✅ All 4 models (RF + SVM + MLP + HGB) work together in the ensemble for best prediction!

Regenerate Pipeline

🔄 Regenerate outputs (run pipeline on trained data)

```
================================================================================
HEALTH PREDICTION PIPELINE - STACKING ENSEMBLE
================================================================================
================================================================================
LOADING AND PREPARING DATA
================================================================================
✓ Dataset loaded: 100 rows, 12 columns
✓ Data sorted chronologically

================================================================================
FEATURE ENGINEERING
================================================================================
✓ Rolling features created (30-min window)
✓ Subgroup feature created: Gender_Sleep_Group
  Unique subgroups: 8

================================================================================
CHRONOLOGICAL SUBGROUP-AWARE SPLIT
================================================================================
✓ Split completed:
  - Training:     56 samples (60%)
  - Validation:   10 samples (10%)
  - Test:         34 samples (30%)

✓ Test subgroup coverage:
Gender_Sleep_Group
Other_Normal     7
Male_Normal      7
Female_Normal    6
Female_Short     4
Other_Short      4
Male_Short       4
Other_Long       1
Male_Long        1

================================================================================
FEATURE PREPARATION
================================================================================

⚠️  Small strata detected (< 6 samples):
    Male_Short_1: 5 samples
    Female_Normal_0: 5 samples
    Other_Normal_0: 5 samples
    Female_Normal_1: 5 samples
    Female_Short_1: 5 samples
    Other_Short_1: 4 samples
    Male_Short_0: 2 samples
    Other_Short_0: 2 samples
    Female_Short_0: 2 samples
✓ SMOTE applied: 56 → 66 samples
✓ Features prepared and scaled
  - Feature count: 14
  - Train shape: (66, 14)
  - Val shape:   (10, 14)
  - Test shape:  (34, 14)
```

```
================================================================================
TRAINING BASE MODELS WITH BAYESIAN OPTIMIZATION (LEVEL-0)
================================================================================

[1/4] Optimizing Random Forest with Bayesian Search...
  Best RF params: OrderedDict({'max_depth': 9, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split
✓ RF - Acc: 0.8000 | F1: 0.8750

[2/4] Optimizing SVM with Bayesian Search...
  Best SVM params: OrderedDict({'C': 0.7756486208064789, 'gamma': 0.5703843027403095, 'kernel': 'rbf'})
✓ SVM - Acc: 0.7000 | F1: 0.8235

[3/4] Optimizing MLP Neural Network with Bayesian Search...
  Best MLP params: OrderedDict({'alpha': 0.0039728931339630255, 'learning_rate_init': 0.0027417584235929055})
✓ MLP - Acc: 0.8000 | F1: 0.8571

[4/4] Optimizing HistGradientBoosting with Bayesian Search...
  Best HGB params: OrderedDict({'learning_rate': 0.29999999999999993, 'max_depth': 8, 'max_iter': 150, 'min_samp
✓ HGB - Acc: 0.7000 | F1: 0.8000

================================================================================
TRAINING META-MODEL (STACKING ENSEMBLE)
================================================================================
✓ Stacking Ensemble trained
  Validation F1-Score: 0.8235
  Selected threshold for meta-model: 0.520 (F1=0.8750)

================================================================================
TEST SET EVALUATION
================================================================================

✓ OVERALL TEST PERFORMANCE:
  - Accuracy:  0.7059
  - F1-Score:  0.8276
  - ROC-AUC:   0.5778

✓ Classification Report:
              precision    recall  f1-score   support

  Low Health       0.00      0.00      0.00         9
 High Health       0.73      0.96      0.83        25

    accuracy                           0.71        34
   macro avg       0.36      0.48      0.41        34
weighted avg       0.53      0.71      0.61        34


================================================================================
🔬 SUBGROUP FAIRNESS ANALYSIS (PER-GROUP THRESHOLDS)
================================================================================

📊 SUBGROUP PERFORMANCE TABLE:
================================================================================
     Subgroup  Sample_Size  Threshold  Accuracy  F1_Score  ROC_AUC  True_High_Health_%
Female_Normal            6       0.52  0.833333  0.909091 0.600000           83.333333
 Female_Short            4       0.52  0.750000  0.857143 0.666667           75.000000
  Other_Short            4       0.52  0.750000  0.857143      NaN          100.000000
 Other_Normal            7       0.52  0.714286  0.833333 0.600000           71.428571
  Male_Normal            7       0.52  0.714286  0.833333 0.600000           71.428571
   Male_Short            4       0.52  0.500000  0.666667 0.250000           50.000000
================================================================================

📈 STATISTICAL SUMMARY (ALL SUBGROUPS):
  - Mean F1-Score:  0.8261
  - Std F1-Score:   0.0829
```

```
  - Best subgroup:  Female_Normal (F1=0.9091)
  - Worst subgroup: Male_Short (F1=0.6667)

📈 ROBUST EVALUATION (≥6 samples per group):
  - Mean F1-Score:  0.8586
  - Std F1-Score:   0.0437
  - Groups included: 3/6

================================================================================
✅ ANSWER TO RESEARCH QUESTION:
================================================================================
✗ NO (Strict) - Model shows UNEQUAL performance across all subgroups.
  (F1 std = 0.0829 ≥ 0.05)

✓ YES (Relaxed) - Model achieves REASONABLE EQUALITY for larger subgroups.
  (F1 std = 0.0437 < 0.10 for groups with ≥6 samples)
  ⚠️  Very small subgroups (<6 samples) show higher variance due to limited data.
================================================================================

📊 Creating visualizations...
✓ Visualizations saved to: /Users/abduljamilsafi/Documents/Second_project/visualizations/subgroup_analysis.png

💾 Saving results...
✓ Subgroup results saved to: /Users/abduljamilsafi/Documents/Second_project/data/final_subgroup_performance.csv
✓ All results saved successfully!

💾 Saving trained model...
✓ Model saved to: /Users/abduljamilsafi/Documents/Second_project/models/stacking_ensemble.pkl
   Features: 14
   Threshold: 0.520

================================================================================
✅ PIPELINE COMPLETED SUCCESSFULLY!
================================================================================
```

✅ Pipeline finished. Refresh the page to see updated results.