Link to Jupyter Notebook: [hospitalreadmissionprediction/eda.ipynb at main · andre-datascience/hospitalreadmissionprediction]()

Summary of Findings on Patient Readmissions

1. Key Statistics & Distributions

- **Age Distribution:** Uniform across the dataset, with no strong trends linked to readmission.
- **Days in Hospital:** Multimodal distribution, indicating different patient subgroups with varying hospitalization lengths.
- **Comorbidity Score:** Distinct peaks suggest common levels of comorbidity among patients.
- **Readmission Rates:** Highly imbalanced dataset; only **18.8%** of patients were readmitted.

2. Correlations & Relationships Between Features

- **Primary Diagnoses with Highest Readmission Rates:**
    1. **Kidney Disease**
    2. **Diabetes**
    3. **COPD**
    4. **Heart Disease**
    5. **Hypertension**
- **Feature Correlations:** No strong linear correlation between individual features and readmission.
- **Readmission vs. Other Features:** No clear trends in scatter plots, suggesting a non-linear relationship.
- **Clusters in Data:** Certain groups emerge based on **days in hospital** and **comorbidity scores**, which may indicate high-risk patient categories.

3. Model Performance & Challenges

- **Severe Class Imbalance:**
    - The model predicts **almost all patients as non-readmitted**, leading to high accuracy but **poor recall** for readmitted patients.
    - **First Logistic Regression Model:**

        - Accuracy: **82.6%**
        - **Completely failed to predict readmissions (Recall for Class 1 = 0.00)**

- **Random Forest Model Performance (After SMOTE Resampling & Hyperparameter Tuning):**
    - Accuracy: **78.9%**
    - Readmission Recall: **3%** → Still **very low**, meaning most actual readmitted cases are missed.
- **Alternative Logistic Regression Model (With Class Balancing & SMOTE):**
    - Accuracy: **49.8%**
    - Readmission Recall: **38%** → Identifies more readmitted patients but at the cost of overall accuracy.

4. Next Steps for Improvement

1. **Further Improve Readmission Recall:**
    - Adjust **SMOTE sampling strategy** to avoid excessive synthetic data.
    - **Tune Random Forest parameters** (lower min_samples_split, adjust class_weight).
2. **Feature Engineering for Better Predictions:**
    - Introduce **interaction terms** (e.g., age × comorbidity_score).
    - Try **non-linear transformations** (e.g., polynomial features).
3. **Explore More Advanced Models:**
    - **Gradient Boosting Models (XGBoost, LightGBM)** → Handle imbalanced data better.
    - **Neural Networks** for capturing complex relationships.