**Birla Institute of Technology &amp; Science, Pilani**

**Work-Integrated Learning Program Division**

**First Semester 2020-2021**

**M.Tech (Data Science and Engineering)**

**Assignment**

**Course No.: DSECF ZC415**

**Dataset Name:** Chronic Kidney Disease

|  |  |
| --- | --- |
| **Abstract**: This dataset can be used to predict chronic kidney disease and it has been collected at a hospital for a period of nearly 2 months. |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 400 | **Area:** | N/A |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 25 | **Date Donated** | 2015-07-03 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes |  |  |

**Source:**

Source:   
Dr.P.Soundarapandian.M.D.,D.M   
(Senior Consultant Nephrologist),   
Apollo Hospitals,   
Managiri,   
Madurai Main Road,   
Karaikudi,   
Tamilnadu,   
India.

**Problem Statement**

1. Identify the factors causing chronic kidney disease.
2. Build a model that can help to determine if a patient is suffering from kidney chronic disease or not.
3. Data Understanding

The data is gathered for two months from patients at a hospital. You need to make utilization of the features presented in the data set for your task. The data set and a document containing the information about the attributes are attached with the assignment problem statement.

Make yourself familiar with these attributes as these might help you in determining the patients with kidney chronic disease.

1. Data preparation and Exploratory Data Analysis

You are supposed to make utilizations of all the appropriate data pre-processing techniques on the given data set. If required, make appropriate assumptions and make it explicitly known while using them in the code or in the presentation. You are required to identify the key factors that influences the presence of chronic kidney disease in a patient. Make appropriate selection of the attributes with sound justification for the same. The data set allows for several new combinations of attributes and attribute exclusions, or the modification of the attribute type (categorical, integer, or real) depending on the purpose of the research.

You are supposed to make use of Python programming language and its libraries to work on this analysis effort.

1. Model building and Evaluation

You are supposed to build a model that predicts if a patient is suffering from the kidney disease or not, provided the several features associated with the delivery personnel’s work are given as input.

Apply the appropriate evaluation techniques in order to determine the accuracy of the predictions made by the model. Think of employing the technique that helps in improving the accuracy of the models along with inclusion of limited number of factors in the model.

Try to obtain a model that can be easily understood and explained but it should not come at the cost of accuracy.

You are supposed to make use of Python’s scikit-learn library for this step. You are free to write your custom algorithm as well provided it help in trying the objective of the use case.

1. Expected Outcomes

The results should consist of

1. The python script file or Jupyter notebook containing all the code for the proposed solution. Write all code in single file only with proper comments and outputs at various places.
2. A presentation which describes the
   * Problem
   * Your understanding of data
   * Pre-processing techniques you have applied
   * Intuition behind Algorithm selection for building model
   * Discussion of results
   * your observations

Convert the presentation in pdf format before submission.