**MAKERERE  UNIVERSITY**

**COLLEGE OF COMPUTING AND INFORMATION SCIENCES**

**(YEAR II) RECESS TERM**

**SOFTWARE DESIGN DOCUMENT**

**FOR: DIABETES PREDICTION MODEL**

(A Case study of Apollo Hospitals, Tamilnadu, India)

By Dr. P. Soundarapandian. MD., D.M (Senior Consult Nephrologist)

A Project Report submitted to the Faculty of Computing and Information Technology for the recess year two 2019 study.

**PROJECT MEMBERS**

|  |  |  |
| --- | --- | --- |
| **NAME** | **REGISTRATION NUMBER** | **STUDENT NUMBER** |
| Kabahinda Rovence | 17/U/4261/PS | 217003233 |
| Meleby Nyika Jino | 17/x/18244/PS | 217019537 |
| Arikosi Michael Okurut | 17/U/2900/PS | 217004064 |
| Mwanje Mike | 17/U/6856/PS | 217004074 |

Supervisor MR KAMULEGEYA GRACE

kougaus@gmail.com

**INTRODUCTION**

Diabetes is a disease in which your blood glucose, or [blood sugar](https://medlineplus.gov/bloodsugar.html), levels are too high. Due to lack of insulin, which is a hormone that helps the glucose get into your cells to give them energy, the glucose stays in your blood thus causing the risk of Diabetes.

According to the 2016 World Health Organization (WHO) Global Report on Diabetes, the prevalence of diabetes and risk related factors has been increasing steadily with the numbers now at 2.7% for males and females at 3.0%, overweight at a total 0f 18.6% and obesity at 3.9% [1].This fact has inspired our model design and the data to be used is obtained from patients’ medical records.

**BACKGROUND**

The dataset (kidney\_disease.csv) published in 2015 was obtained through the UCL Machine Learning Repository website [2]. The source of the data is Apollo hospital in India provided by Dr.P.Soundarapandian.M.D.,D.M (Senior Consultant Nephrologist).

This dataset is properly structured having 400 rows and 25 real features (14 numeric variables and 11 categorical variables).

The features to be majorly used for this model are; Blood pressure (bp in mm/Hg), Blood glucose Random which is the level of glucose in the patient’s blood (bgr in mgs/dl), Sugar levels in the patient (su - (0,1,2,3,4,5)), Blood urea (bu in mgs/dl) and Diabetes Mellitus(dm - (yes,no)).

From the data provided, we intend to predict the Diabetes prevalence based on blood glucose levels, sugar levels, blood urea, and blood pressure.

Eventually, we can determine patients who potentially have diabetes and their age group.

**DATA ANALYTICS APPROACH (DATA PIPELINE)**

**Data pipeline**

**Data loading**

Load kidney\_disease.csv into jupyter notebook.

**Data inspection and cleaning**

Fill all columns except id to address missing values.

Label encode (convert string data to 0’s and 1’s).

Feature scale column su to avoid outliers

**Data classification**

Split dataset into train and test data.

Feed training data into machine learning algorithms to classify(KNN and Logistic Regression)

**Visualization**

Plot regression lines, bar graphs, histograms among others to show relationship and correlations in features

**Model presentationand Deployment**

Reports for supervisor’s evaluation.

Model demonstration

Model roll out for use by medical personnel

**Model evaluation**

Use Precision metrics and recall performance metrics to evaluate the model efficiency.

Choosing most efficient algorithm.

**Loading the dataset.**

The dataset is loaded into the jupyter notebook using python pandas and numpy libraries.

**Data wrangling/ cleaning.**

Under this step there a series of activities are carried out on the data. These include;

* **Filling Missing values**:After loading the dataset, all columns except id hadmissing values. By using the pandas function “.isnull().any()”, a list of the columns with missing values implied by ‘NaN’ was returned. We intend to addressthis problemusing backward fill approach for all the columns except id.
* **Label encoding**: For columns dm, pc, rbc, pcc, ba, htn, cad, appet, and ane. We intend to convert them to 0’s and 1’s for easy manipulation.
* **Feature scaling**:For column su, we intend to do feature scaling so as to eliminate outliers thereby increasing correlation and also increasing the accuracy of the data.

**Visualization.**

We shall placethe cleaneddataset into a visual context so that patterns, trends and correlations. This is intended to expose certain characteristics that might be overlooked by mere eyeballing of the data.

Python has multiple great graphing libraries that come packed with lots of different features to create interactive, live or highly customized plots. The following are a few popular plotting libraries that we intend to use to visualize our dataset (kidney\_disease.csv)

* [**Matplotlib:**](https://matplotlib.org/)This is for us to create basic graphs like line charts,regression curves, bar charts, histograms, scatter plots that we shall use to visualize our dataset.
* **Pandas:** we intend to use pandas library because it provides the best data structures data frames that are easy to manipulate and analyze.
* **Seaborn.** This is under matlab language also helpful in drawing of graphs such as bar graphs, line graphs, regression curves that we shall use to visualize our data set.

**Data classification:**

Classification method consists of a two-steps process and they are training and testing. The kidney\_disease.csv dataset is split into training data and test data. Here, we shall import the ***train\_test\_split*** from ***Sklearn sublibrary\_model\_selection*** so as to split the into training and test sets.

In the training process the model is to classify the training data creating some rules within. In the second step, the training sets are analyzed using the set rules, and the model is tested on how precise, accurate and the sensitive the resulted classes are when compared to the actual known classes. The testing step in the classification method is less computationally expensive compared to the training step as the data in the training set is substantially larger than the training step.We shall use the following machine learning algorithms;

1. **KNearest Neighbor(K-NN):** Classifies the data points according to a voting system, and uses the nearest neighbors to classify the data [3].

***Significance of K-NN***:

One of the simplest classification methods; high convergence speed; proved to be a good method in various datasets including chronic illnesses [3].

1. **Logistic Regression (LG):** A special case of Linear regression. It assumes binary response variable which disturbs the normality assumption of regression models [4].

***Significance of LG:***Able to create simple probabilistic formula to carry out classification. The method was used in several healthcare literatures for classification purposes including Diabetes [4].

Therespective performance results shall be obtained and the best algorithm will be selected.

**Evaluation.**

For our classification problem (Diabetes prediction), predictions shall be made. These predictions shall be performed on unseen data using cross-validation and multiple cross-validation. We shall use classification accuracy (the number of correct predictions made divided by the total number of predictions made, multiplied by 100 to obtain a percentage) and average classification accuracy to measure the accuracy of the predictions.

Classification accuracy, typically, is not enough information to make a conclusive decision on the efficiency of the model. Therefore, we shall carry out **Precision measures** and **Recallperformance measures** to perform final evaluation for diabetes prediction model.

**Model Presentation and Deployment**

To practically transform our model design for presentation to oursupervisor and lecturers, we intend to use the following:

* Sharing detailed written of the development and implementation of the model.
* Demonstration of the working model
* Deploy the model for use by medical personnel and patients.

**SOFTWARE REQUIREMENTS SPECIFICATION**

**Introduction**

This is the description of the whole kidney\_disease.csv dataset pipeline, who the intended users are, what they will use the pipeline for and use case diagrams that describe the interactions the pipeline must provide to different users of the system.

**Intended users of the data pipeline**

The intended users of this document and pipeline include:

* System admin
* Users(student, medical personnel, supervisor)

**Use of the pipeline to different users.**

***System admin***

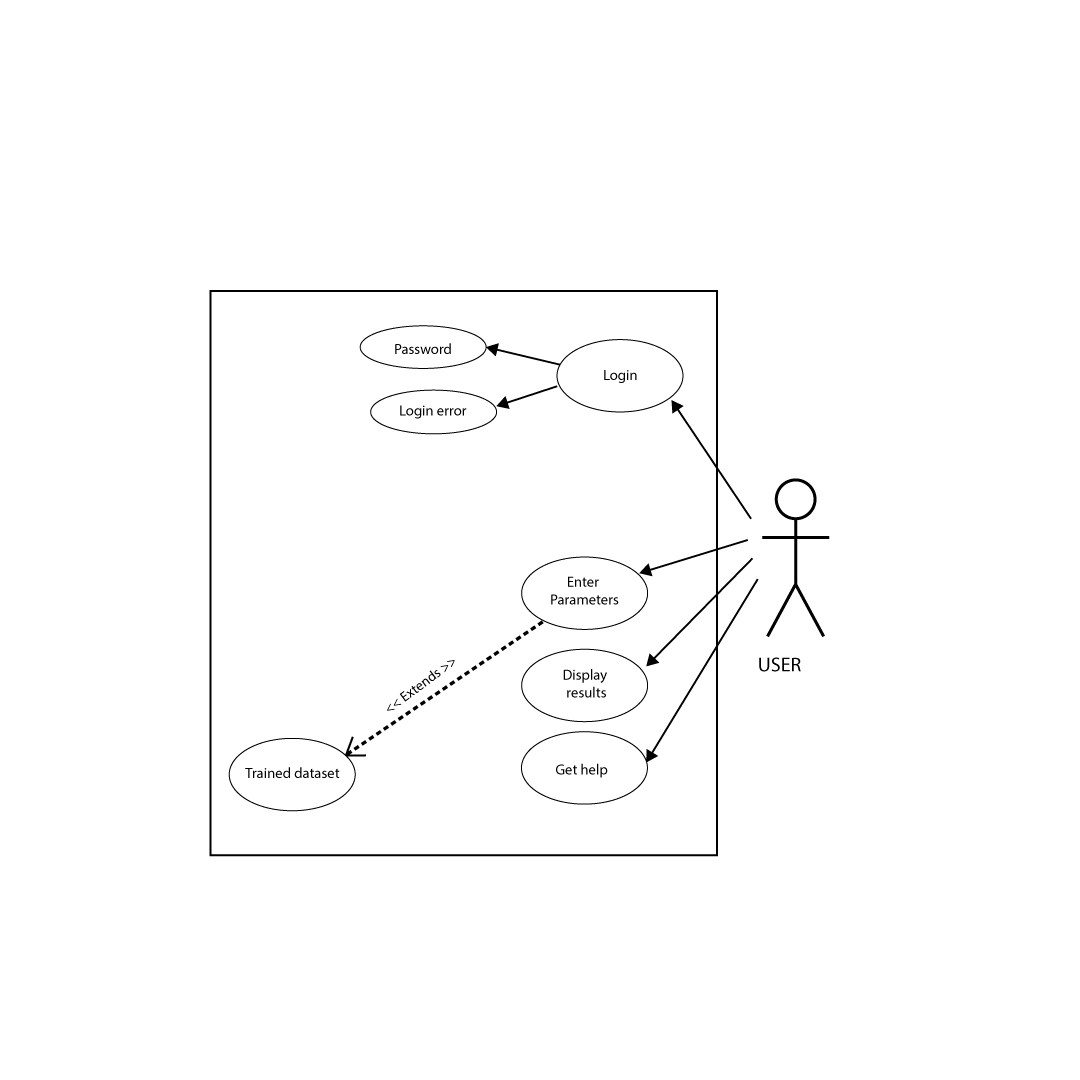
* He/ She will use the pipeline to fine tune the system parameters, train the model and address any errors to ensure the different users have a smooth interaction with the system.

***Users***

* *Student –* will use the data pipeline to gain a proper understanding of the data science concepts undertaken to accomplish the diabetes detection task.
* *Medical personnel –* will use the data pipeline to understand how the model will use the provided features to determine if a person has diabetes or not.
* *Supervisor –* will use the data pipeline to verify if the model was created in the right way to attain the best results as a way of assessing the student.

***USE CASE DIAGRAMS FOR THE PIPELINE***

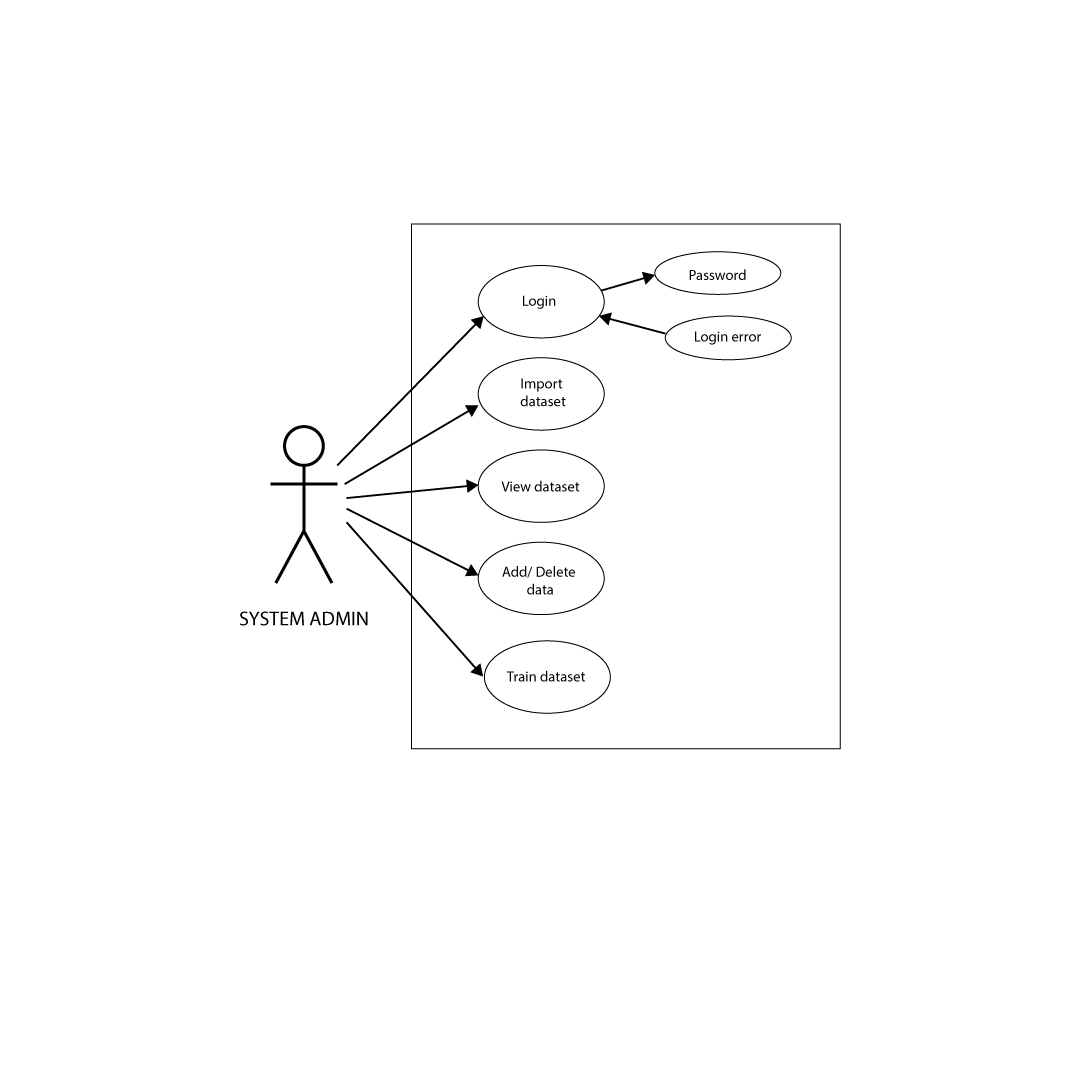
1. **User**



**Description**

* The user requests login in to the system.
* The user credentials are verified before login. If credentials are correct, the user successfully accesses the system. If not the, the login error message is displayed.
* The user can enter required parameters like blood glucose levels, blood pressure, blood urea, and sugar levels to be used by analyzed by the model
* The parameters are passed to the trained dataset for analysis.
* The results are output are then displayed to the user
* The user can request for help on how to user the model and instruction are returned accordingly.

1. **System administrator**



**Description**

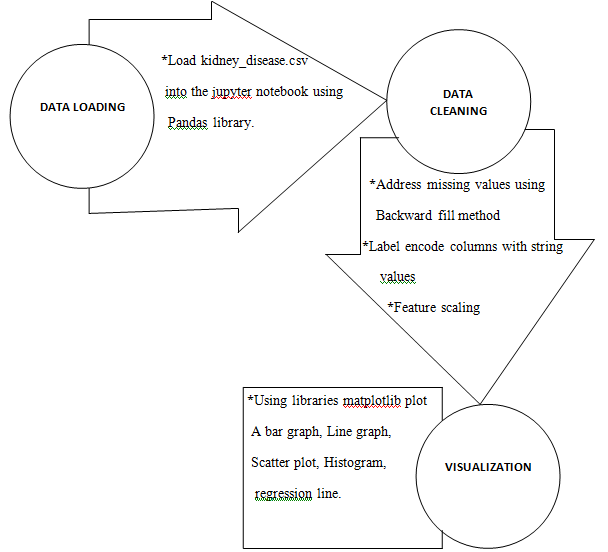
* The user requests login in to the system.
* The user credentials are verified before login. If credentials are correct, the user successfully accesses the system. If not the, the login error message is displayed.
* The systems admin imports the dataset compatible with the model
* The data in the dataset can then be viewed by the user.
* The system admin can add new rows to the dataset, delete rows from the dataset and modify existing data in the dataset.
* The admin can then use the data in the dataset to train the predictive analysis model.

**SOFTWARE DESIGN DOCUMENT**

**Introduction**

The System Design Document consists of the whole data pipeline from data sources all the way to visualization, a description of all the key components in this data pipeline, what it is, what each component does and how it does it. It also includes the description of different python libraries that we intend to use during the implementation of our project and the reasons as to why we chose to use these particular libraries. It also includes the visualization components for our pipeline, the visualizations we intend to use and why we think they are the best among other visualization tools.

**Diagram showing data pipeline up to visualization**



*Source: diagram generated by group members using Microsoft word*

**Data loading**

We intend to load the kidney\_disease.csv dataset into the jupyter notebook using *pandas library*, a data manipulation library.

This we shall do by importing pandas library first by using the statement “*import pandas as pd*” and so after this statement, the jupyter notebook will understand that from then every time we type pd, we are referring to pandas. After importing pandas we shall read the dataset into a pandas data structure that is; a dataframe (*df*) using *read\_csv()* function so that every time we run the code we do not have to run the *read\_csv()* function but instead simply call data frame (df).

And for us not to flood the whole screen with data, we decided that we use functions *df\_read.tail()* and *df\_read.head().* These help to automatically generate the last and first 5 elements of the data frame respectively.

***Why use pandas and then pandas dataframe?***

Pandas library provides data structures (*pandas dataframes* and *pandas series*) and data analysis tools for different functions.

A pandas data frame is a 2 or (more dimensional) data structure basically a table with rows and columns. The columns have names and the rows have indices

Just like excel files, pandas data frames provide various functionalities that will be used in the analysis, changing and extracting of valuable information from the kidney\_disease.csv dataset.

Data frames also provide various column helper functions which shall extremely be useful for extracting valuable information from the column for example; *unique* - provides unique elements from a column by removing duplicates, *mean* - provide the mean value of all the items in the column and among others

.

**Data cleaning/ wrangling**

Our data cleaning process shall involve fine tuning the kidney\_disease.csv dataset to transform it to a state that will help us attain the best possible results in the grand scope of the diabetes detection problem. Since our data set has missing values, we shall address this problem by using **interpolation**for the numeric columns and **backward fill** for the nominal (columns with string values) columns. The interpolation method uses the interpolate() function which fills in a missing value basing on a specified domain of values around it.

**Backward fill** on the other hand uses bfill() function on available values in a column to fill in precedent missing values in the same column. Both functions utilize the pandas library and NumPy libararies and the major reasons for the use of these two libraries will be discussed in the tail of this section.

The next step of our data cleaning process shall be **label encoding**. This means mapping different string values of a feature to distinct numeric values and it shall be achieved using the *.map()* function in the *numpy library* which lets a user specify strings and assign them to a distinct corresponding numeric values. This operation will be carried out on the rbc(reb blood cells), pc(pus cell), pcc(pus cell count), ba(bacteria), htn(hypertension) and dm(diabetes mellitus**)** features of the kidney\_disease.csv dataset to encode their nominal values to binary 0’s and 1’s since their values vary between normal or abnormal, present or non-present and yes or No.

The final phase shall involve handling outliers in each column by feature scalingthe numeric columns of wc , bgr and bu which stand for white blood cell count, blood glucose and blood urea respectively whose values span over a wide range. This step will help our machine learning algorithm K-Nearest Neighbor (KNN) to perform better in the diabetes disease prevalence problem and this shall be made possible by the help of the *scikit learn library* using the *Robust Scaler function* which works by removing the median and using the interquartile range (Q1-Q3) of the values hence it focuses on the parts where the bulk of the data is.

***Reasons for using the libraries mentioned above:***

* **Pandas–**it is an open source and high performance library that is easy to learn, easy to use and also easy to maintain. It also will us provide a multitude of functions for data manipulation, analysis and modelling hence allow us to focus more on developing an accurate model for the diabetes disease prevalence model rather than spending time programming functions that have already been coded and tested. Pandas also provides a fast and efficient data frame object for data manipulation, less relevant columns such as pc(pus cell) from our kidney\_disease dataset can be deleted/ dropped . Pandas also has widespread community support so in case of any difficulties concerning usage of some its functions then we can get help online.
* **Numpy –**it is the fundamental package for scientific computing with python. It will provide us with a wide variety of sophisticated functions to carry out mathematical computations and analysis on our kidney\_disease.csv dataset. It is also widely used hence there is support available online in case of challenges when using it during our implementation.
* **Scikit learn** – it shall provide us with a range of learning algorithms through a consistent interface in python. It will help us to model data and do feature selection to identify meaningful attributes from which to create our best possible model. It will help us in parameter tuning to get the most accuracy out of our model and also in splitting our dataset to train our model.

**Data visualization**

This section is a brief description of how we intend to help our users (medical personnel, supervisor, data scientists) and the system administrator to understand the significance of the kidney\_disease.csv dataset by placing it in a visual context. In other words, it describes the movement from raw data that is; kidney\_disease.csv dataset to meaningful insights and learning

In the first place, we shall have to import ***matplotlib*** into the workspace (jupyter). This matplotlib library has a module called ***pyplot*** which is used to create figures, create a plotting area within the figure, plot some lines or dots within the figure, decorate plots among other functionalities.

We shall import the pyplot module into our workspace by adding a statement; “ *from matplotlib import pyplot as plt* ” or “*import matplotlib.pyplot as plt*” incase matplotlib is already imported

***Why use matplotlib.pyplot?***

* It supports a very wide variety of plots and graphs for example scatter plots, regression lines, histograms and bar graphs.

***Why plot the graphs mentioned above?***

* Scatter plot
* Barplot
* Histogram
* Regression line

*Scatter Plot*: this describes the value of any ­­­­two sets of data on two dimensions. Each dot represents an observation. It is relevant in our diabetes disease prevalence problem because it is good for the study of the relationship between two variables. In addition, it shall provide more information about our data by using colors or shapes to show groups or any other third variable

*Histogram*: This is made up of columns plotted on a graph. Usually there is no space between these columns. This shall be used to show the skewdness of the data

*Bar plot:* This is made up of columns plotted on a graph also just like a histogram. Usually there’s space between the columns. This shall be used to show the distribution of the data

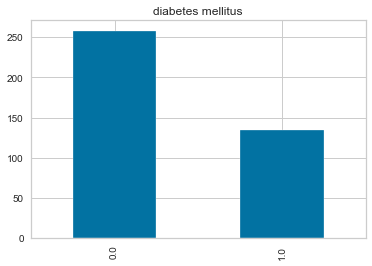
*Regression line*: This shall be used for prediction and forecasting where its use has a significant overlap with machine learning. It shall be used to understand which among the independent variables are related to the dependent variable and to explore the forms of these relationships.

**IMPLEMENTATION REPORT**

**Introduction**

This document is a brief description of what kinds of visualization tools were used during our diabetes prevalence analytics pipeline,what insights each tool gives us into our data. Some of these insights might be already known (but perhaps not yet proven) while others are completely new and surprising and also key recommendations to the key stakeholders on what these insights mean and what they should be used for.

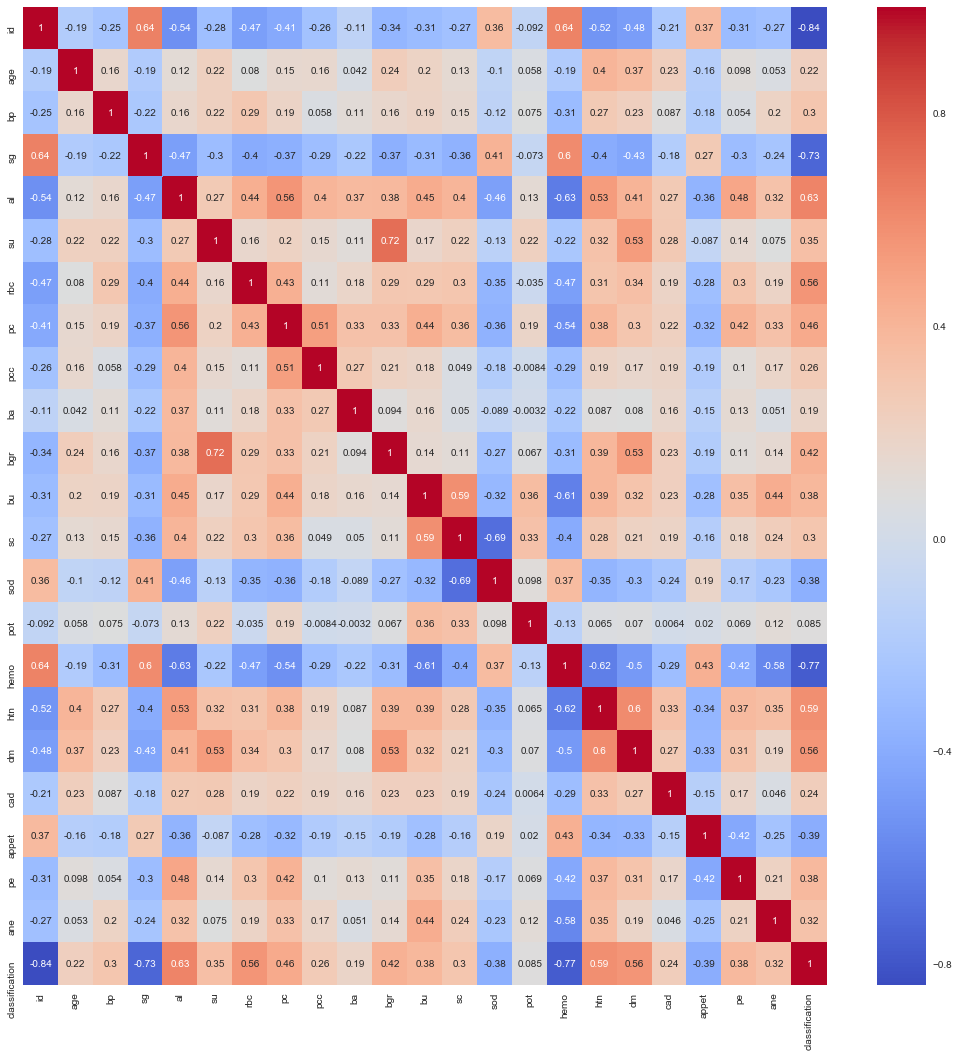
To begin with is a bar plot which is very good and well known for comparing categorical data



From the context of our diabetes prevalence prediction, the bar plot above shows that of 400 patients whose records were taken at Apollo Hospitals, 134 of them have diabetes mellitus while 258 patients did not have diabetes .

Since our dataset has around 20 features and 400 rows. A good way to quickly check correlations among these 20 features is by using heat map.

**A correlation heatmap showing correlation among all the features with diabetes mellitus (dm)**

****

Colour palletes are used to represent what kindof correlation exist between which features. In this case; Red means positive while blue means negative. The stronger the color, the larger the correlation magnitude.

Checking out correlations of other features with diabetes mellitus (dm) from the heat map above, it is found out that features age, bp, al, su, rbc, pc, pcc, ba, bgr, bu, sc, htn, cad, pe, ane and classification have a positive correlation unlike the rest of the features.

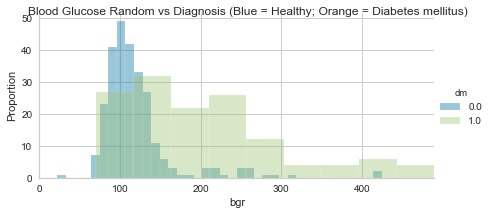
Among the features with the correlation su, bgr ,htn, class have a stronger correlation with diabetes mellitus(dm) which is not the case with other features.

Now, to decide which features to drop, the heatmap above was analysed and those features with the lowest correlation with diabetes mellitus(dm) in our diabetes prevalence prediction were dropped.

Therefore, the features id, sg, sod, pot, appet and hemo fall under this category hence were dropped before proceeding with more visualization of the kidney\_disease.csv dataset.

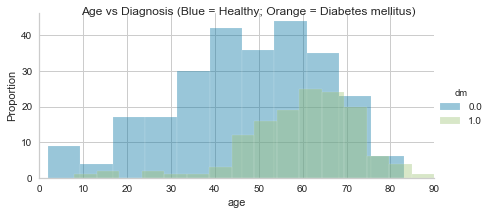
The reason for dropping of the above features will further enable our future model to easily learn with the most relevant features (features that have a positive correlation with diabetes mellitus (dm)) and thereafter evaluate its performance using the three performance metrics of precision, recall and first score.

**A histogram showing of blood glucose random ( bgr ) against diagnosis**



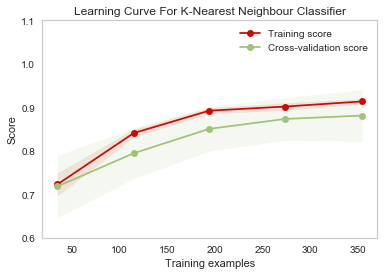
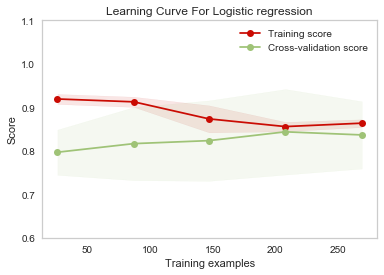
From this visualization, it is clear that the higher the blood glucose levels, the higher the chances of the patient bearing diabetes mellitus.

**A histogram showing of age against diagnosis**



From the visualization it can be said that the more a patient ages, the higher the chances of acquiring diabetes mellitus. Typically a person from around the age of 45 years are in greater risk.

**Learning Curve**



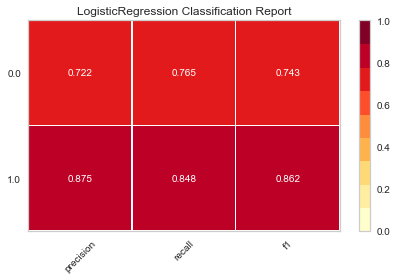
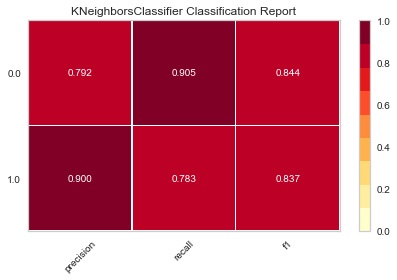
The Training and validation score for **Logistic Regression** Model converges to a low score (value ~ 0.8). The gap between the curves is bigger, when all the training data is used. Therefore, we will not benefit from adding more data to train the model since it has a high bias characteristic. This curve depicts an under-fit model; model is not able to learn efficiently on training data.

We will probably have to use an estimator or a parametrization of the current estimator that can learn more complex concepts (i.e. has a lower bias).

For **K Nearest Neighbor Classifier**, the training score is greater than the validation score for the maximum number of training samples and converges towards a higher accuracy score; this implies that the model would benefit from adding more training data to increase generalization. This curve is also a good-fit, and has a low bias and high variance characteristic.

Therefore, this means that the K-NN model is more efficient than the Logistic Regression model.

**YellowBrick Classification report**



The classification reports shows a representation of the main classification metrics (precision, recall and f1 score) on per-class basis (1.0, 0.0). Thus giving a more detailed and deeper intuitive summary of the classifier behavior over global accuracy.

**Logistic Regression**has unbalanced classification metrics for its classes. Axis for Class 1.0 is redder than 0.0 thus it is has higher precision, recall and f1 score. This imbalance weakens the overall performance of this model.

**K Nearest Neighbor** classifierhas relatively balanced classification metrics for its classes. Axis for Class 1.0 and 0.0 have strong precision, recall and f1 score. This strengthens the overall performance of this model.

Overall, K Nearest Classifier is redder than Logistic Regression hence it is a. more efficient model selection

**Deployed API with Flask on localhost Server.**

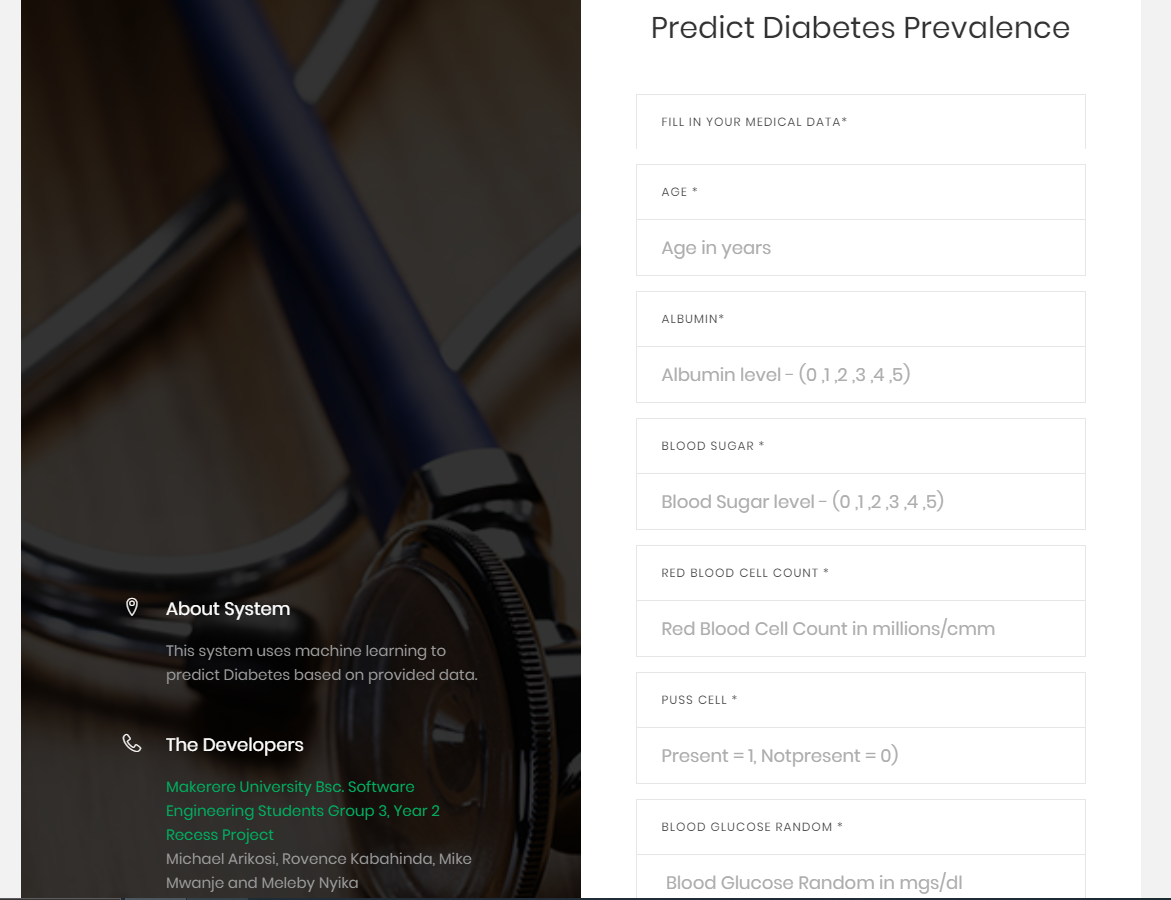
****

Figure 1 User Requested to enter required data to run the Diabetes Test Prediction

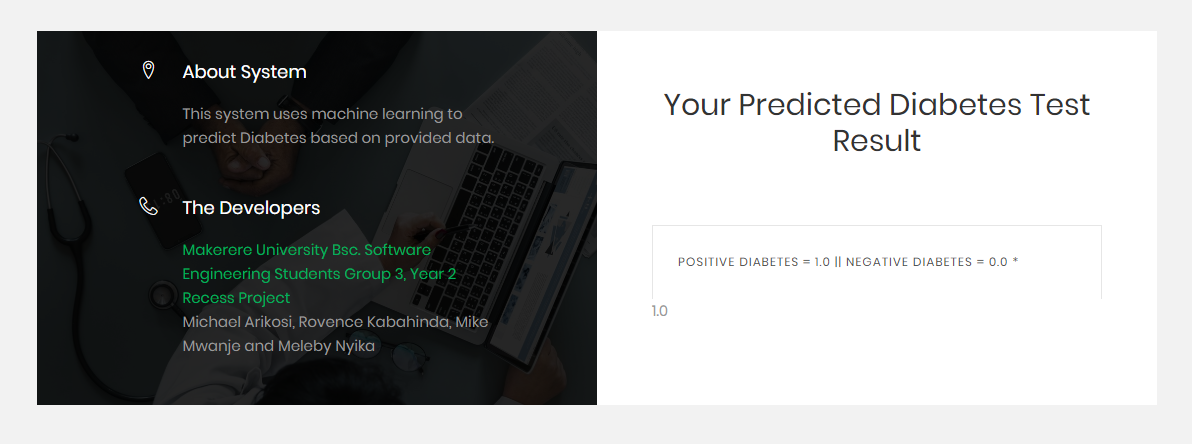
****

Figure 2 Results returned to the user after prediction API has finished processing

**Recommendations**

1. Add more observations (rows)to the dataset to improve overall accuracy and the predictive strength of the model as it has more data from which to learn.
2. Add more complete observations to the dataset having nonull cells. This reduces the negative implications of backward filling for null values in the data. Backward filling causes uncertain distribution of data, which may skew the accuracy of the model. More accurate and real data helps to build a more realistic and accurate model.
3. We recommend dropping the Null rows when the dataset grows in size and has numerous observations with completely filled features. This will increase accuracy and predictive strength by working with only real and accurate data.
4. Increase the balance between the classification features (dm) i.e. Ratio of (0.0:1.0) in the. This reduces the negative implications of oversampling data in K- Neighbor Classifier model as it tries to balance the classification Classes.
5. Drop features that are irrelevant to the training of the model. These features will skew the accuracy as they cause a high bias trade-off in the learning of the model. These features have a very low correlation coefficient relative to (dm) ie less than 0.3.

**REFERENCES:**

1. Prevalence of Diabetes in Uganda on the Rise, World Health Organization, 2016 [online]<http://who.int/diabetes/global-report/en/>
2. Chronic\_Kidney\_Disease Data Set, “UCI,” 21 11 2016. [online]. Available: https:\\www. archive.ics.uci.edu.
3. Tomar, D., and Agarwal, S., 2013, “A survey on Data Mining approaches for Healthcare,” International Journal of Bio-Science and Bio-Technology, 5(5), 241-266.
4. Yeh, I. C., and Lien, C. H., 2009, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications,” 36(2), 2473-2480.