**MACHINE LEARNING DETAILS**

**FEATURE ENGINEERING DETAILS**

1. **pyodbc:** We used pyodbc package in python 3.6 for connecting to the SQL server to retrieve the echo data stored in our database.
2. **pandas:** We used python 3.6 pandas library for data manipulation and analysis.
3. **re:** We used regular expression in python for text processing [cpython/re.py at 3.10 · python/cpython · GitHub](https://github.com/python/cpython/blob/3.10/Lib/re.py)
4. **nltk:** We used ‘english’ stopwords from nltk library and ‘tokenize’ to achieve tokenization in python 3.6.
5. **sklearn:** We used ‘countvectorizer’ to create a matrix of token counts. max\_df and min\_df has default settings. We ignored terms that appear in more than 100% of the echo reports and less than one echo report. We used ‘fit\_transform’ for scaling and used ‘TfidfTransformer’ to transform a count matrix to a ‘tf-idf’ representation. ‘labelencoder’ is used to encode the target values in python 3.6.
6. **imblearn:** We used imblearn library in python 3.6 and utilized ‘smotenn’ for oversampling and ‘EditedNearestNeighbours’ for under sampling the minority class. The sampling strategy is set to ‘all’

**MACHINE LEARNING IMPLEMENTATION DETAILS**

1. **StratifiedKFold:** We used StratifiedKFold obtain training and test data with 5 folds to obtain equal number of training and test sets. The ‘random state’ is set to 42 and ‘shuffle’ is set to ‘true’ in python 3.6.
2. **classification\_report:** ‘classification\_report’ from ‘sklearn.metrics’ is leveraged to define the classification report with accuracy score in python 3.6.
3. **LinearSVC:** ‘linear SVC’ is imported from sklearn.svm to train the SVM model in python 3.6.
4. **RandomForestClassifier: ‘**RandomForestClassifier’ is imported from sklearn. ensemble to train the ‘Random Forest’ model in python 3.6.
5. **XGBClassifier:** ‘XGBClassifier’ is imported from ‘xgboost’ to train the extreme gradient boosting model in python 3.6.

**DATA RESAMPLING RESULTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Outcomes of Interest** | **Stratified Folds** | **Training set size** | **Test set sample size** |
| **Mitral Regurgitation** | Fold 1,2 | 85,441 | 21,361 |
| Fold 3,4,5 | 85,442 | 21,360 |
| **Mitral Stenosis** | Fold 1 to 5 | 153,304 | 38,326 |
| **Aortic Regurgitation** | Fold 1,2,3,4 | 115,347 | 28,837 |
| Fold 5 | 115,348 | 28,836 |
| **Aortic Stenosis** | Fold 1 | 146,372 | 36,594 |
| Fold 2,3,4,5 | 146,373 | 36,593 |
| **Flow Gradient across Aortic Stenosis** | Fold 1,2 | 66,913 | 16,729 |
| Fold 3,4,5 | 66,914 | 16,728 |
| **Diastolic Function** | Fold 1,2 | 39,485 | 9,872 |
| Fold 3,4,5 | 39,486 | 9,871 |
| **Ejection Fraction (Numeric)** | Fold 1 to 5 | 34,572 | 8,643 |
| **Ejection Fraction (Ranges)** | Fold 1,2,3 | 94,218 | 23,555 |
| Fold 4,5 | 94,219 | 23,554 |
| **Ejection Fraction (Qualitative)** | Fold 1,2,3 | 506 | 127 |
| Fold 4,5 | 507 | 126 |

**STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR MITRAL REGURGITATION**

**A table with numbers and letters

Description automatically generated**

**STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR MITRAL STENOSIS**

**A table with numbers and symbols

Description automatically generated**

**STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR AORTIC REGURGITATION**

**A table with numbers and symbols

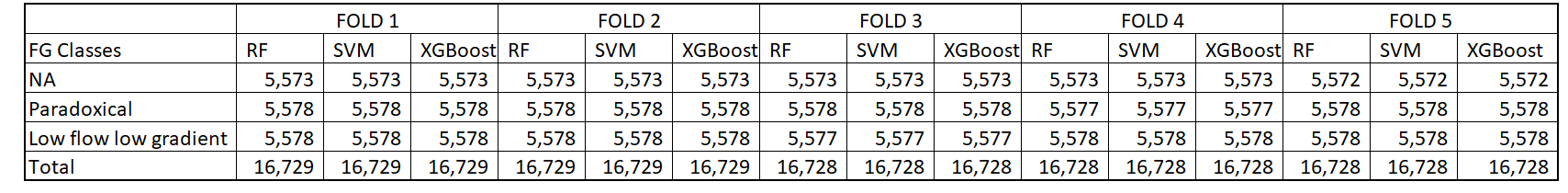
Description automatically generated**

**STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR AORTIC STENOSIS**

**A grid of numbers and numbers

Description automatically generated**

**STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR FLOW GRADIENT OF AORTIC STENOSIS**

****

**STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR DIASTOLIC FUNCTION**

**A table with numbers and symbols

Description automatically generated**

**STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR EJECTION FRACTION (NUMERIC)**

**A screen shot of a table

Description automatically generated**

**STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR EJECTION FRACTION (RANGES)**

**A table with numbers and letters

Description automatically generated**

**STRATIFIED SUPPORT SAMPLES FOR EACH CLASS IN A BALANCED TEST SET OBTAINED USING SMOTE – ENN METHOD IN FIVE FOLDS FOR EJECTION FRACTION (QUALITATIVE)**

**A grid of numbers and symbols

Description automatically generated**