Capstone Project 1 In-depth Analysis

Supervised classification algorithms were used to predict whether a person has cervical cancer or not. Biopsy was considered as target.

1. Preprocessing data

Before analyzing, standardization is performed to the datasets, since it is a common requirement for many machine learning estimators. After standardization, all the features with numeric data are with zero mean and unit variance.

2. Classification metrics

For medical data like this cervical cancer dataset, we should consider the correct diagnosis, only the total accuracy is not enough for evaluating an algorithm. So I employed the following classification metrics.

	Diagnosis	
	positive	negative
Test outcome positive	TP (True positive)	FP (False positive)
Test outcome negative	FN (False negative)	TN (True negative)

Term	Formula
Accuracy	(TP + TN)/(P+N)
Recall	TP/(TP+FN)
Precision	TP/(TP+FP)
F-measure	(2 x recall x precision) / (recall+precision)

Confusion Matrix were used to present the TP, FP, FN and TN of the prediction. Classification Report were used to show the precision, recall, F1-score and support for each class.

3. Compare Machine Learning Models

The following classification algorithms were used to analyze this dataset:

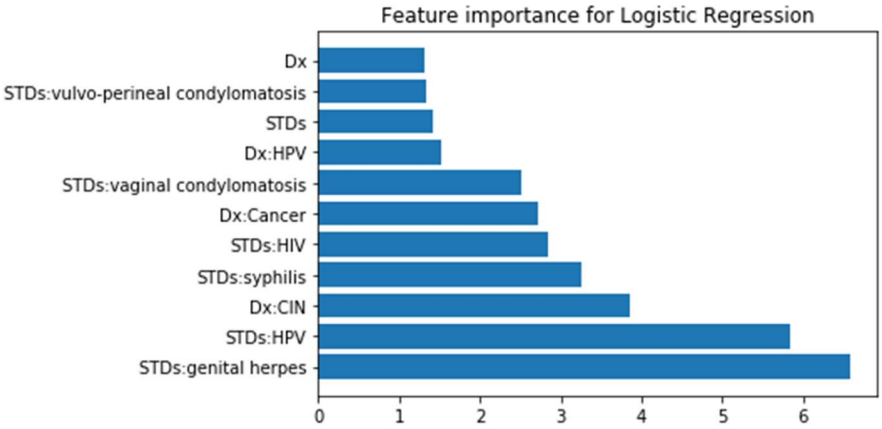
- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Decision Tree (DR)
- Random Forest (RF)

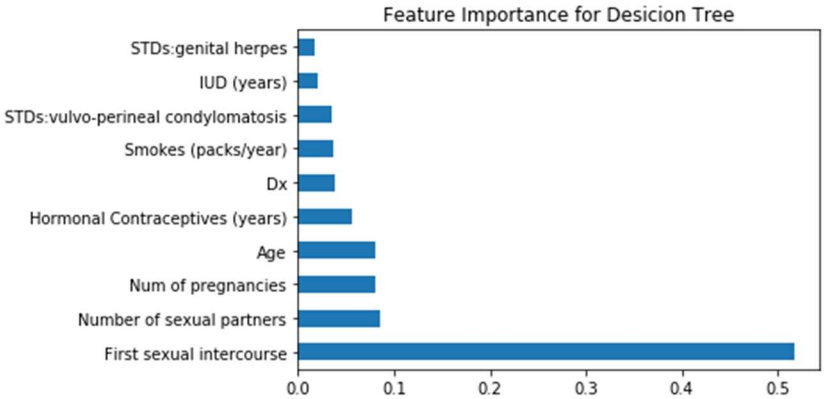
After tuning the parameters using cross validation (Recall was used as the scoring parameter, because correctly detecting more people with cancer is the most important. I also used f1 as the scoring parameter, similar result has gotten). Although the accuracy seems ok, all four classifiers suffered with low Recall, precision and F1 scores. Decision tree and random forest perform better than the other two models in terms of recall, precision and F1.

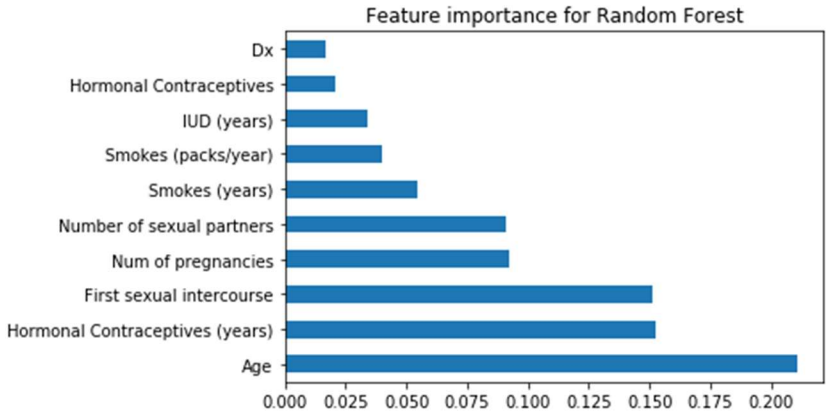
Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.942	0.913	0.900	0.953
Recall	0.111	0.222	0.444	0.333
Precision	0.333	0.200	0.222	0.6
F1	0.167	0.211	0.296	0.429

3.1 Feature Selection

Keeping irrelevant features can result in overfitting and decreasing accuracy. Feature selection could reduce overfitting, improve accuracy and reduce training time. So for each classifier (except for the SVM with rbf kernel), the most important features for the prediction were identified. The models were trained with the most important 10 features to improve the performance of the classifiers.







Classifier /Performance	Logistic regression	Decision Tree	Random Forest
Accuracy	0.948	0.907	0.948
Recall	0.111	0.667	0.222
Precision	0.5	0.316	0.5
F1	0.182	0.429	0.308

After feature selection, decision tree classifier was improved with higher recall, precision and F1. The accuracy of logistic regression also increased. The performance of random forest did not change much.

4. Class Imbalance

To address the imbalanced classes' problem, the following three techniques were used.

4.1 Oversample minority class (people with cancer)

Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.744	0.779	0.907	0.959
Recall	0.556	0.333	0.222	0.222
Precision	0.111	0.0857	0.182	1
F1	0.185	0.136	0.200	0.363

The recall improved for logistic regression and SVM, but the accuracy decreased for these two classifiers. Decision tree and random forest did not perform better than using the original data.

4.2 Undersample majority class (people without cancer)

Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest

Accuracy	0.599	0.738	0.517	0.639
Recall	0.444	0.555	0.667	0.777
Precision	0.0588	0.109	0.0698	0.104
F1	0.104	0.182	0.126	0.184

This technique did not perform well, because the sample size is too small.

4.3 Generate synthetic samples

Imblearn's SMOTE (Synthetic Minority Oversampling Technique) was used to generate synthetic samples. This technique uses a nearest neighbors algorithm to generate new and synthetic data.

Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.779	0.756	0.860	0.936
Recall	0.556	0.333	0.222	0.222
Precision	0.128	0.0769	0.105	0.333
F1	0.208	0.125	0.143	0.267

Similar results has gotten to the oversampling technique. The recall improved for logistic regression and SVM, but the accuracy decreased for these two models. Decision tree and random forest did not perform better than using the original data.

5. Summary

All classifiers suffered with low Recall, precision and F1 scores. It appears for this particular dataset Decision Tree with feature selection has the best overall performance.

6. Ongoing Works

I will perform feature selection for the upsampled and synthetic samples. I will employ ROC Curves and Precision-Recall curves to further evaluate the classifiers. Other algorithms, such as XGboost will be tried.