**MAKERERE  UNIVERSITY**

**COLLEGE OF COMPUTING AND INFORMATION SCIENCES**

**(YEAR II) RECESS TERM**

**CONCEPT PAPER FOR: DIABETES PREDICTION**

**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 |

**PROJECT COORDINATOR**: MR. KAMULEGEYA GRACE

**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 presentation and 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 had missing values. By using the pandas function “.isnull().any()”, a list of the columns with missing values implied by ‘NaN’ was returned. We intend to address this problem using 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 place the cleaned dataset 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. **K Nearest 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].

The respective 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 **Recall** **performance measures** to perform final evaluation for diabetes prediction model.

**Model Presentation and Deployment**

To practically transform our model design for presentation to our supervisor 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.

**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.