**Project1. Test result prediction**

**Dataset Information:**

Each column provides specific information about the patient, their admission, and the healthcare services provided, making this dataset suitable for various data analysis and modeling tasks in the healthcare domain. Here's a brief explanation of each column in the dataset.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Definition | Example |
| **Name** | This column represents the name of the patient associated with the healthcare record. | Patrick Parker  Charles Horton  Patty Norman |
| **Age** | The age of the patient at the time of admission is expressed in years. | 41  82  55 |
| **Gender** | Indicates the gender of the patient, either "Male" or "Female." | Male  Female |
| **Blood Type** | The patient's blood type, which can be one of the common blood types (e.g., "A+", "O-", etc.). | AB+  A+  O- |
| **Medical Condition** | This column specifies the primary medical condition or diagnosis associated with the patient, such as "Diabetes," "Hypertension," "Asthma," and more. | Arthritis  Hypertension  Diabetes |
| **Date of Admission** | The date on which the patient was admitted to the healthcare facility. | 8/20/2020  3/22/2021  5/16/2019 |
| **Doctor** | The name of the doctor responsible for the patient's care during their admission. | Robin Green  Patricia Bishop  Brian Kennedy |
| **Hospital** | Identifies the healthcare facility or hospital where the patient was admitted. | Boyd PLC  Wheeler, Bryant and Johns  Brown Inc |
| **Insurance Provider** | This column indicates the patient's insurance provider, which can be one of several options, including "Aetna," "Blue Cross," "Cigna," "UnitedHealthcare," and "Medicare." | Aetna  Cigna  Blue Cross |
| **Billing Amount** | The amount of money billed for the patient's healthcare services during their admission. This is expressed as a floating-point number.  The minimum value is 1000 and the maximum is 50000. | “<1000” and “>50000”  22522.363384853266  39593.43576  13546.817249364824 |
| **Room Number** | The room number where the patient was accommodated during their admission. | “>0”  180  161  384 |
| **Admission Type** | Specifies the type of admission, which can be "Emergency," "Elective," or "Urgent," reflecting the circumstances of the admission. | Emergency  Urgent  Elective |
| **Discharge Date** | The date on which the patient was discharged from the healthcare facility, is based on the admission date and a random number of days within a realistic range. | 8/23/2020  4/15/2021  6/2/2019 |
| **Medication** | Identifies a medication prescribed or administered to the patient during their admission. Examples include "Aspirin," "Ibuprofen," "Penicillin," "Paracetamol," and "Lipitor." | Aspirin  Lipitor  Ibuprofen |
| **Test Results (Target column)** | Describes the results of a medical test conducted during the patient's admission. Possible values include "Normal," or "Abnormal," indicating the outcome of the test. | Abnormal  Normal |

**Requirements**

1. **Preprocessing:** Before building your models, you must ensure the dataset is clean and ready-to-use.

* Deal with empty cells or duplicate records.
* Convert the used categorical columns to numerical columns using one of the encoding techniques.
* Deal with wrong data “for example billing amount>50000 which exceeds its range”.
* Apply feature scaling (normalization) for variables, if necessary.
* Modify data in the wrong format if found).

1. **Feature Selection and Extraction:**
   * Visualize correlation.
   * Using any feature selection or feature extraction methods. (bonus).
2. **Make Classification Models (Training and testing steps):**
   * Train using Logistic Regression, SVM, and Decision Tree (ID3) models.
3. **Model Evaluation:** The evaluation metrics should be calculated on the testing set based on your split size.
   * Accuracy.
   * Confusion matrix.
   * Precision & recall.
   * F1-score.