**README File for Kemri Hackathon Project**

**Project Title:**

**Data-Driven Solution for Space Availability in Health Facilities of Murang'a County**

**Context:**

Health stakeholders in Murang'a County are facing significant challenges due to the high demand for quality health services, which current facilities are unable to meet due to space constraints. Identifying and obtaining land for the construction of additional health facilities is a critical issue. To address this, a data-driven solution is required to analyze existing data and facilitate decision-making regarding space availability for new health infrastructure.

**Objective:**

The primary goal of this project is to develop a Machine Learning-based classification model to identify areas within Murang'a County where there is space available for the development of new health facilities. This model will help in recommending suitable areas based on factors influencing space availability.

**Data Description:**

The dataset includes various attributes related to the health facilities in Murang'a County. Below is the detailed data dictionary:

* **SUBCOUNTY**: Name of the sub-county.
* **FACILITY\_CODED\_IDENTIFY**: Unique identifier for the facility.
* **YEAR\_2011**: Number of people who visited the health facility in 2011.
* **YEAR\_2012**: Number of people who visited the health facility in 2012.
* **YEAR\_2013**: Number of people who visited the health facility in 2013.
* **AGENCY**: Ownership of the facility (Options: GOK – County Government, CBO – Community Based Organization).
* **KEPH\_LEVEL**: Classification of the health facility by the national government (Options: Dispensary (Level 2), Health Center (Level 3), Sub-County Hospital (Level 4), County Hospital (Level 5/6)).
* **SPACE1**: Availability of space (land) for future expansion (Options: Yes, No).
* **FUNCTIONALITY2**: Whether the facility offers the basic minimum health services as prescribed by national GoK policy (Options: Below expectation, Fair, Good).
* **STATUS3**: Physical status of the health facility (Options: Bad, Fair, Good).
* **Land\_elevation**: Physical land elevation relative to mean sea level (Options: 1200+ m above mean sea level, <1200m above mean sea level).
* **Available\_capacity\_to\_offer\_prescribed\_health\_services**: Capacity to offer prescribed health services comparing operational outputs to available capacity (Options: No capacity, Minimal capacity, Sufficient installed capacity).

**Project Structure:**

1. **Data Collection**: Compile the provided dataset and ensure all relevant attributes are included.
2. **Data Cleaning**: Handle missing values, outliers, and inconsistencies in the dataset.
3. **Exploratory Data Analysis (EDA)**: Understand the data distribution, relationships between variables, and key drivers influencing space availability.
4. **Feature Engineering**: Create and transform variables that will enhance the performance of the classification model.
5. **Model Building**: Develop and train a classification model to predict space availability.
6. **Model Evaluation**: Assess the performance of the model using appropriate metrics and validate its accuracy.
7. **Recommendation System**: Based on the model's predictions, recommend suitable areas within the county for new health facilities.
8. **Documentation and Reporting**: Document the process, findings, and provide a comprehensive report with actionable insights.

**Usage Instructions:**

1. **Setup Environment**: Ensure you have Python and necessary libraries installed (e.g., pandas, scikit-learn, matplotlib, seaborn).
2. **Load Data**: Import the dataset into your working environment.
3. **Data Preprocessing**: Execute data cleaning and preprocessing scripts to prepare the data for analysis.
4. **Run EDA**: Perform exploratory data analysis to understand data characteristics and distributions.
5. **Feature Engineering**: Apply feature engineering techniques to create meaningful predictors for the model.
6. **Model Training**: Train the classification model using the prepared dataset.
7. **Model Evaluation**: Evaluate the model using test data and check its accuracy and reliability.
8. **Generate Recommendations**: Use the trained model to predict space availability and recommend suitable areas for new health facilities.

**Dependencies:**

* Python 3.7+
* pandas
* numpy
* scikit-learn
* matplotlib
* seaborn

**Contact Information:**

For further information or queries regarding this project, please contact:

* **Data Scientist**: Levi Bushuru
* **Email**: bushurumark@gmail.com
* **Phone**: +254704084567 or +254704135291

**Acknowledgements:**

We would like to thank the health stakeholders in Murang'a County for providing the data and support needed for this project. Special thanks to the Kemri Hackathon organizers for facilitating this initiative.