# GOOD/BAD CAR PURCHASE FOR AUTO DEALERS

Final Presentation for I526



Nihar Khetan


Ghanshyam Malu


Xiao Liang

# Problem Statement

When we go to buy a second hand car we expect to get a good car and dealers want profits as well.

How?

**Aim**

Trying to predict that whether a  bought by a dealer at an auction is a good buy or not

# Data Preprocessing

**Removed Redundant Features**

**Handled Null/Missing Values**

Continuous Data - Took Average

Discrete Data - Created new Category NULL

**Removed Features with More Than 95% Missing Values**

**Normalized All Continuous Values**

Data is from kaggle

# Generated Balanced Datasets

Oversampling

SMOTE Tech

# Feature Selection

**Expert Knowledge**

1

VehOdo
VehicleAge”
“MMRCurrentAuctionCleanPrice”,
“MMRCurrentRetailAveragePrice”,
“WarrantyCost”.

**Chi Square Ranks**

Unbalanced Data :
Best score for All features

Balanced data :
17 Features

2

**Recursive Feature Elimination**

3

MMRAcquisitionAuctionAveragePrice,
MMRAcquisitonRetailCleanPrice,
MMRCurrentAuctionAveragePrice,
MMRCurrentAuctionCleanPrice,
"WarrantyCost"

# Evaluation Criteria

## Balanced Data

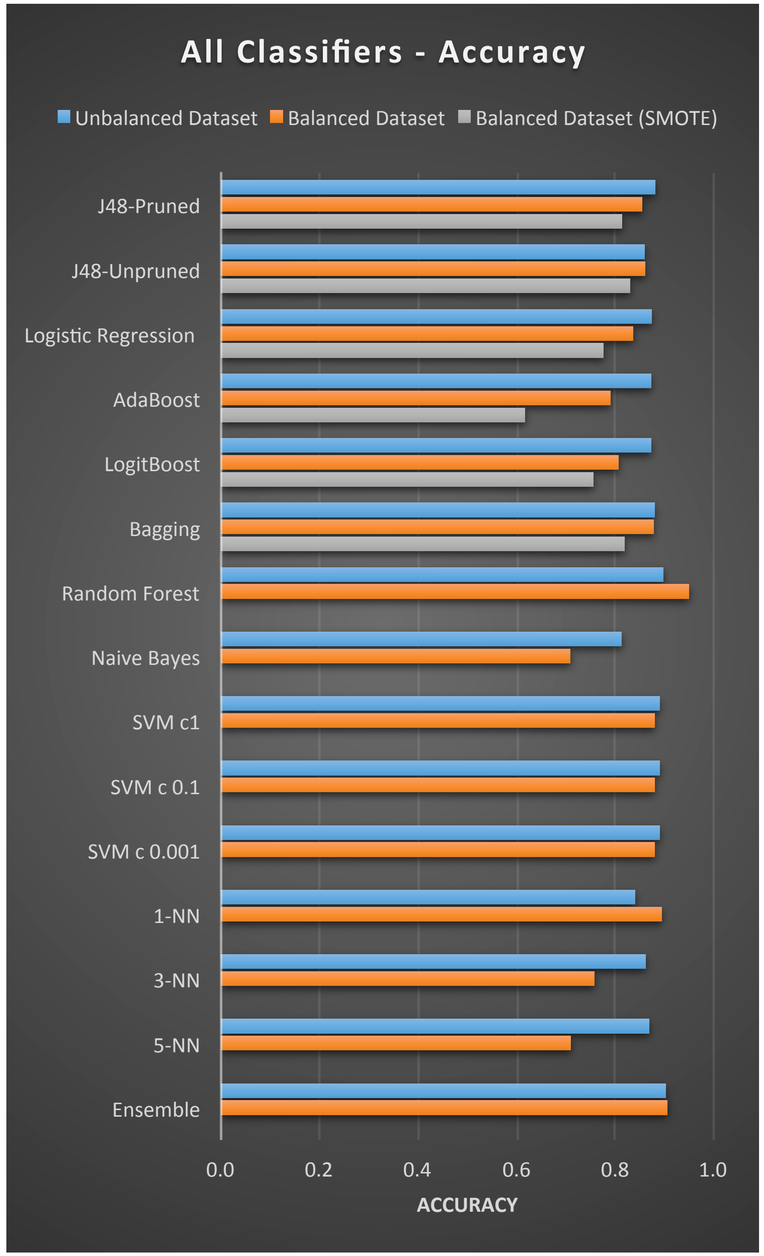
Accuracy

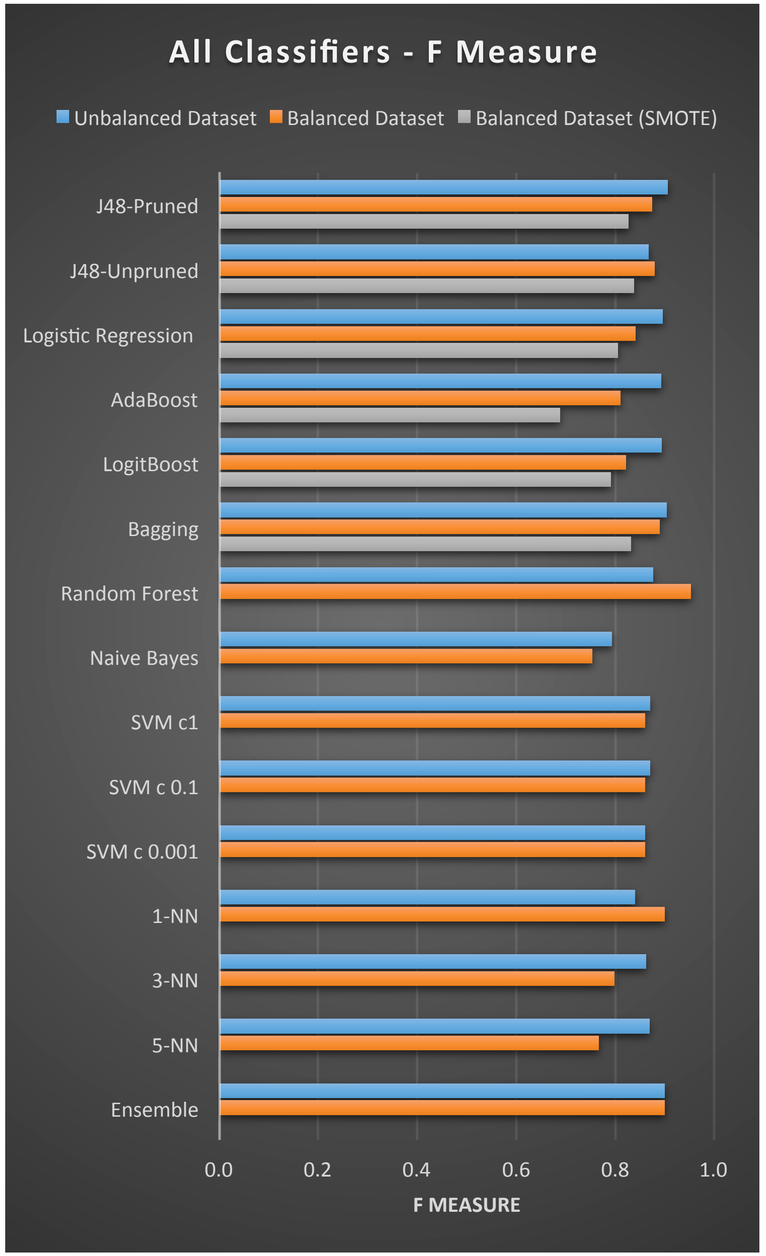
## Unbalanced Data

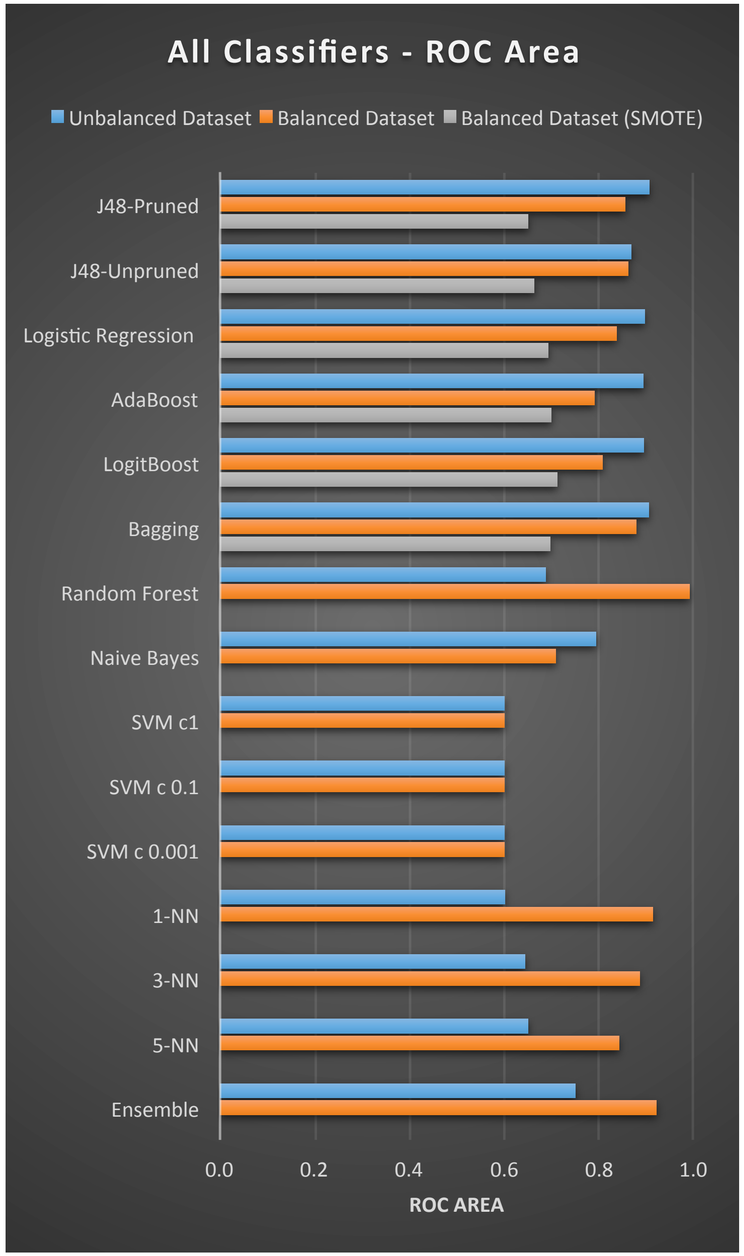
F1 Score

ROC AUC

# Results & Analysis

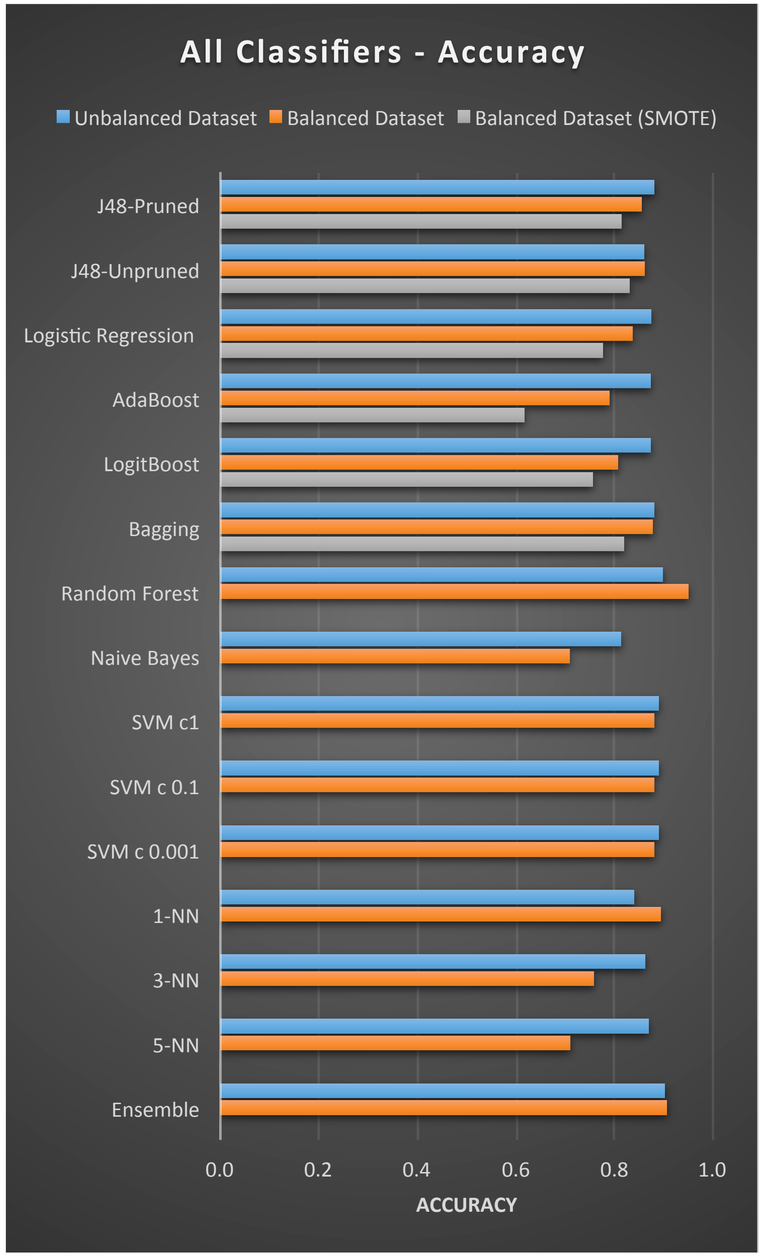


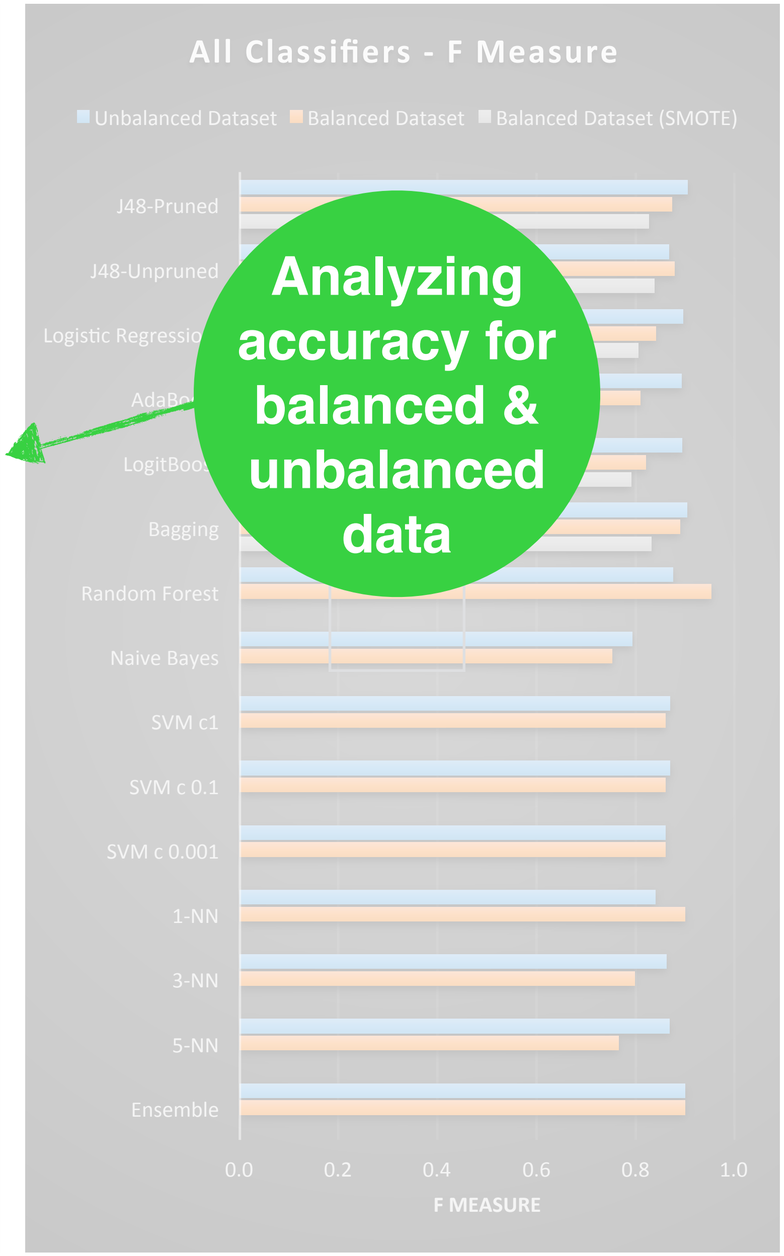



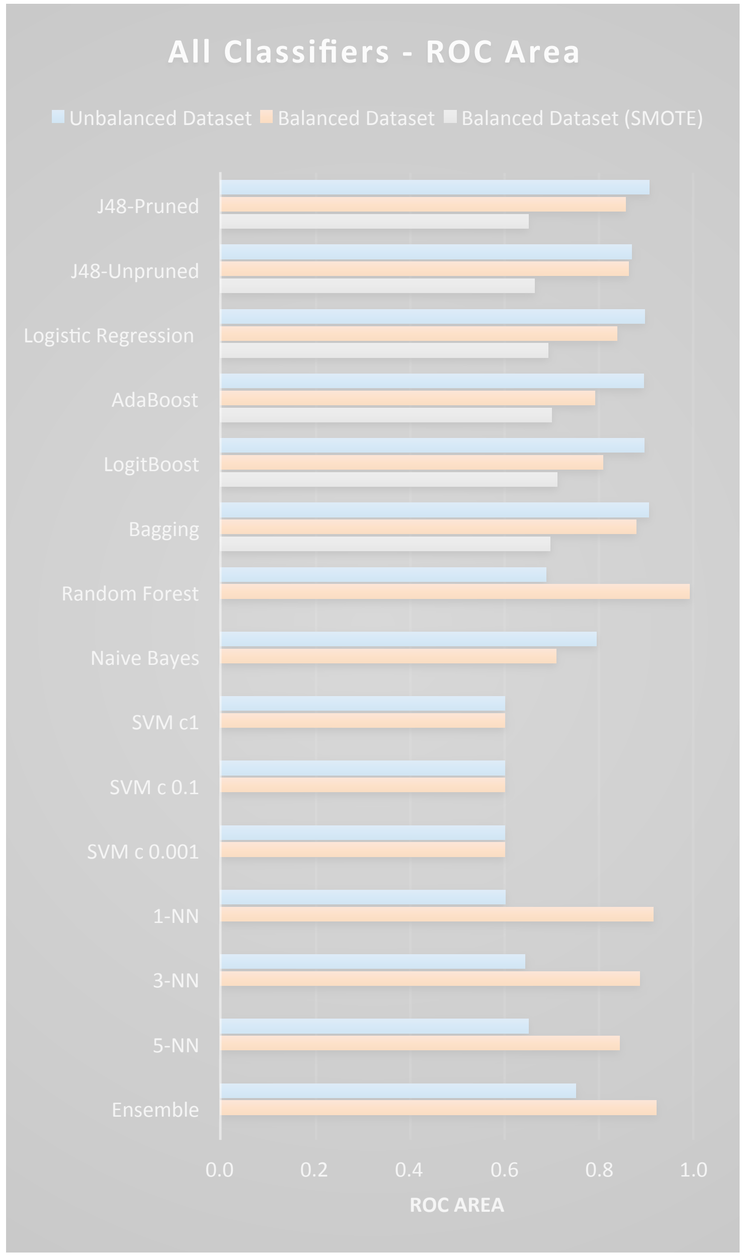
# Results & Analysis





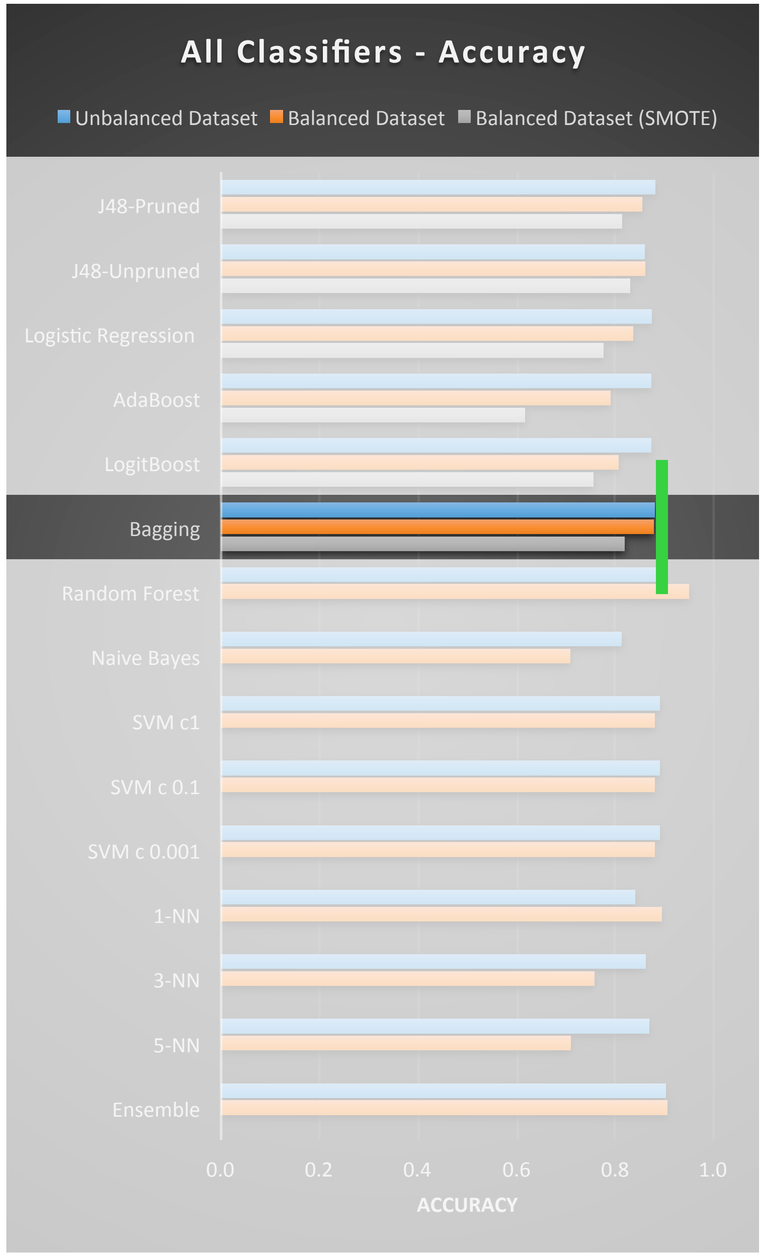



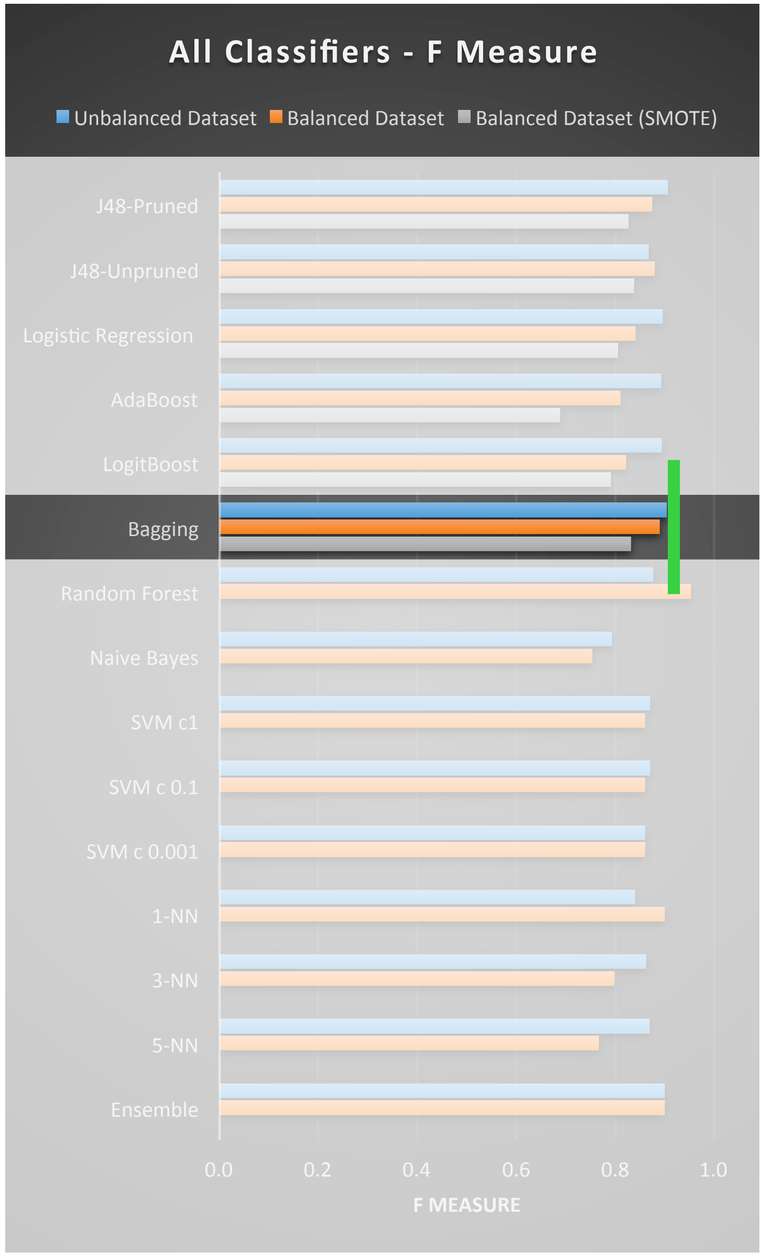

Analyzing accuracy for balanced & unbalanced data

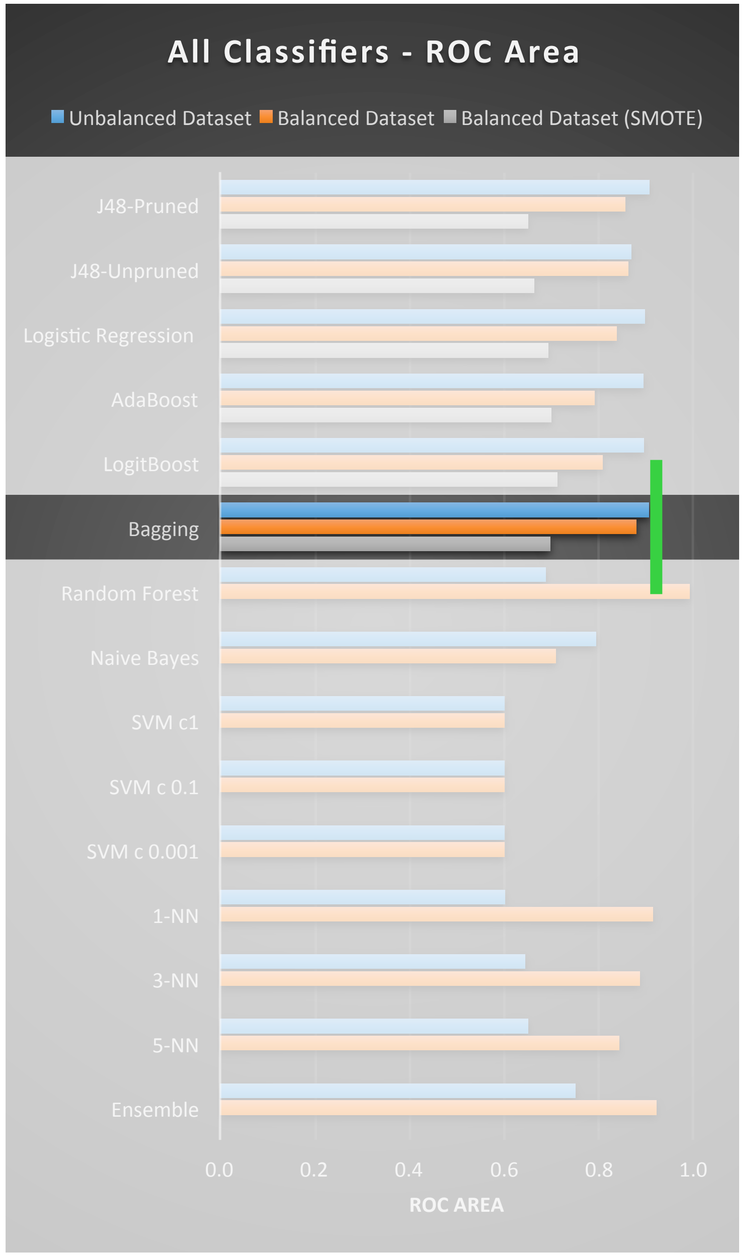


# Results & Analysis



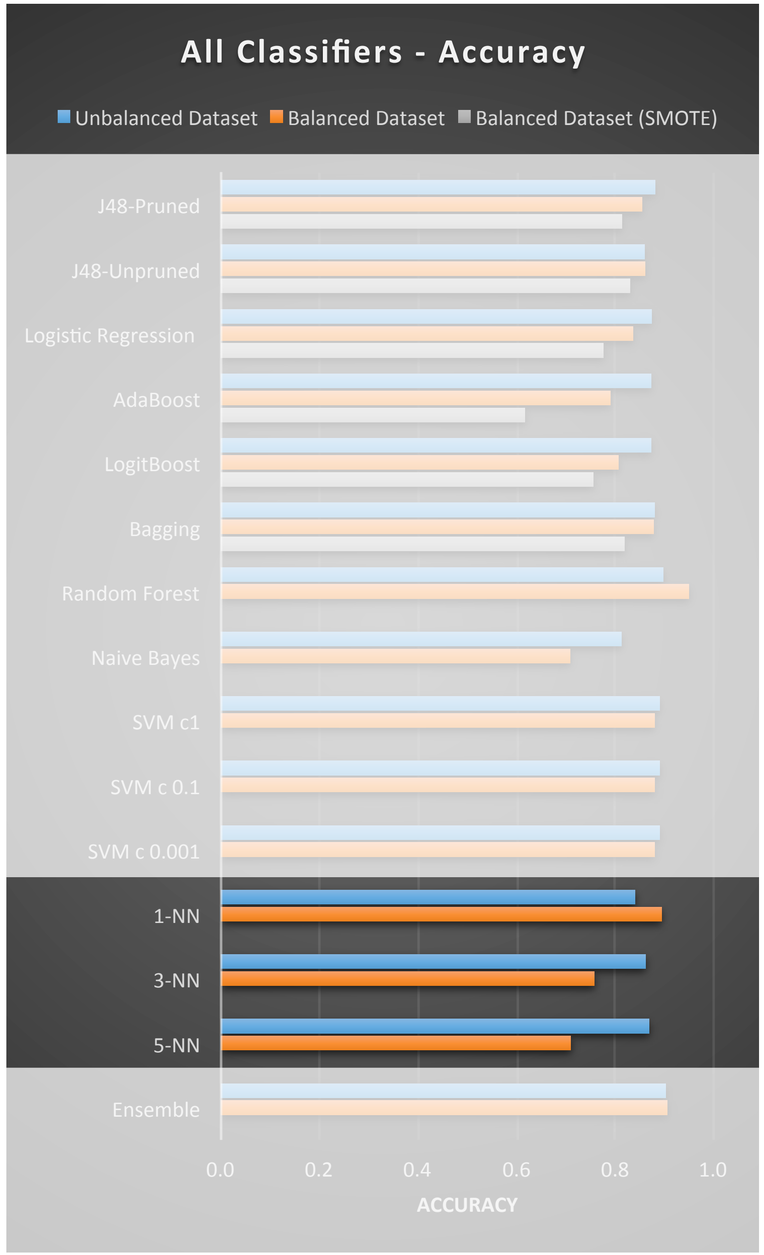


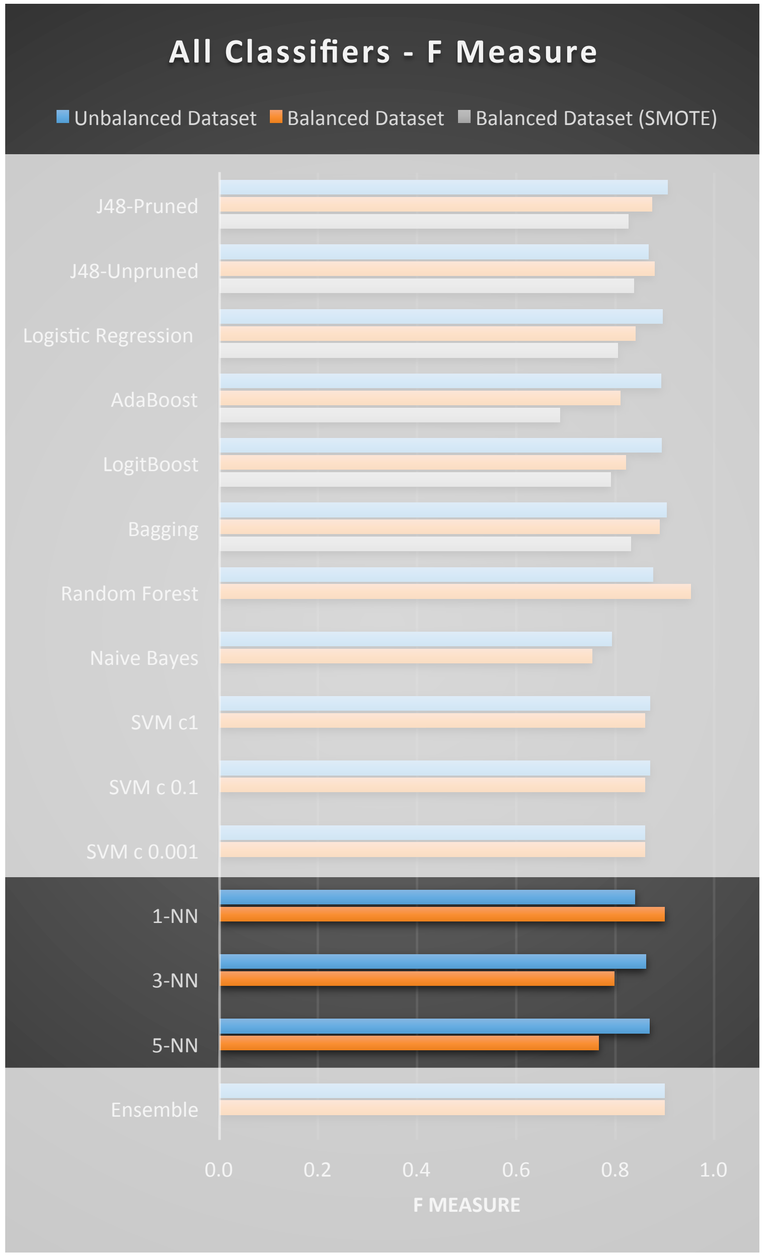



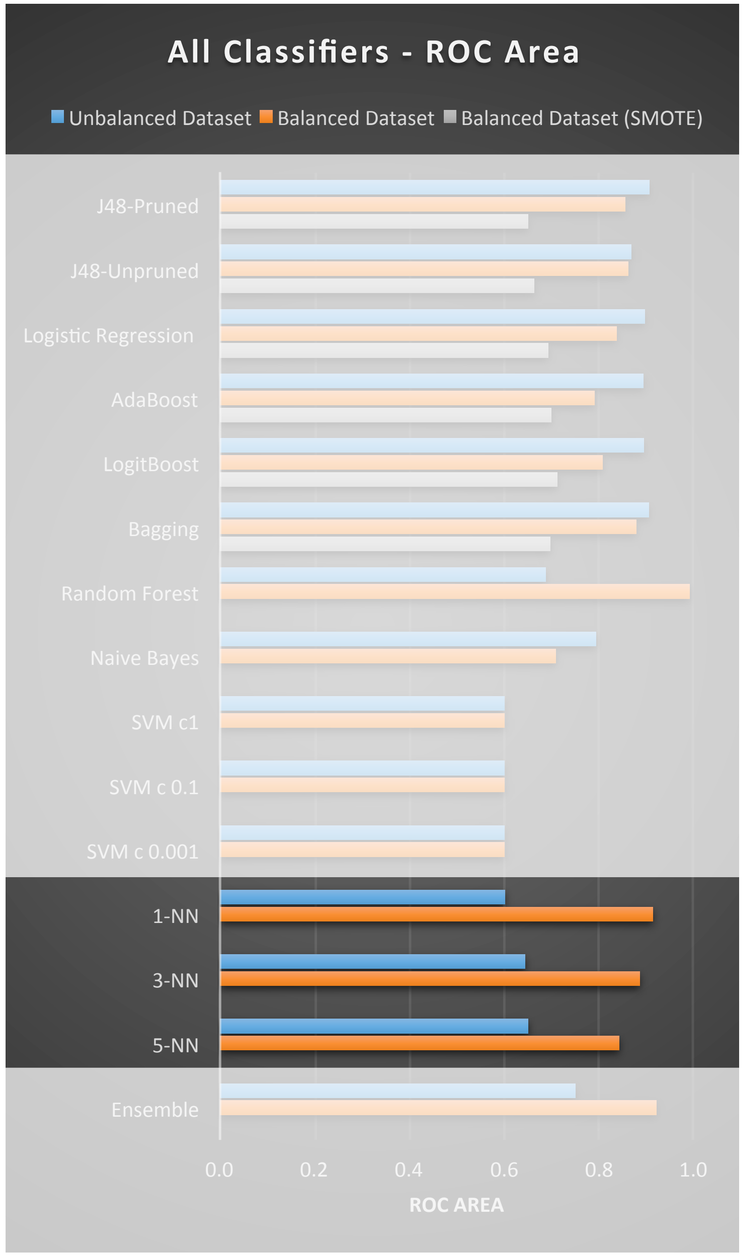



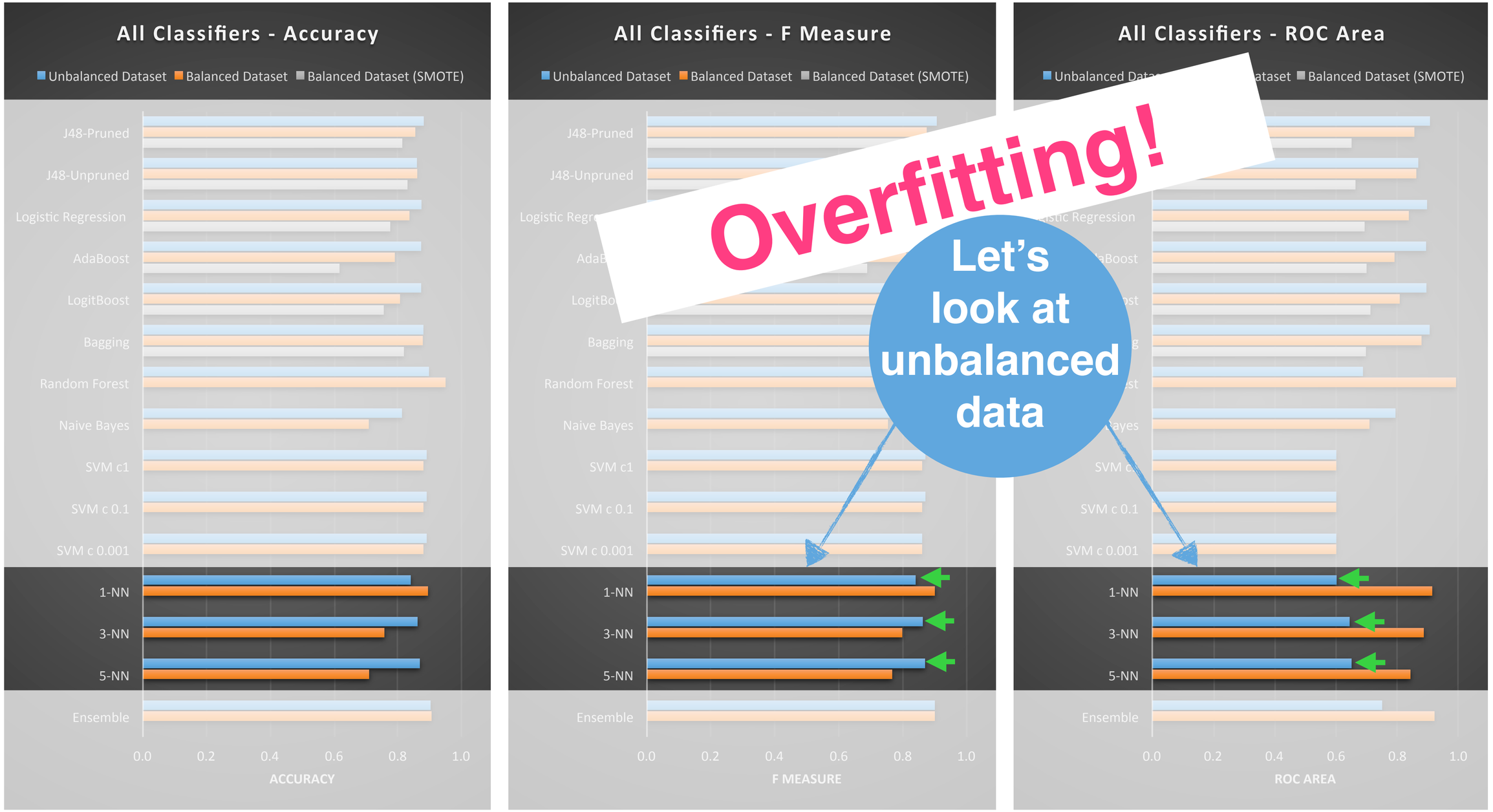
# Results & Analysis





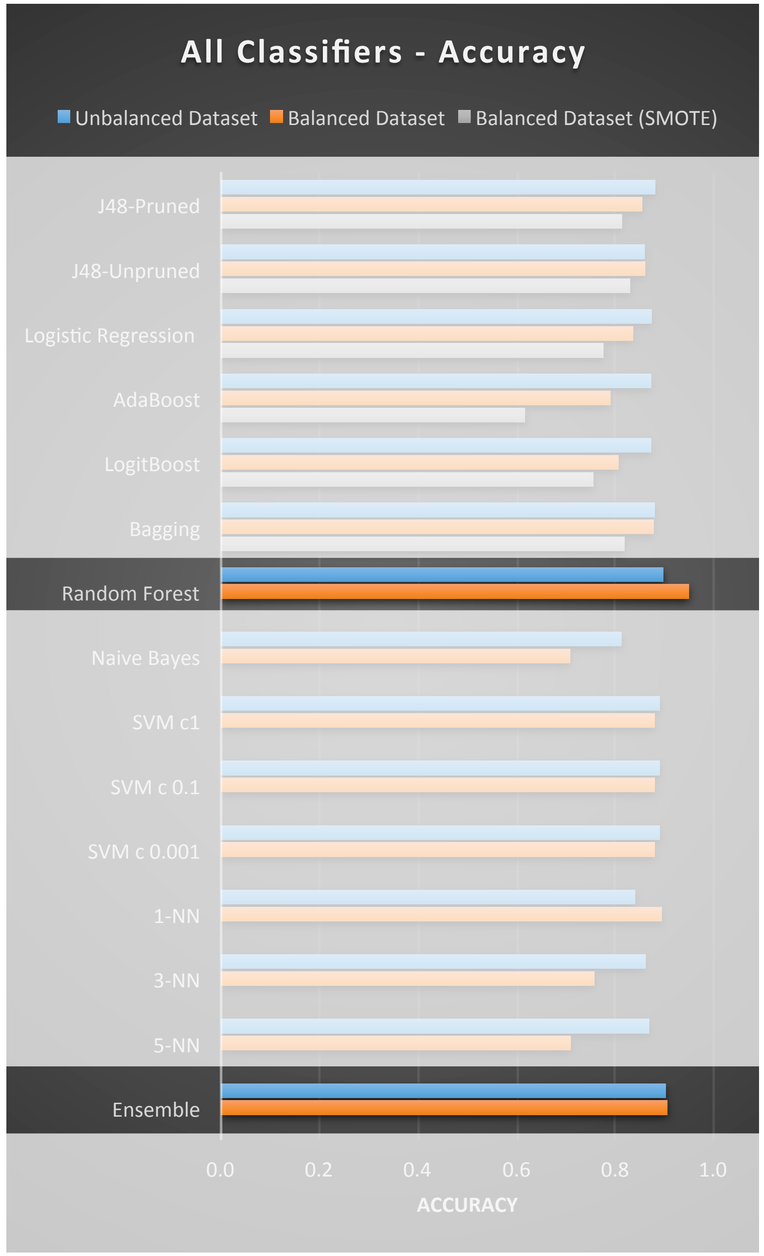



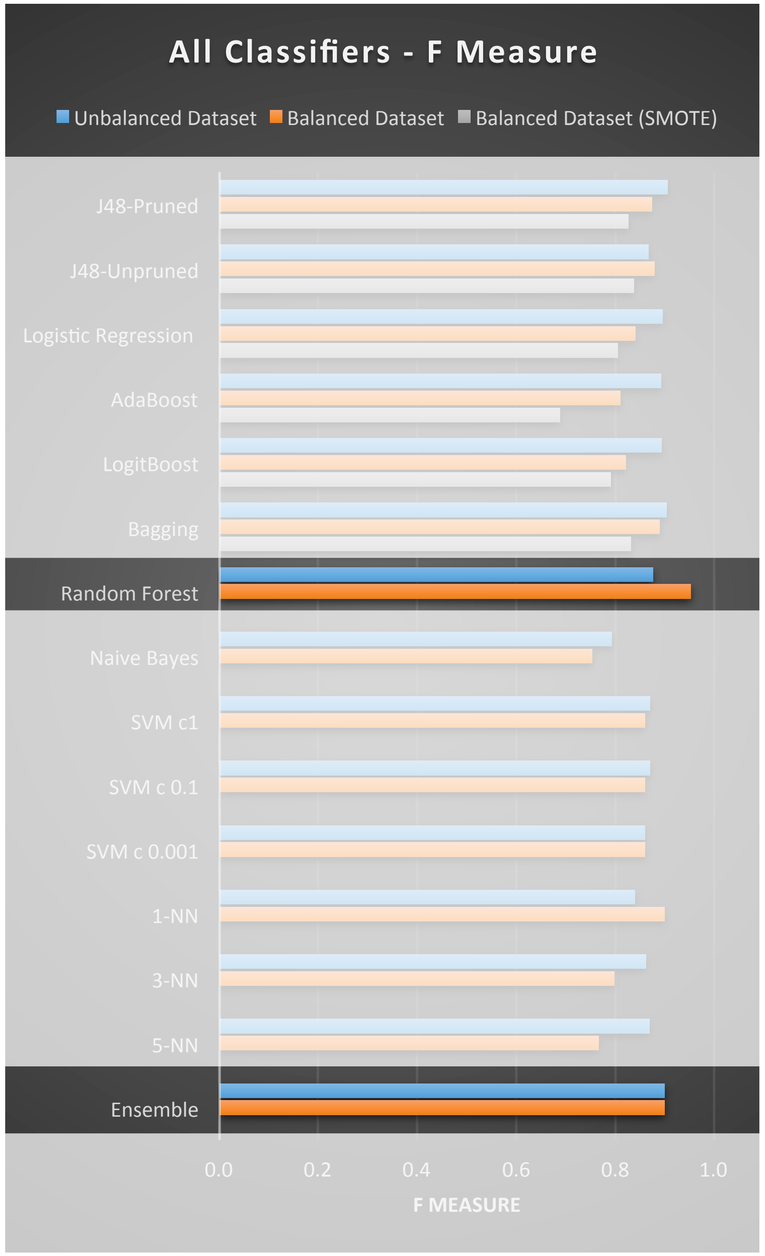

All Classifiers - Accuracy
Unbalanced Dataset
Balanced Dataset
Balanced Dataset (SMOTE)
J48-Pruned
J48-Unpruned
Logistic Regression
AdaBoost
LogitBoost
Bagging
Random Forest
Naive Bayes
SVM c1
SVM c 0.1
SVM c 0.001
1-NN
3-NN
5-NN
Ensemble
0.0
0.2
0.4
0.6
0.8
1.0
ACCURACY
All Classifiers - F Measure
F MEASURE
All Classifiers - ROC Area
ROC AREA
Overfitting!
Let's look at unbalanced data

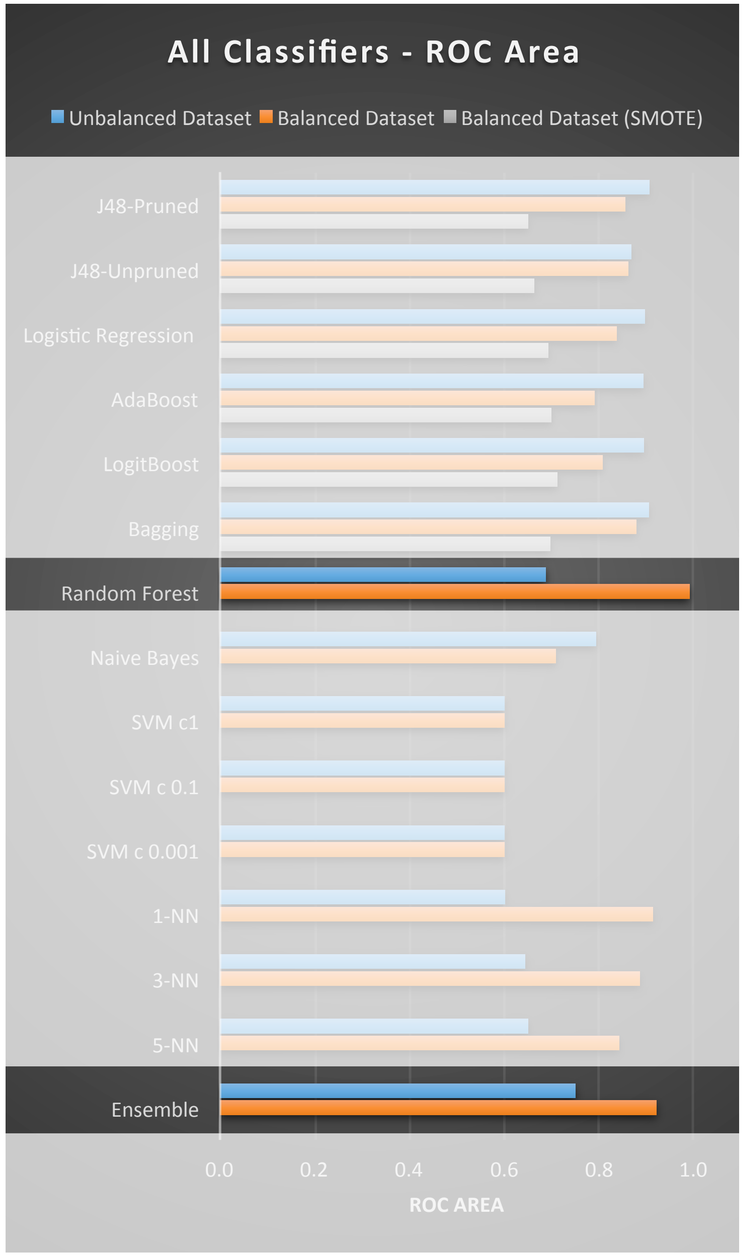

All Classifiers - Accuracy
Unbalanced Dataset
Balanced Dataset
Balanced Dataset (SMOTE)
J48-Pruned
J48-Unpruned
Logistic Regression
AdaBoost
LogitBoost
Bagging
Random Fo
Naive
SVM
SVM c 0.1
SVM c 0.001
1-NN
3-NN
5-NN
Ensemble
0.0
0.2
0.4
0.6
0.8
1.0
ACCURACY
All Classifiers - F Measure
F MEASURE
All Classifiers - ROC Area
ROC AREA
Let's look at balanced data
Performs very Well!

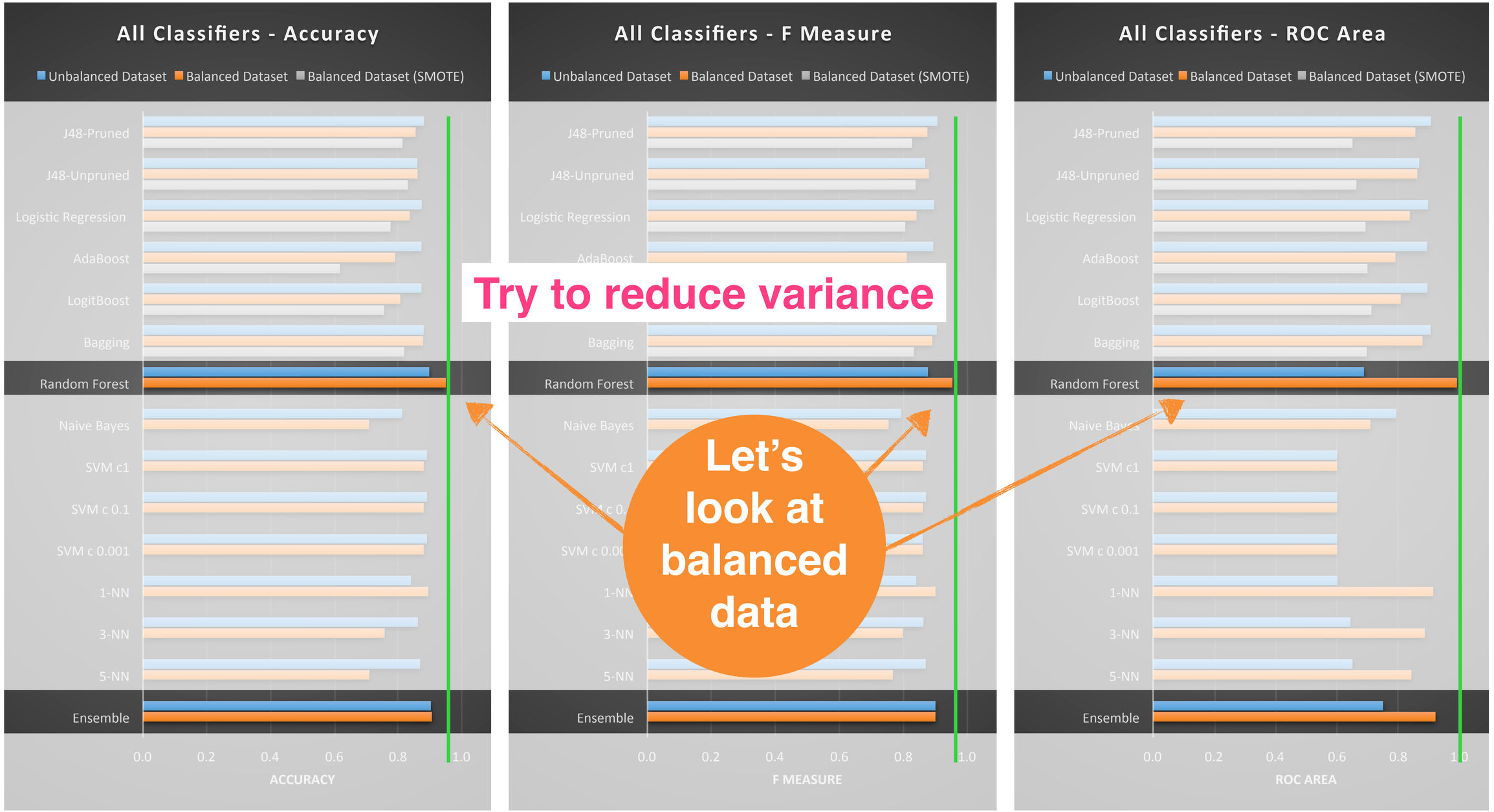
# Results & Analysis





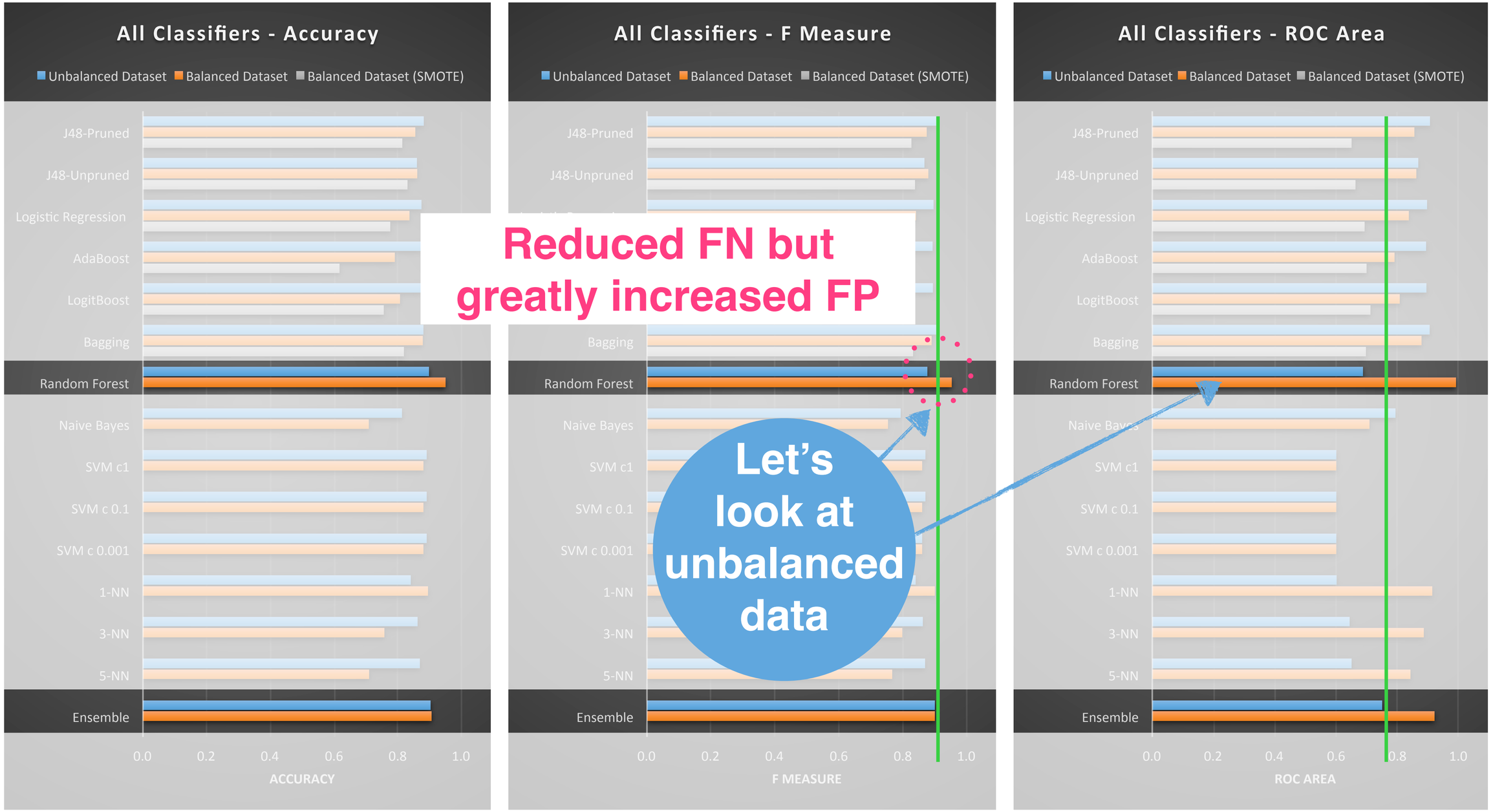




All Classifiers - Accuracy
Unbalanced Dataset
Balanced Dataset
Balanced Dataset (SMOTE)
J48-Pruned
J48-Unpruned
Logistic Regression
AdaBoost
LogitBoost
Bagging
Random Forest
Naive Bayes
SVM c1
SVM c 0.1
SVM c 0.001
1-NN
3-NN
5-NN
Ensemble
0.0
0.2
0.4
0.6
0.8
1.0
ACCURACY
All Classifiers - F Measure
F MEASURE
All Classifiers - ROC Area
ROC AREA
Try to reduce variance
Let's look at balanced data

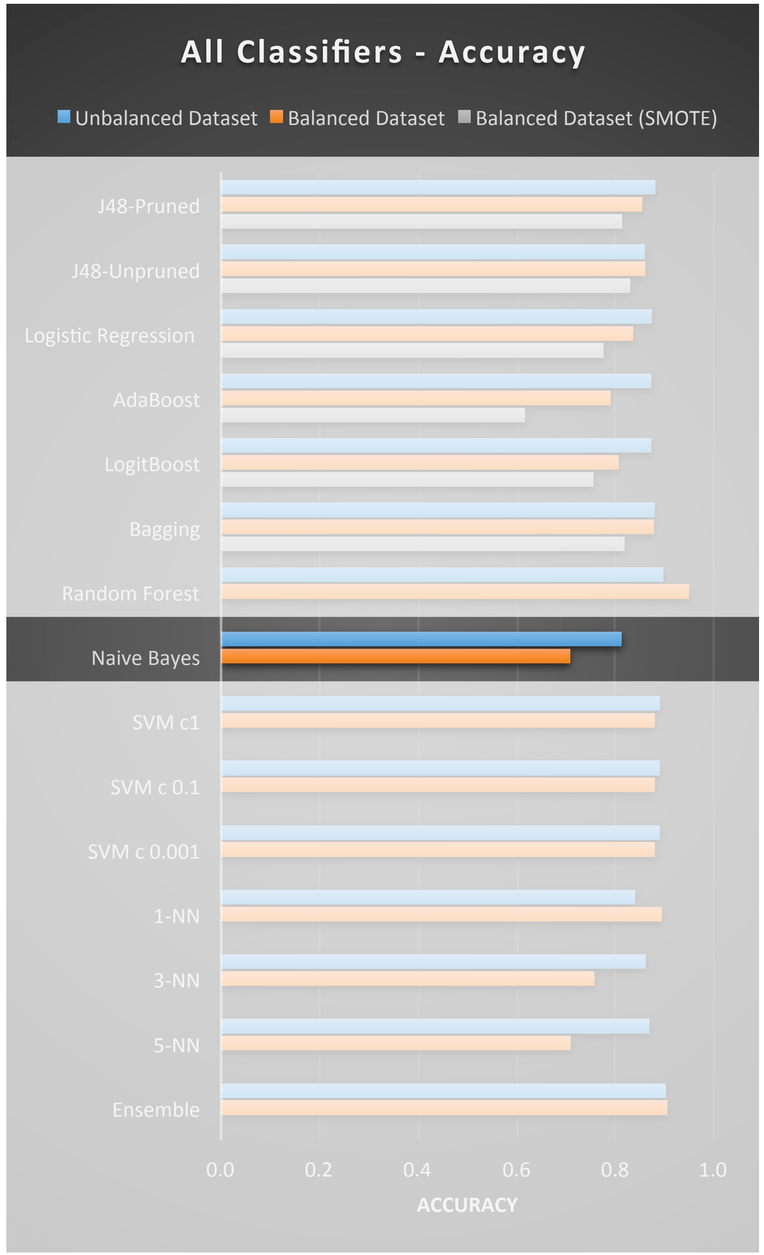
# Results & Analysis

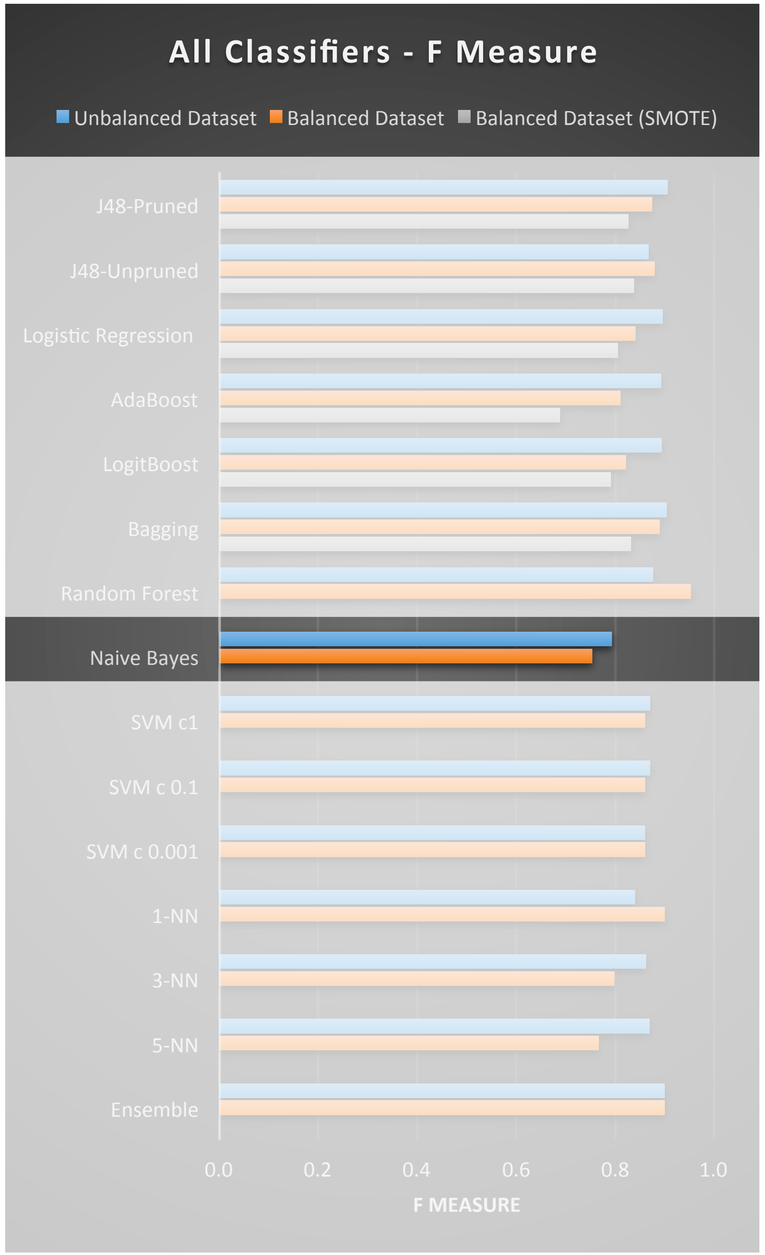


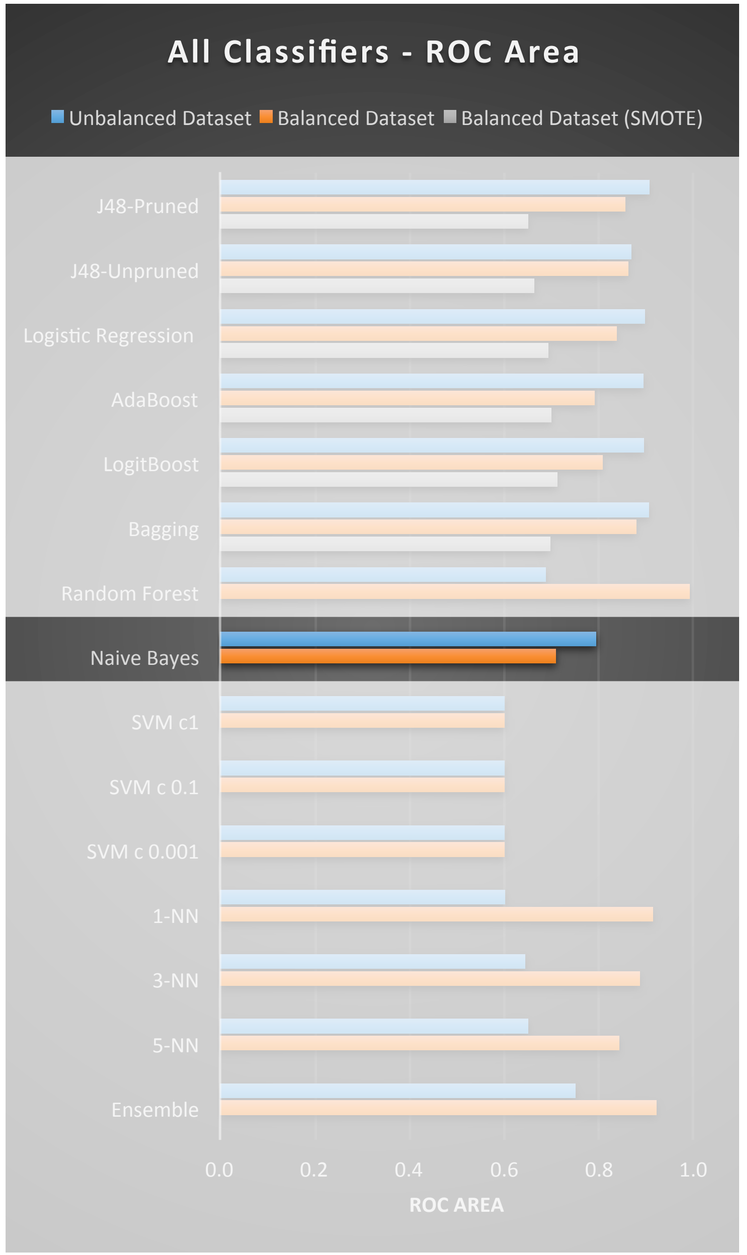


# Results & Analysis



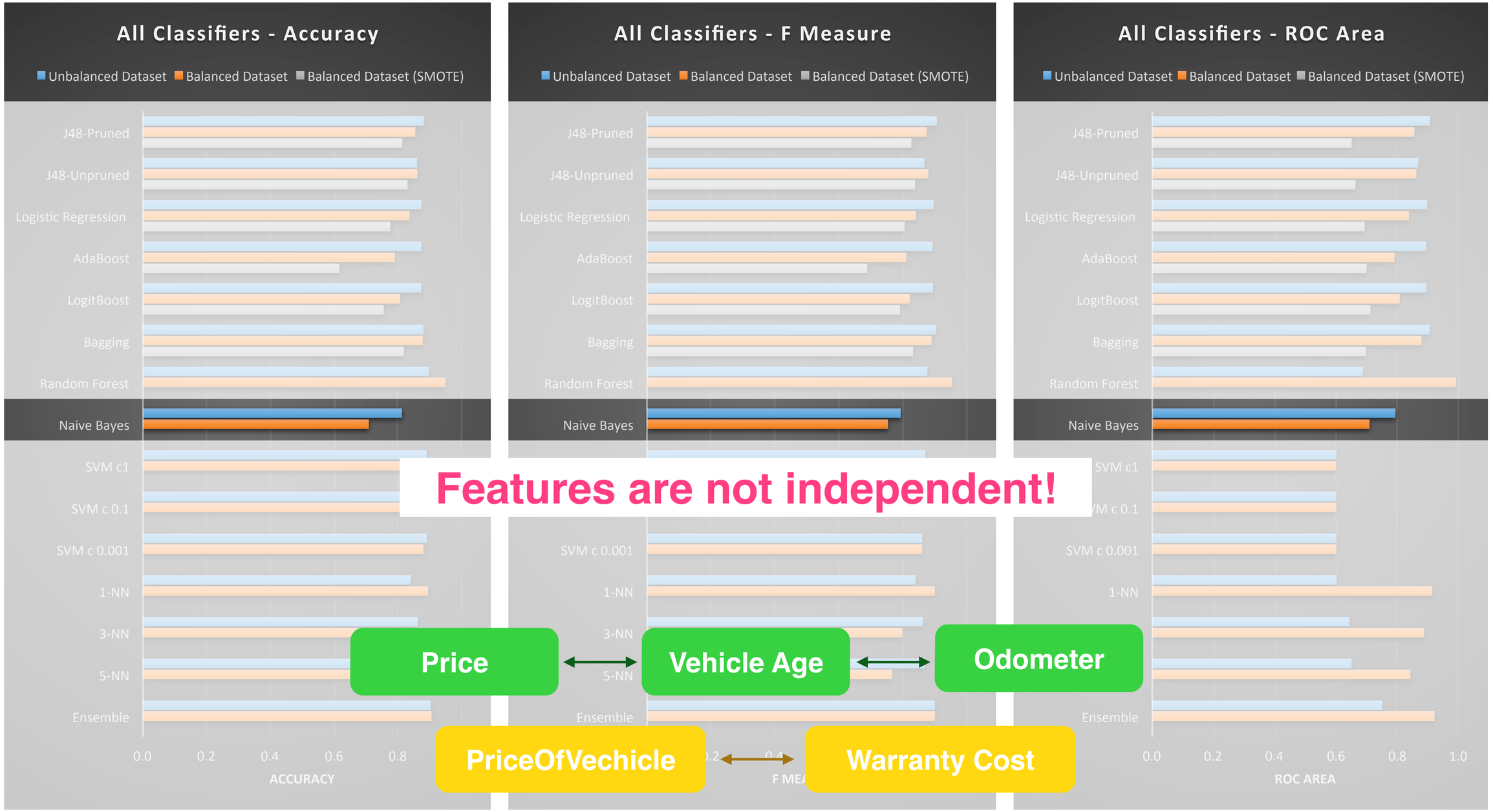


# Results & Analysis





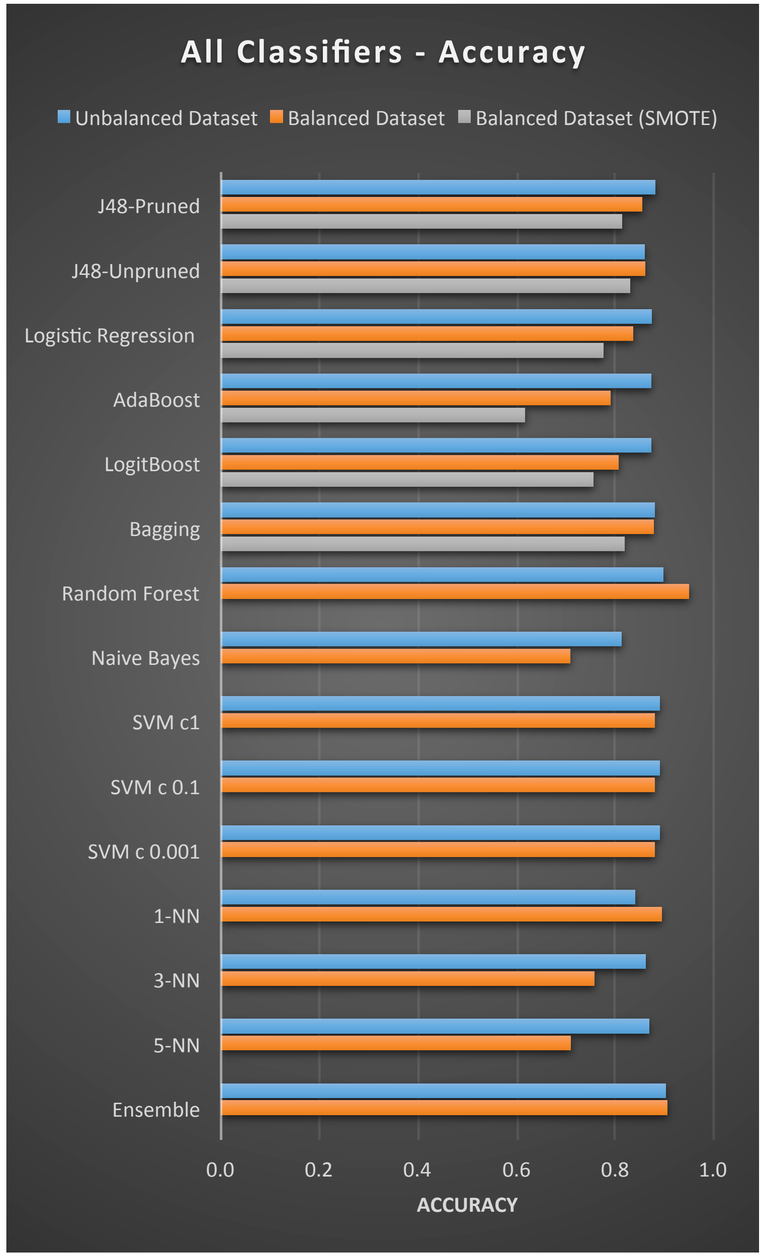

**Features are not independent!**

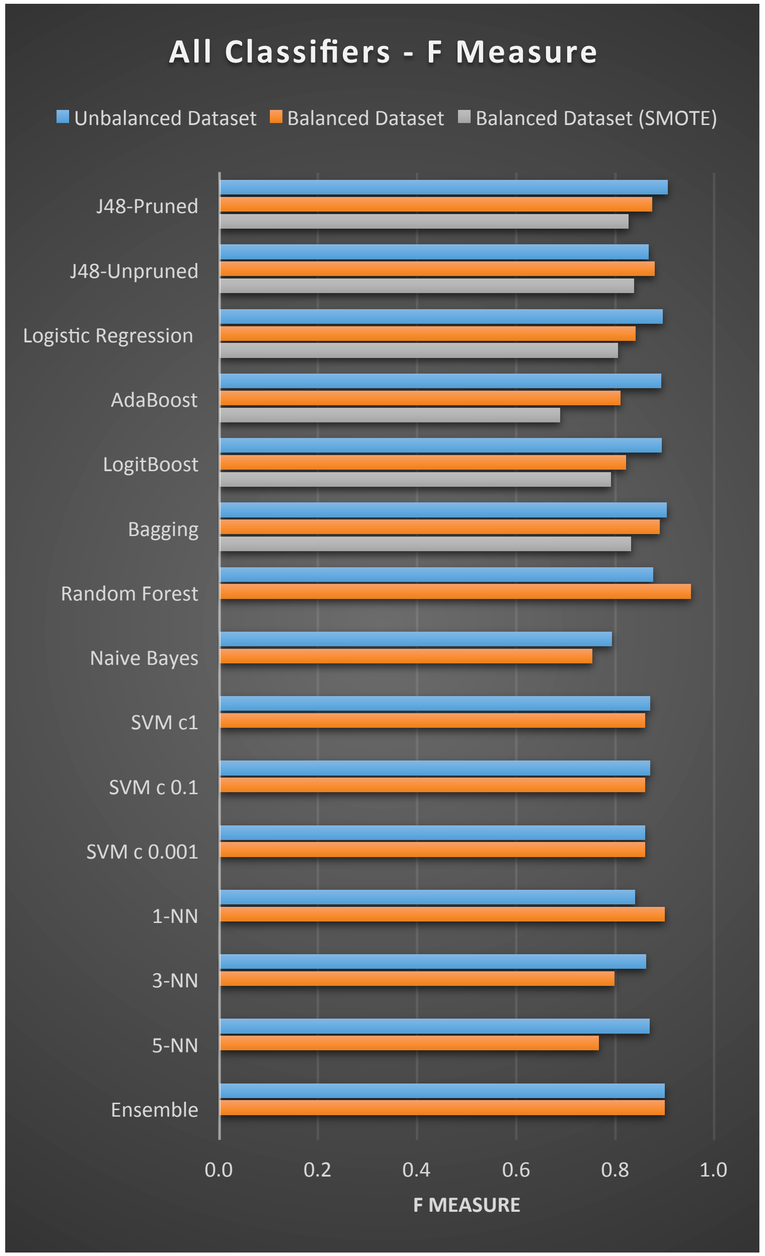
**Price** ⟷ **Vehicle Age** ⟷ **Odometer**

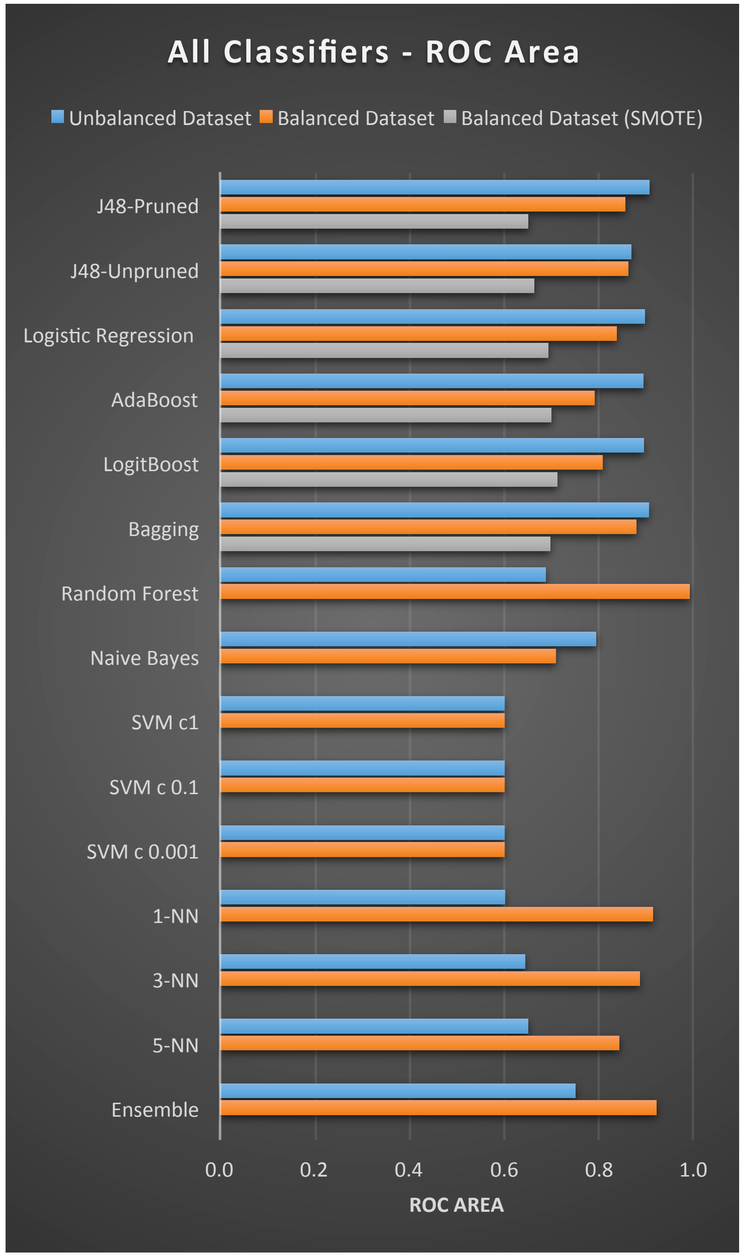
**PriceOfVechicle** ⟷ **Warranty Cost**

# Oversampling with Replacement > SMOTE

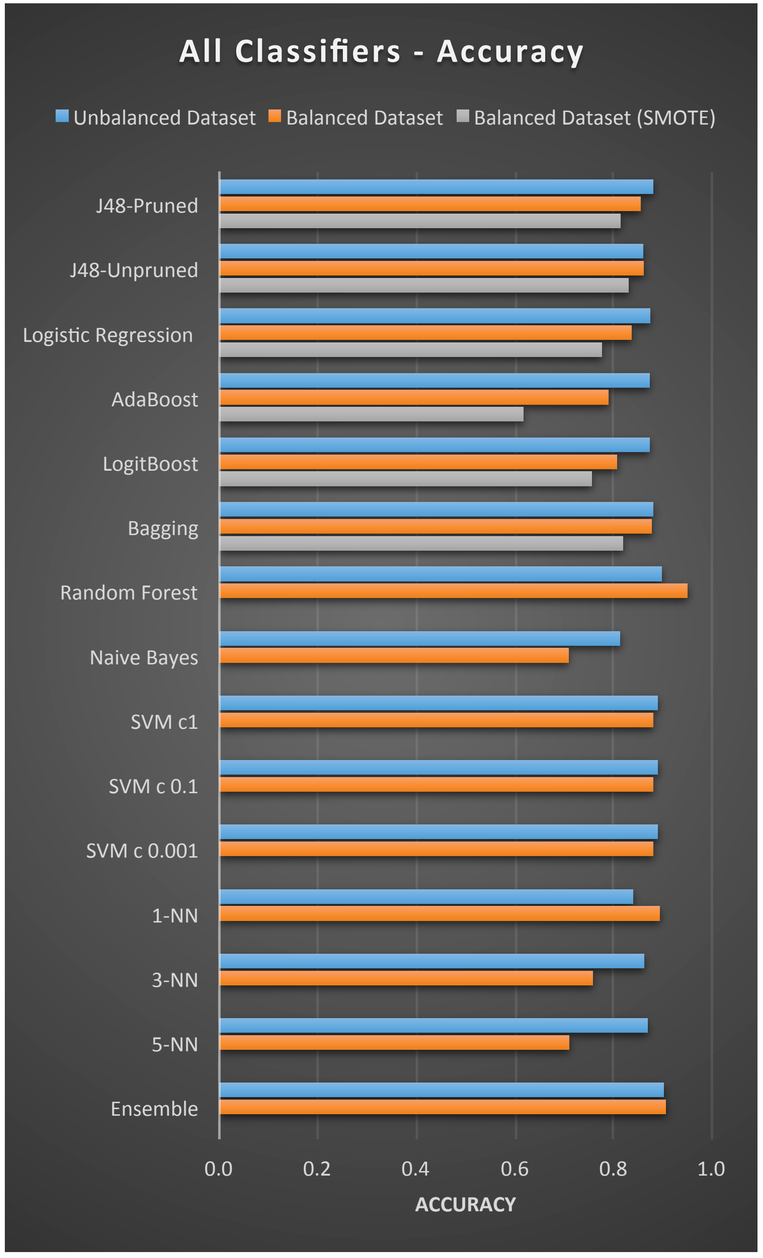




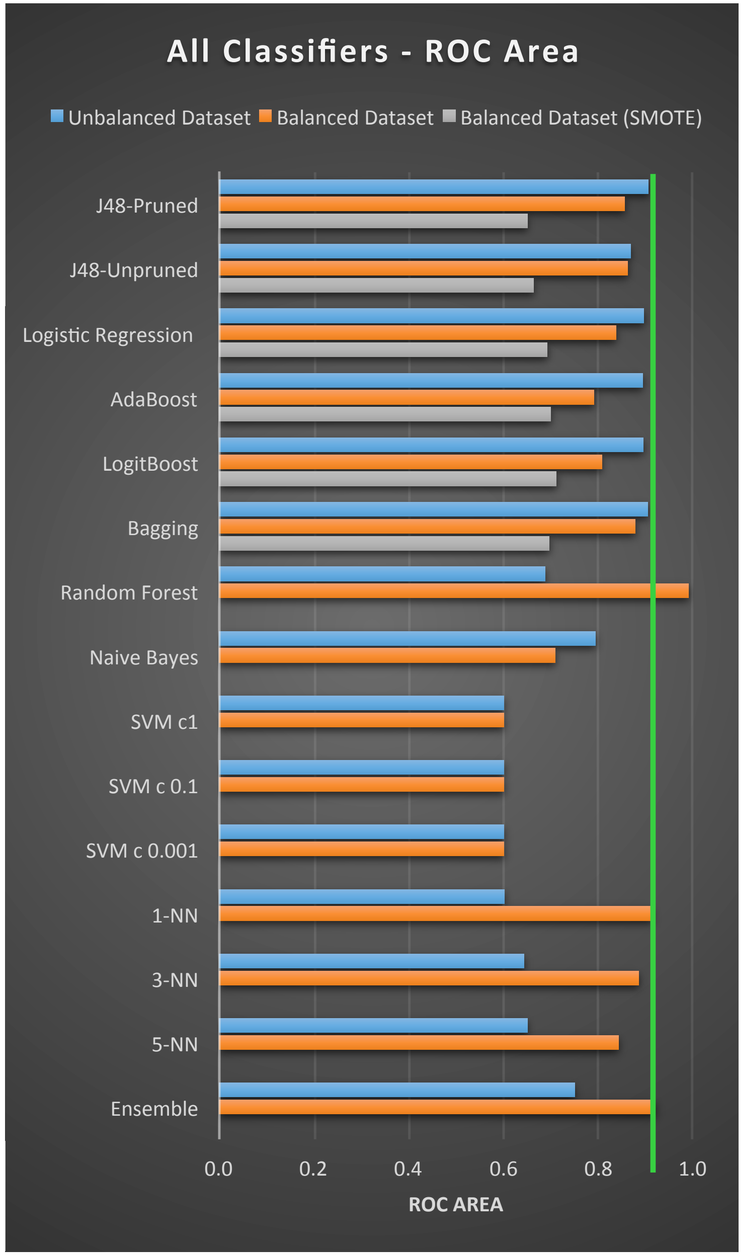



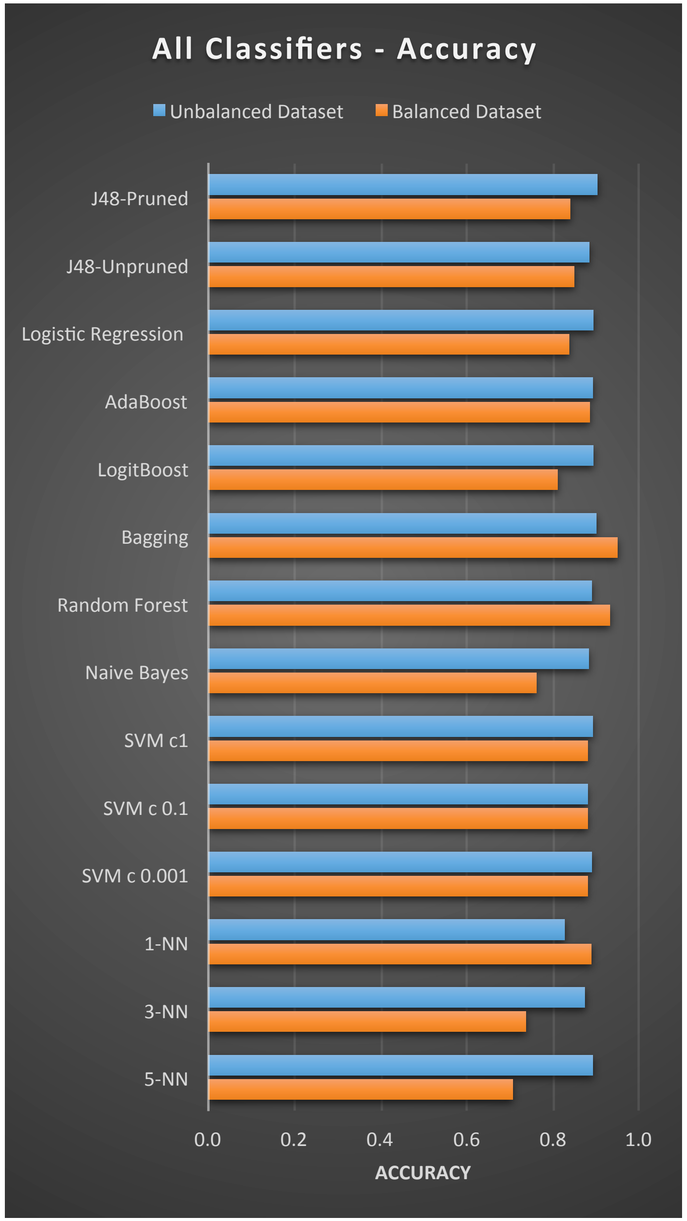
# Best Classifiers — Logistic Regression & Decision Tree

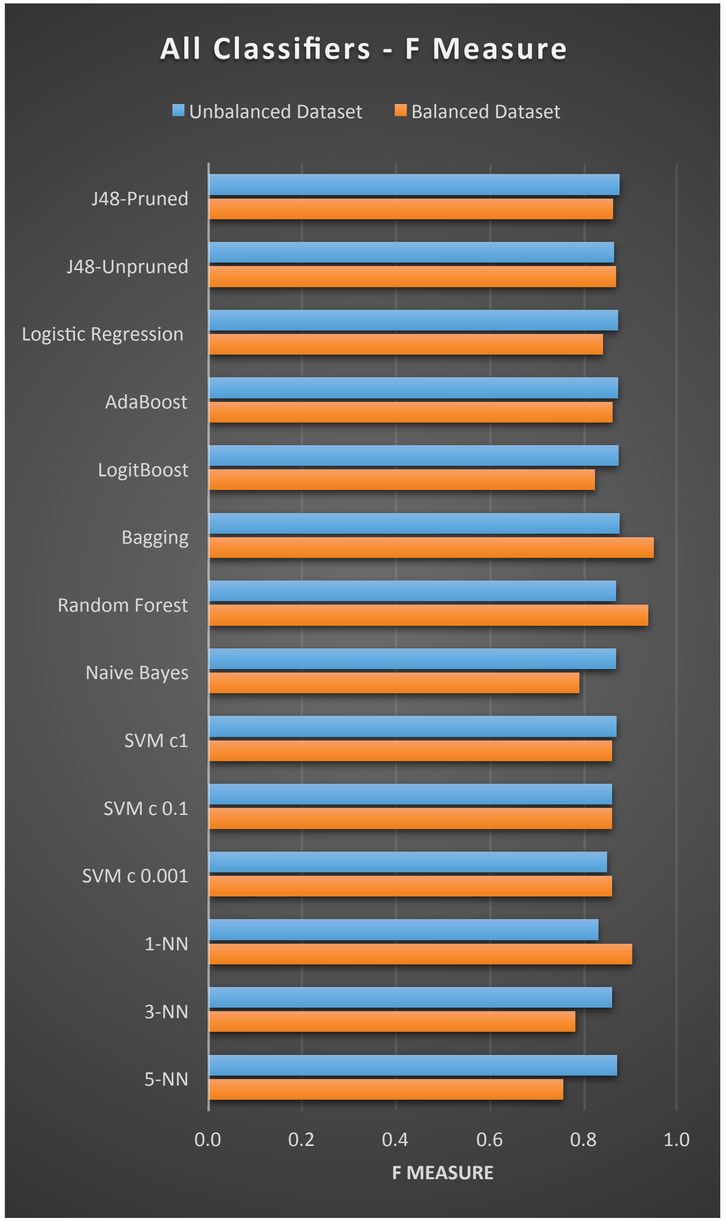


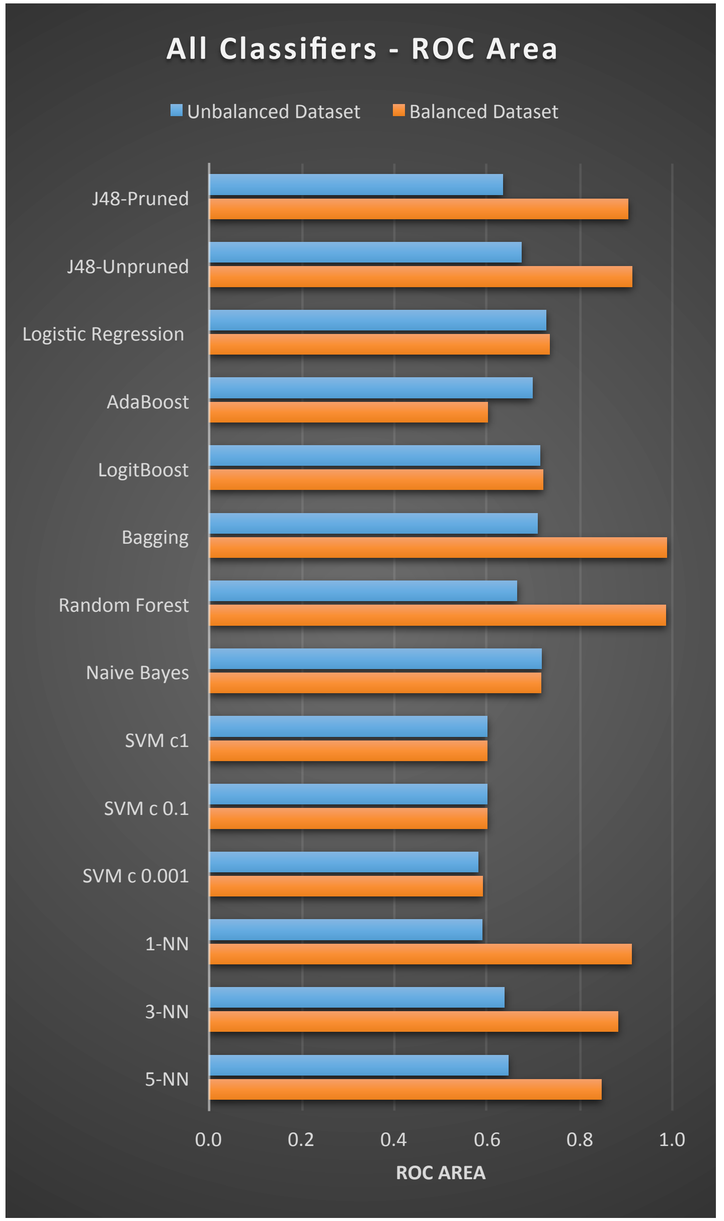




# Expert Handpicked Features



## Key Learnings

- Important features
    - VehOdo
    - VehAge
    - MMRCurrentAuctionCleanPrice
    - MMRCurrentRetailAveragePrice
    - WarrantyCost
    - Wheel type

- In our dataset classifiers work better on Unbalanced data rather than Balanced data

- Decision tree (J48 ) and logistic regression perform best

- Oversampling with replacement work better than SMOTE

- Undersampling the majority class did even WORSE