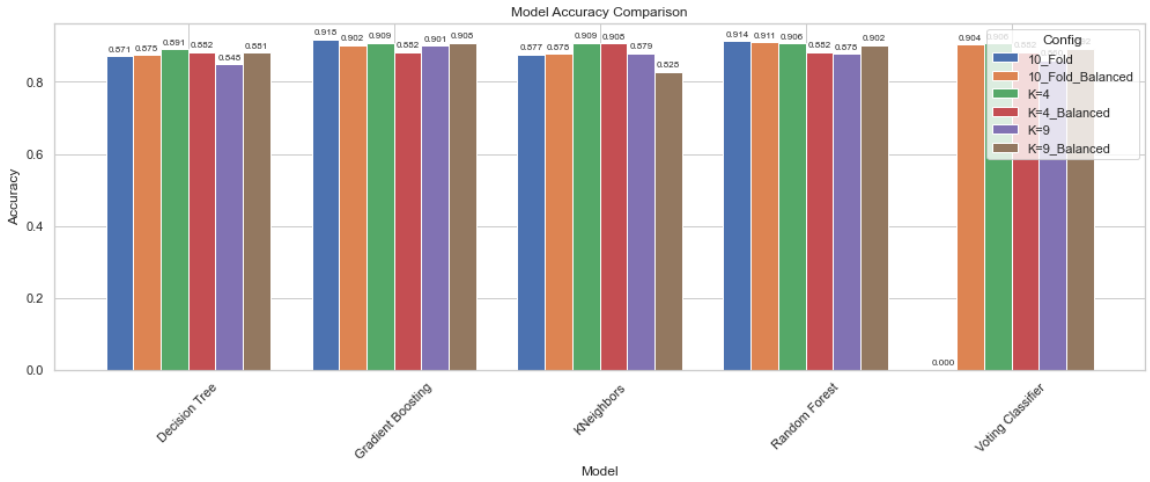

Figure 54 Accuracy Comparison For infertility Dataset

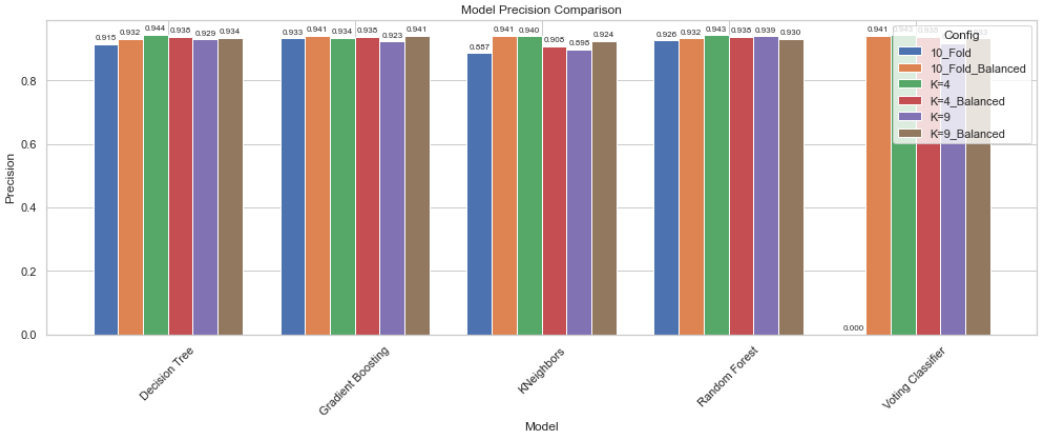

Figure 55 Precision Comparison For infertility Dataset

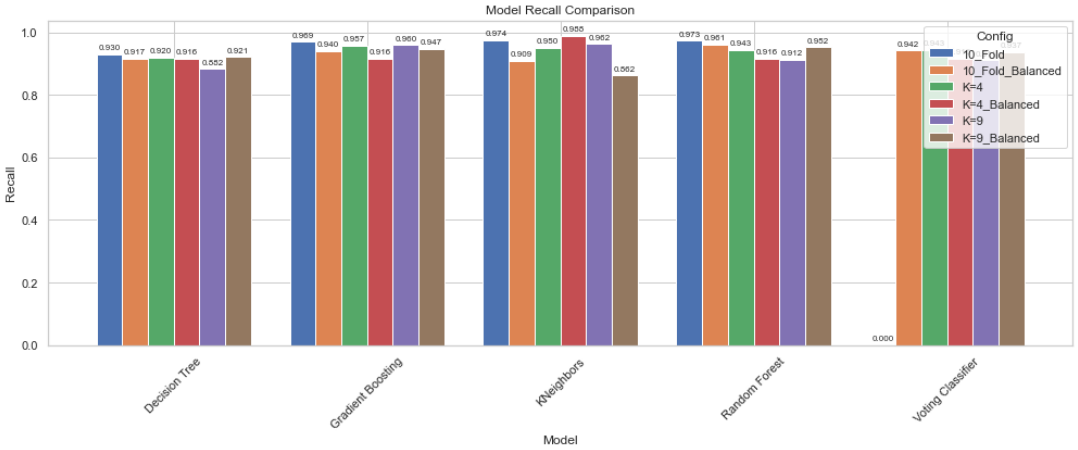

Figure 56 Recall Comparison For infertility Dataset

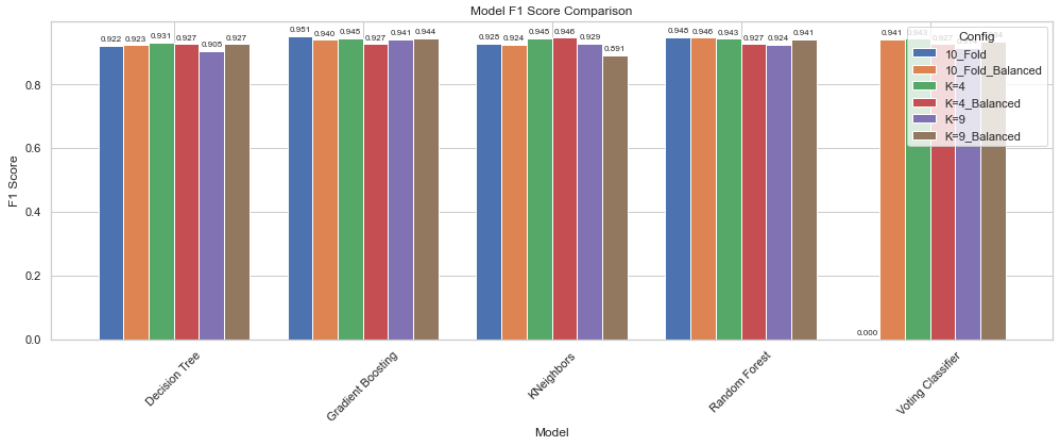

Figure 57: F1Score Comparison For infertility Dataset

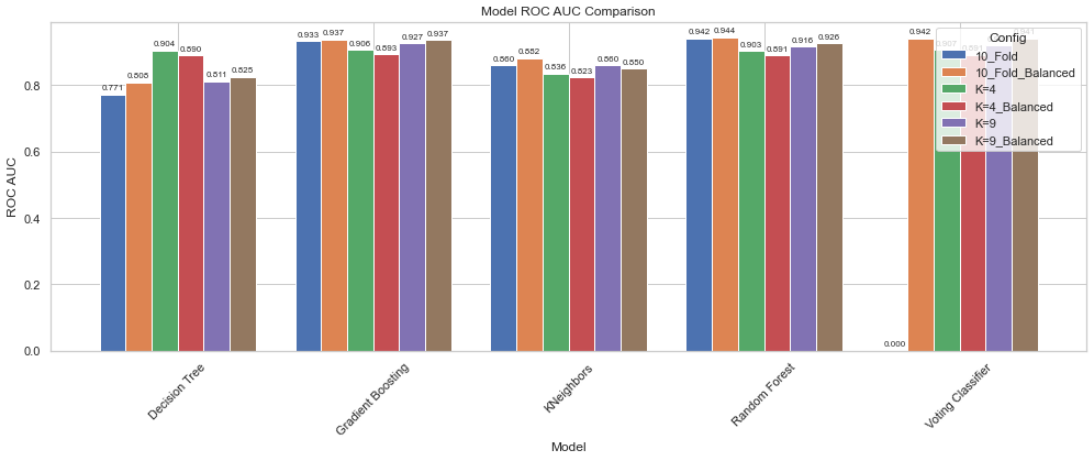

Figure 58 Roc AUC Comparison For infertility Dataset

Conclusion

Reference

1. Johnson, N. P., Hummelshoj, L., Adamson, G. D., Keckstein, J., Taylor, H. S., Abrao, M. S., et al. (2022). “World Endometriosis Society consensus on the classification of endometriosis.” *Human Reproduction*, vol. 37, no. 2, pp. 143–153.
2. Teede, H. J., Misso, M. L., Costello, M. F., Dokras, A., Laven, J., Moran, L., et al. (2018). “Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome.” *Fertility and Sterility*, vol. 110, no. 3, pp. 364–379.
3. Lim, S. S., Kakoly, N. S., Tan, J. W. J., Fitzgerald, G., Bahri Khomami, M., Joham, A. E., et al. (2019). “Metabolic syndrome in polycystic ovary syndrome: A systematic review, meta-analysis and meta-regression.” *Obesity Reviews*, vol. 20, no. 2, pp. 339–352.
4. https://www.kaggle.com/datasets/michaelanietie/endometriosis-dataset/data
5. https://www.kaggle.com/datasets/michaelanietie/endometriosis-dataset/data
6. https://www.kaggle.com/datasets/michaelanietie/endometriosis-dataset/data