Machine Learning Assignment:

Using Supervised Machine Learning to Predict In-Hospital Mortality Aim To apply learned data exploration and machine learning skills to design and implement an end-to-end machine learning pipeline for predicting a binary outcomes from a highly dimensional real-world dataset. Indicative Timetable The following table shows a breakdown of the activities for this week to help you in prioritizing your time schedule. The actual time needed to complete each task is dependent on your implementation/analysis speed and personal circumstances. The timetable provided below is merely representative of the average time needed to complete each task and the time needed for each relative to other tasks. The timetable is also provided to ensure that you know that the assignment is time-consuming and requires thoughtful planning to complete all required tasks. Activity Tentative Duration Data Cleaning 2 Days Data Aggregation 3-4 Days Data Exploration & Visualization 2 Day Classifier Implementation &Hyper parameter Tuning4-5 Days Model Evaluation 2 Days Analysis and Reflection 2 Days Writing 3 Days Assignment Details Now that you have been equipped with the skills to use different Machine Learning algorithms, you will have the opportunity to practice and apply it to a practical problem using real-world hospital data extracted from the MIMIC-III database. The dataset consists of measurements of 25 laboratory test results and vital signs for 2670patients recorded over 48 hours. In this assignment, you will complete and submit a Jupiter notebook containing an end-to-end supervised learning pipeline using the XG Boost algorithm to predict 30-day in-hospital mortality from aggregates of patient vital signs and laboratory test results given in thePneumoniaTimeSeries.csv dataset. Recap of Supervised Learning The majority of practical machine learning uses supervised learning. In supervised learning, an algorithm learns a function that maps input variables (X) to an output variable (y), i.e. y = f(X). The goal is to approximate the real underlying mapping so that when the algorithm is supplied with unseen data, it can predict the output variables (y) for the new samples. Supervised learning is called so because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. During supervised training, we know the correct answers; the algorithm iteratively makes predictions on the training data and is corrected by making updates. Learning stops when the algorithm achieves an acceptable level of performance. Supervised learning problems can be further grouped into regression and classification.• Classification: A classification problem is when the output variable is a category, such as “red” and “blue” or “disease” and “no disease.”• Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight.”