**Data Science Project Proposal**

**Title:** Inpatient Analysis & Predicting Length of Stay

**Project prepared by:** Diaa Aldein Alsayed Ibrahim Osman

**Prepared for:** Epsilon AI Institute

**Project Overview:** The aim of this data science project is to perform a comprehensive analysis of inpatient data for the state of New York in 2015, and develop a predictive model for the length of stay. The project will leverage advanced data analytics and machine learning techniques to provide valuable insights for healthcare providers and policymakers.

**Project Objectives:**

1. **Data Exploration and Analysis:** Conduct exploratory data analysis (EDA) to understand the distribution and relationships among various patient and hospital-related features. Explore factors affecting the length of stay, including patient demographics, severity of illness, payment types, and hospital locations.
2. **Predictive Modeling:** Build and validate a predictive model for length of stay using machine learning algorithms. The model will help hospitals and healthcare professionals estimate patient stays more accurately, allowing for better resource allocation and improved patient care.
3. **Feature Importance:** Determine the most influential factors affecting length of stay, providing actionable insights to reduce hospital costs and improve efficiency.
4. **Visualization:** Create informative visualizations to present the findings and make complex healthcare data accessible to stakeholders.

**Dataset Description:**

Data obtained from <https://www.kaggle.com/datasets/jonasalmeida/2015-deidentified-ny-inpatient-discharge-sparcs/data>

**About Dataset**

Public Health Data This is the public dataset made available at <https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8> by the Dept of Health of New York state. The following description can be found at that page:

The Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified File contains discharge level detail on patient characteristics, diagnoses, treatments, services, and charges. This data file contains basic record level detail for the discharge. The de-identified data file does not contain data that is protected health information (PHI) under HIPAA. The health information is not individually identifiable; all data elements considered identifiable have been redacted. For example, the direct identifiers regarding a date have the day and month portion of the date removed.

The data is unclean, has missing values, and contains 2.35 million rows and 37 columns. It may not be necessary to include all instances and features to achieve the goal of this project.

Dataset info table below:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2346760 entries, 0 to 2346759

Data columns (total 37 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Health Service Area 2343849 non-null object

1 Hospital County 2343849 non-null object

2 Operating Certificate Number 2343849 non-null float64

3 Facility Id 2343849 non-null float64

4 Facility Name 2346760 non-null object

5 Age Group 2346760 non-null object

6 Zip Code - 3 digits 2342333 non-null object

7 Gender 2346760 non-null object

8 Race 2346760 non-null object

9 Ethnicity 2346760 non-null object

10 Length of Stay 2346760 non-null object

11 Type of Admission 2346760 non-null object

12 Patient Disposition 2346760 non-null object

13 Discharge Year 2346760 non-null int64

14 CCS Diagnosis Code 2346760 non-null int64

15 CCS Diagnosis Description 2346760 non-null object

16 CCS Procedure Code 2346760 non-null int64

17 CCS Procedure Description 2346760 non-null object

18 APR DRG Code 2346760 non-null int64

19 APR DRG Description 2346760 non-null object

20 APR MDC Code 2346760 non-null int64

21 APR MDC Description 2346760 non-null object

22 APR Severity of Illness Code 2346760 non-null int64

23 APR Severity of Illness Description 2346648 non-null object

24 APR Risk of Mortality 2346648 non-null object

25 APR Medical Surgical Description 2346760 non-null object

26 Payment Typology 1 2346760 non-null object

27 Payment Typology 2 1584414 non-null object

28 Payment Typology 3 701190 non-null object

29 Attending Provider License Number 2343849 non-null float64

30 Operating Provider License Number 1733912 non-null float64

31 Other Provider License Number 71336 non-null float64

32 Birth Weight 2346760 non-null int64

33 Abortion Edit Indicator 2346760 non-null object

34 Emergency Department Indicator 2346760 non-null object

35 Total Charges 2346760 non-null object

36 Total Costs 2346760 non-null object

dtypes: float64(5), int64(7), object(25)

memory usage: 662.5+ MB

**Data Sources:** The primary data source for this project is the "Hospital Inpatient Discharges - SPARCS De-Identified" dataset provided by the New York State Department of Health. This dataset contains information on patient demographics, hospital characteristics, diagnoses, treatments, and more.

**Methodology:**

1. **Data Preparation & Preprocessing:** Clean, transform, and preprocess the dataset to ensure data quality and consistency.
2. **Exploratory Data Analysis & Visualization:** Analyze the dataset to uncover patterns, relationships, and correlations among various features.

Create informative visualizations, such as correlation heatmaps and scatter plots, to present key insights in an accessible manner.

1. **Feature Engineering:** Create new features or modify existing ones to improve the predictive model's performance.
2. **Model Building:** Employ machine learning algorithms, such as regression, decision trees, to build a predictive model for length of stay.
3. **Model Evaluation:** Assess the model's performance using appropriate evaluation metrics, including mean absolute error (MAE) and root mean squared error (RMSE).
4. **Expected Outcomes:**

* A predictive model for length of stay that can be integrated into healthcare systems.
* Identification of key factors influencing patient stays, helping hospitals allocate resources more efficiently.
* Insights into trends and patterns in inpatient data, aiding healthcare providers and policymakers in decision-making.

**Budget and Resources:** The project will require access to relevant data sources, data science tools (Python, Jupyter notebooks, VS Code), GitHub, and Streamlit for model deployment.

**Conclusion:** This data science project aims to provide valuable insights into inpatient data for New York State in 2015 and develop a predictive model for length of stay. The results will enable healthcare providers to make more informed decisions, optimize resource allocation, and enhance patient care.