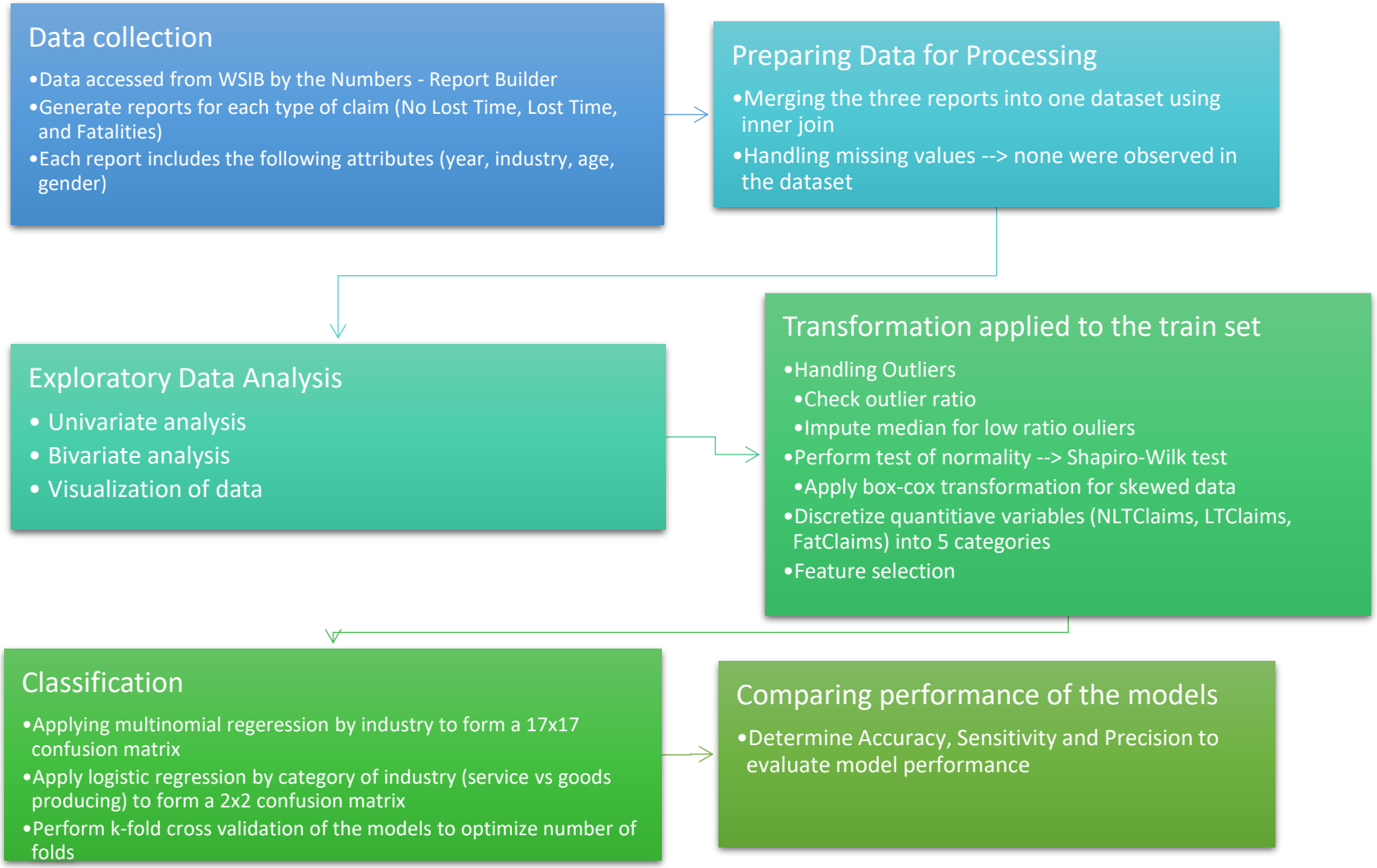
Data collection
•Data accessed from WSIB by the Numbers - Report Builder
•Generate reports for each type of claim (No Lost Time, Lost Time, and Fatalities)
•Each report includes the following attributes (year, industry, age, gender)
Preparing Data for Processing
•Merging the three reports into one dataset using inner join
•Handling missing values --> none were observed in the dataset
Exploratory Data Analysis
• Univariate analysis
• Bivariate analysis
• Visualization of data
Transformation applied to the train set
•Handling Outliers
•Check outlier ratio
•Impute median for low ratio ouliers
•Perform test of normality --> Shapiro-Wilk test
•Apply box-cox transformation for skewed data
•Discretize quantitiave variables (NLTClaims, LTClaims, FatClaims) into 5 categories
•Feature selection
Classification
•Applying multinomial regeression by industry to form a 17x17 confusion matrix
•Apply logistic regression by category of industry (service vs goods producing) to form a 2x2 confusion matrix
•Perform k-fold cross validation of the models to optimize number of folds
Comparing performance of the models
•Determine Accuracy, Sensitivity and Precision to evaluate model performance