Assignment 1 – Research Practice and Ethics

Research Question and the Objectives – the WHAT?

Research Question

- Scope of the research lies in predicting the status of the person having the Chronic Kidney Disease using Artificial Neural Networks, Support Vector Machines and K-Nearest Neighbour Machine Learning Algorithms.

Research Objectives

- **Comparative analysis** of different machine learning algorithms for the above problem.
- **Handling the imbalance issue** in the dataset for improving the performance.
- **Improving the performance** of the obtained best machine learning model through Hyper Parameter Optimization.

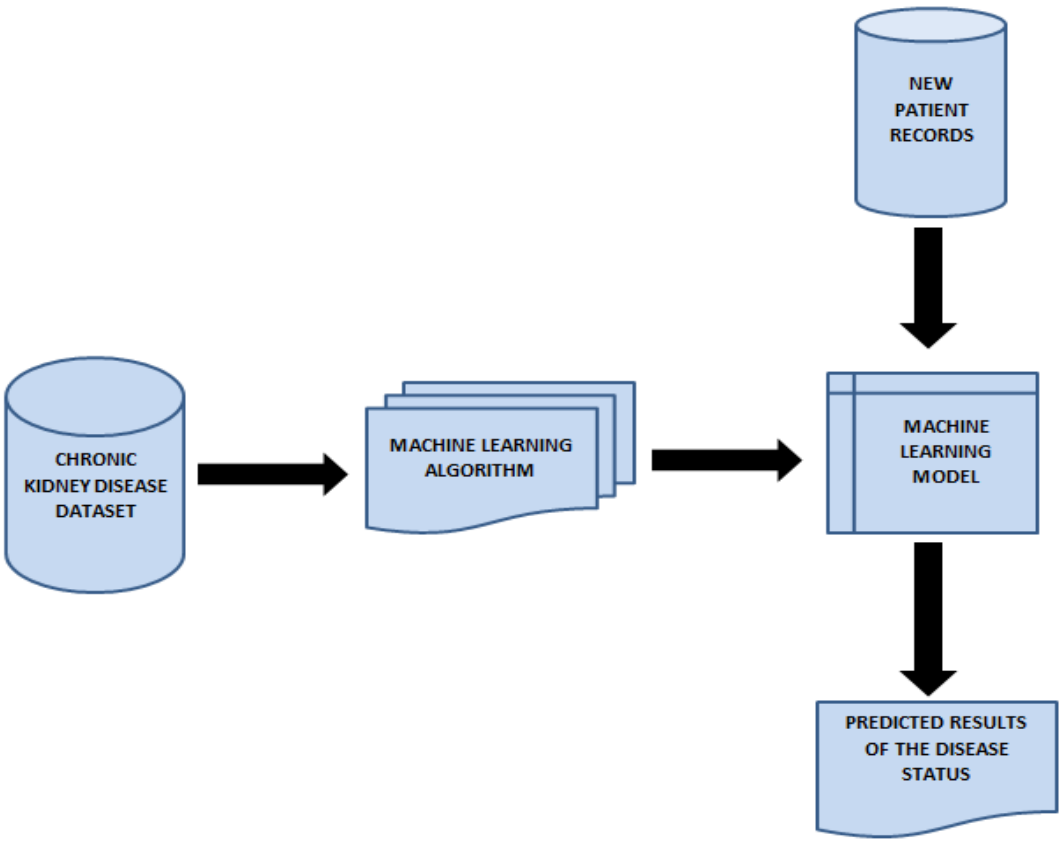
Context of the Research Question- the WHY?

- This Machine Learning Model will facilitate the timely diagnosis of the patients predicted with the chronic kidney disease.
- Early treatment of the patients detected with CKD through my model will
  - ✓ Cure kidney failures
  - ✓ Prevent the situation of the patient from getting worse
- A real-time machine learning system could be deployed in the hospitals as the tool for the doctors for the prevention of the development of the chronic kidney disease in the patient.
- Will save the time as well as the costs of the tests that are needed to be evaluated for the detection of the kidney disease.
- Previous techniques don't handle the **problem of imbalance** in the data which tend to reduce the performance of the medical datasets.
- Previous work also doesn't focus on **optimizing the model** in order to further escalate the model detection accuracy.

Basic Methodology – the HOW?

- This is the **supervised machine learning** problem as the kidney dataset is labeled which means that for every individual patient record we have the associated class – can have the disease or not have the disease.
- It is a **binary classification problem** as we are only having the 2 classes as the output.

High level Machine Learning Work Flow for my problem



Exhaustive Machine Learning Pipeline for Kidney Disease Prediction

Dataset Link - Chronic Kidney Disease Data Set from UCI machine learning repository

https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

Implementation Language – Python 3.7 **& Libraries Used** – Scikit Learn, Seaborn, Pandas

Integrated Development Kit – Jupyter Notebook for presenting the results of each step and PyCharm for debugging.

1) **Exploratory Data Analysis –** Summarizing the main characteristics of the dataset through visualization, correlation and feature engineering.
2) **Data Preparation –** Outlier detection, Missing values, Feature selection and Handling imbalance.
3) **Building and Evaluating Model –** Comparative analysis of performance of machine learning algorithms like Neural Network, SVM and KNN. Evaluating through the metrics such as confusion metrics and area under curve (AUC).
4) **Fine Tuning and Hyper-parameter Optimization –** Techniques of hyper parameter optimization such as grid search and weight optimization in case of neural nets would be considered while dealing with this research problem.