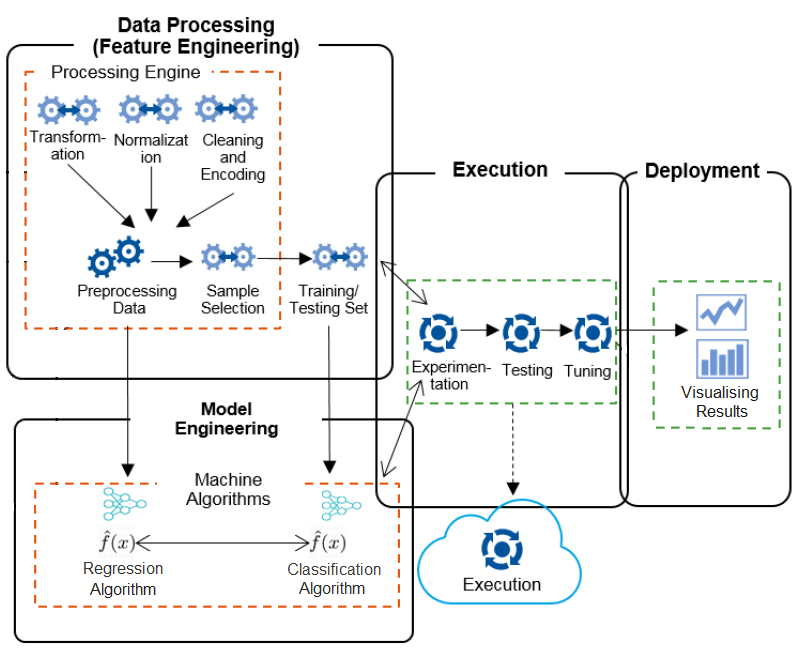
Cohort 12

NLMS Team

Project Architecture and Design Specifications



Using NLMS national patient data set, the purpose of this project will be to test a hypothesis that certain factors help explain variations in the probability of mortality. Statistical models such as Cox proportional hazards (Cox PH) regression analysis, Logistic regression and Random Forest classification will be used to predict probability of mortality.

Project will be broken down and executed in the following stages to achieve the above stated objectives:

**Stage 1: Data Processing**. Available data set in CSV format will undergo data analysis and wrangling to remove outliers, normalize data and infer missing variables when possible or if appropriate. Relational database management system (RDBMS) SQLite will be used at this stage to facilitate working with data.

**Stage 2: Model Engineering**. At this stage Cox PH, Logistic regression and Random Forest classification models will be developed using Python libraries and packages, such as Pandas, NumPy, Scikit-Learn, and training data set will be fitted to the models. Best fitting model will be trained using train part of the dataset, i.e. 80%, and then cross-validated by predicting mortality outcomes for the test part of the data, i.e. remaining 20% of the patient records.

**Stage 3: Execution**. Following model engineering stage, the team will focus on experimenting with multiple combinations of the dependent variables to be used in the models, to identify the most relevant set of variables yielding highest accuracy score for the models’ prediction performance. Testing models consistent with hypotheses and tuning of the models will be performed at this stage.

**Stage 4: Model Deployment & Visualization**. Tableau visualization will be created based on selected model to plot dependency of mortality on a number of statistically significant variables. Visualization of geographical distribution of patient population, distribution by age, gender, income etc. will also be considered.

**Project Toolkit:**

**Stage 1:**

* SQLite

**Stage 2 & 3:**

* Python 3.6
* NumPy
* Pandas
* Scikit-Learn

**Stage 4:**

* Matplotlib
* Tableau Public