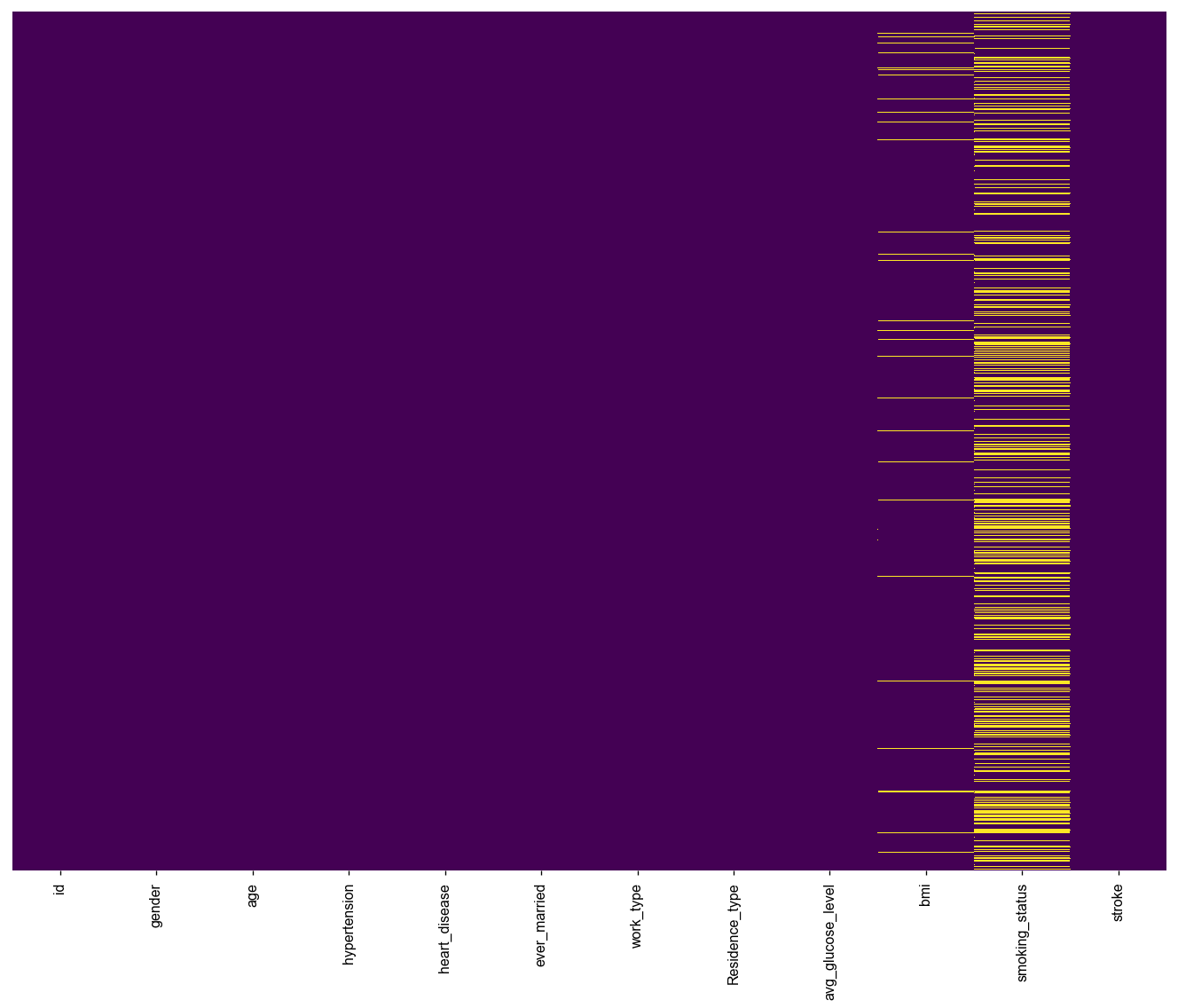
Stroke Clinical Predictive Model

## Clinical Predictive Model classifier for Stroke: Stroke-CPM is a tool that takes information available about a patient and their observed predictive factors and makes a prediction regarding their diagnosis and causal factors.



After checking for missing values, smoking status and bmi had missing values which could lead to drawing an inaccurate inference about the stroke classification

#### Multiple Imputation for Missing Data: To prevent bias in prediction, missing values are predicted and imputed into the stroke dataset. Using MICE package, Predictive Mean Matching and Poly Regression techniques are used to impute data missing at random from smoking status and bmi

#### Exploratory Analysis and Class Imbalance: Stroke class from exploratory graph showed imbalance in stroke outcome, hypertension predictor and heart disease. Fitting a model on this imbalanced dataset would lead to overfitting whereby the trained model crams the train set data.

In order for us to predict the minority class in the imbalanced variable, the factor variables need to be encoded to vector types to enable prediction through oversampleing of minority class.

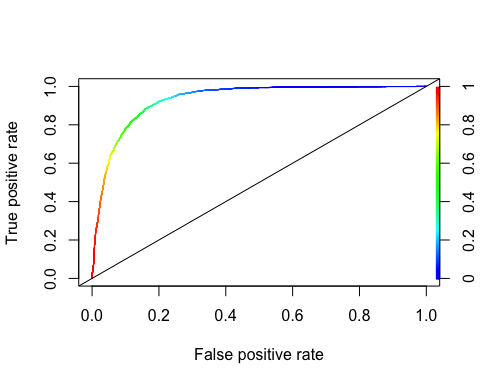
To predict the minority class in stroke, the train set data is exposed to SMOTE package to use K-Nearest Neighbor for predicting minority class.

Cross Validation:For good predictions of stroke, a model must be well calibrated and have high discrimination. The input data was divided into two groups, in order to fit the data in one group and validate it in the other. This helps to reducing training bias.

#### Model Training - Logistic Regression: To classify patients who have stroke or do not have stroke and to check for contributing causal factors to stroke, a binary classifier is built to predict the cases of stroke. State-of-the-art binary classifier Logistic Regression is applied. As per the TRIPOD guidelines, we would need to report this stroke clinical predictive model in sufficient detail to allow for reproducability on a totally dufferent dataset.

Used stepwise AIC technique, the trained model for stroke has dropped the variables BMI and Reesident Type as they are not contributing factors that would lead to stroke. Stroke classification is non-preedicttive from the above 2 variables, and they have non-significant p-values which are > 0.05.

#### Discrimination and Calibration of Logistic Regression Classifier: Using the Discrimination, we check if trained CPM has a simultaneously high sensitivity and specificity. A receiver operator characteristic (ROC) curve is ploted using sensitivity against ‘1-specificity’ across the full range of potential cutpoints at 93%.



The ROC Curve shows further to the top left of the graph, this indicates a great test.

To determine how far to the top left of the graph the curve is, the Area Under the Curve (AUC) is calculated. So, the AUC for the Logistic Regression Classifier is 0.92 which represents a near perfect discrimination.

Calibration using Hosmer and Lemeshow goodness of fit test shows significant p-value which indicates model validity

#### Conclusion

Overall logistic regression was selected as the best model to predict if a patient can have stroke or not. To achieve this and reduce the potential of bias and oveerfitting of trained model, imbalanced data are sampled because it is a common hurdle in healthcare. Also performed multiple imputation to predict missing data of smoke status in other ways as well