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

**Work-Integrated Learning Program Division**

**First Semester 2019-2020**

**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 | **Number of Web Hits:** | 132630 |

**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.

## Evaluation Matrix

|  |  |  |
| --- | --- | --- |
| S No | Criteria |  |
| 1 | Data Understanding and Preparation along with EDA (40%) | * Data quality issues are identified and addressed * Appropriate data pre-processing measures are applied wherever applicable * Any notable exceptions are reported in form of comments, wherever appropriate * Attempt in right direction to find out contributing factors * Right set of visuals are used for univariate and bivariate data analysis * Meaningful insights are derived and presented in effective manner |
| 2 | Model building and evaluation (40%) | * Right data mining task is identified * Train and test data derived and used properly * Appropriate data mining technique is used for the model building * Model parameters are fine tuned to improve the model accuracy * Appropriate technique is used to identify the factors contributing to the model accuracy * Model evaluation is done based on the appropriate measures and criteria |
| 3 | Effective Story telling through the presentation (10%) | * The presentation has proper structure, not too big, not too small. Elaborate the important points in more precise manner. * Has focus on the problem to be solved * Talks about the factors contributing to the issue along with right kind of proofs * Explaining the observations visually where visuals are showcasing the facts * The recommendations / suggestions are spelt out clearly. * Assumptions are specified at right places. |
| 4 | Code readability and organization (5%) | * Code is executing, no syntax errors * No customizations is needed to execute the code * Code is simple and augmented with proper comments wherever required * Built-in functions / libraries are used wherever possible * Repeated code is moved into functions and used appropriately when required * Long code snippets are broken down into small parts and made available as functions to increase the modularity of the code * Appropriate variable names are used to improve the readability of the code |
| 5 | Overall utilization of the concepts learnt in the course (5%) | * Appropriate steps are carried out in the data preparation stage which involves the concepts learnt in the class for the same * Various concepts from the data mining process are paid enough attention while developing the model |