Capstone Project

Project proposal

1. Group description

1.1. Group name

Tdata

1.2. Students names, background and target industry if any

Names

Bariki Elilaki
Florence Mayo
Innocent Ngowi
Lucas Katisho
Raymond Alick

Background
Software Developers

Target Industry
Government, Insurance Companies , Medical center

1.3. Group structure: roles and responsibilities

Student	Roles	Responsibilities
Bariki	- Database Design and Administration - Data engineering	- Designs the SQL database in MySQL To handle the Data sets from the five years and data engineering
Florence	- Team Organiser - Project Documentation - Data Preprocessing	- Setups time and workflow - Documenting project outcomes and deliverables
Innocent	- EDA - Data Preprocessing	- EDA and data Cleaning
Lucas	- Data engineering - Supervised Machine learning	- Explore data and develop some models to do predictions

Raymond	- Data engineering - Supervised Machine learning	- Explore data and develop some models to do predictions

2. Why do we want to develop a data science project?

2.1 Objective: what problem do you want to solve? What questions are you trying to answer? How will you **measure the success** of your analysis from a business/user perspective?

Objectives

- To perform data preprocessing from Fixed Width File To a relational Database.
- To perform Exploratory Data Analysis on the datasets.
- To develop a supervised machine learning model to predict Probability of admission into the NICU for newborns.

Measure the success

- The successful creation of a model for predicting Admission to The NICU.

2.2. Scope of application: what population and timeframe will your analysis/model be applied to or used for?

- Population:The Project will only cover the last five(5) years (2014-2018) of Dataset from Center for Disease Control and Prevention(CDC)

- Timeframe: Two weeks

3. How do you translate the objective and scope in terms of data?

3.1. What **dataset**(s) do you plan to use? Initial description: source, granularity, number of observations, variables list…

Dataset

- Birth Data Files from 2014 to 2018
- Source: From the Center For Disease Control And Prevention
- No of observation: 5 fixed width files of different five year each of at least 200
- Variables : Will be selected or created from more than 200 variables

3.2. What **data treatment and analysis** do you plan? Data Aggregation, target variable definition, tools, analysis/machine learning, ...

Data preparation

- Download the zipped files from https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm#Downloadable
- Convert the fixed width .txt files to .csv files
- Design the needed database tables ready to import generated csv files
- Import generated .csv files to the sql database

Target variable

- AB_NICU (Admission to NICU)

Tools

- Python for data manipulation
- SQL Database for data storage (Mysql/MariaDB)
- Pycharm, Visual Studio Code for text editing of code.
- Sequal Pro,PhpMyadmin and MySql Workbench for easy interaction with mysql.

Analysis

- Exploratory data analysis: univariate and bivariate analyses of variables against AB_NICU (Admission to NICU).

4. Project plan

Project plan and schedule