Member 2: Modeling.

---

1. Model Selection and Justification

We selected Random Forest as our model because it offers a strong balance between predictive performance and interpretability, which is essential in the healthcare domain. Hospital readmission prediction involves both numerical features (e.g., lab results, age, length of stay) and categorical variables (e.g., diagnosis codes, insurance types), and Random Forest handles mixed data types efficiently.

Additionally, Random Forests are relatively robust to missing data and outliers — common challenges in EHR and insurance claims datasets. Importantly, they also provide feature importance scores, allowing clinicians and hospital administrators to understand which factors contribute most to patient readmission risk. This makes Random Forest a practical choice for building trust while delivering accurate predictions.

2. Data Splitting Strategy

To ensure our model generalizes well and avoids overfitting, we will split the dataset into three subsets:

- Training set (60%): Used to train the Random Forest model on historical patient data.
- Validation set (20%): Used during model development to fine-tune hyperparameters such as the number of trees ($n\_estimators$) and maximum depth (max_depth).
- Test set (20%): Used only once — after model tuning — to evaluate final performance and simulate real-world deployment.

This 60/20/20 split helps ensure that our model performs well not just on training data but also on unseen patients, which is critical in a high-stakes setting like healthcare.

3. Hyperparameter Tuning

For our Random Forest model, two key hyperparameters we will tune are:

1. $n\_estimators$: This defines the number of decision trees in the forest. A higher number generally improves model performance by reducing variance, but increases computational cost and training time. We'll tune this to balance accuracy with efficiency.
2. max_depth: This sets the maximum depth allowed for each decision tree. A deeper tree can capture more complex patterns, but risks overfitting. Limiting the depth

helps the model generalize better, especially with noisy or imbalanced medical data.

◇ Part 2: Case Study.

1. Model Justification for Hospital Scenario
In a hospital setting, Random Forest is well-suited for predicting 30-day readmissions due to its ability to handle large, messy datasets like EHRs and insurance claims. It can account for complex patterns across clinical and demographic variables (e.g., comorbidities, age, lab results), which are often nonlinear.

Moreover, feature importance from the model helps clinicians understand which factors contribute to each prediction, supporting accountability and decision-making. This is essential in healthcare, where transparency can impact patient safety, trust, and regulatory compliance.

2. Confusion Matrix

	Predicted: Readmit	Predicted: No Readmit
Actual: Readmit	80	20
Actual: No Readmit	30	70

3. Precision and Recall Calculation

- Precision = 80 / (80 + 30) = 0.727 or 72.7%
- Recall = 80 / (80 + 20) = 0.8 or 80%

These metrics indicate that the model is effective at identifying patients at risk of readmission (recall) while maintaining a reasonable false-positive rate (precision), which is important in ensuring efficient but targeted clinical interventions.

◇ Part 3: Support.

In healthcare, interpretability often matters as much as accuracy, because clinical decisions must be explainable to doctors, patients, and stakeholders. While high-performance models like neural networks may offer better accuracy, their "black-box" nature makes it hard to trust or understand their outputs.

We chose Random Forest because it strikes a strong balance — it performs well on complex medical data while still offering interpretability through feature importance scores. This allows clinicians to understand the key drivers behind predictions and helps ensure accountability in patient care. In a sensitive, high-stakes domain like healthcare, this trade-off makes Random Forest a responsible and effective choice.