**Predicting Patient Readmission**

# **Introduction:**

The aim of this project is to predict if a patient will be readmitted to the hospital after being discharged. Hospital readmissions are costly and may suggest that the patient’s condition is not improving as expected. By predicting which patients are more likely to be readmitted, doctors and healthcare providers can take early action to prevent this, which can improve patient care and reduce hospital costs.

This project uses a dataset that includes information about patients' age, gender, medical history, treatments they received, and whether they were readmitted. By analyzing this data, we can build a model that helps identify patients who are at high risk of returning to the hospital. This can help doctors make better decisions about patient care and improve overall healthcare outcomes.

# **Models Used**

1. **Logistic Regression**:
   * A simple, interpretable model.
   * Provides insights into the relationship between patient features and the likelihood of readmission.
   * Useful as a baseline to compare with more complex models.
2. **Decision Tree Classifier**:
   * Creates a tree-like structure to make predictions based on feature importance.
   * Easy to visualize and interpret.
   * Highlights key factors influencing patient readmission.
3. **Random Forest Classifier**:
   * Combines multiple decision trees to reduce overfitting and improve accuracy.
   * Identifies the most critical features contributing to readmission risk.
4. **K-Nearest Neighbors (KNN)**:
   * Predicts outcomes by comparing patients to similar cases in the dataset.
   * Simple and intuitive but may struggle with larger datasets.
5. **Support Vector Machine (SVM)**:
   * Effective for handling high-dimensional data.
   * Finds the optimal boundary to separate classes, offering robust performance.

# **Workflow**

1. **Data Preprocessing**:
   * Cleaned the data by removing unnecessary columns and handling missing values.
   * Categorical features were converted to numerical values using one-hot encoding.
   * Numerical data was normalized for uniform scaling.
2. **Data Preparation**:
   * The dataset was split into training (80%) and testing (20%) sets.
3. **Model Selection and Training**:
   * Selected five machine learning models based on interpretability and performance.
   * Each model was trained on the training dataset.
4. **Model Evaluation**:
   * Models were evaluated using accuracy, classification reports, and confusion matrices.
   * Cross-validation was used to ensure robust comparisons.

# **Dataset Overview**

The dataset includes several features that influence hospital readmissions:

| **Feature** | **Description** |
| --- | --- |
| Patient\_ID | Unique identifier for each patient (excluded from model training). |
| Age | Patient’s age; older individuals are at higher risk of readmission. |
| Admission\_Type | Type of admission (e.g., emergency, elective). |
| Diagnosis | Patient’s primary diagnosis during admission. |
| Num\_Lab\_Procedures | Number of lab tests performed. |
| Num\_Medications | Number of medications prescribed. |
| Num\_Outpatient\_Visits | Number of outpatient visits; indicative of chronic conditions. |
| Num\_Inpatient\_Visits | Number of previous hospitalizations. |
| Num\_Emergency\_Visits | Number of emergency room visits; often correlates with severe health issues. |
| Num\_Diagnoses | Total number of diagnoses. |
| A1C\_Result | Blood sugar control level; critical for diabetic patients. |
| Pollution\_Index | Environmental pollution level; higher values may exacerbate chronic illnesses. |
| Temperature | Average environmental temperature, affecting conditions like heart disease or asthma. |

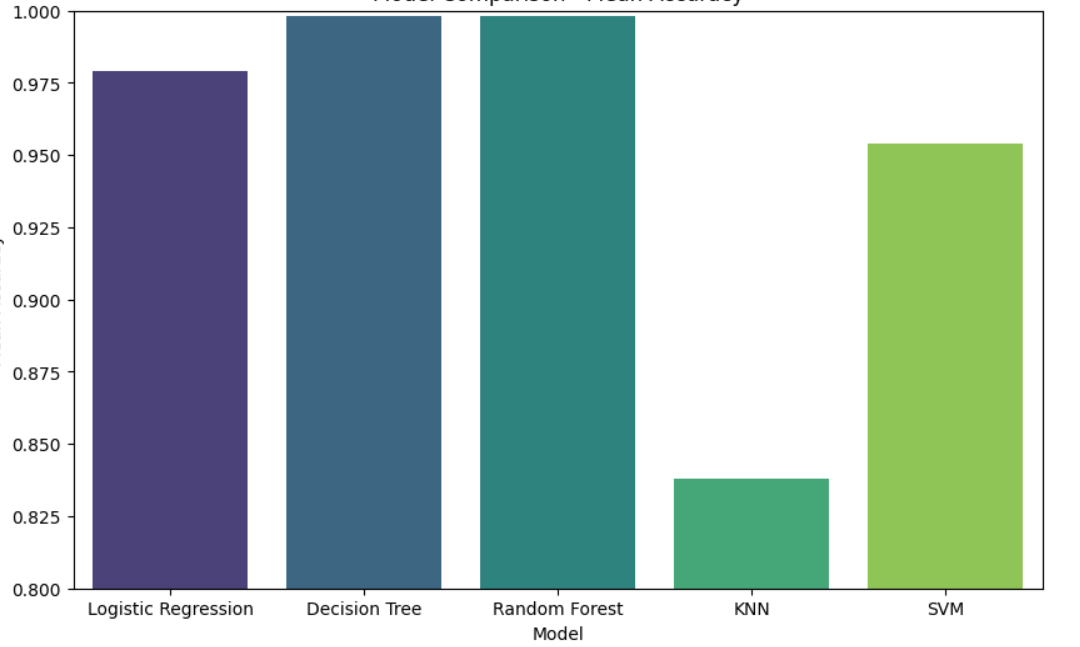
# **Key Steps**

1. **Model Comparison**:
   * Models were evaluated using train-test splits and cross-validation.
   * Accuracy and standard deviation metrics were used to compare performance.
2. **Visualization**:
   * Confusion matrices and accuracy bar plots helped illustrate model performance.
3. **Best Model**:
   * Random Forest emerged as the most accurate and stable model, with a high mean accuracy and low standard deviation during cross-validation.

# **Results**

| **Model** | **Mean Accuracy** | **Standard Deviation** |
| --- | --- | --- |
| Logistic Regression | 98.2% | 0.75% |
| Decision Tree | 99.8% | 0.24% |
| Random Forest | 99.9% | 0.20% |
| KNN | 82.9% | 3.77% |
| SVM | 95.3% | 1.08% |

Random Forest and Decision Tree performed the best, with Random Forest offering better generalizability.



# **How We Created the Data:**

In creating this dataset, we chose specific features that are likely to impact hospital readmission risk, using ranges informed by other medical datasets to ensure realism and relevance. By building the dataset ourselves, we had the flexibility to control and design each feature according to our project goals.

For example:

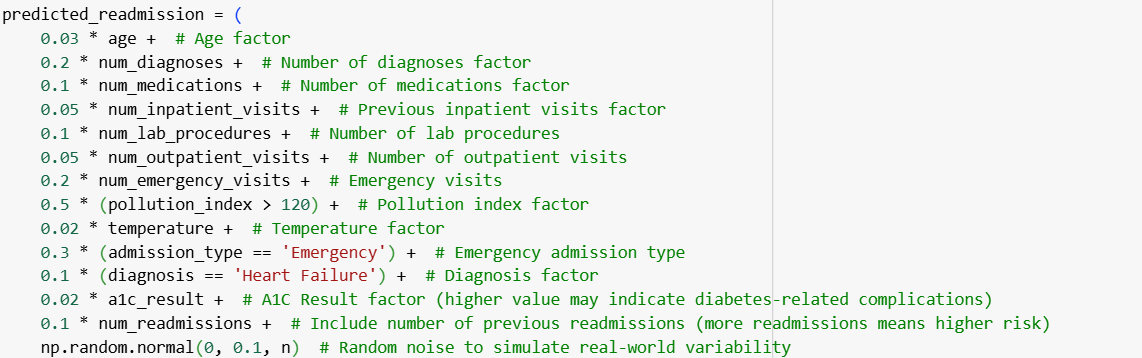
* We created ages based on an average of 65 years old (with some variation).
* We set the pollution index and temperature to realistic values to simulate different environmental conditions.

**How We Calculated Readmission Risk:**

We used a formula to predict the likelihood of a patient being readmitted based on the features in the dataset. This prediction score was calculated by giving different weights to each feature (i.e., age, number of diagnoses, etc.) and adding them together. The more important a feature is, the higher its weight in the formula.

The formula looks like this:

Predicted Readmission=



Once we got the predicted score, we adjusted it to make sure it was between 0 and 1 for easier comparison across patients.

* ***References***:
* Internet Citation: How CMS Measures the "30-Day All Cause Rehospitalization Rate" on the Hospital Compare Web Site. Content last reviewed March 2013. Agency for Healthcare Research and Quality, Rockville, MD. https://www.ahrq.gov/patient-safety/settings/hospital/red/toolkit/redtool-30day.html

* Wang, S., & Zhu, X. (2021). Predictive Modeling of Hospital Readmission: Challenges and Solutions. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1–1. doi:10.1109/tcbb.2021.3089682
* Ramirez, J. C., & Herrera, D. (2019). Prediction of diabetic patient readmission using machine learning. 2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI). doi:10.1109/colcaci.2019.8781796