# 732A92 Text Mining
# Multi-class classification of clinical specialities from medical transcriptions

Farhana Chowdhury Tondra (farch587)

2022-03-10

# Abstract

Clinical text-based data has been explored to classify diseases based on symptoms and to classify medical documents. Supervised and unsupervised learning methods have been used in these works and researchers have published results supporting both methods. The recorded clinical transcriptions typically belong to a multitude of medical specialities and are stored in a large database. Manual parsing of such database is tedious and, in some cases, not feasible. However, an automated system utilizing a well-trained classification model can perform the job with minimal human interaction. The primary focus of this project is to classify medical transcriptions using four well known supervised classification techniques, namely Multinomial Naive Bayes (MultinomialNB), Linear Support Vector Classifier (LinearSVC), Logistic Regression and Random Forest. Text data was normalized and tokenized by Term Frequency- Inverse Document Frequency (TF-IDF) matrix. To improve the classification work, four re-sampling methods (RandomOverSampler, SMOTE, RandomUnderSampler, SMOTENN) and a feature reduction method, truncatedSVD was used before applying classification methods. Among the used four models, LinearSVC achieved highest accuracy $(81\%)$ after applying RandomOverSampler for the medical transcriptions data.

# Contents

# List of Figures

# List of Tables

# Introduction

Clinical text data is the bridge between health care system and medical research. The detailed text of medical history of patients are effective for retrieving clinical assessments and clinical decision making. Natural language processing (NLP) is a branch of artificial intelligence (AI) which is used to perform text data identification and topic classification. Medical text based on clinical assessments and patients disease history is taken by a general physician or a nurse for further assessment or escalating to the specialist doctor of relevant department in a hospital. Natural language processing can facilitate automation of the process. A model can be trained based on stored medical descriptions and the new transcriptions can be redirected to the appropriate specialists based on text analysis. This project work is dedicated to learn clinical data properties and find the association between different specialized sectors and the transcriptions. It aims to classify different medical specialities using multi-class classification techniques. Below steps on the input data were applied in the work :

* Data preparation and processing: The processing of dataset consists of balancing the data, removal of NULL values and filtering punctuation marks, numbers and stop words.
* Word Vectorization: The text-vectorized method used in this project is term frequency-inverse document frequency.
* Classifications : Two baseline classification methods such as Multinomial Naive Bayes and Logistic Regression, a kernel based classifier, namely Linear Support Vector Classifier and a decision tree based ensemble learning method, namely Random Forest Classifier has been used to classify the text vectors with grid search cross validation. Afterwards, two classification methods was selected based on their performance; input data was re-sampled and applied to the selected classification methods to further improve the results.
* Evaluation: The performance of all classifiers is compared based on their accuracy, precision, recall, and f1-score.

# Theory

Normalization of text data, which refers to conversion of the raw text into a convenient and standard form, is the first step of any text classification analysis. There are many forms of normalization of text data; the below sections describe the normalization process used in this work.

## 1 Tokenization

Sentence segmentation breaks text data into single separated sentences whereas tokenization is the process of separating words or characters from a text into small units called tokens. In this process it removes punctuation, single characters, numbers, articles etc. An example of tokenization can be seen as follows :

**Before Tokenization :** *'2-D M-MODE: , ,1. Left atrial enlargement with left atrial diameter of 4.7 cm.,2. Normal size right and left ventricle.,3. Normal LV systolic function with left ventricular ejection fraction of 51%.,4. Normal LV diastolic function.,5. No pericardial effusion.,6. Normal morphology of aortic valve, mitral valve, tricuspid valve, and pulmonary valve.,7. PA systolic pressure is 36 mmHg.,DOPPLER: , ,1. Mild mitral and tricuspid regurgitation.,2. Trace aortic and pulmonary regurgitation.'*

**After Tokenization :** *'MODE', 'left', 'atrial', 'enlargement', 'left', 'atrial', 'diameter', 'cm', 'normal', 'size', 'right', 'leave', 'ventricle', 'Normal', 'LV', 'systolic', 'function', 'left','ventricular', 'ejection', 'fraction', 'Normal', 'LV', 'diastolic', 'function', 'pericardial', 'effusion', 'normal', 'morphology', 'aortic', 'valve', 'mitral', 'valve', 'tricuspid', 'valve', 'pulmonary', 'valve', 'PA', 'systolic', 'pressure', 'mmHg', 'mild', 'mitral', 'tricuspid', 'regurgitation', 'trace', 'aortic', 'pulmonary', 'regurgitation'*

## 2 Word or Character Vectorization

To train the data using a machine learning algorithm, data needs to be processed to have row as instances and columns as features/ predictor variables. Tokenized text data are not suitable for machine learning. Therefore, a vector representations or the numerical transformation from the tokenized data is prepared. The process is referred to as vectorization or feature extraction. There are several vectorization techniques but among them TF-IDF is one of the popular methods in NLP domain. This bag-of-word representation (TF-IDF) only describes a document in a stand-alone fashion, not taking into account the context of the data whereas the relative frequency of tokens (TF) in the document takes into account the context of the documents. In the aforementioned transcription text, tokens such as *left*, *atrial*, and *normal* appear more frequently in documents that goes under the clinical speciality **Cardiovascular / Pulmonary**, while other tokens like *mitral*, *cm*, *diameter* are less important. TF–IDF normalizes the frequency of tokens in a document with respect to the rest of the data. The term frequency of a particular text can be a boolean value or the count of the text, whereas inverse document frequency is a measure of whether a term is common or rare in a given document corpus. Both the TF and IDF are scaled logarithmically to prevent bias of terms that appear much more frequently relative to other terms. Equation 2.1 describes the relative frequency of term $t$ in document $d$:

$$tf(t,d) = 1 + log f_{t,d} \tag{2.1}$$

where, $tf(t,d)$ is the relative frequency of term $t$ in document $d$.

Similarly, the inverse document frequency of a term given the set of documents can be logarithmically scaled as follows:

$$idf(t,d) = log 1 + N/n_t \tag{2.2}$$

where, N is the number of documents and $n_t$ is the number of occurrences of the term in all documents. TF–IDF is then computed completely as:

$$tfidf(t,d,D) = tf(t,d) * idf(t,d) \tag{2.3}$$

If the ratio of the idf log function is greater or equal to 1, the TF–IDF score is always greater than or equal to zero. If the TF-IDF of a token is close to 1 the token is more valuable to the text and vice versa [10].

# 3 N-grams

The extraction of single word (or char) or feature from text can be done by tokenization. A variable containing frequency counts of a single word in each text is called a unigram whereas a two-word sequence is called bigrams. $n$ can have higher values also. A high value of $n$ in a n-gram model results in large feature space imposing high computational overhead.

* Unigrams : 'left', 'ventricular', 'cavity', 'size', 'wall', 'thickness'
* Bigrams : 'left ventricular', 'cavity size', 'wall thickness'
* Trigrams : 'left ventricular cavity', 'size wall thickness'

# 4 Classification Methods

## 4.1 Multinomial Naive Bayes

A multinomial distribution is used to find probabilities of a input data based on its feature counts when there is a possibility to have multiple outcomes. In NLP, the multinomial naive bayes classification uses word frequency in documents where a document is an ordered sequence of word events, drawn from the vocabulary V. The conditional probability of a document $d_i$, given a class $c_j$, is a product of the probability of each observed word in the corresponding class. Assuming each document $d_i$ is drawn from a multinomial distribution of words with as many independent trials as the length of $d_i$, $N_{it}$ being the count of the number of times word $w_t$ occurs in document $d_i$, the conditional probability is given by Equation 2.4 from [11]

$$P(d_i|c_j;\theta) = P(|d_i|)|d_i|! \prod_{t=1}^{|V|} \frac{P(w_t|c_j;\theta)^{N_{it}}}{N_{it}!} \tag{2.4}$$

where $\theta$ is the mixture model parameter.
The probability of occurring word $w_t$ in given class $c_j$ is given by Equation 2.5

$$\hat{\theta}_{w_t|c_j} = P(w_t|c_j;\hat{\theta}_j) = \frac{1 = \sum_{i=1}^{|D|} N_{it} P(c_j|d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} P(c_j|d_i)} \tag{2.5}$$

where $D$ is the set of labeled training documents $\{d_1, ..., d_D\}$.
and the class prior parameters, $\hat{\theta}_{c_j}$ is expressed by Equation 2.6 from [11]

$$\hat{\theta}_{c_j} = P(c_j|\hat{\theta}) = \frac{\sum_{i=1}^{|D|} P(c_j|d_i)}{|D|} \tag{2.6}$$

### 4.2 Support Vector Machine

In NLP, Support Vector Machine (SVM) is capable of separating text documents into respective classes. SVM determines a hyperplane by transforming training set with the help of mathematical functions called **kernels** if the problem is not linearly separable. Types of kernels are linear, sigmoid, radial basis function, non-linear, polynomial, etc. To find the points that lie closest each class, the distance between the dividing plane and the points are measured. The points are called support vectors and the distance between the points and the dividing line is called margin. The aim of an SVM algorithm is to maximize this margin. The hyperplane becomes optimal if the margin reaches its maximum limit. By maximizing the distance, future data points will be classified more accurately. According to Hotho et al. [7] when the data is textual there is a small effect in accuracy of classification based on the choice of kernels but if the feature space is high dimensional, then the precision and recall depends on the choice of kernels. In this work, Linear Support Vector Classification with kernel **linear** has been used to solve multi-class classification problem.

### 4.3 Logistic Regression

Logistic regression is another supervised classification method where the target variable $y$ can only take discrete values for a given set of features $X$. A logistic function is a sigmoid function with Equation 2.7 [3]

$$f(x) = \frac{L}{1 + \exp^{-k(x-x_0)}} \tag{2.7}$$

where, $x_0$ is the $x$ value of the sigmoid's midpoint, $L$ is the curve's maximum value, and $k$ represents the logistic growth rate or steepness of the curve.
For classification of medical specialities using transcriptions multinomial logistic regression has been used [13].

### 4.4 Random Forest Classification

A random forest is a classifier $h(x, \Theta_k)$, consisting of a collection of de-correlated decision trees where each tree casts a unit vote for the most popular class at input vector $x$. For classification problems, a splitting criterion **entropy** is used to specify the lower bound on the length of a random variable's bit representation. The entropy is calculated at each internal node of a decision tree and is given by Equation 2.8:

$$E = -\sum_{i=1}^{c} p_i \times log(p_i) \tag{2.8}$$

where, $c$ is the number of unique classes and $p_i$ is the prior probability of each given class. Since the final decision is made from averaging the trees, Random Forest is known as en ensemble learning method.

## 5 Performance Metrices

The performance measures of classification model includes different methods [14], namely accuracy, precision, recall and f1-score. These measures are defined using the below quantities:

- TP = True Positive, classifier predicts positives that are actually positive.
- FN = False Negative, classifier predicts negatives that are actually positive.
- TN = True Negative, classifier predicts negatives that are actually negative.
- FP = False Positive, classifier predicts positives that are actually negative.

### 5.1 Accuracy :

Accuracy is the fraction of data that are classified correctly and the total number of data.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.9}$$

### 5.2 Precision :

Precision is the fraction of actual positives and total predicted positives.

$$Precision = \frac{TP}{TP + FP} \tag{2.10}$$

### 5.3 Recall :

Recall or sensitivity is the measure of the proportion of actual positives and total actual positive.

$$Recall = \frac{TP}{TP + FN} \tag{2.11}$$

### 5.4 F1-score :

F1 score is the measure to balance between precision and recall.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.12}$$

All these measures will come forward to analyze the performance of the classifiers in the upcoming chapters.

# Data

The data is collected from **Kaggle**[2] but it was originally scrapped from **MTSamples** that has a big collection of transcribed medical reports [12]. The transcribed medical transcription contains patients' clinical reports and observations that were collected by many users and hereby the authenticity can not be verified.

MTsamples contains 6 columns (**descriptions**, **transcription title**, **transcription keywords**, **transcriptions and medical speciality**) where **medical speciality** is the target variable. Among all fields **sample medical transcriptions** was used to classify the speciality fields.

## 1 Data Preprocessing

The data was cleaned by removing empty rows and null values based on the target variable (4966 rows). It was explored and some labels such as **Discharge Summary, Consult - History and Phy., SOAP / Chart / Progress Notes, Emergency Room Reports, Office Notes, Letters** were removed since they do not qualify as clinical specialities. To balance the dataset, few other labels containing less than 10 samples were also removed. The classes **Surgery, Radiology, General Medicine** were conflicting with many specialities as they share similar clinical terminologies. Likewise **Orthopedic and Neurology and Neurosurgery** were missclassifed due to their common medical terms. Thus **Surgery, Radiology, General Medicine and Orthopedic** has been removed from the data. Figure 3.1 depicts the barplot of the clinical specialities.

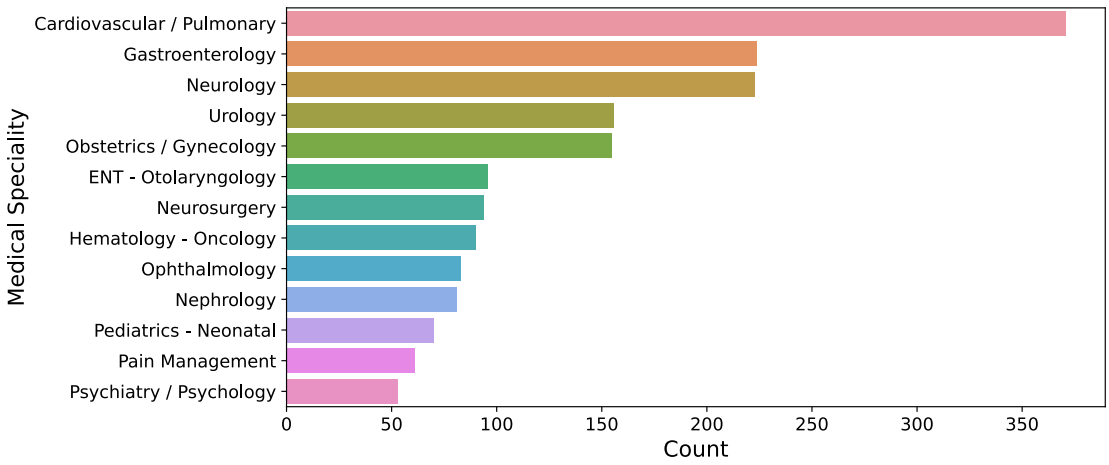


Figure 3.1: Barplot of Clinical Specialities

## 2 Text Pre-processing

Transcription texts contained stop words, punctuation marks, capital letters, numbers, words etc. NLP technique based **Spacy** [6] package has been used to tokenize the words into their base form and remove stop words, punctuation marks, white spaces and non-alphabetic text. Few words such as **procedure, diagnosis, description, preoperative, past, subjective** are also removed from the text as they do not symbolize any medical terms. Table 3.1 represents data after tokenization.

Table 3.1: The pre-processed dataframe

| Transcription | Medical specialty | Preprocessed |
|---|---|---|
| 2-D M-MODE: , ,1. Left atrial enlargement wit... | Cardiovascular/ Pulmonary | [MODE, left, atrial, enlargement, left, atrial... |
| 1. The left ventricular cavity size and wall ... | Cardiovascular/ Pulmonary | [left, ventricular, cavity, size, wall, thickn... |
| 2-D ECHOCARDIOGRAM,Multiple views of the heart... | Cardiovascular/ Pulmonary | [ECHOCARDIOGRAM, multiple, view, heart, great,... |

## 3 Data Re-sampling

It is evident from Figure 3.1 that the data is skewed, some categories are much smaller than the top categories. These can impose low prediction accuracy towards the minority classes as the machine learning algorithms assume balanced dataset. To reduce imbalance, re-sampling method such as over-sampling or under-sampling can be used. The former one adds data to the minority class and latter removes data from the majority class. There are few other methods to maximize the classification task that includes combination of over and under sampling and ensemble methods. In this work, methods mentioned in Table 3.2 has been utilized to maximize the accuracy of classification [9].

Table 3.2: Re-sampling Methods

| Category | Method Name | Methodology |
|---|---|---|
| Over-sampling | RandomOverSampler | Over sampling the minority classes by randomly picking samples with replacement. |
| Over-sampling | SMOTE | Over sampling the minority classes by creating "synthetic" data along the line segments of the k-minority class nearest neighbors instead of randomly picking samples with replacement. |
| Under-sampling | RandomUnderSampler | It under samples the majority class(es) by randomly picking samples with or without replacement. |
| Combination of Over and Under-sampling | SMOTEENN | It works combinely over-sampling using SMOTE and under-sampling using Edited Nearest Neighbours. |

## 4 Feature Reduction

It can be advantageous to use dimensionality reduction to improve classification accuracy. The truncated SVD method performs linear dimensionality reduction by means of truncated singular value decomposition (SVD). However, it does not center the data before computing the singular value decomposition. The method **TruncatedSVD** from **sklearn** has been applied for feature reduction [13].

# Method

This project work has explored four different classification methods; the best hyper parameters of each method has been selected using **gridsearchCV** method as depicted in Figure 4.1. The package **train_test_split** from **sklearn** was used to split data in stratified way into train and test set keeping 30% data in test set. The train set was converted to a collection of raw documents to a matrix of TF-IDF features by using the function **TfidfVectorizer**. The feature extraction using TF-IDF reflects how important a word is to a text in a document. After evaluating the results with test data set, two classification methods were chosen for further improvement. Re-sampling methods were employed to reduce the data imbalance problem and there after classification methods were executed followed by **truncated SVD** to do feature reduction (500 features reduced from 2000 features).

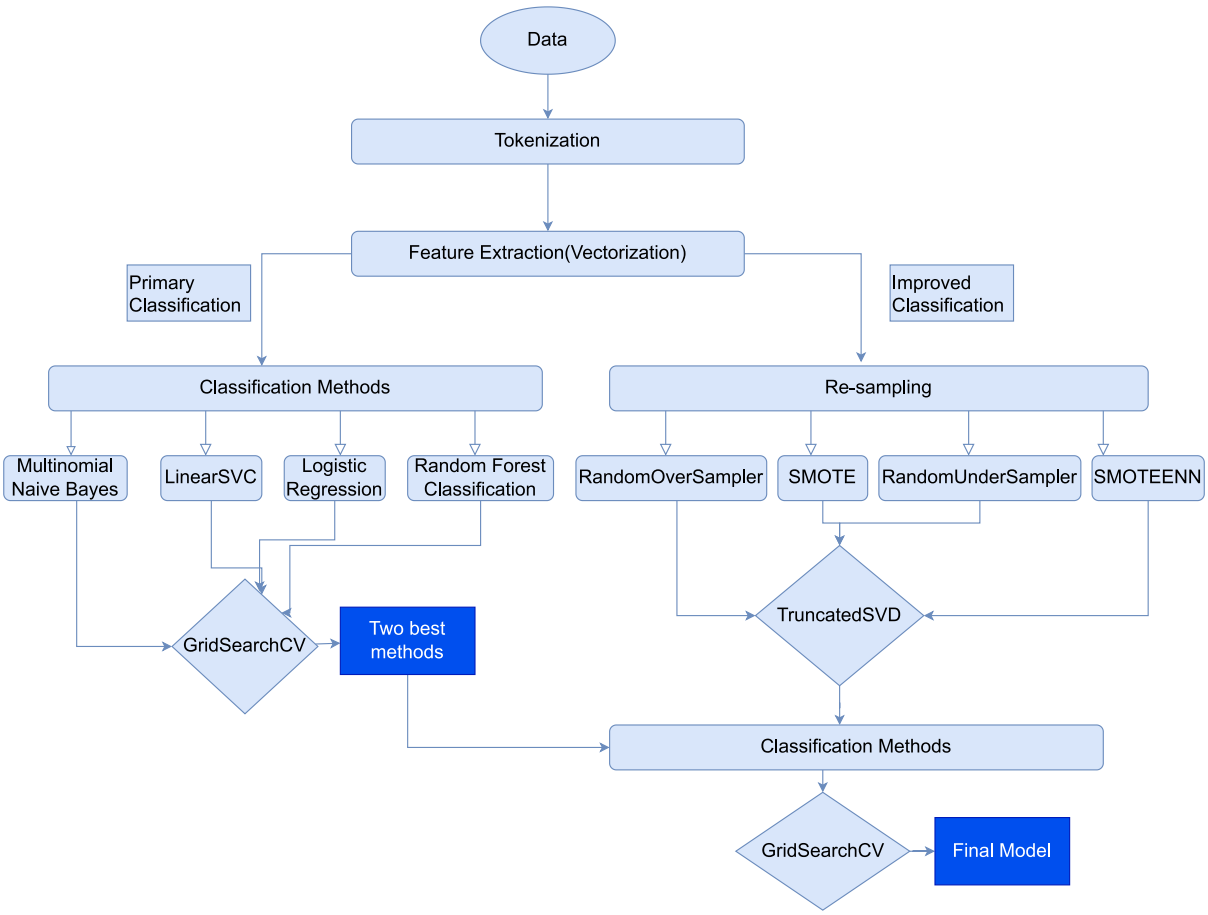


Figure 4.1: Work Flow of this project

# 1 Multinomial Naive Bayes classifier

Multinomial Naive Bayes classifier counts the occurrences and thus is suitable for discrete features classification. **MultinomialNB** from **sklearn** was implemented with grid of alpha ($\alpha$) values ranging from 0 to 1, where $\alpha = 0$ refers to no smoothing and $\alpha = 1$ refers to laplace smoothing. Fit_prior was checked for both True and False. Table 4.1 contains detaild hyper parameter selection.

Table 4.1: Hyper-parameters of TfidfVectorizer and MultinomialNB

| ngram_range | analyzer | max_df | min_df | max_features | alpha | fit_prior |
|---|---|---|---|---|---|---|
| (1,2) | char | 0.75 | 5 | 2000 | 0.05 | True |
| (1,3) | word | 0.95 | | | 0.1 | False |
| | | | | | 0.2 | |

# 2 LinearSVC

For the second model **LinearSVC** from **sklearn** was applied to classify data based on loss function using $kernel = Linear$. The chosen hyper parameters **loss, class_weight, C** is listed in Table 4.2. The rationale of choosing linear kernel was that it has more flexibility for choosing the penalties and loss functions. The rationale of choosing SVM was that it separates most text categorization problem linearly [8].

Table 4.2: Hyper-parameters of TfidfVectorizer and LinearSVC

| ngram_range | analyzer | max_df | min_df | max_features | loss | C |
|---|---|---|---|---|---|---|
| (1,2) | char | 0.75 | 5 | 2000 | squared_hinge | 0.2 |
| (1,3) | word | 0.95 | | | | 0.3 |
| | | | | | | 0.5 |

# 3 Logistic Regression

The third model has been trained using logistic regression. **LogisticRegression** from **Sklearn** was used with penalty **elasticnet** normalization and solver **saga**. The **multi_class** option is set to **multinomial** to solve the multiclass classification and **l1_ratio** is kept in between 0 and 1. Details are listed in Table 4.3.

Table 4.3: Hyper-parameters of TfidfVectorizer and Logistic Regression

| ngram_range | analyzer | max_df | min_df | max_features | penalty | l1_ratio | solver | multi_class |
|---|---|---|---|---|---|---|---|---|
| (1,2) | char | 0.75 | 5 | 2000 | elasticnet | 0.5 | saga | multinomial |
| (1,3) | word | 0.95 | | | | 0.6 | | |

# 4 Random Forest Classification

Finally, the last method random forest classification algorithm has been applied to the data for finding

the optimal decision trees. The hyperparameters in Table 4.4 was considered to find the best parameter for the final model.

Table 4.4: Hyper-parameters of TfidfVectorizer and Random Forest Classifier

| ngram_range | analyzer | max_df | min_df | max_features | n_estimators | max_depth | min_samples_split | min_samples_leaf |
|---|---|---|---|---|---|---|---|---|
| (1,2) | word | 0.75 | 5 | 2000 | 50 | 10 | 30 | 5 |
| (1,3) | char | 0.95 | | | 100 | 50 | 50 | |

# Results

The outcome from the project can be described as follows.

## 1 Primary Classification results from GridSearchCV

### 1.1 Multinomial Naive Bayes classifier

Multinomial Naive Bayes classifier's best score is 0.77 obtained from gridsearchCV. The hyper-parameters for multinomialNB resulting in best performance are **alpha = 0.2, fit_prior=True**. For tfidfvectorizer **word** is chosen as analyzer, the **ngram_range** is selected to **(1, 2)** which takes both unigrams and bi-grams for analysis. For buidling the vocabulary ignore terms a float value of **0.95** is selected for **max_df** and an int value of 5 is selected as cut-off. The classification report obtained from the model is presented on Table 5.1.

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| Cardiovascular / Pulmonary | 0.83 | 0.93 | 0.88 | 106 |
| ENT - Otolaryngology | 0.92 | 0.66 | 0.77 | 35 |
| Gastroenterology | 0.78 | 0.74 | 0.76 | 68 |
| Hematology - Oncology | 0.52 | 0.52 | 0.52 | 23 |
| Nephrology | 0.75 | 0.20 | 0.32 | 30 |
| Neurology | 0.70 | 0.84 | 0.76 | 63 |
| Neurosurgery | 0.65 | 0.71 | 0.68 | 28 |
| Obstetrics / Gynecology | 0.94 | 0.69 | 0.80 | 48 |
| Ophthalmology | 0.97 | 0.94 | 0.95 | 32 |
| Pain Management | 0.83 | 0.88 | 0.86 | 17 |
| Pediatrics - Neonatal | 0.41 | 0.54 | 0.46 | 24 |
| Psychiatry / Psychology | 0.76 | 0.81 | 0.79 | 16 |
| Urology | 0.59 | 0.76 | 0.67 | 38 |

Table 5.1: Classification report of Multinomial NB Classifier

### 1.2 LinearSVC

Linear support vector machine performs well when there are lot of features. It takes less computational time and less parameters therefore it is faster than any other kernel methods for training data. The best CV score obtained was 0.80 from the gridsearchCV and the best hyper-parameter C which controls the strength of the regularization was found to be 0.2. The selected loss function was **squared_hinge**. Rest of the vectorization hyper-parameters were similar to MultinomialNB described in 1.1 except for building the vocabulary ignore terms where a float value of 0.75 was selected for **max_df**.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Cardiovascular / Pulmonary | 0.81 | 0.95 | 0.87 | 106 |
| ENT - Otolaryngology | 0.90 | 0.80 | 0.85 | 35 |
| Gastroenterology | 0.79 | 0.90 | 0.84 | 68 |
| Hematology - Oncology | 0.56 | 0.39 | 0.46 | 23 |
| Nephrology | 0.71 | 0.33 | 0.45 | 30 |
| Neurology | 0.68 | 0.86 | 0.76 | 63 |
| Neurosurgery | 0.70 | 0.57 | 0.63 | 28 |
| Obstetrics / Gynecology | 0.95 | 0.85 | 0.90 | 48 |
| Ophthalmology | 0.97 | 0.97 | 0.97 | 32 |
| Pain Management | 0.88 | 0.82 | 0.85 | 17 |
| Pediatrics - Neonatal | 0.75 | 0.38 | 0.50 | 24 |
| Psychiatry / Psychology | 0.85 | 0.69 | 0.76 | 16 |
| Urology | 0.72 | 0.89 | 0.80 | 38 |

Table 5.2: Classification report of Linear SVC

### 1.3 Logistic Regression

Logistic Regression uses one vs rest method or the cross-entropy loss method for solving multi-class classification problem. In this project **multi_class** option was set to **multinomial** for using cross-entropy loss method. The optimization algorithm **saga** solver was applied with the **elasticnet** penalty. Mixing parameter **l1_ratio** was set by gridsearchCV to 0.5 which resulted in the best score of 0.75. The vectorization hyper-parameters were similar to MultinomialNB described in 1.1 except for the **ngram_range** which was chosen as (1, 3) that takes unigrams, bigrams and trigrams.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Cardiovascular / Pulmonary | 0.66 | 0.97 | 0.79 | 106 |
| ENT - Otolaryngology | 0.93 | 0.71 | 0.81 | 35 |
| Gastroenterology | 0.80 | 0.87 | 0.83 | 68 |
| Hematology - Oncology | 0.62 | 0.35 | 0.44 | 23 |
| Nephrology | 0.80 | 0.40 | 0.53 | 30 |
| Neurology | 0.70 | 0.84 | 0.76 | 63 |
| Neurosurgery | 0.72 | 0.46 | 0.57 | 28 |
| Obstetrics / Gynecology | 0.95 | 0.81 | 0.88 | 48 |
| Ophthalmology | 0.97 | 0.94 | 0.95 | 32 |
| Pain Management | 0.80 | 0.71 | 0.75 | 17 |
| Pediatrics - Neonatal | 0.78 | 0.29 | 0.42 | 24 |
| Psychiatry / Psychology | 1.00 | 0.56 | 0.72 | 16 |
| Urology | 0.76 | 0.89 | 0.82 | 38 |

Table 5.3: Classification report of Logistic Regression

### 1.4 Random Forest Classification

Random forest classification method comprises of a bunch of decision tress of which are trained using random sub-sample of features (words or documents) and average them to improve the predictive accuracy and control over-fitting. The drawback of the random forest is computational time for hyper-parameter tuning. This method achieved best CV score of 0.74 from the best hyper-parameters such as maximum depth of a tree is 50 and minimum number of samples to split an internal node is 30, the

minimum number of samples at a leaf node was 5 and the number of trees in the forest model was 100. Rest of the vectorization hyper-parameters are similar to MultinomialNB described in 1.1.

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| Cardiovascular / Pulmonary | 0.66 | 0.97 | 0.79 | 106 |
| ENT - Otolaryngology | 0.90 | 0.77 | 0.83 | 35 |
| Gastroenterology | 0.78 | 0.87 | 0.82 | 68 |
| Hematology - Oncology | 0.50 | 0.17 | 0.26 | 23 |
| Nephrology | 0.60 | 0.20 | 0.30 | 30 |
| Neurology | 0.75 | 0.75 | 0.75 | 63 |
| Neurosurgery | 0.72 | 0.64 | 0.68 | 28 |
| Obstetrics / Gynecology | 0.95 | 0.81 | 0.88 | 48 |
| Ophthalmology | 0.91 | 0.97 | 0.94 | 32 |
| Pain Management | 0.76 | 0.76 | 0.76 | 17 |
| Pediatrics - Neonatal | 0.67 | 0.17 | 0.27 | 24 |
| Psychiatry / Psychology | 0.85 | 0.69 | 0.76 | 16 |
| Urology | 0.67 | 0.87 | 0.76 | 38 |

Table 5.4: Classification report of Random Forest

### 1.5 Comparison of Primary Classification results

A comparison of all four models based on classification results are shown in Table 5.5 representing the best-score from gridsearchCV, accuracy, precision, recall and f1-score.

| Classification Tech | Best Score from CV | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Multinomial NB | 0.77 | 0.75 | 0.74 | 0.71 | 0.71 |
| Linear SVC | 0.80 | 0.79 | 0.79 | 0.72 | 0.74 |
| Logistic Regression | 0.75 | 0.77 | 0.81 | 0.68 | 0.71 |
| Random Forest | 0.74 | 0.75 | 0.75 | 0.66 | 0.68 |

Table 5.5: Performance comparison based on measures

However, it is evident that Linear SVC and Logistic regression has better classification rate than Multinomial NB and Random Forest for this data. This leads to the secondary step of classification using re-sampling and dimentionality reduction with Linear SVC and Logistic regression for further analysis.

## 2 Classification after re-sampling

Four different types of re-sampling methods has been applied with Linear SVC and Logistic Regression methods.

### 2.1 LinearSVC

Table 5.6 depicts the improvement of LinearSVC after imposing different re-sampling methods.

| Classification | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| RandomOverSampler (ROS) | 0.81 | 0.79 | 0.79 | 0.79 |
| SMOTE | 0.80 | 0.78 | 0.78 | 0.78 |
| RandomUnderSampler (RUS) | 0.75 | 0.71 | 0.76 | 0.73 |
| SMOTEENN | 0.73 | 0.74 | 0.78 | 0.74 |

Table 5.6: Classification results of LinearSVC based on different re-sampling techniques

Below table contains detailed classification report of all the medical specialities,

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| Cardiovascular / Pulmonary | 0.87 | 0.86 | 0.86 | 112 |
| ENT - Otolaryngology | 0.86 | 0.83 | 0.84 | 29 |
| Gastroenterology | 0.83 | 0.82 | 0.83 | 67 |
| Hematology - Oncology | 0.60 | 0.56 | 0.58 | 27 |
| Nephrology | 0.63 | 0.71 | 0.67 | 24 |
| Neurology | 0.84 | 0.78 | 0.81 | 67 |
| Neurosurgery | 0.69 | 0.86 | 0.76 | 28 |
| Obstetrics / Gynecology | 0.91 | 0.89 | 0.90 | 47 |
| Ophthalmology | 0.92 | 0.92 | 0.92 | 25 |
| Pain Management | 0.93 | 0.78 | 0.85 | 18 |
| Pediatrics - Neonatal | 0.48 | 0.62 | 0.54 | 21 |
| Psychiatry / Psychology | 0.83 | 0.94 | 0.88 | 16 |
| Urology | 0.82 | 0.77 | 0.79 | 47 |

Table 5.7: Classification report of ROS re-sampled LinearSVC

The classes that were classified well before are now having few more false negatives **Cardiovascular / Pulmonary, ENT - Otolaryngology** . It is also notable that classes who had lower recall rate **Pediatrics - Neonatal, Neurosurgery** have achieved higher recall rate than precision. Though **Neurology, Pain Management, Urology** has better precision and recall rate but the precision rates are higher than recall rates. **Hematology - Oncology** has higher rate of both false negatives and false positives.

The confusion matrix found from the fitted model of ROS re-sampled LinearSVC model is shown in Figure 5.1,

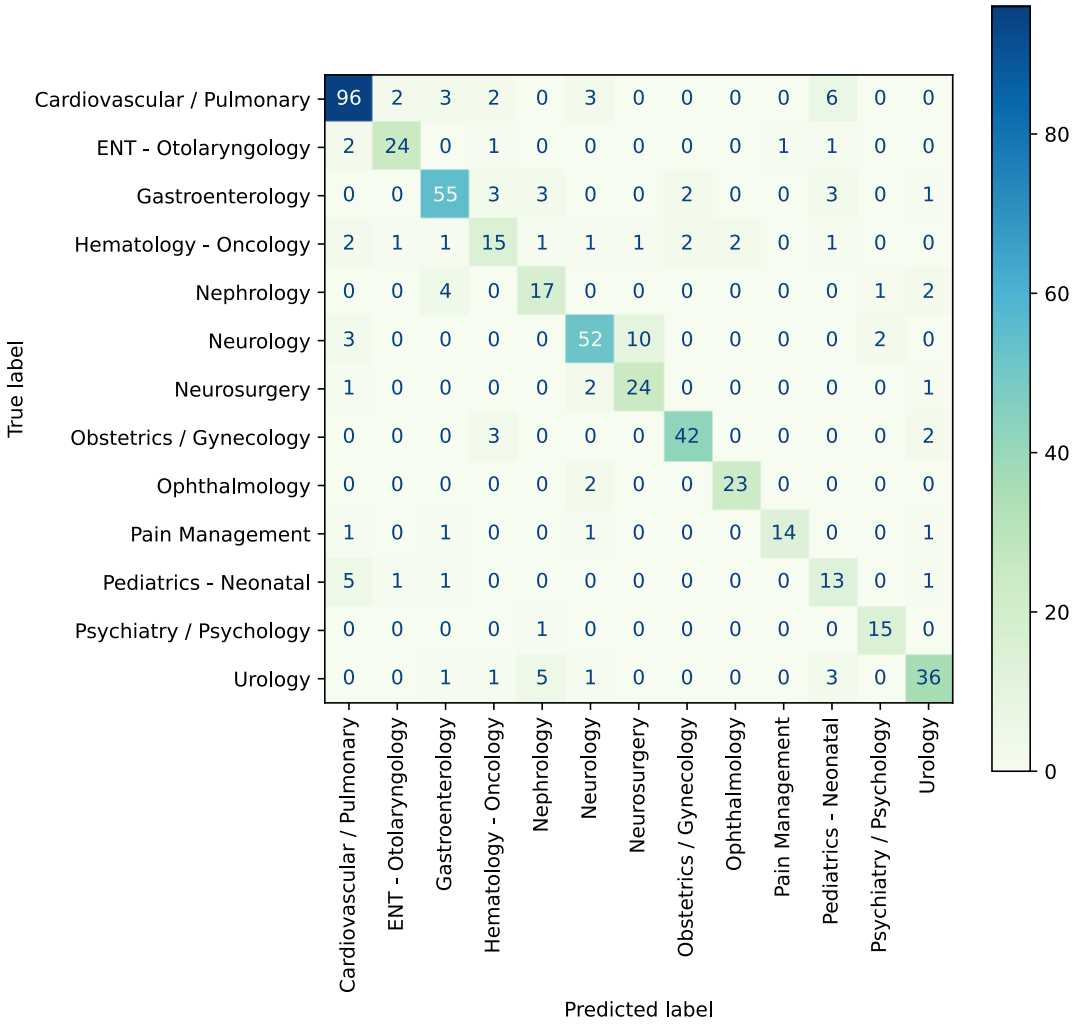


Figure 5.1: Confusion matrix of ROS re-sampled LinearSVC model

## 2.2 Logistic Regression

Table 5.8 depicts the comparison of the revised logistic regression followed by different re-sampling methods.

| Classification | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| RandomOverSampler (ROS) | 0.80 | 0.78 | 0.79 | 0.78 |
| SMOTE | 0.79 | 0.77 | 0.78 | 0.77 |
| RandomUnderSampler (RUS) | 0.71 | 0.69 | 0.74 | 0.70 |
| SMOTEENN | 0.70 | 0.72 | 0.77 | 0.72 |

Table 5.8: Classification results of Logistic Regression based on different re-sampling techniques

Table 5.9 shows the classification between different clinical specialities based on medical transcriptions.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Cardiovascular / Pulmonary | 0.86 | 0.83 | 0.85 | 112 |
| ENT - Otolaryngology | 0.78 | 0.86 | 0.82 | 29 |
| Gastroenterology | 0.81 | 0.85 | 0.83 | 67 |
| Hematology - Oncology | 0.58 | 0.56 | 0.57 | 27 |
| Nephrology | 0.62 | 0.67 | 0.64 | 24 |
| Neurology | 0.83 | 0.78 | 0.80 | 67 |
| Neurosurgery | 0.72 | 0.93 | 0.81 | 28 |
| Obstetrics / Gynecology | 0.89 | 0.85 | 0.87 | 47 |
| Ophthalmology | 0.91 | 0.84 | 0.87 | 25 |
| Pain Management | 1.00 | 0.78 | 0.88 | 18 |
| Pediatrics - Neonatal | 0.54 | 0.62 | 0.58 | 21 |
| Psychiatry / Psychology | 0.80 | 1.00 | 0.89 | 16 |
| Urology | 0.83 | 0.72 | 0.77 | 47 |

Table 5.9: Classification report of ROS re-sampled Logistic Regression

The class **Hematology - Oncology** has now low recall rate and precision rate and is less than 60% referring to many false negatives. Moreover **Neurology, Urology, Pain Management** has also lower recall rates than precision rates even though the precision rate is much better than previous models. Rest of the classes are classified considerably well. To know better, we look down the confusion matrix of Figure 5.2 obtained from this model.

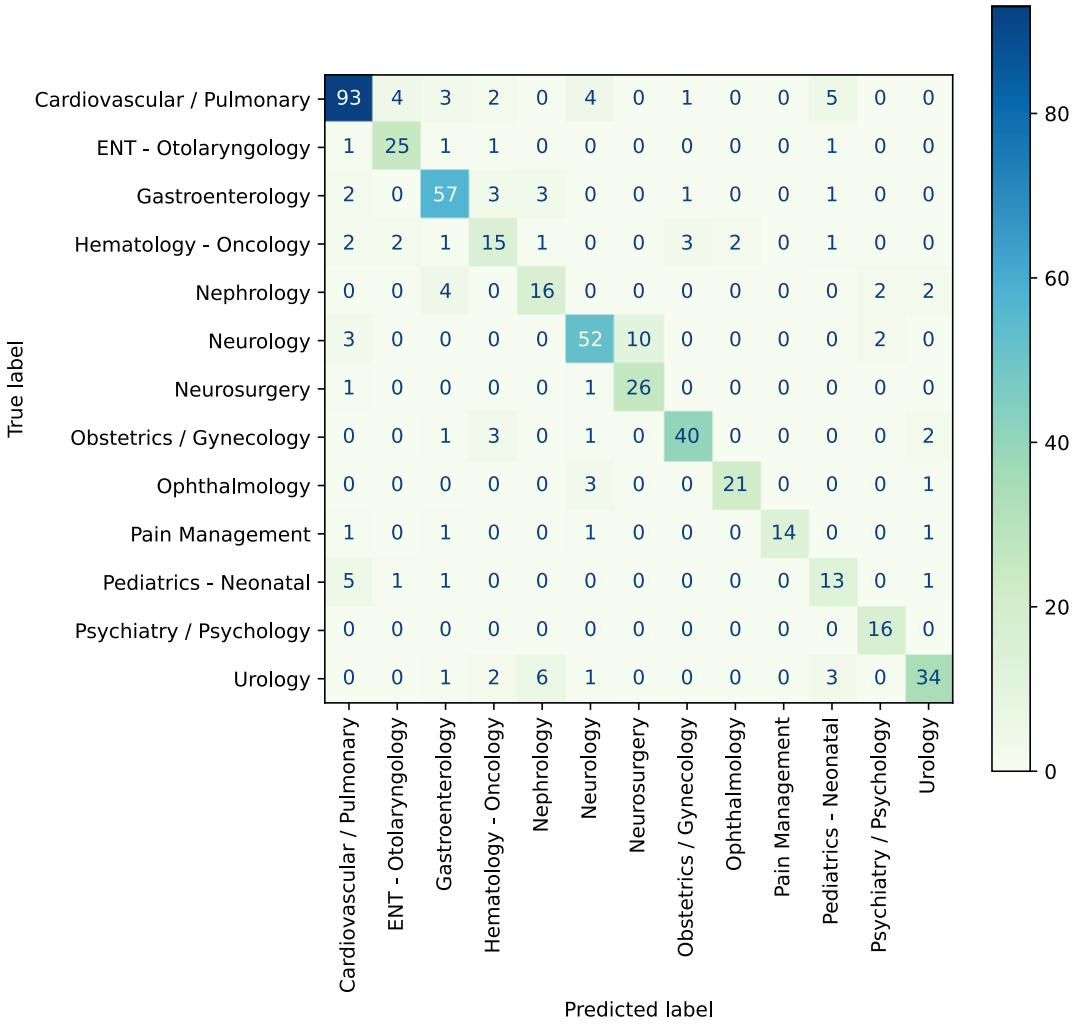


Figure 5.2: Confusion matrix of ROS re-sampled Logistic Regression

# Discussion

Accuracy is the primary performance measure metric to observe a classification model. However, it is the ratio of correctly predicted values to the total values which gives proper direction if false positive and false negatives are almost same. But for skewed data or imbalanced data other measures such as precision, recall and f1-score plays vital role for evaluating a classification model.

According to Table 5.1 overall classification of different clinical specialities are satisfactory except **Nephrology** class. However, further investigation reveals some of the classes being not correctly classified. The precision, recall and f1_score are not commensurate to each other. For example, precision for **ENT - Otolaryngology** is 92% whereas recall and f1_score are 66% and 77% respectively. Hence, it is evident that there are some words that were not classified to the right categories.

Furthermore, the classification report of Linear SVC presented in Table 5.2 shows the recall and f1_score for **ENT - Otolaryngology** are 80% and 85% respectively which seems to be an improvement over MultinomialNB. However, few other classes such as **Hematology - Oncology, Nephrology, Pediatrics - Neonatal, Neurosurgery** seems to have more false negatives as they have lower recall than precision rate.

Table 5.3 shows the classification of different specialities using Logistic Regression. It can be seen that the overall performance has not improved compared to previous methods. Moreover, **Hematology - Oncology, Nephrology, Pediatrics - Neonatal, Neurosurgery, Psychiatry / Psychology** classes contains more false negatives than before.

Researchers claimed that medical textual data can achieve high accuracy from random forest classification [4]. Random forest classification is also capable of handling imbalaned data but for this dataset some of the classes **Hematology - Oncology, Nephrology, Pediatrics - Neonatal** are still getting lower recall value than precision. According to Table 5.4, **Cardiovascular / Pulmonary** acquired 66% precision rate whereas in LinearSVC and MultinomialNB it had 83% precision rate. The low precision from random forest and logistic regression refers that a very few of the transcriptions are classified in the true class even though the recall is high.

According to the classification reports, none of the methods are able to classify some of the classes. An imbalanced data set can be a reason for this problem. From Table 5.6 it is observed that RandomOverSampler (ROS) re-sampling technique is able to classify correctly than other re-sampling methods. According to the Figure 5.1, **Cardiovascular / Pulmonary** has been classified approximately well without re-sampling but after re-sampling some of the transcriptions are erroneously classified to different classes, notably **Pediatrics - Neonatal**. It has been reported that premature birth with chronic lung disease are at risk of cardiovascular sequelae [1]. Therefore, the transcriptions that are classified as **Pediatrics - Neonatal and Cardiovascular / Pulmonary** can be in both classes. Around 10 **Neurology** documents are classified as **Neurosurgery**. We know that neurology is a nonsurgical specialty and its corresponding surgical specialty is neurosurgery [16]. Therefore, it has become difficult to classify these classes properly.

Referring to Table 5.8, re-sampling method **ROS** has achieved better performance than other techniques in Logistic Regression similar to **LinearSVC**. The confusion matrix from Figure 5.2 gives us a clear view of why we were getting lower recall rate in **Neurology and Urology** even after re-sampling shown in table 5.9. Around 10 **Neurology** documents are classified as **Neurosurgery**. As mentioned earlier [16], neurology is a nonsurgical specialty and its corresponding surgical specialty is neurosurgery. Thus it is not separable by using text classification methods. It will require more domain knowledge to separate them. The other class **Hematology - Oncology** has been predicted in almost all classes except **Pain Management, Psychiatry / Psychology, and Urology** . We can relate this to fact that those predicted specialities can refer a patient to Hematology for cancer investigation. **Urology** can be related to **Nephrology** too as both of them relates to Urinary System.

However, according to a recent study Convolutional Neural Networks (CNN) has proved that it can classify clinical text much better than random forest and support vector machine [15]. But the drawback is it is applicable for only binary classification. Therefore, this project did not explored CNN. Authors published a comparative study of different re-sampling methods in predicting students' performance using random forest, k-NN, artificial neural network, XG-boost, support vector machine (radial basis function), decision tree, logistic regression, and naive bayes. From their study they found random forest classifier performed best using SVM-SMOTE as a re-sampling method [5]. In the first phase of this work, the ensemble re-sampling had an unsatisfactory performance compared to base re-sampling methods, thus not included in the report and found sufficient to use a variation of over and under sampling and mix of over and under sampling methods.

# Conclusion

Precision is the ratio of correctly predicted positive data to the total predicted positive data whereas the ratio of correctly predicted positive data to the total data in true class is called recall. High precision means low false positive rate and high recall means low false negative rates. According to Table 7.1, the precision rate for logistic regression is 81%. However, after re-sampling the precision rate downgraded to 78%. On the contrary, the recall rate from logistic regression was 68%, which increased to 79% after re-sampling. An increase of 10% recall rate reduced the number of false negatives. This is an improvement of our model. A good classification consists of high recall and high precision which can be assured by both linearSVC and Logistic Regression after re-sampling by RandomOverSampler.

| Classification | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| LinearSVC | 0.79 | 0.79 | 0.72 | 0.74 |
| RandomOverSampler re-sampled LinearSVC | 0.81 | 0.79 | 0.79 | 0.79 |
| Logistic Regression | 0.77 | 0.81 | 0.68 | 0.71 |
| RandomOverSampler re-sampled Logistic Regression | 0.80 | 0.78 | 0.79 | 0.78 |

Table 7.1: Performance comparison based on measures

F1-score is the weighted average of precision and recall which means it takes false positives and false negatives for measurement and balance the bias. As the data was very skewed so we took f1 score as a final measure. Before and after re-sampling LinearSVC achieved higher f1-score than Logistic Regression.

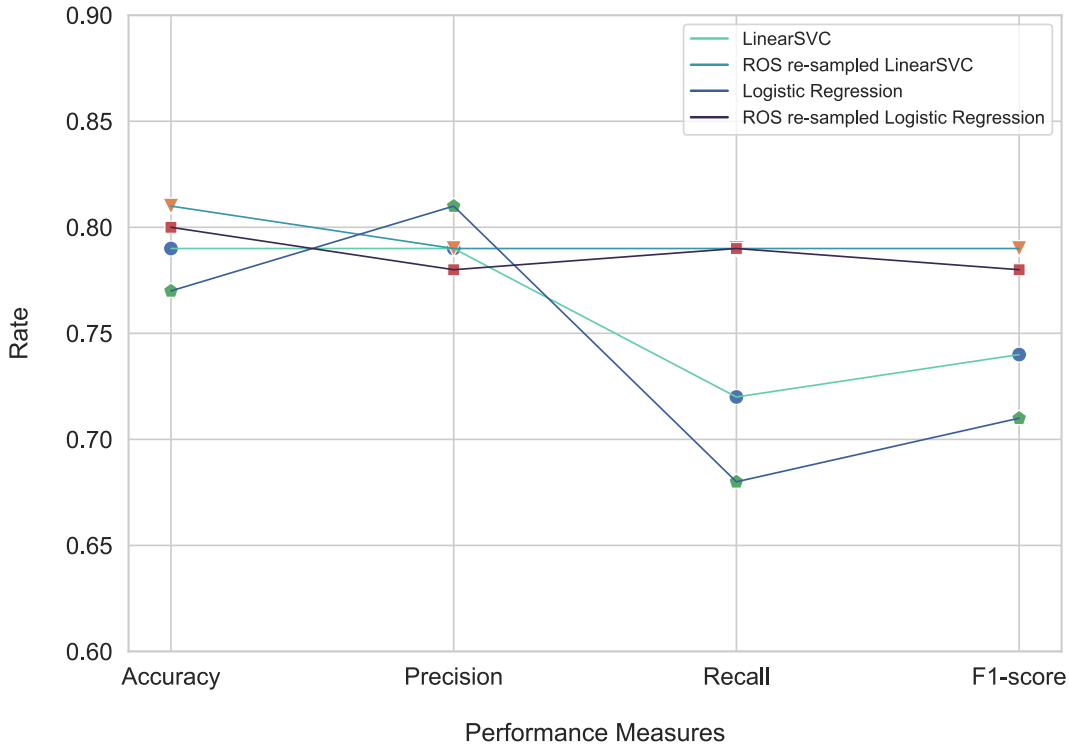


Figure 7.1: Comparison of Classification Metrices

Our observations in this project are different compared to the previous works. We found Linear SVC to be the best performing model instead of Random Forest. Many works has shown that SMOTE is capable of better re-sampling than any other methods whereas in our case RandomOverSampler outperformed it. However, none of the works are exactly similar to our project. The data was for a multiclass classification and it also lacks authenticity. Therefore, a new data set of medical transcriptions can be useful to classify medical specialities using this approach. This will enhance the acceptance of the work. Nonetheless, I have learnt that the reshuffling some categories (add or remove some categories such as **Hematology - Oncology**) can change classification of the speciality categories. There are few kaggle works [2] based on this data and those works included vectorization before train test split which violates the model training rule. It carries the bias to the validation dataset (classification accuracy can reach up to 95%). During this project I learned that feature reduction does not improve classification rather saves computational time. However, this is not covered into the results section due to length restriction. Additionally, **trigrams** can improve the classification but it is computationally expensive and therefore, was not suitable for this work. The **countvectorizer** transformation can be a choice instead of **tfidfvectorizer** for text vectorization for further improvement. The improvement can also be possible by manually removing irrelevant words from the text corpus.

# Bibliography


[1] ABMAN, S. H. Monitoring cardiovascular function in infants with chronic lung disease of prematurity. *Archives of Disease in Childhood - Fetal and Neonatal Edition 87*, 1 (2002), F15–F18.

[2] BOYLE, T. Medical transcriptions, 2018. [Online; accessed 10-March-2022].

[3] CONTRIBUTORS, W. Logistic regression — Wikipedia, the free encyclopedia, 2020. [Online; accessed 10-March-2022].

[4] ELZEHEIRY, H. A., BARAKAT, S., AND REZK, A. Different scales of medical data classification based on machine learning techniques: A comparative study. *Applied Sciences 12*, 2 (2022).

[5] GHORBANI, R., AND GHOUSI, R. Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access 8* (2020), 67899–67911.

[6] HONNIBAL, M., AND MONTANI, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. [Online; accessed 10-March-2022], 2017.

[7] HOTHO, A., NÜRNBERGER, A., AND PAASS, G. A brief survey of text mining. *LDV Forum 20* (2005), 19–62.

[8] JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98* (1998), C. Nédellec and C. Rouveirol, Eds., Springer Berlin Heidelberg, pp. 137–142.

[9] LEMAÎTRE, G., NOGUEIRA, F., AND ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research 18*, 17 (2017), 1–5. [Online; accessed 10-March-2022].

[10] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008.

[11] MCCALLUM, A., AND NIGAM, K. A comparison of event models for naive bayes text classification, 1998.

[12] MTSAMPLES. Medical transcriptions, 2018. [Online; accessed 8-March-2022].

[13] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., ET AL. Scikit-learn: Machine learning in python. *Journal of machine learning research 12*, Oct (2011), 2825–2830.

[14] SOKOLOVA, M., JAPKOWICZ, N., AND SZPAKOWICZ, S. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. vol. Vol. 4304, pp. 1015–1021. [Online; accessed 10-March-2022].

[15] WANG, Y., SOHN, S., LIU, S., SHEN, F., WANG, L., ATKINSON, E. J., AMIN, S., AND LIU, H. A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making 19*, 1 (Jan 2019), 1.

[16] WIKIPEDIA CONTRIBUTORS. Neurology — Wikipedia, the free encyclopedia, 2022. [Online; accessed 10-March-2022].