NKI Breast Cancer Dataset 1

Source: https://data.world/deviramanan2016/nki-breast-cancer-data
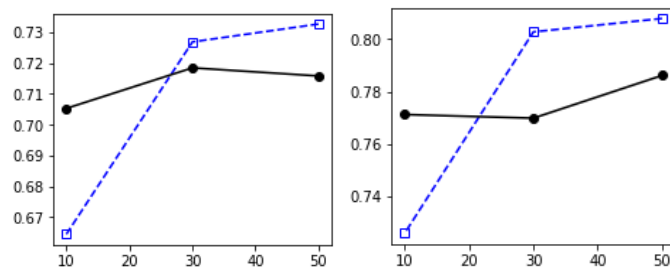
-Sample Size: 272

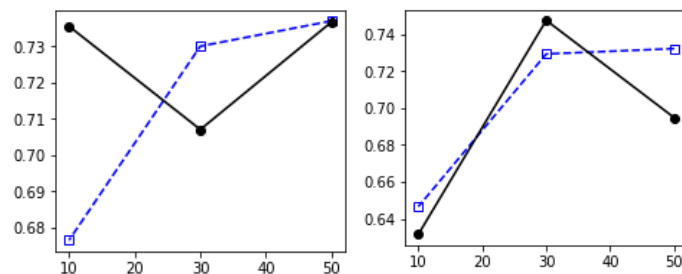-Attributes: 252

-196:76 of 0:1

-Information: This dataset contains information on the symptoms and flares present in different breast cancer patients. The binary dependent variable is represented by eventual death of the patient within the extended course of this study.

For Test Size 0.5 Rescaled
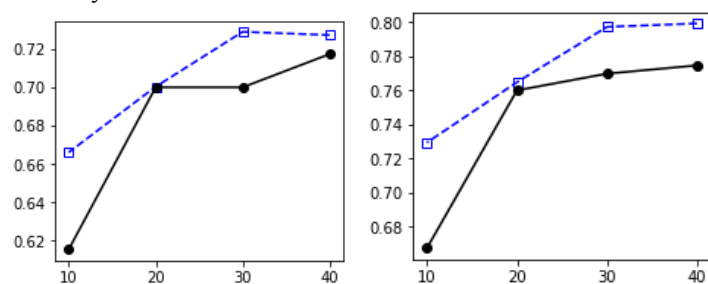
Mean Accuracy and Area Under the Curve:
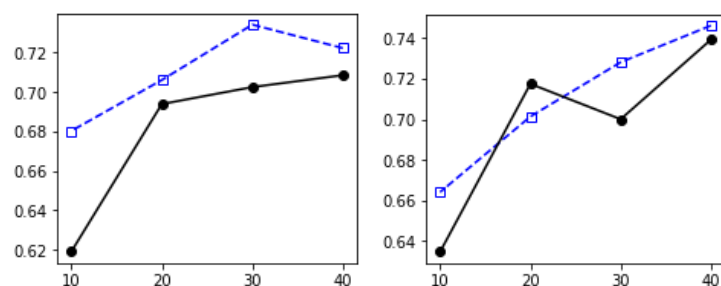

Precision and Recall:

For Test Size 0.6 Rescaled
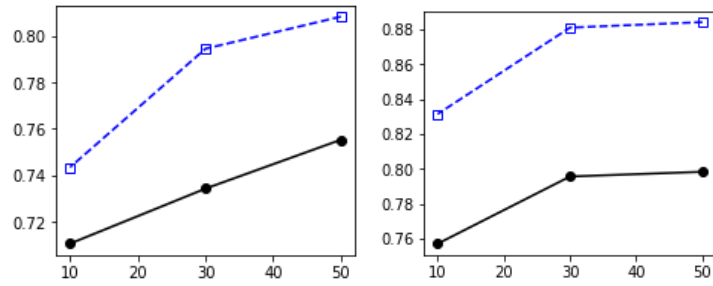
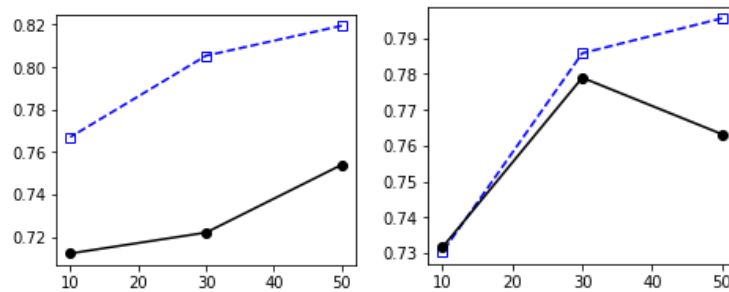Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.5 Not scaled
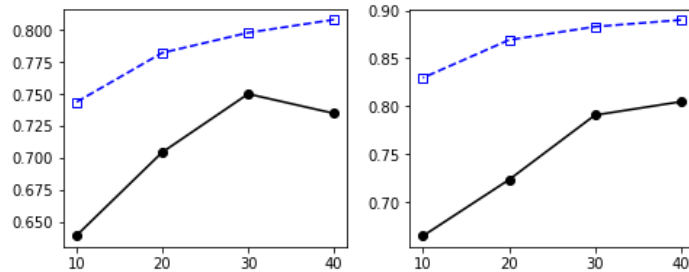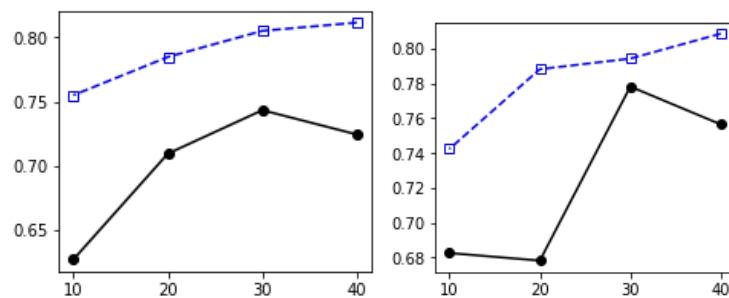Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled
Accuracy and Area Under the Curve:

Precision and Recall:

Peptide and Cancer Dataset 2
Source:
-Sample Size: 699
-Attributes: 58
-676:23 of 0:1
-Information: This dataset includes information on the association between a variety of factors including elevated plasma levels of natriuretic peptides in cancer patients with the eventual death of these patients, the binary dependent variable.

For Test Size 0.5 Rescaled (Only .5 and .3 Train Sizes)

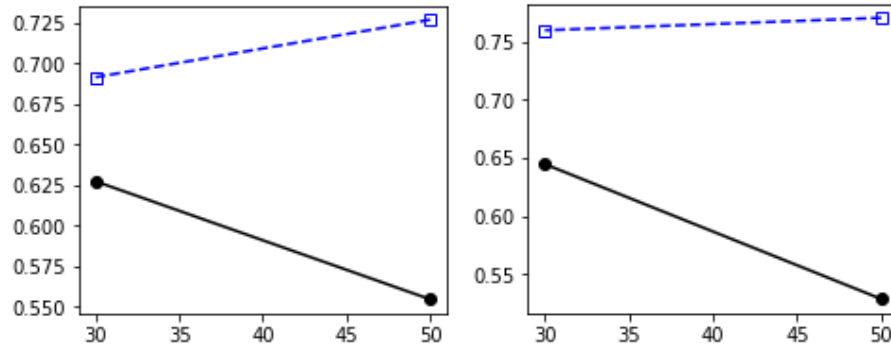Mean Accuracy and Area Under the Curve:
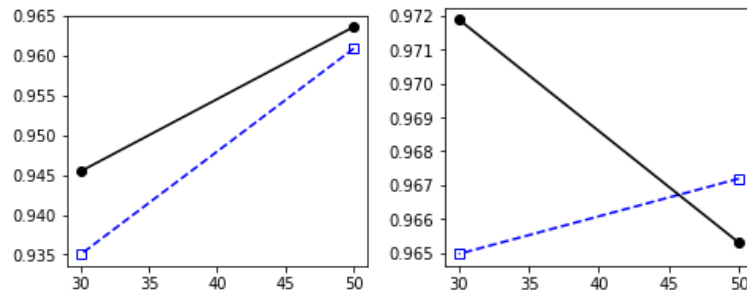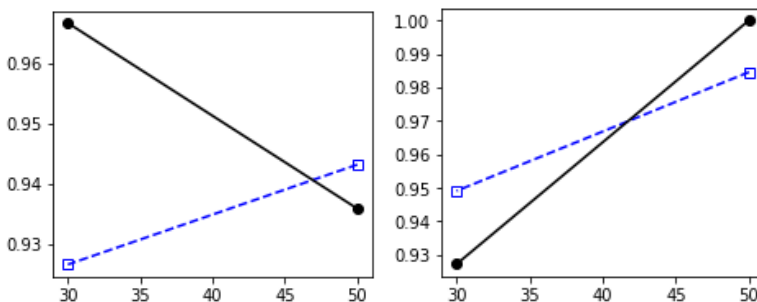
Precision and Recall:

For Test Size 0.6 Rescaled (error with logistic regression--- keeps plotting same y values as before, and reports error with dataset, even though it works for ER)
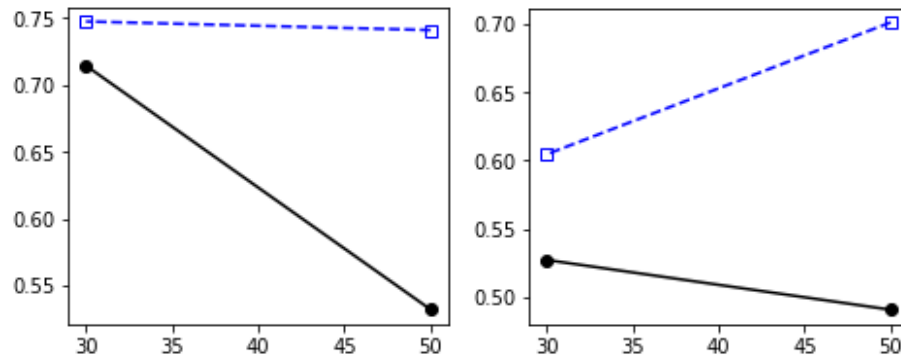
Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.5 Not scaled (Only .5 and .3 Train Sizes)
Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

Accuracy and Area Under the Curve:

Precision and Recall:

Antiseptic for Amputation Dataset 3

Source: http://users.stat.ufl.edu/~winner/data/lister.dat

-Sample Size: 75

-Attributes: 4

-22:53 of 0:1

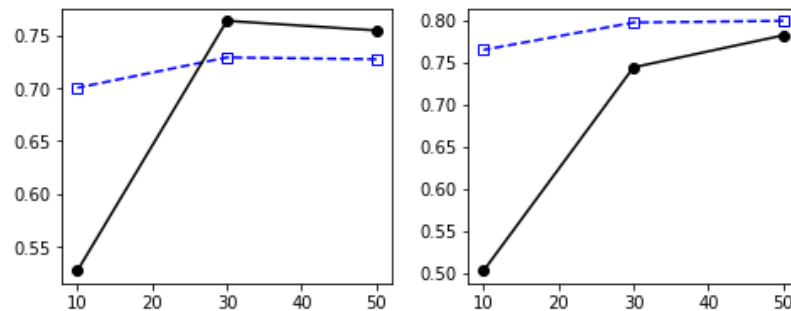-Max ACC: ER: 0.75454545; LR: 0.73045455

-Max AUC: ER: 0.78181818; LR: 0.7463306
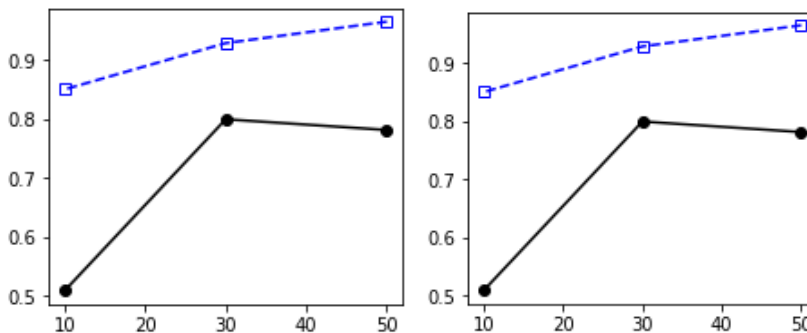
-Max Prec: ER: 0.74300699; LR: 0.71563223

-Max Recall: ER: 0.78181818; LR: 0.79

-Information: This data contains information on different amputees, and the binary dependent variable represents if antiseptic was able to successfully act as a treatment or not.

Mean Accuracy and Area Under the Curve:
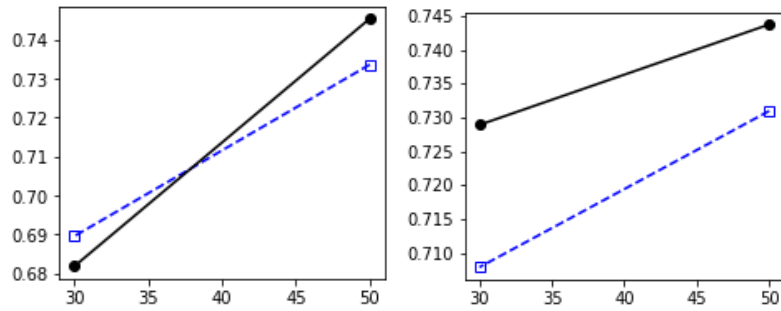

Precision and Recall:

For Test Size 0.6 Rescaled (error with logistic regression---keeps plotting same y values as before, and reports error with dataset, even though it works for ER)
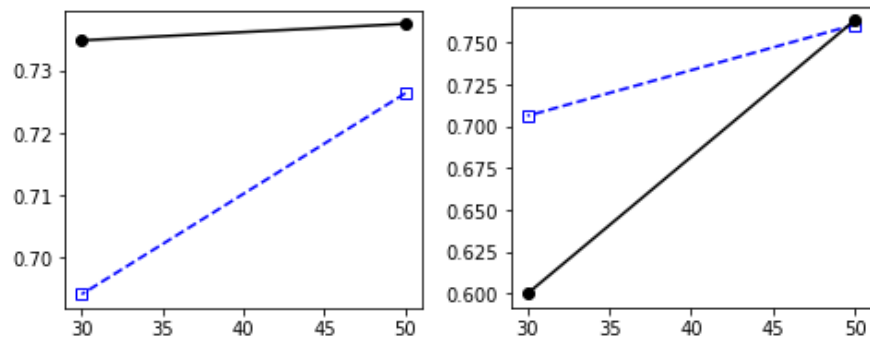
Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.5 Not scaled
Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

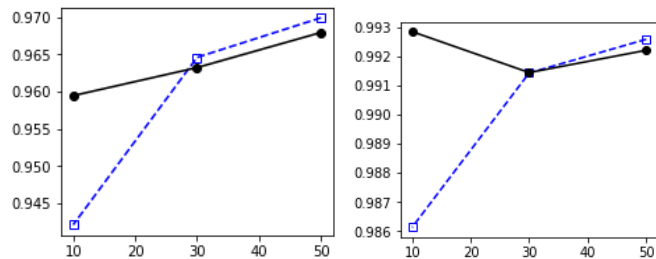Accuracy and Area Under the Curve:

Precision and Recall:

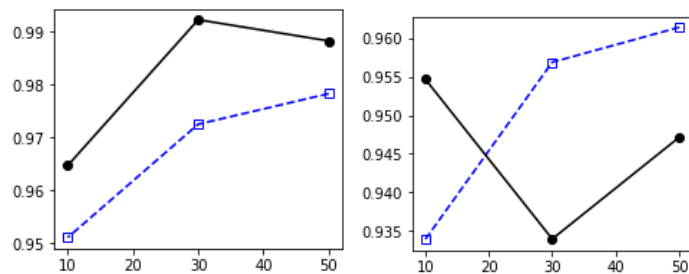-Sample Size: 569
-Attributes: 30
-357:212 of 0:1
-Information: This dataset provides information on a variety of details of lab work done on patients with breast
tumors, and then classifies them as either benign or malignant. In adjusting the dataset to work with the code, benign
tumors were set as 0 and malignant as 1.
Mean Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Rescaled
Accuracy and Area Under the Curve:
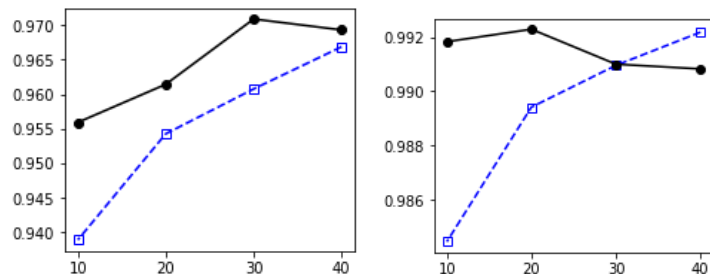
Precision and Recall:
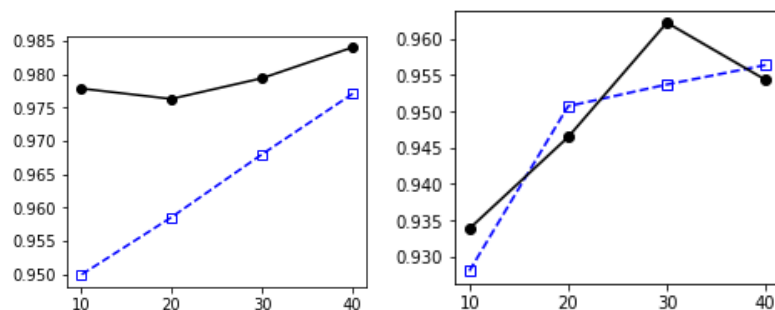
For Test Size 0.5 Not scaled
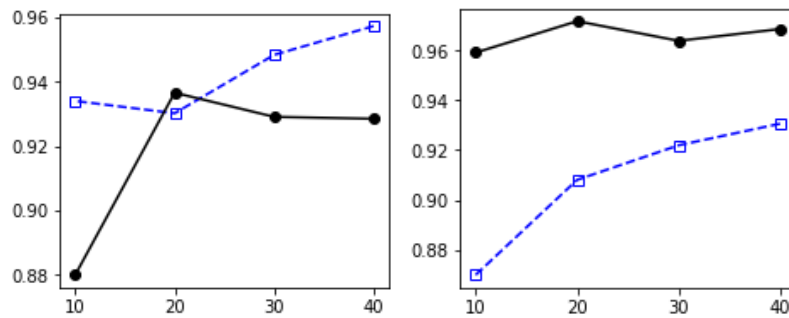Accuracy and Area Under the Curve:
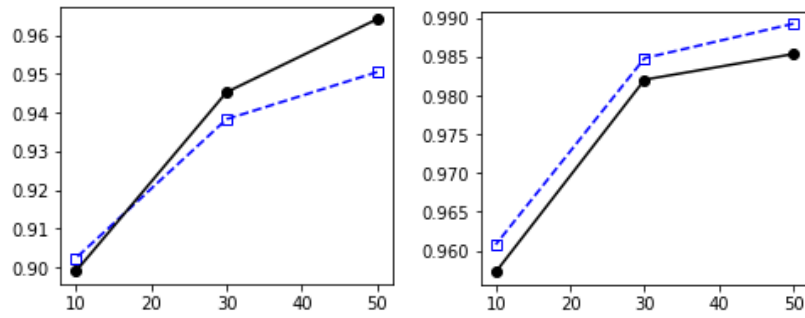
Precision and Recall:
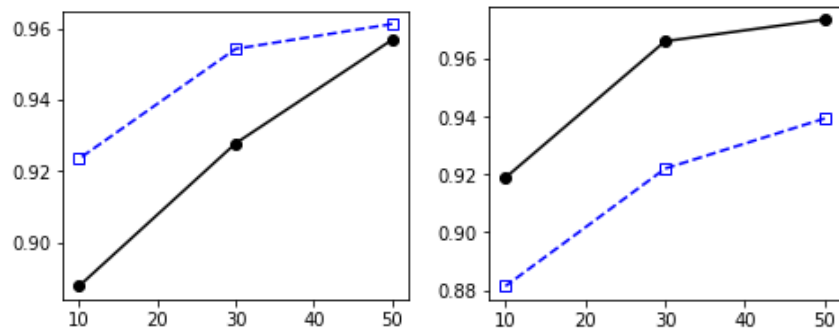
For Test Size 0.6 Not scaled
Accuracy and Area Under the Curve:
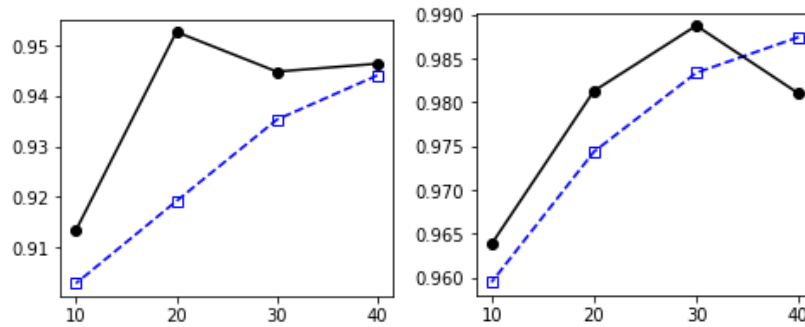
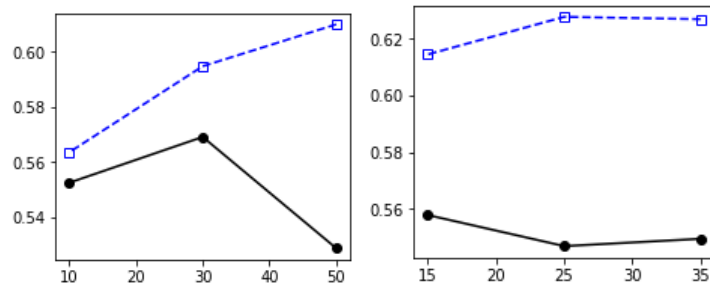Precision and Recall:

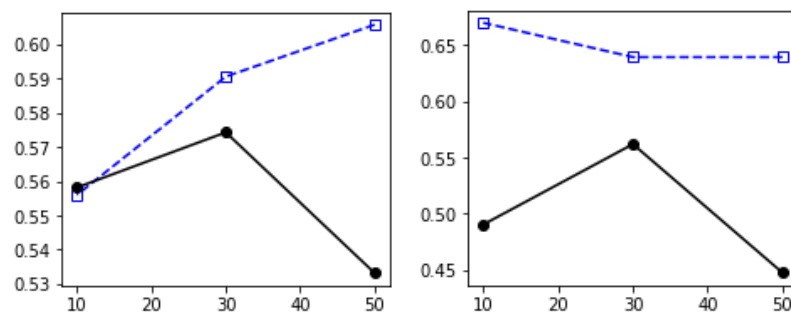## Heat-Related Illness Dataset 5

Source:

-Sample Size: 3175 ; -Attributes: 34 -3090:85 of 0:1

-Information: This dataset includes information on a novel early risk assessment tool for detecting, the binary dependent variable, clinical outcomes in patients with heat related illness.
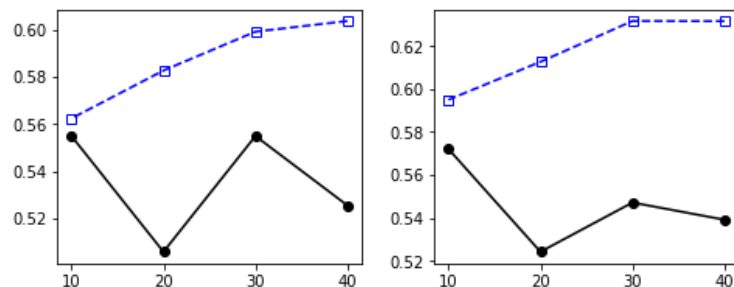
Mean Accuracy and Area Under the Curve:
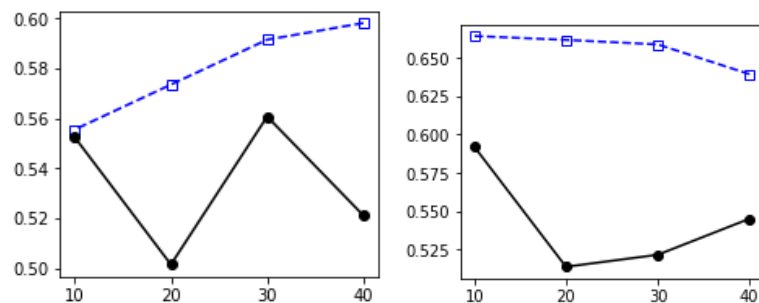

Precision and Recall:

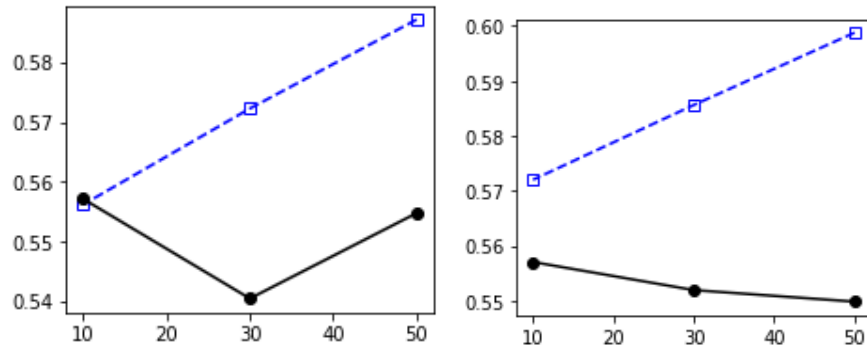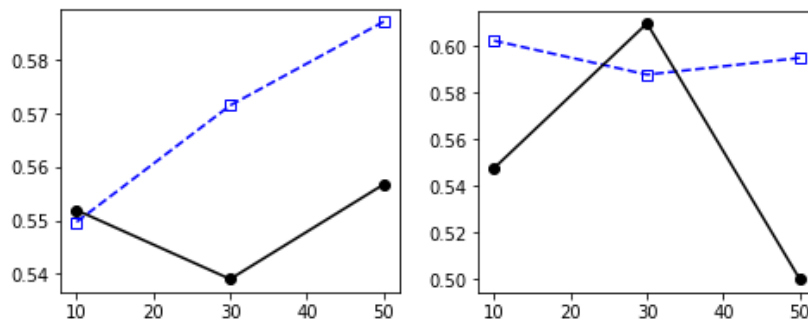**For Test Size 0.6 Rescaled**

Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.5 Not scaled

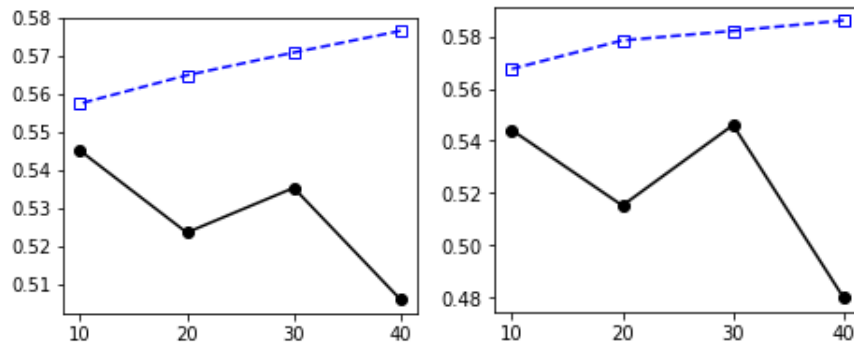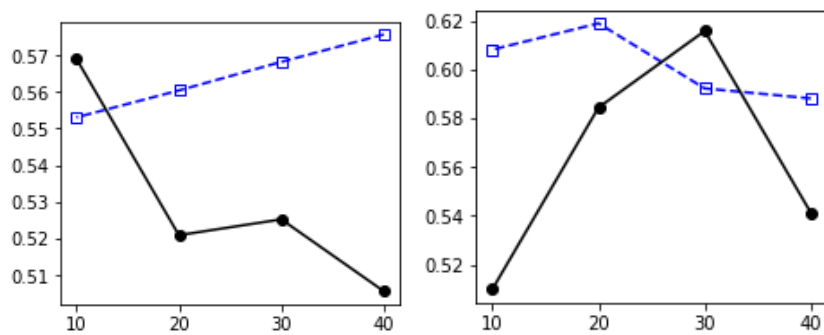Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

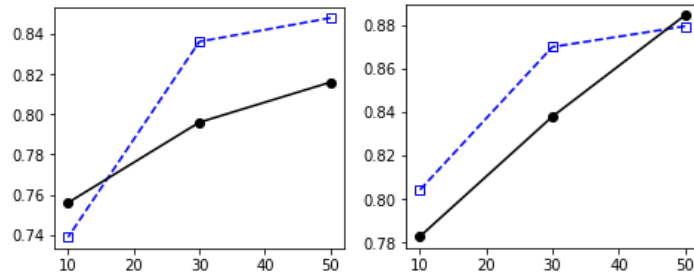Accuracy and Area Under the Curve:

Precision and Recall:

LMW PAHs and Pulmonary Health Dataset 6
Source: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84649
-Sample Size: 341 -Attributes: 17   -290:51 of 0:1
-Information: This dataset includes information on low molecular weight polycyclic aromatic hydrocarbons and their effect on pulmonary health. Factors include secondhand smoke, environmental tobacco smoke, activation of mitogen activated protein kinases, with a binary variable measuring treatment response.
Mean Accuracy and Area Under the Curve:
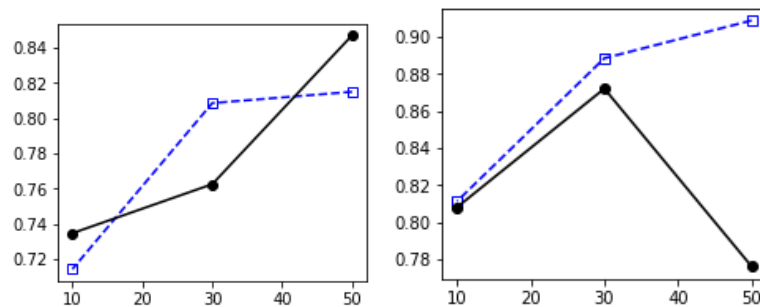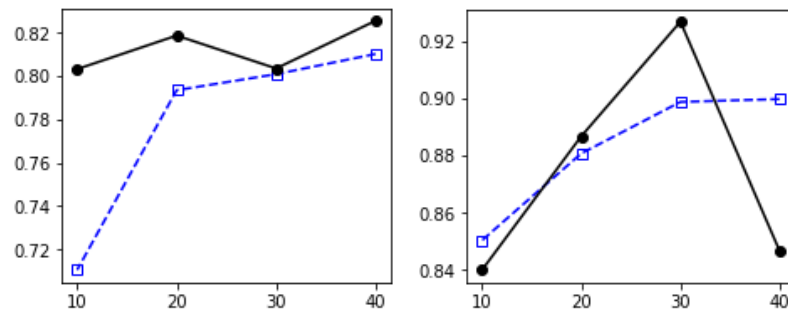

Precision and Recall:

For Test Size 0.6 Rescaled
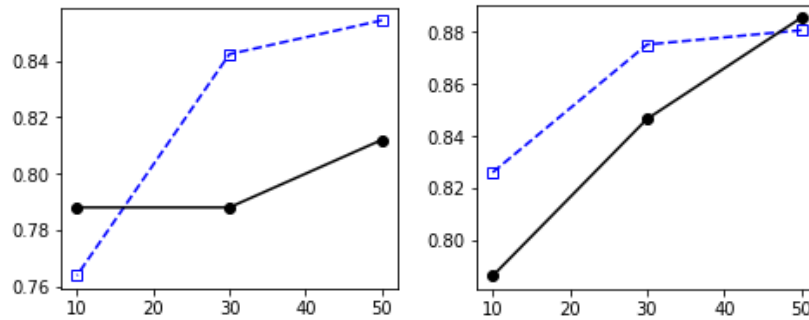Accuracy and Area Under the Curve:
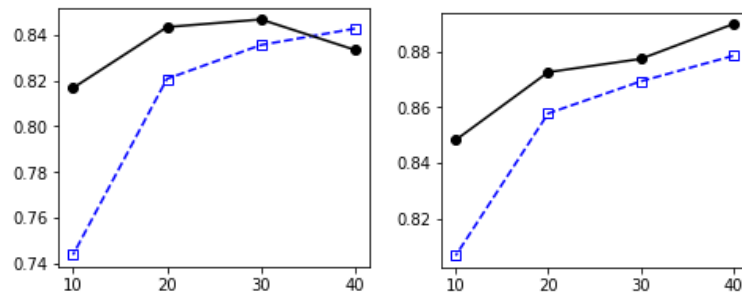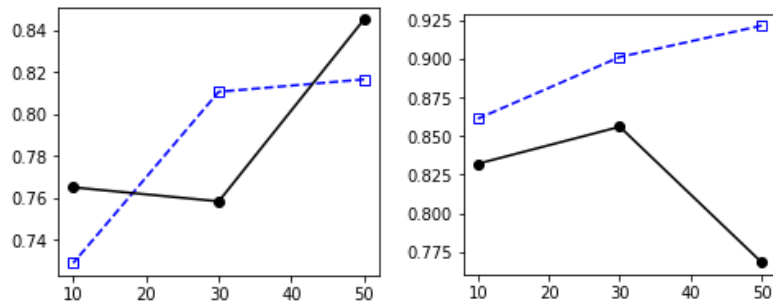

Precision and Recall:

For Test Size 0.5 Not scaled

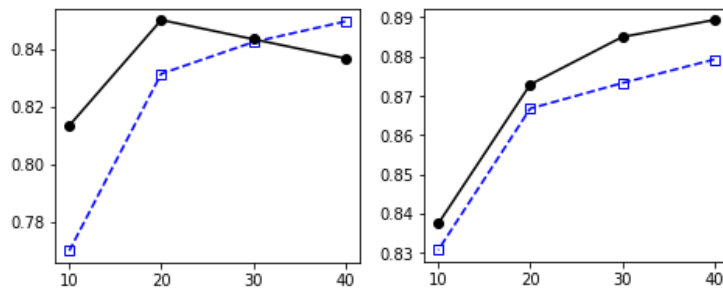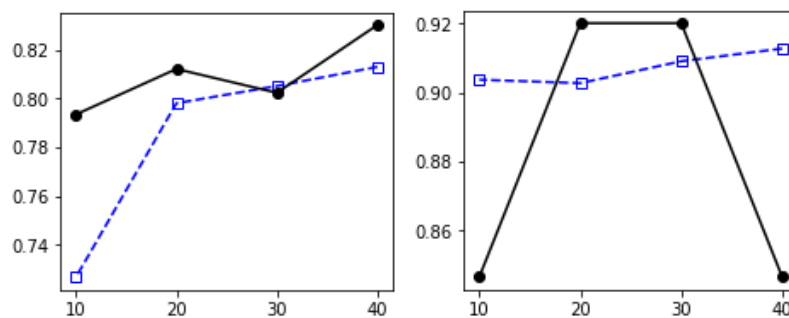Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled
Accuracy and Area Under the Curve:

Precision and Recall:

**Paradoxical Reactions in Buruli Ulcer Dataset 7**

-Sample Size: 241   -Attributes: 36   -178:63 of 0:1

-Information: This dataset includes information on the predictors of paradoxical reactions which have been hypothesized to be significant in being associated with buruli ulcers. The binary dependent variable is the presence of paradoxical reactions.
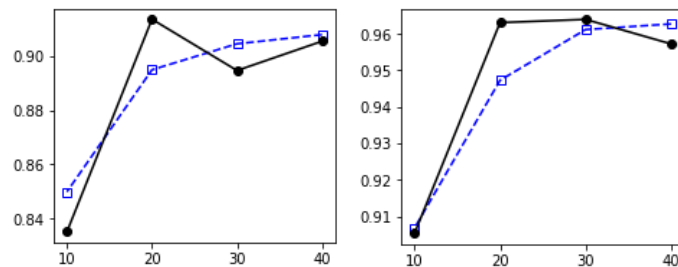
Mean Accuracy and Area Under the Curve:
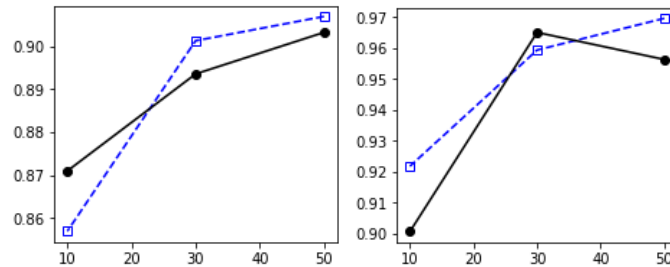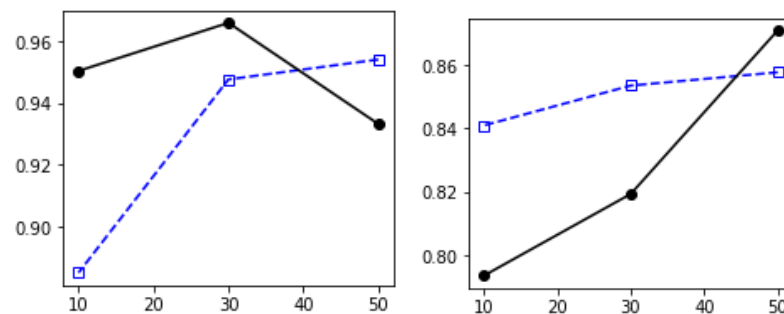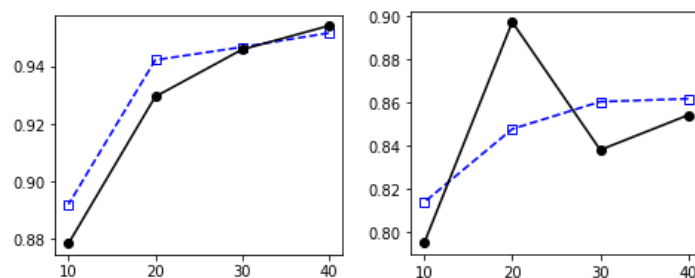

Precision and Recall:

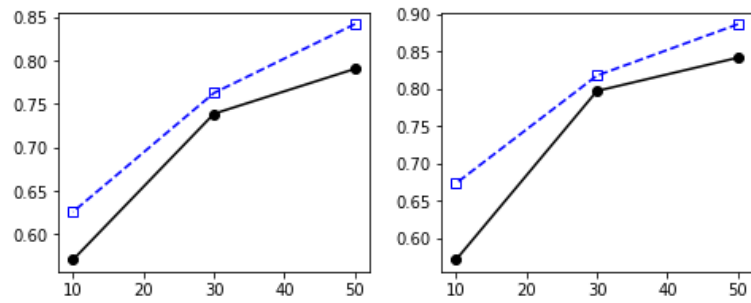**For Test Size 0.6 Rescaled**

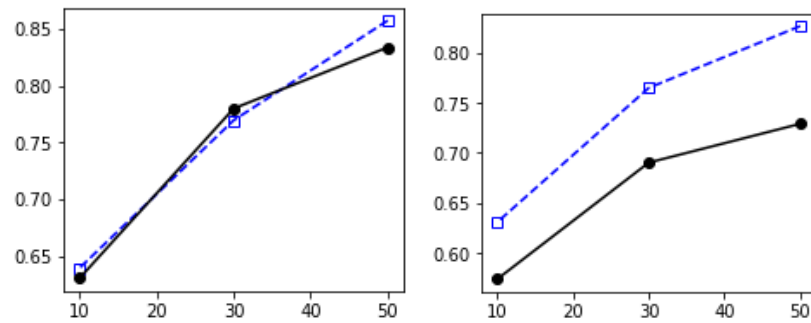Accuracy and Area Under the Curve:

Precision and Recall:

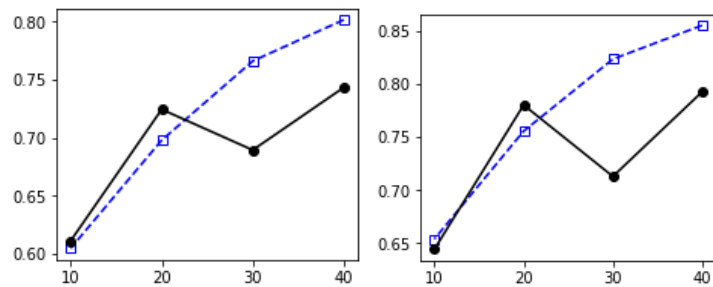**For Test Size 0.5 Not scaled**

Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled
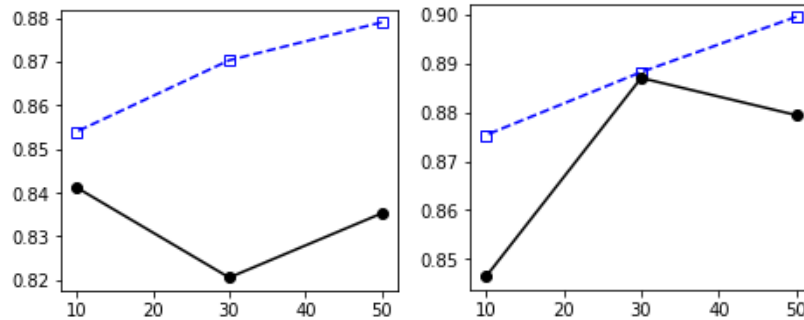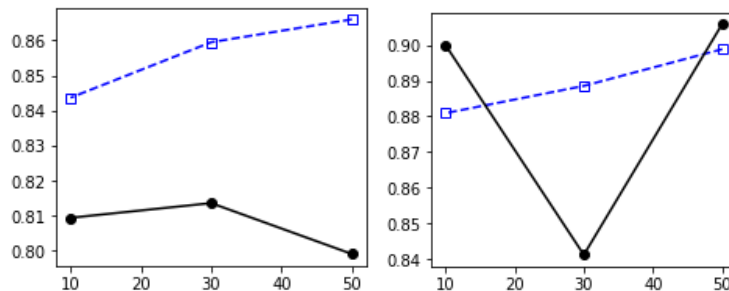Accuracy and Area Under the Curve:

Precision and Recall:
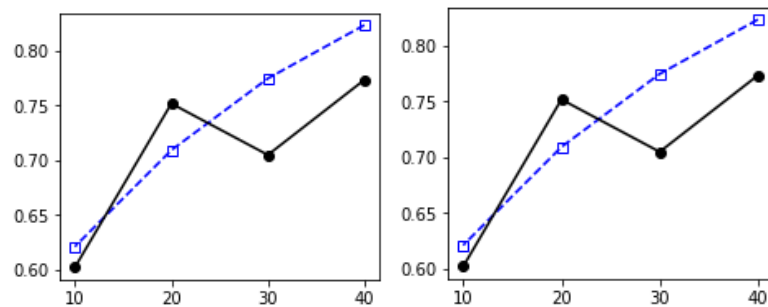
-Sample Size: 3947   -Attributes: 79   -68:3879 of 0:1
-Information: This dataset contains information on an NIH study of adolescent to adult health understanding the relationship between respondents' social, economic, psychological, amongst other factors and their relationship with social environments and later on achievement outcomes in adulthood. The binary independent variable is Year 1 course level.

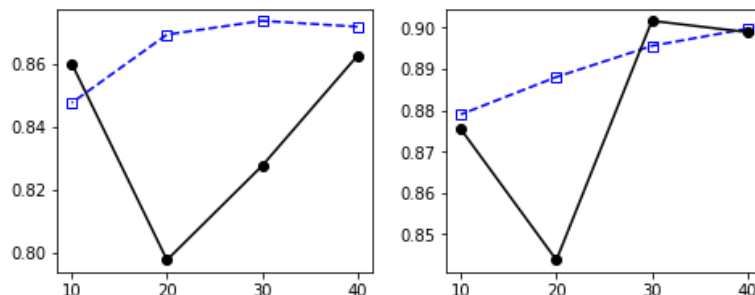Mean Accuracy and Area Under the Curve:
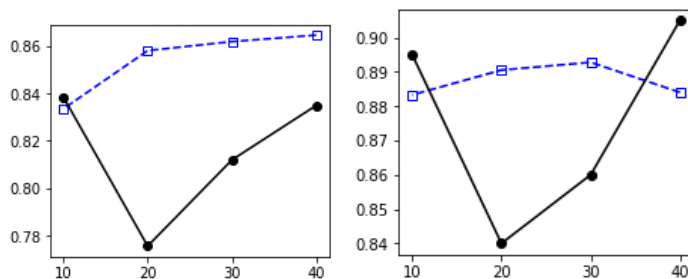

Precision and Recall:

For Test Size 0.6 Rescaled
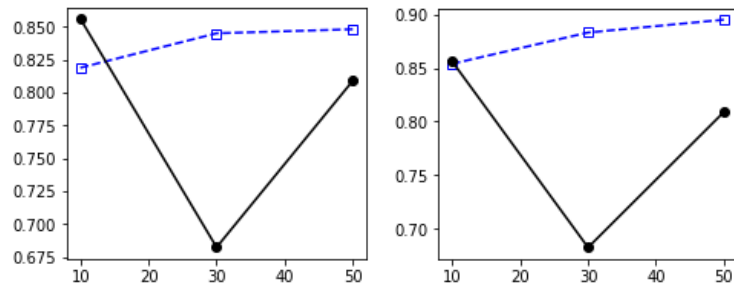Accuracy and Area Under the Curve:
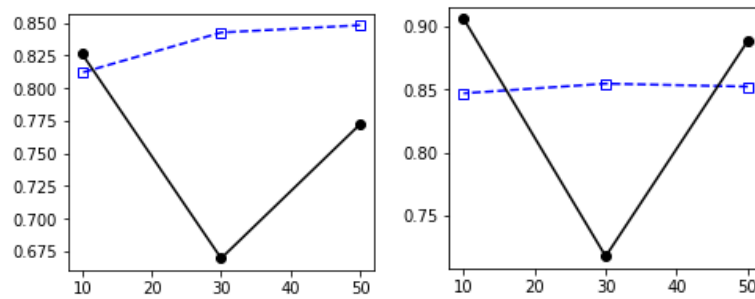

Precision and Recall:

For Test Size 0.5 Not scaled
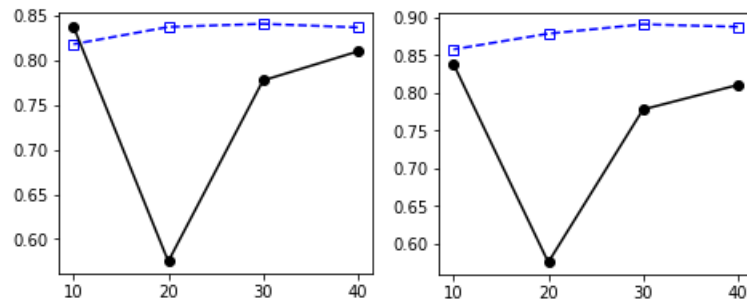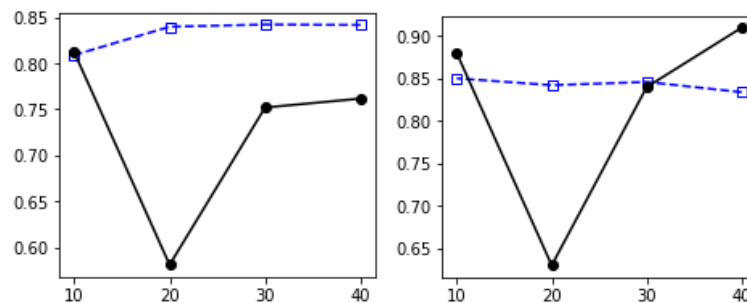Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled
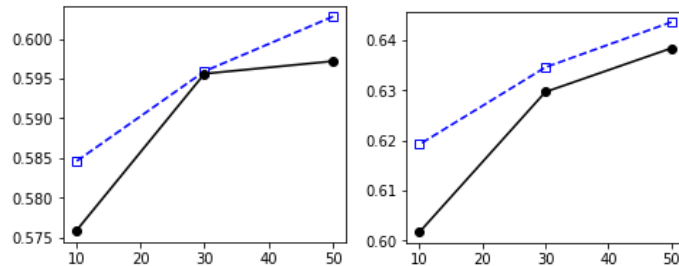Accuracy and Area Under the Curve:

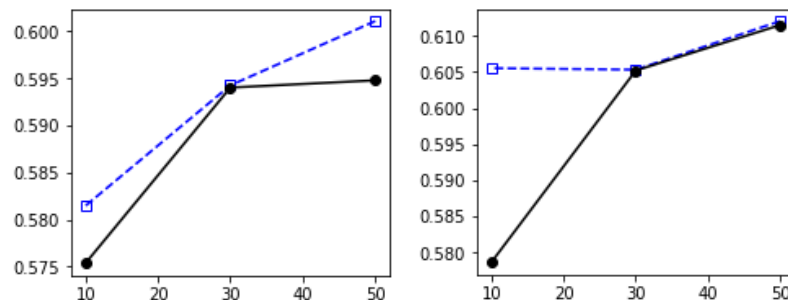Precision and Recall:

HIV Stigma Dataset 9

-Sample Size: 10665   -Attributes: 37 -1774:8891 of 0:1
-Information: This dataset includes information on the application of an HIV related stigma scale to undergraduates in mainland china with the binary variable focused on being willingness to engage in treatment even in the context of an environment with stigma, though based on the details of the study, it seems to be independent with the relationship with other variables being associative.
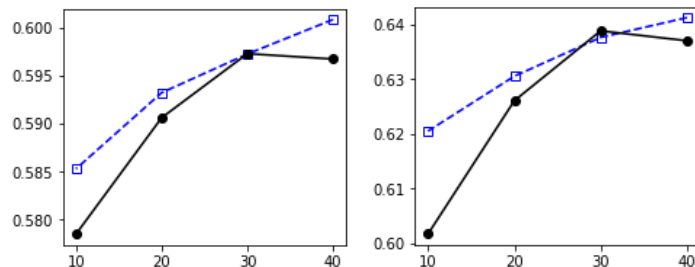
Mean Accuracy and Area Under the Curve:
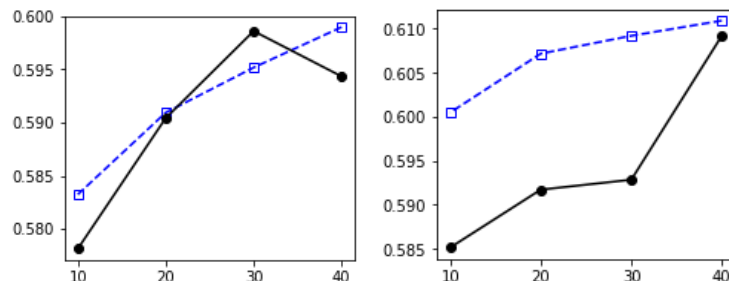

Precision and Recall:

For Test Size 0.6 Rescaled
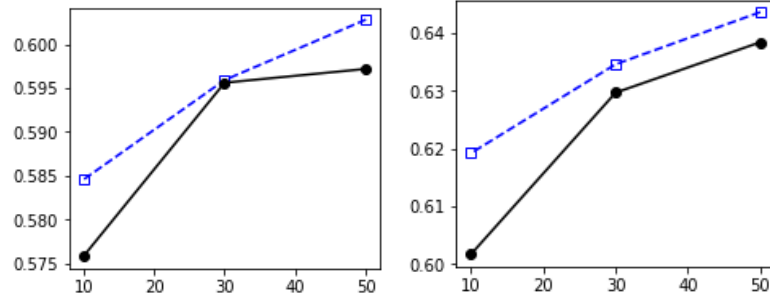
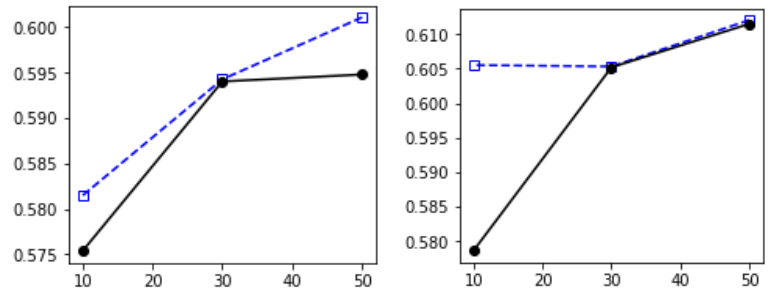Accuracy and Area Under the Curve:

Precision and Recall:

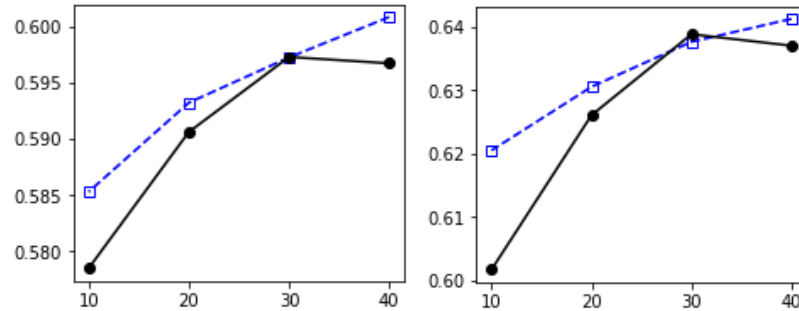**For Test Size 0.5 Not scaled**
Accuracy and Area Under the Curve:

Precision and Recall:

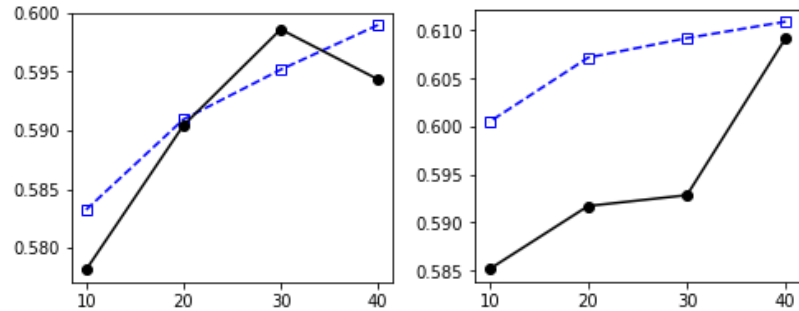**For Test Size 0.6 Not scaled**
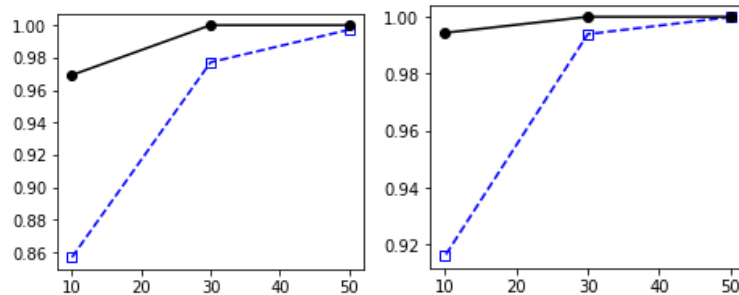Accuracy and Area Under the Curve:

Precision and Recall:

Erectile Function Dataset 10
Source: https://datadryad.org/resource/doi:10.5061/dryad.0b58t3v
-Sample Size: 1004   -Attributes: 54   -899:105 of 0:1
-Information: This dataset focuses on the relationship between erectile function and ejaculatory function with factors that impact the binary dependent variable, erectile functioning, although there was another dependent variable present in the dataset of ejaculatory function.
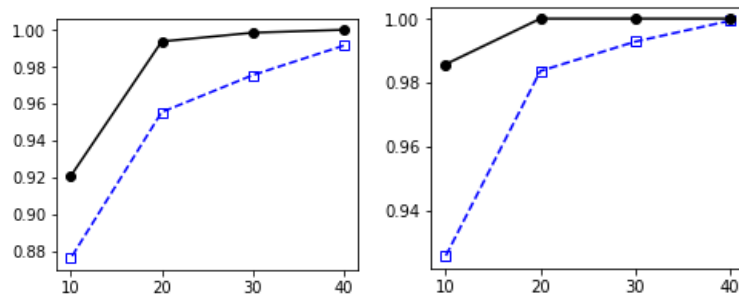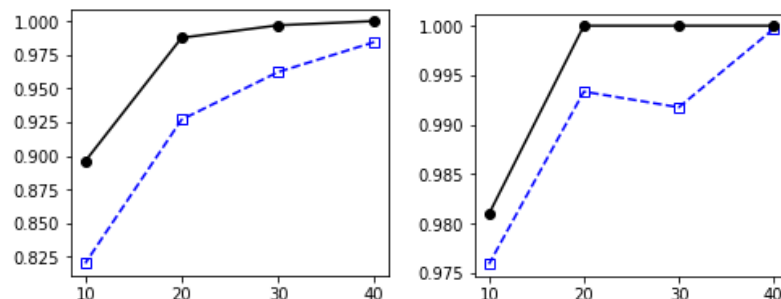Mean Accuracy and Area Under the Curve:
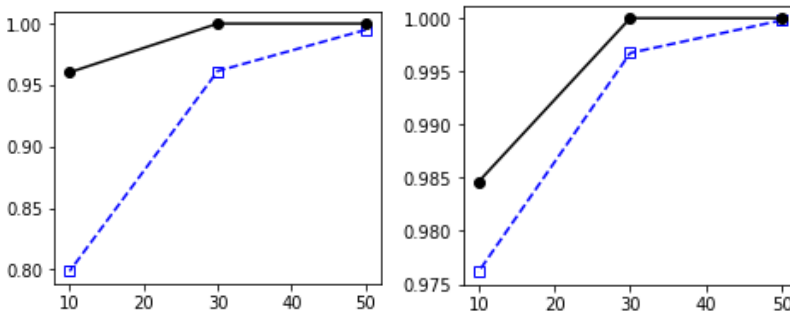
Precision and Recall:

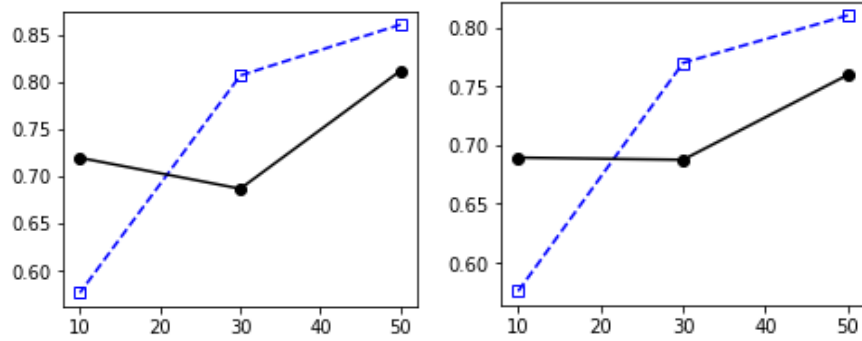**For Test Size 0.6 Rescaled**
Accuracy and Area Under the Curve:
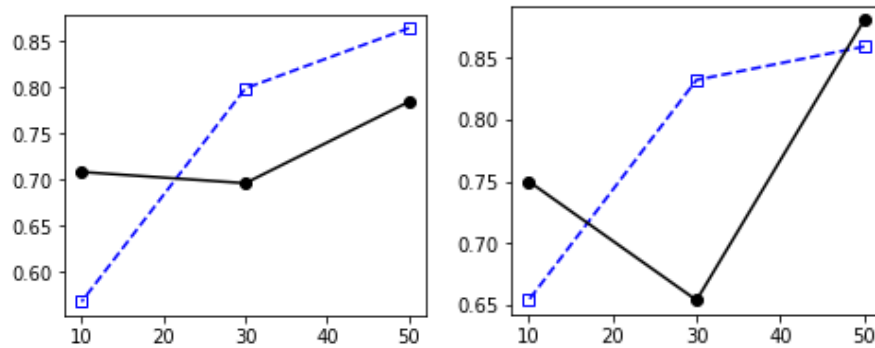

Precision and Recall:

For Test Size 0.5 Not scaled
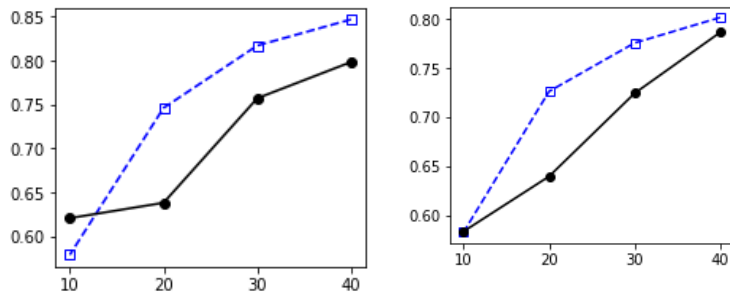Accuracy and Area Under the Curve:

Precision and Recall:

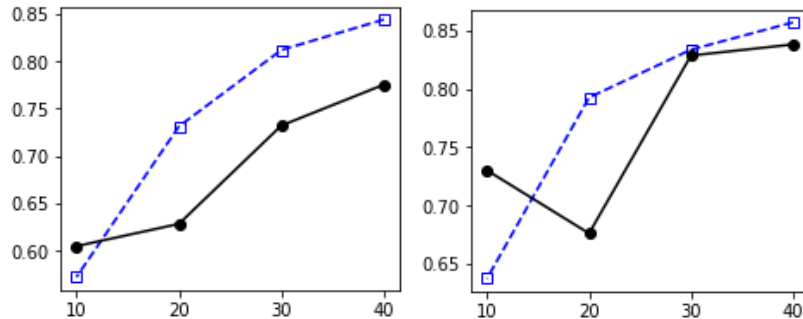**For Test Size 0.6 Not scaled**
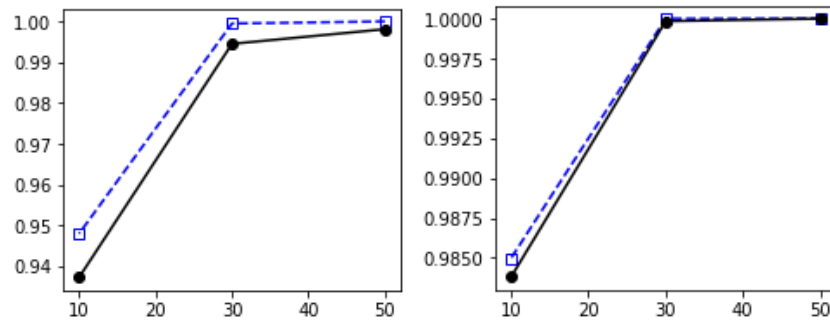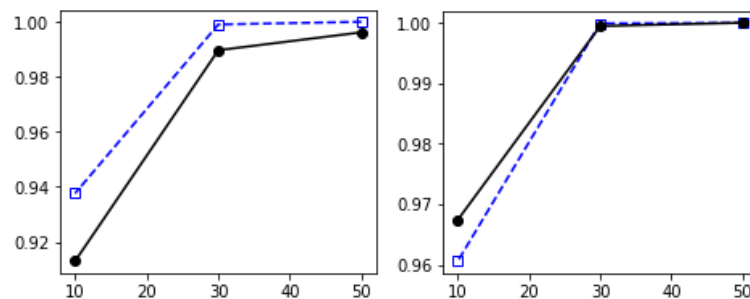Accuracy and Area Under the Curve:

Precision and Recall:

-Sample Size: 1588 -Attributes: 118 -864:724 of 0:1
-Information: This dataset contains information on tobacco use and the relationship between multiple independent variables including public opinion on smoking regulations and the binary variable focused on for this figure, health status of individuals.
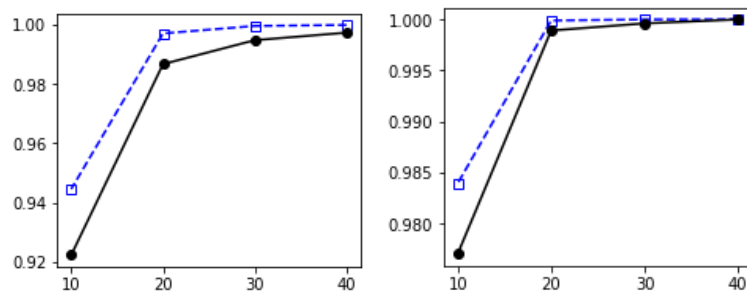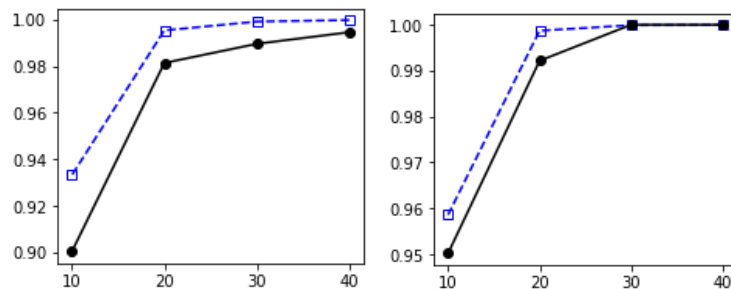Mean Accuracy and Area Under the Curve:

Precision and Recall:

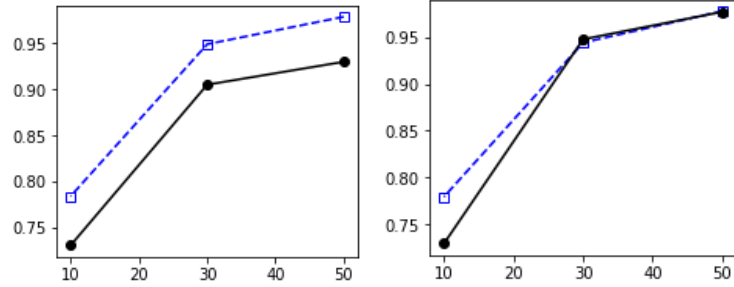**For Test Size 0.6 Rescaled**
Accuracy and Area Under the Curve:

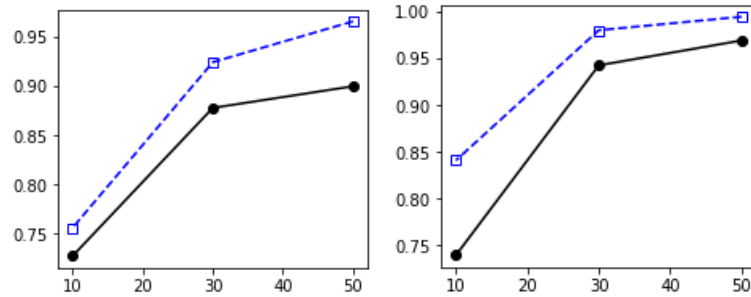

Precision and Recall:

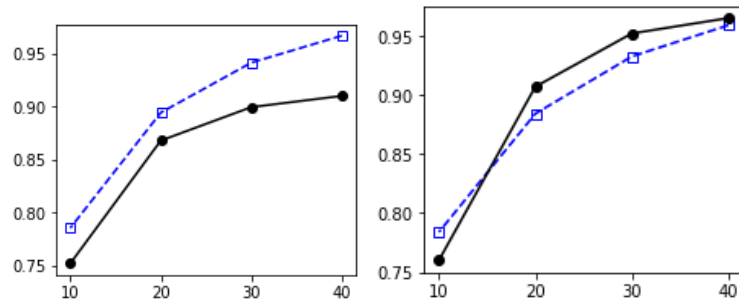**For Test Size 0.5 Not scaled**
Accuracy and Area Under the Curve:

Precision and Recall:

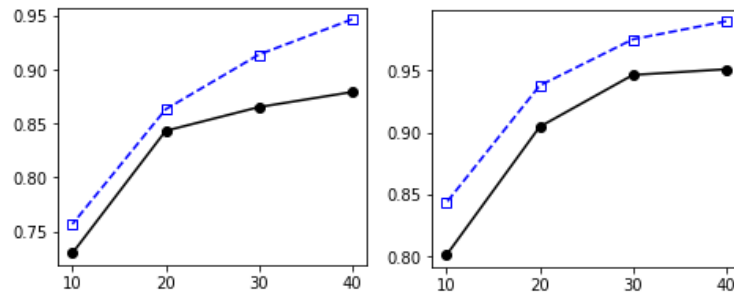**For Test Size 0.6 Not scaled**
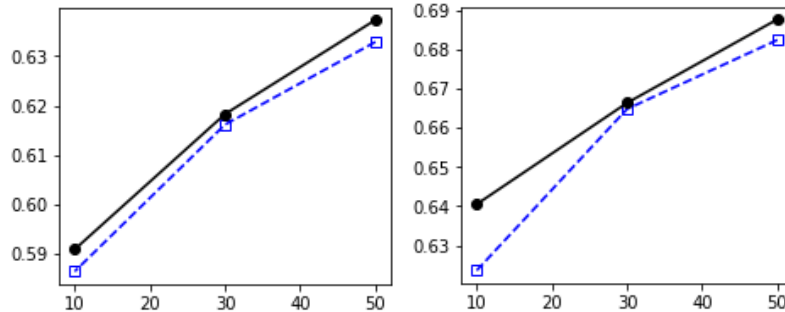Accuracy and Area Under the Curve:

Precision and Recall:

Coagulation Dataset 12

Source: https://datadryad.org/resource/doi:10.5061/dryad.3ds63

-Sample Size: 1498   -Attributes: 22   -994:504 of 0:1

-Information: This data contains information on anticoagulant therapy and its effectiveness at reducing risk of death for patients with sepsis and coagulation dysfunction. The binary dependent variable is whether or not the patient died after treatment.

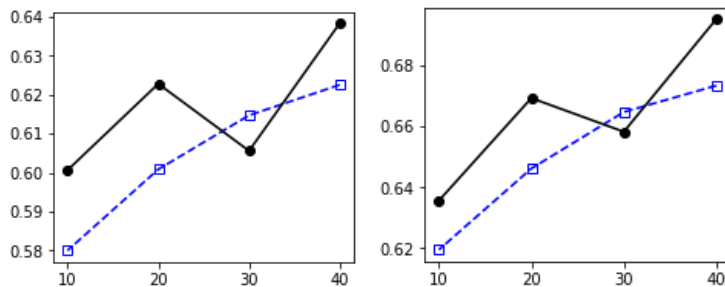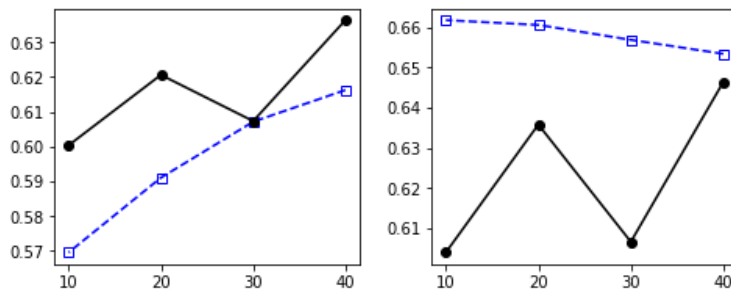Mean Accuracy and Area Under the Curve:
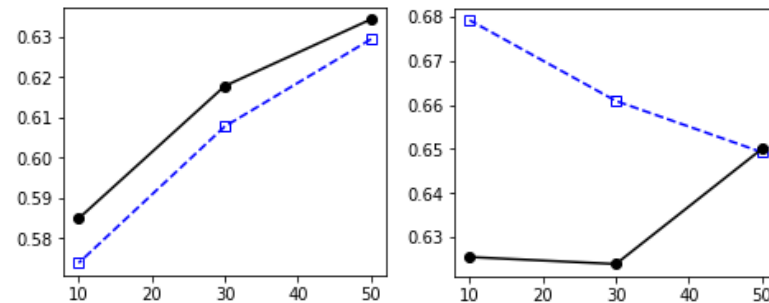
Precision and Recall:

For Test Size 0.6 Rescaled
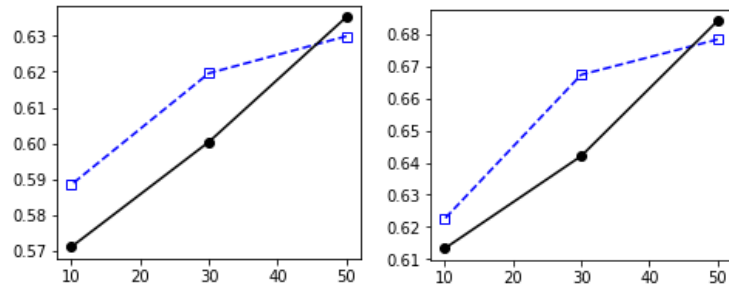
Accuracy and Area Under the Curve:
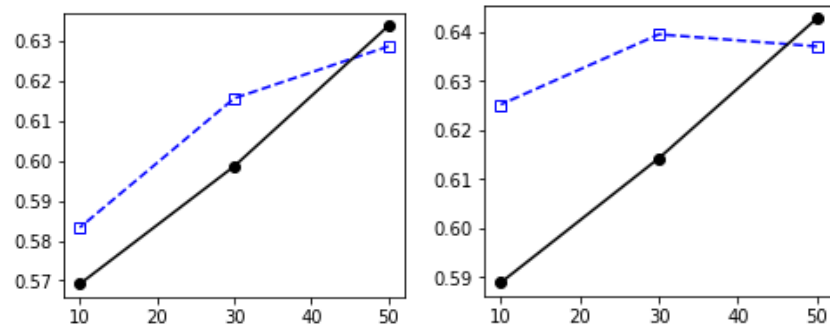

Precision and Recall:

For Test Size 0.5 Not scaled
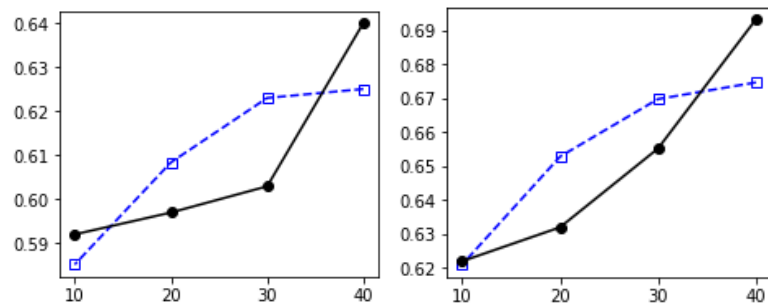Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled
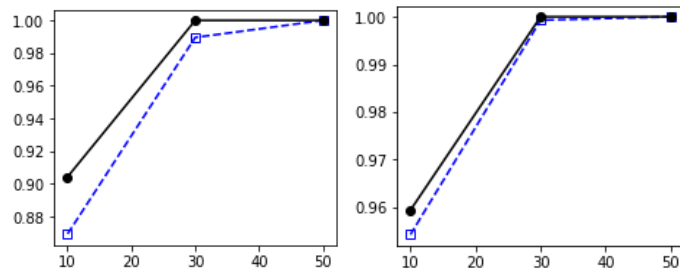Accuracy and Area Under the Curve:

Precision and Recall:
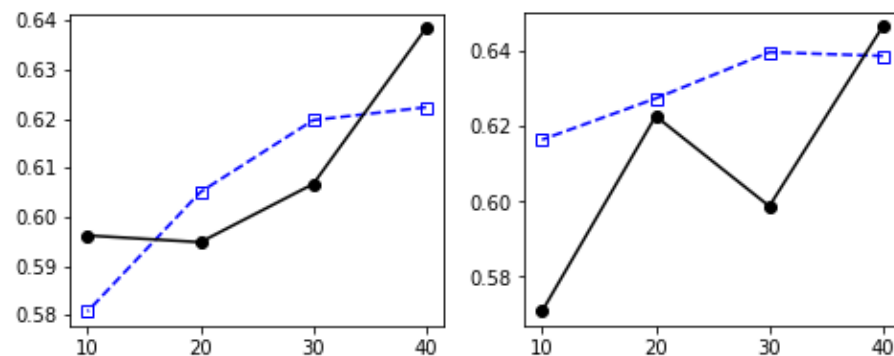
Anemia Dataset 13
Source: https://datadryad.org/resource/doi:10.5061/dryad.br52k
-Sample Size: 249 -Attributes: 41   -200:49 of 0:1
-Information: This dataset contains information on anemia and its association with different impaired clinical outcomes in diabetic chronic kidney disease. The binary dependent variable is whether or not patients died in response to hepcidin-25, the key hormone of iron-metabolism.
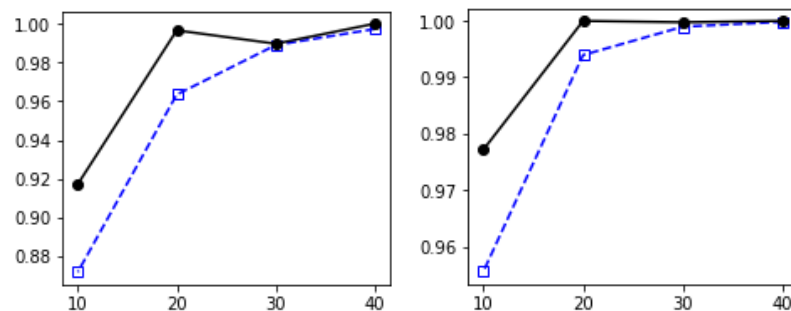Mean Accuracy and Area Under the Curve:
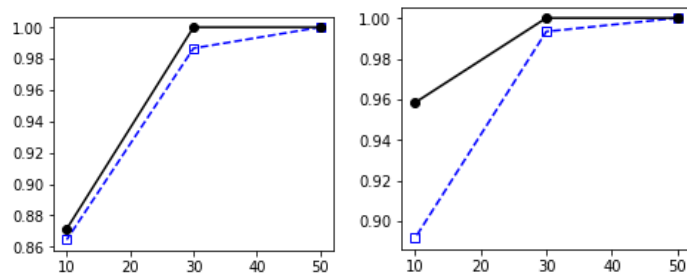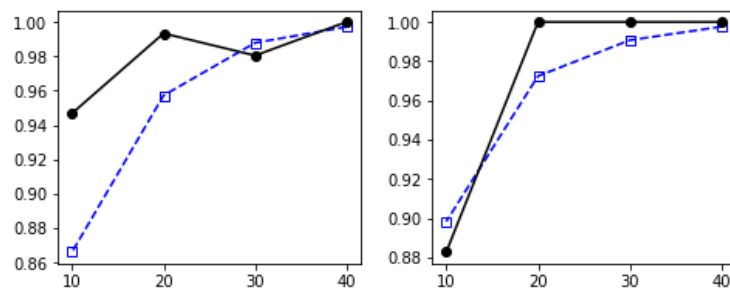

Precision and Recall:

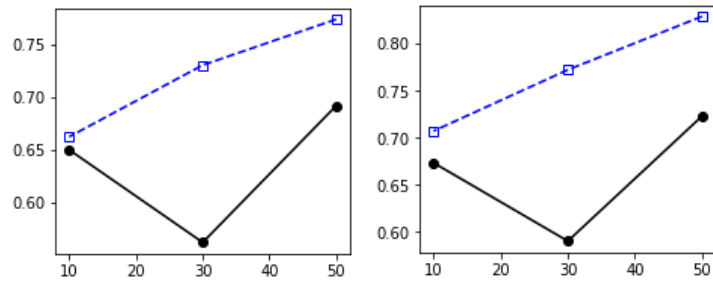**For Test Size 0.6 Rescaled**
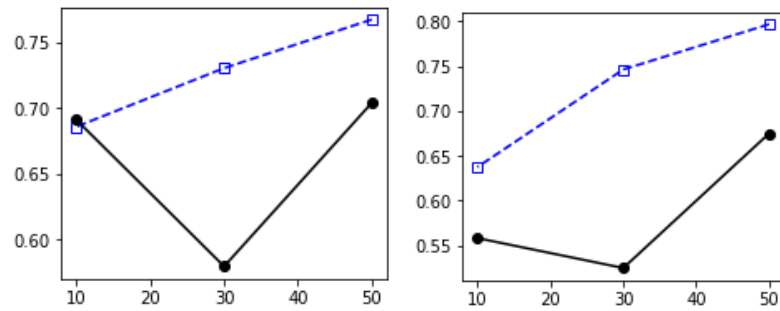Accuracy and Area Under the Curve:

Precision and Recall:

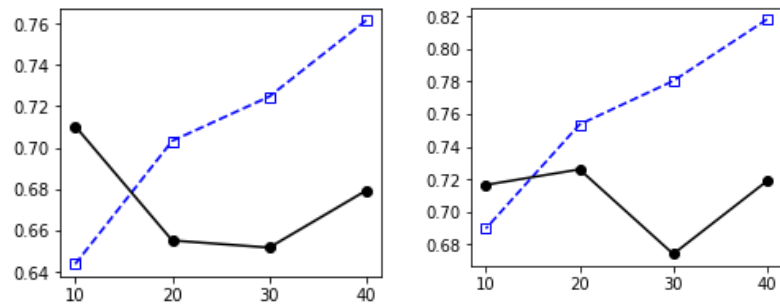**For Test Size 0.5 Not scaled**

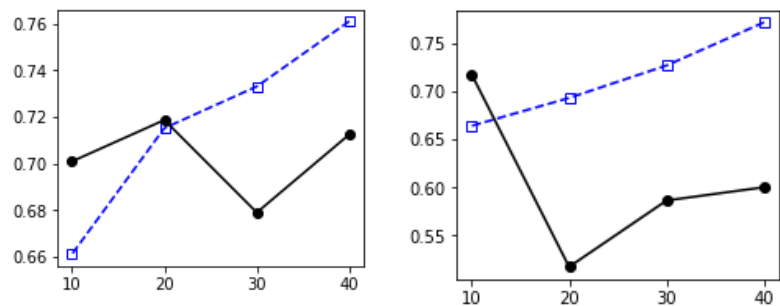Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

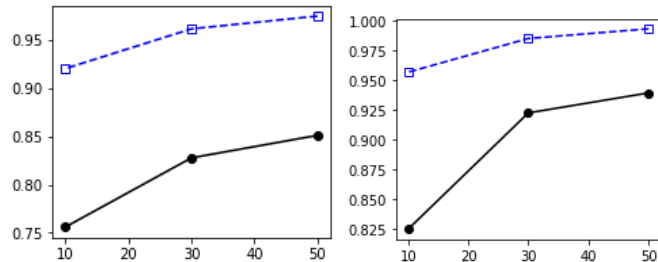Accuracy and Area Under the Curve:

Precision and Recall:

Cervical Cancer Dataset 14

Source: https://data.world/uci/cervical-cancer-risk-factors/workspace/file?
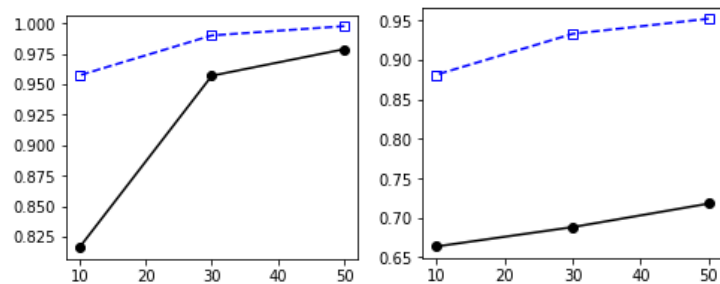filename=risk_factors_cervical_cancer.csv

-Sample Size: 858   -Attributes: 27   -377:481 of 0:1

-Information: This dataset provides information on a variety of risk factors contributing to cervical cancer with the binary independent variable being contraceptive use across patients.
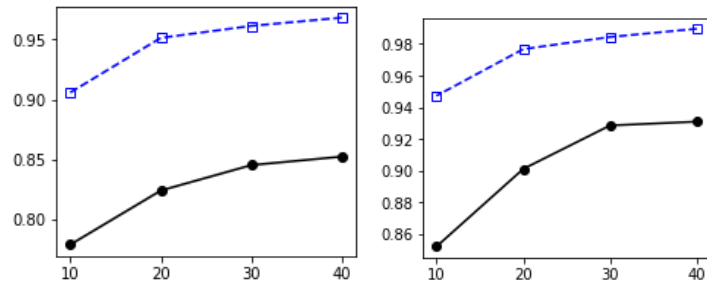
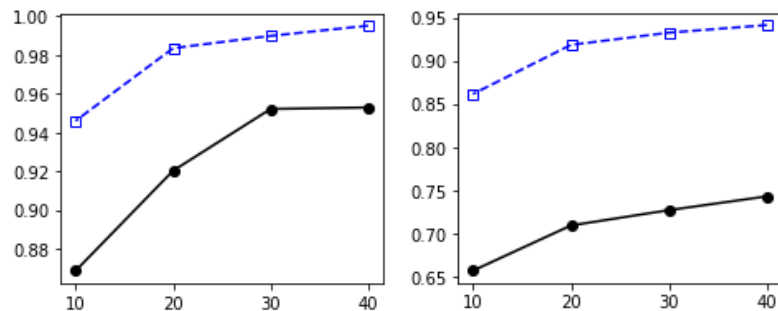Mean Accuracy and Area Under the Curve:

Precision and Recall:

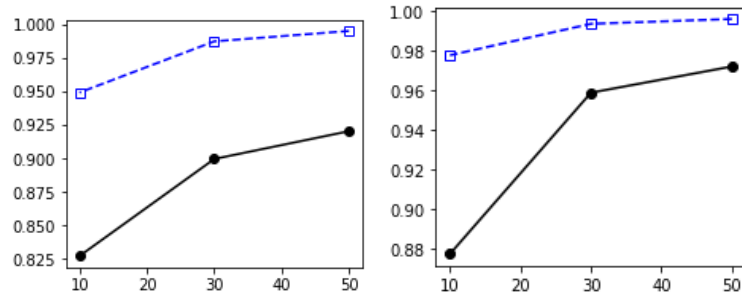**For Test Size 0.6 Rescaled**

Accuracy and Area Under the Curve:

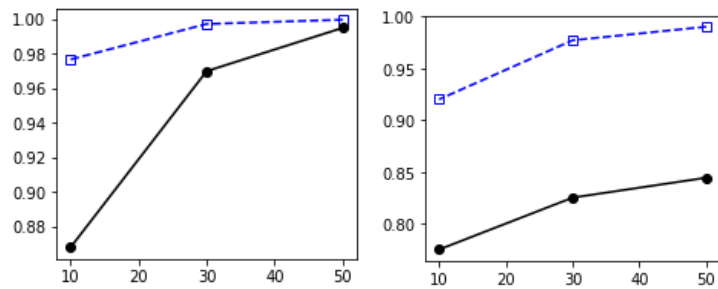Precision and Recall:

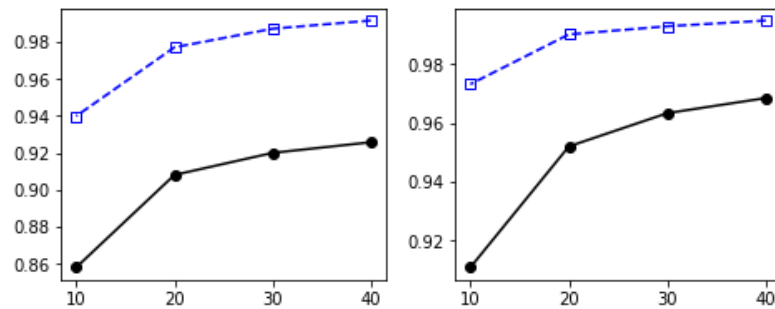**For Test Size 0.5 Not scaled**
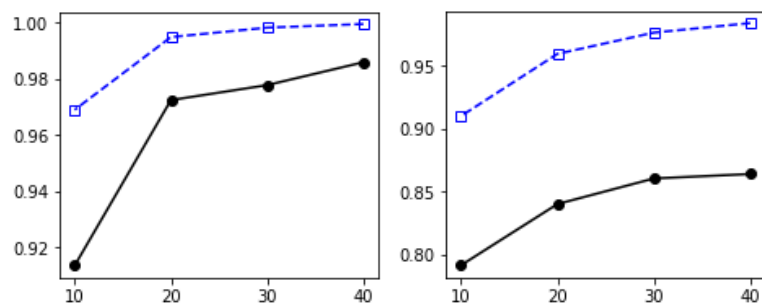
Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

Accuracy and Area Under the Curve:

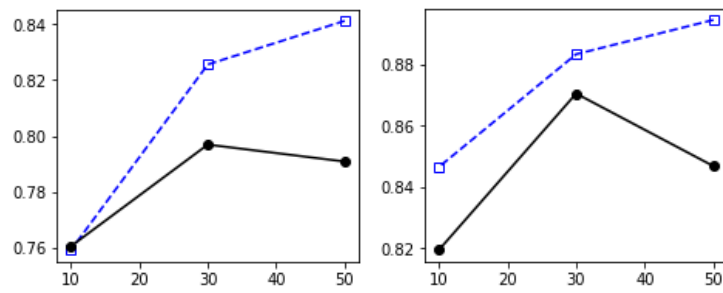Precision and Recall:

Mental Health Dataset 15

Source:

-Sample Size: 776   -Attributes: 35 -67:709 of 0:1

-Information: This dataset includes information on the association between prior psychiatric associations and diagnostic measures and the binary variable of a depression diagnosis in medical students in Mexico. While the dataset didn't specify if this was a dependent variable, it clearly has a strong associative relationship with other factors in the study,
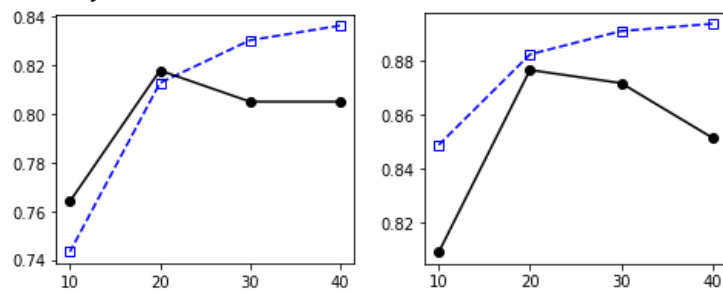
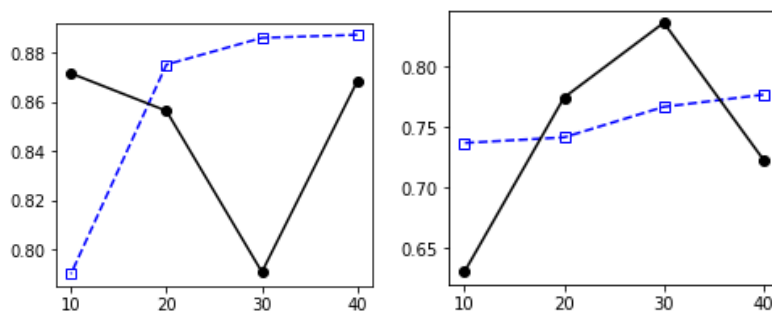Mean Accuracy and Area Under the Curve:
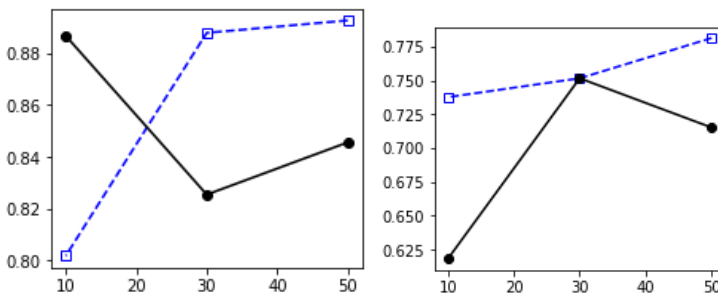
Precision and Recall:

For Test Size 0.6 Rescaled

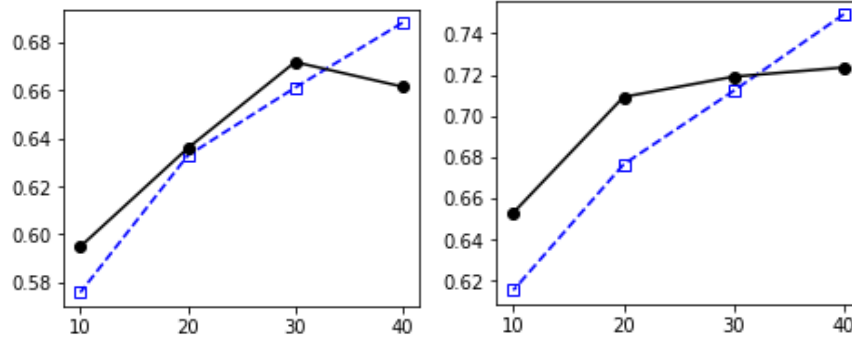Accuracy and Area Under the Curve:

Precision and Recall:

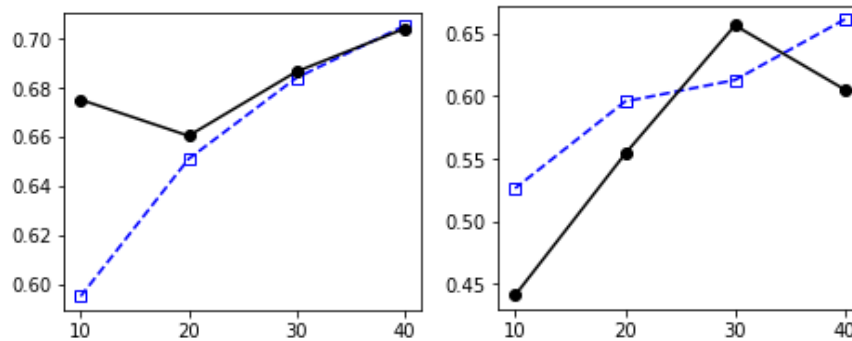**For Test Size 0.5 Not scaled**

Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

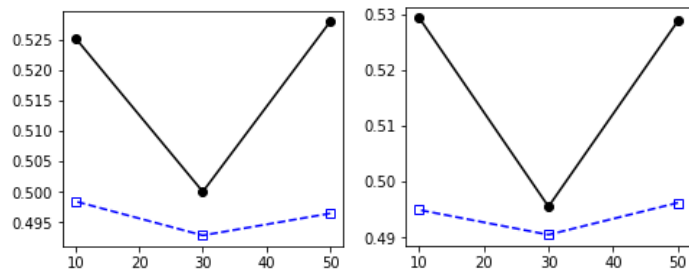Accuracy and Area Under the Curve:

Precision and Recall:

Dementia Dataset 16

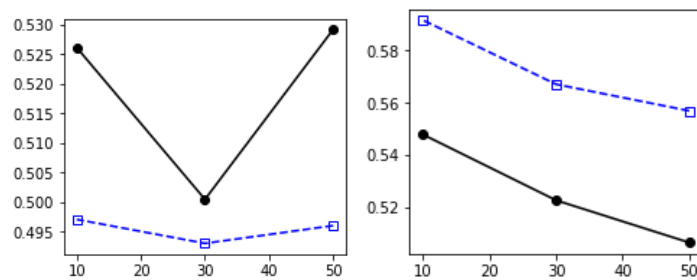Source: https://figshare.com/articles/Dataset-2019_xlsx/8273006/1

-Sample Size: 800   -Attributes: 12   -223:578 of 0:1

-Information: This dataset contains information on biological factors contributing to the development of the binary dependent variable, Dementia (AD and FTD) or Parkinson's Disease.
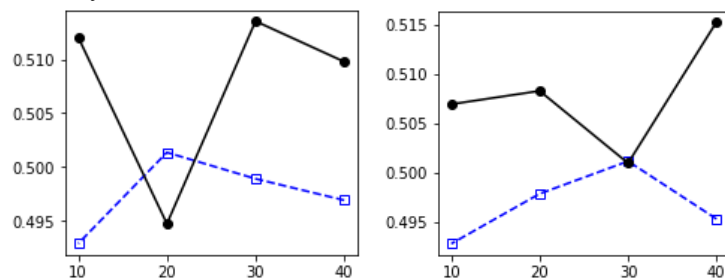
Mean Accuracy and Area Under the Curve:
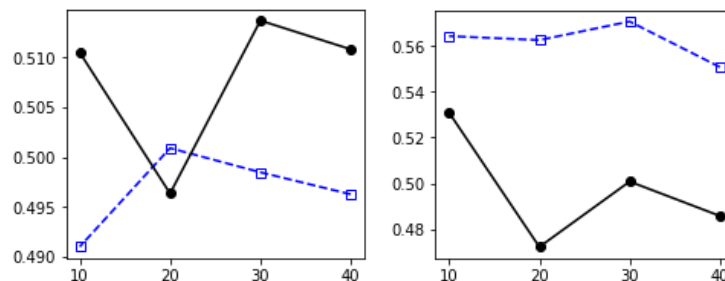

Precision and Recall:

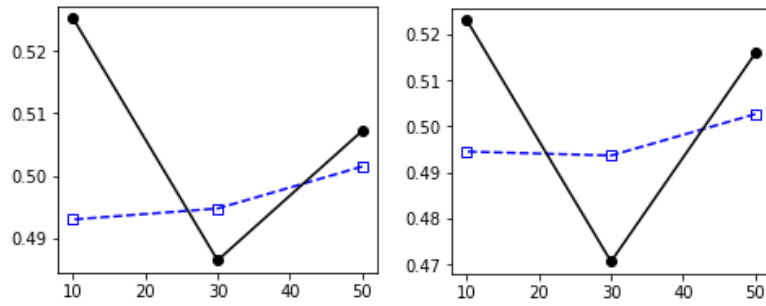**For Test Size 0.6 Rescaled**

Accuracy and Area Under the Curve:

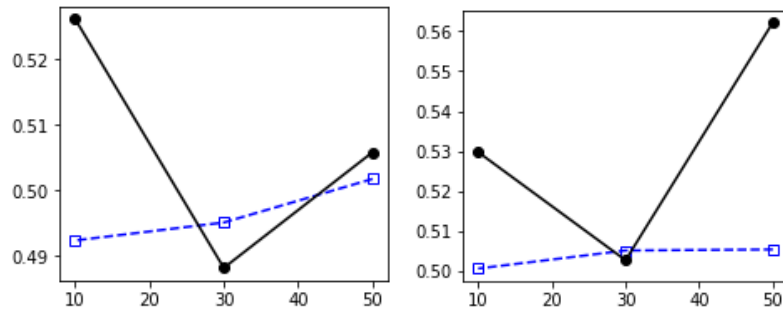Precision and Recall:

For Test Size 0.5 Not scaled

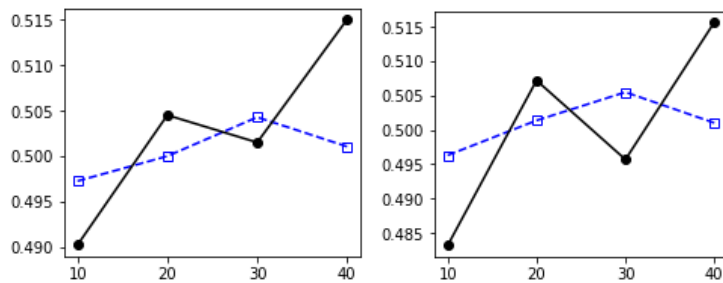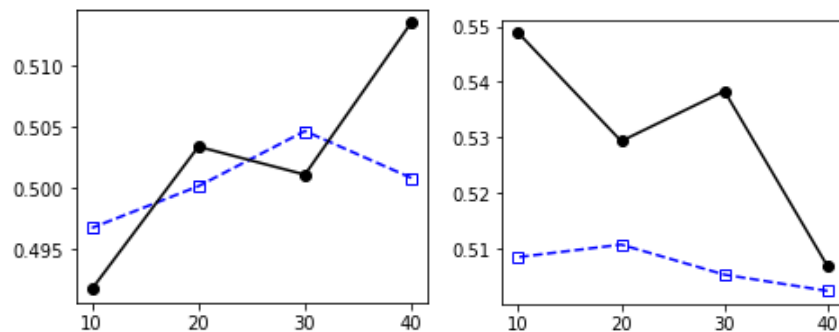Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

Accuracy and Area Under the Curve:

Precision and Recall:

Sepsis and Cancer Patients Dataset 17
Source: https://datadryad.org/resource/doi:10.5061/dryad.6qk05
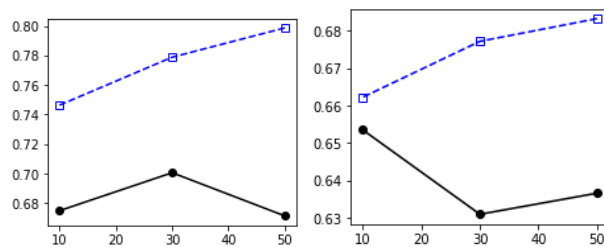-Sample Size: 352
-Attributes: 70
-142:210 of 0:1
-Information: This dataset provides information comparing the mortality rate (the binary dependent variable)
between cancer patients with or without sepsis, comparing multiple factors and lab results in comparing what could
contribute to the mortality rate.
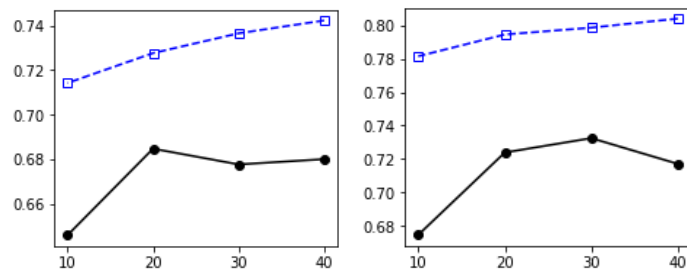
Mean Accuracy and Area Under the Curve:
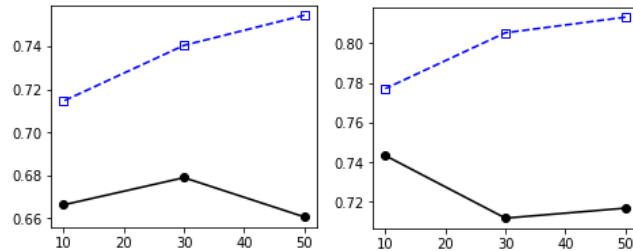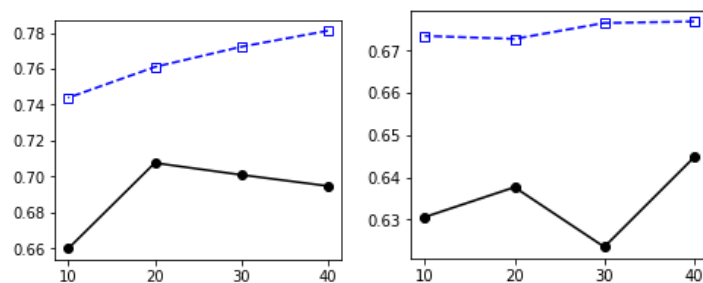

Precision and Recall:

For Test Size 0.6 Rescaled
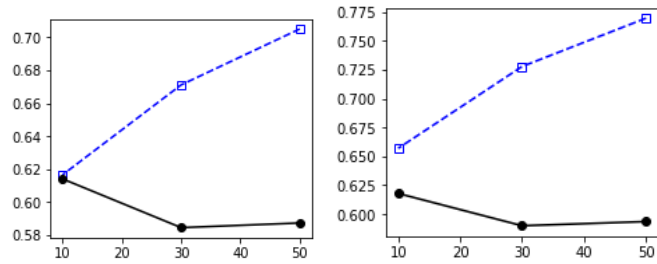Accuracy and Area Under the Curve:
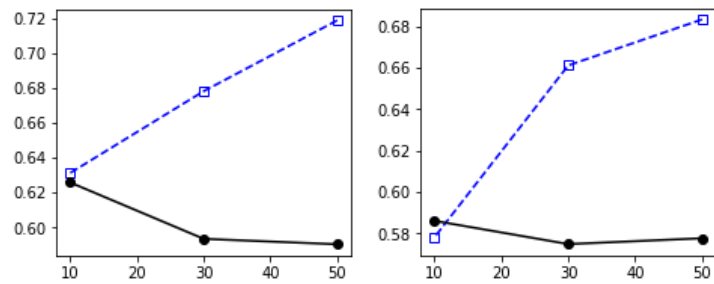

Precision and Recall:

For Test Size 0.5 Not scaled
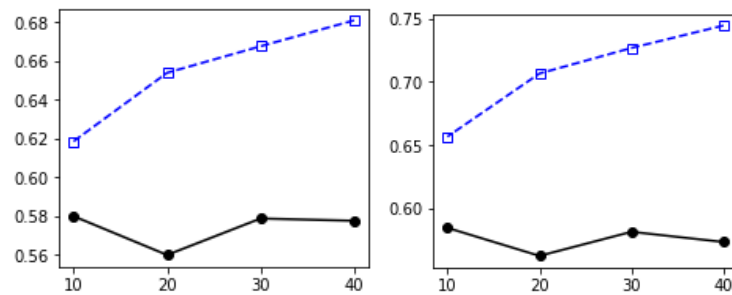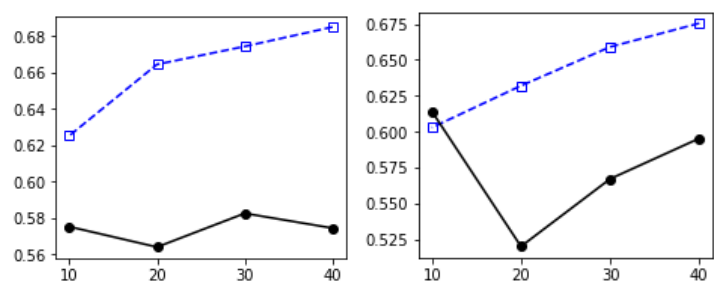
Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

Accuracy and Area Under the Curve:

Precision and Recall:

Breast Cancer Coimbra Dataset 18
Source: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra
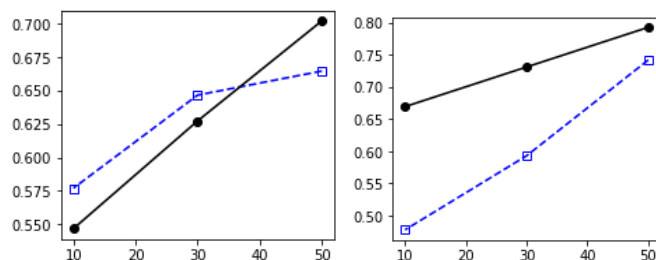-Sample Size: 117
-Attributes: 9
-64:53 of 0:1[1]
-Information: This dataset includes comparisons based on different biomarkers of breast cancer. These predictors are all quantitative, affecting the binary dependent variable of whether or not there is presence or absence of breast cancer.
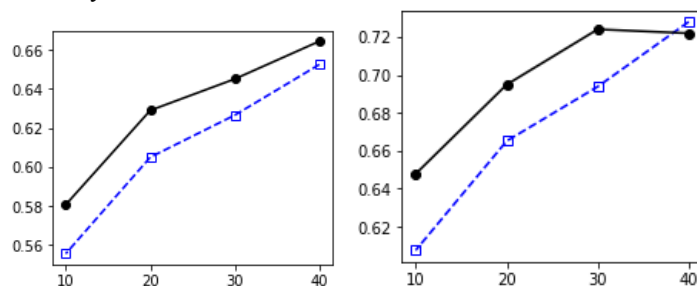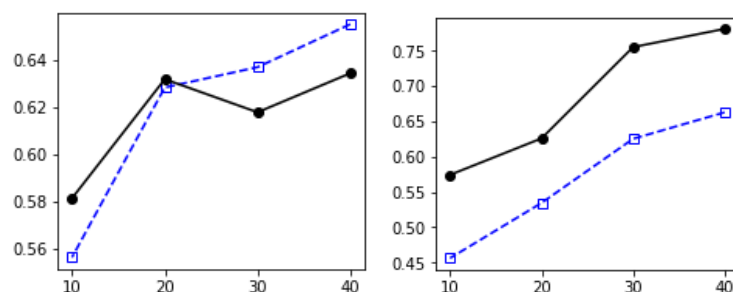Mean Accuracy and Area Under the Curve:
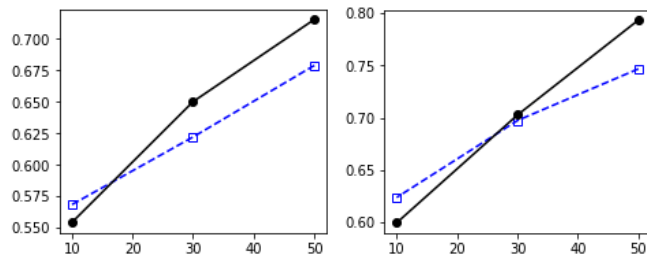
Precision and Recall:

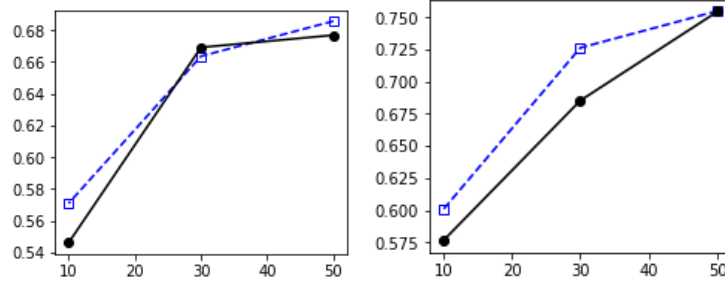**For Test Size 0.6 Rescaled**
Accuracy and Area Under the Curve:
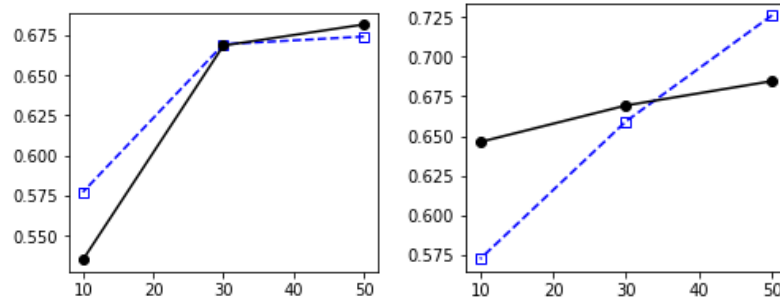

Precision and Recall:

---

1 Was represented as 2:1 in dataset, so all of the 2's were replaced with 0.

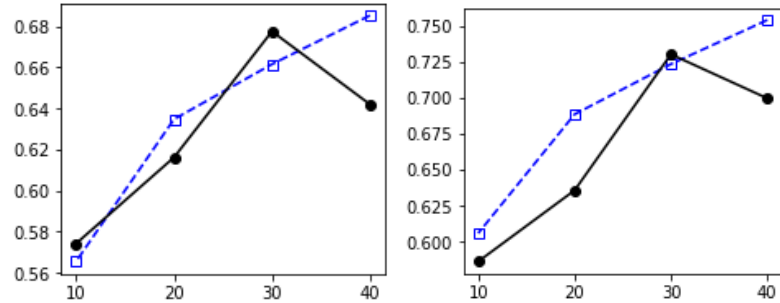**For Test Size 0.5 Not scaled**
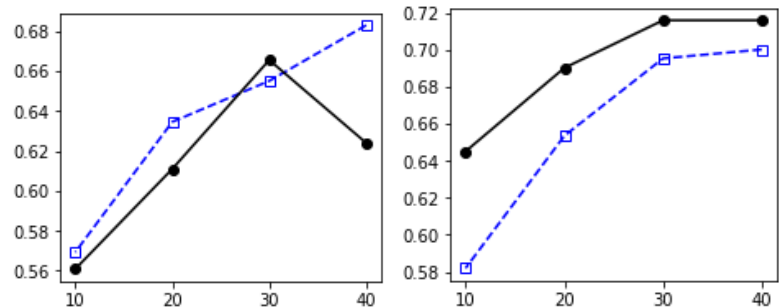Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

Accuracy and Area Under the Curve:

Precision and Recall:

Cryotherapy Dataset 19
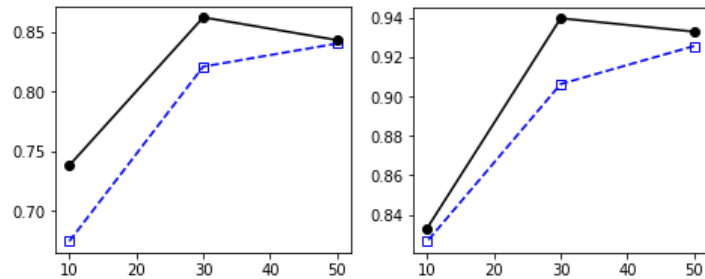Source: https://archive.ics.uci.edu/ml/datasets/Cryotherapy+Dataset+
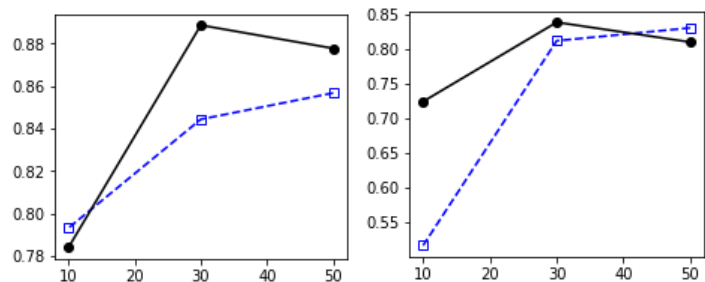-Sample Size: 90
-Attributes: 6
-43:47 of 0:1
-Information: This dataset contains information about the wart treatment results of 90 patients using cryotherapy and the dependent binary variable represents the success of the therapy.
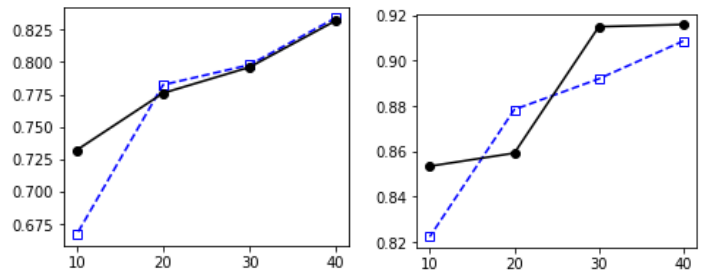Accuracy and Area Under the Curve:
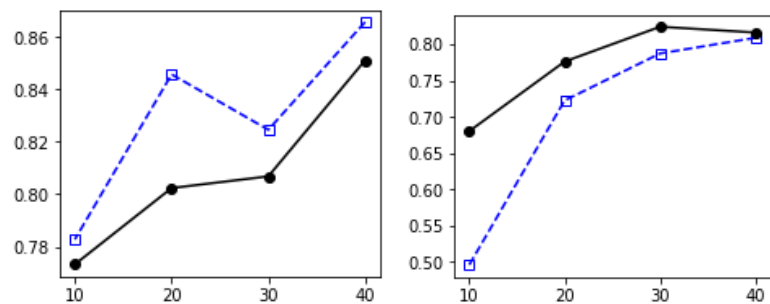

Precision and Recall:

For Test Size 0.6 Rescaled
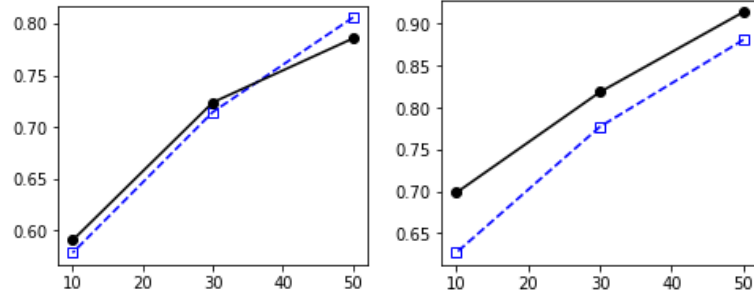Accuracy and Area Under the Curve:
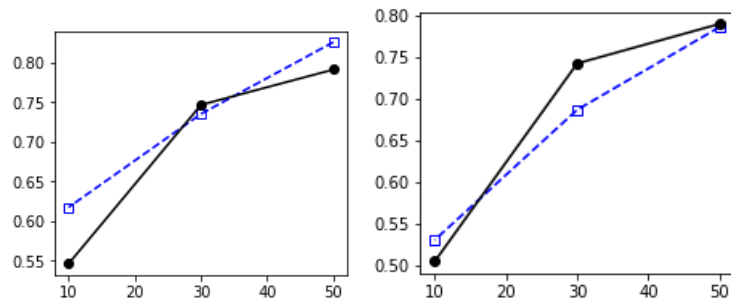

Precision and Recall:

For Test Size 0.5 Not scaled
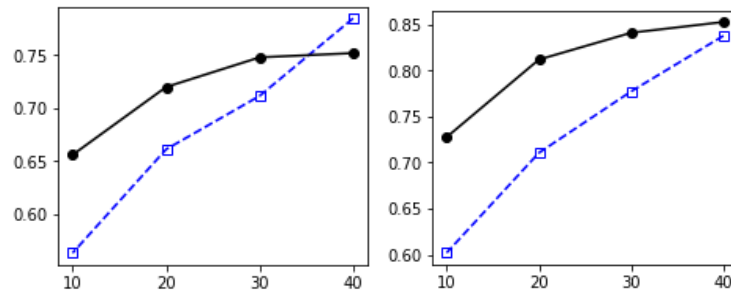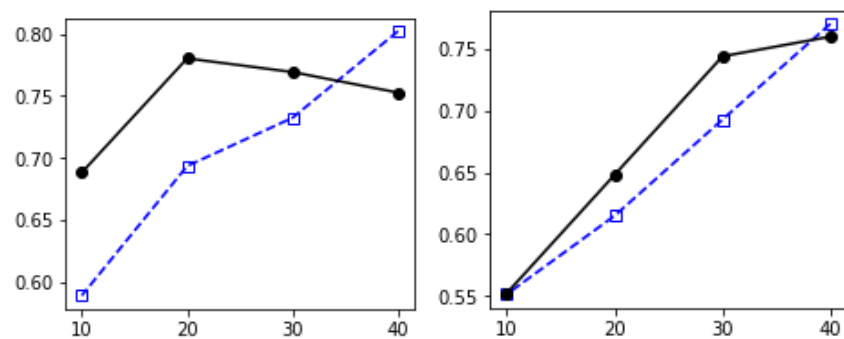Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

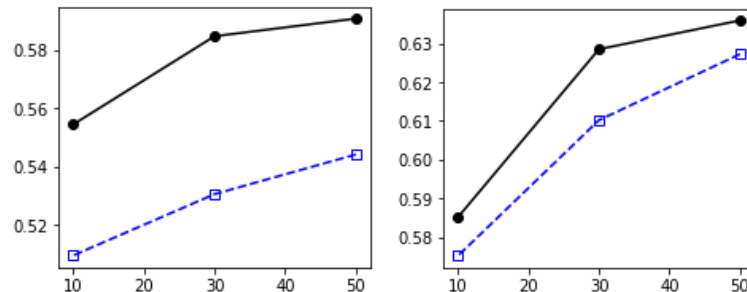Accuracy and Area Under the Curve:

Precision and Recall:

Somerville Dataset 20

Source: https://archive.ics.uci.edu/ml/datasets/Somerville+Happiness+Survey

-Sample Size: 143   -Attributes: 6   -65:78 of 0:1

-Information: This dataset contains information on a variety of public health focused factors, including the price of housing, the relationship with local law enforcement, and public works, with a binary dependent variable representing if a sample is happy or unhappy.
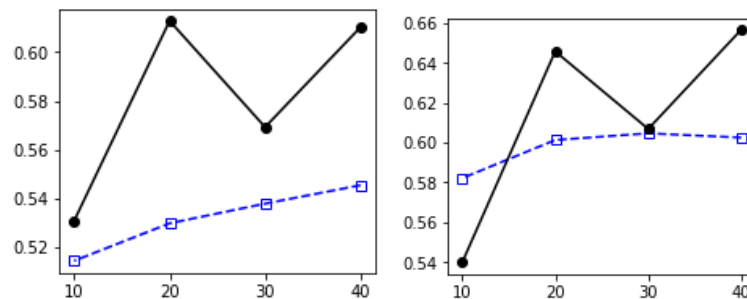
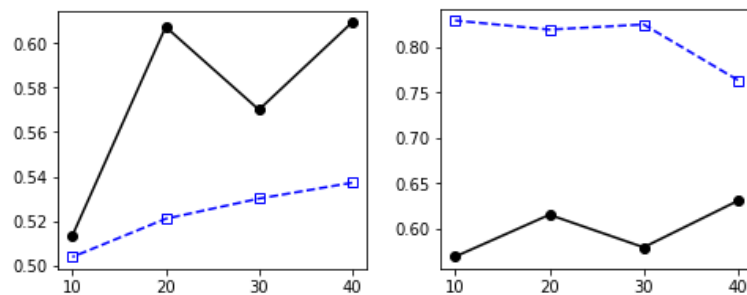Accuracy and Area Under the Curve:
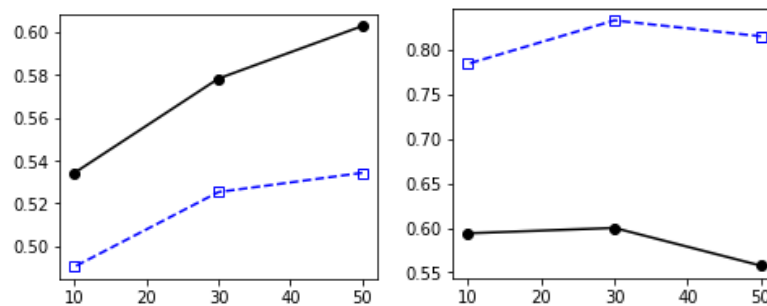
Precision and Recall:

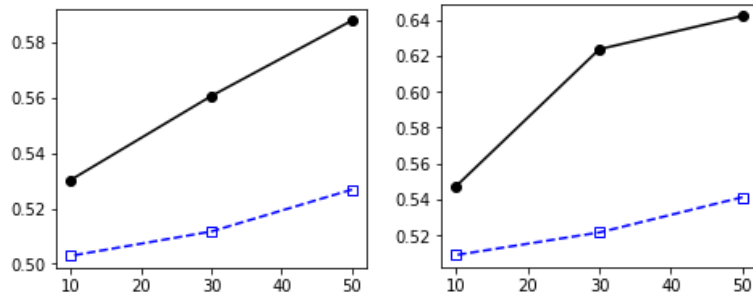**For Test Size 0.6 Rescaled**

Accuracy and Area Under the Curve:

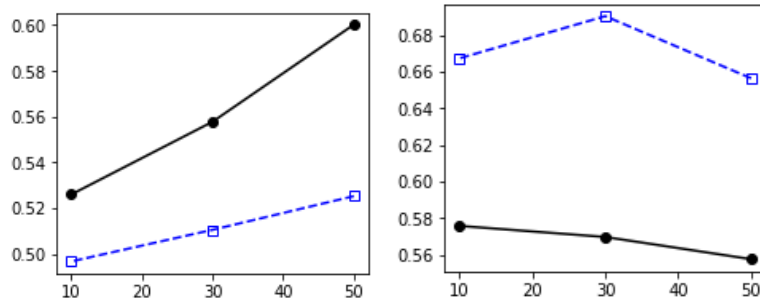Precision and Recall:

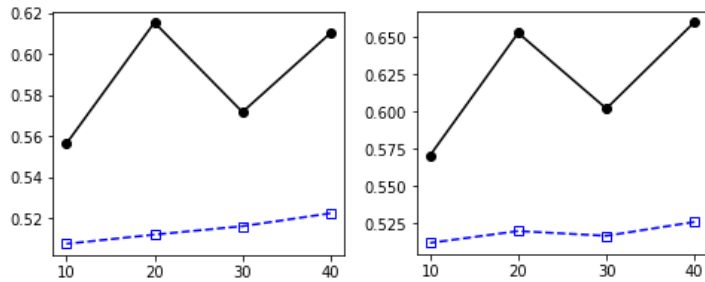**For Test Size 0.5 Not scaled**

Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

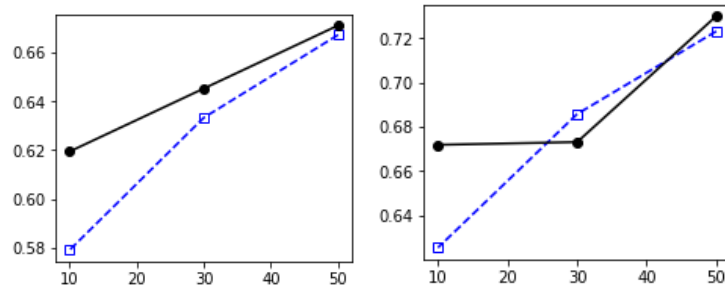Accuracy and Area Under the Curve:
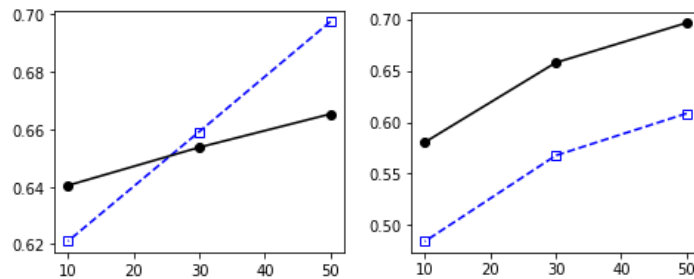

Precision and Recall:
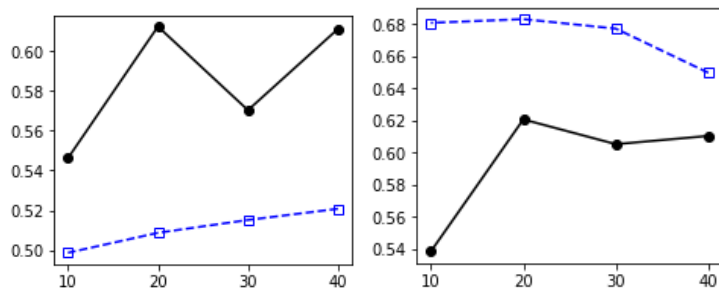
-Sample Size: 165   -Attributes: 49   -56:109 of 0:1
-Information: This dataset contains information on hepatocellular carcinoma patients and whether or not they have other potential diseases, such as HIV. Many of the variables are presented in binary, with the dataset ultimately also assessing if patients have symptoms as a binary variable.

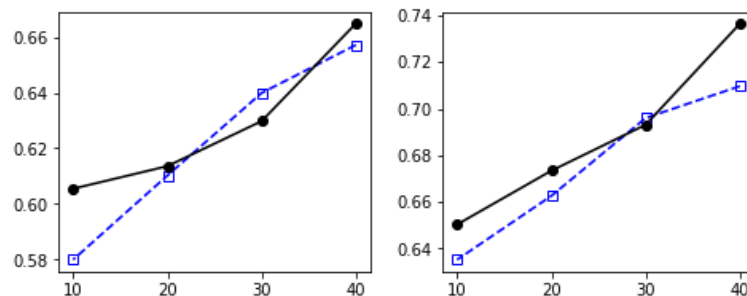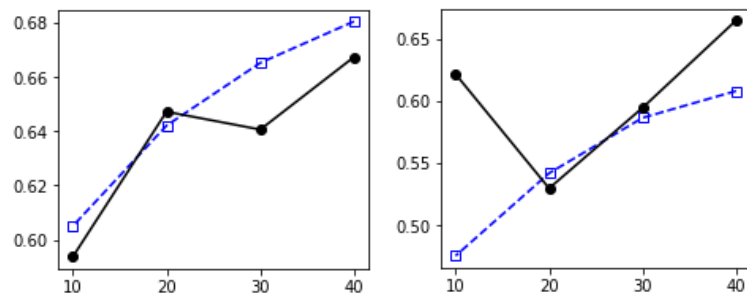Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Rescaled
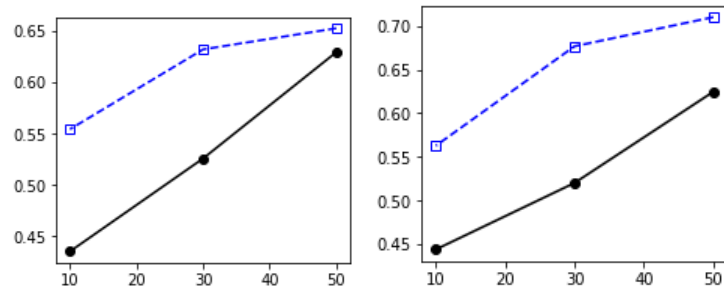
Accuracy and Area Under the Curve:
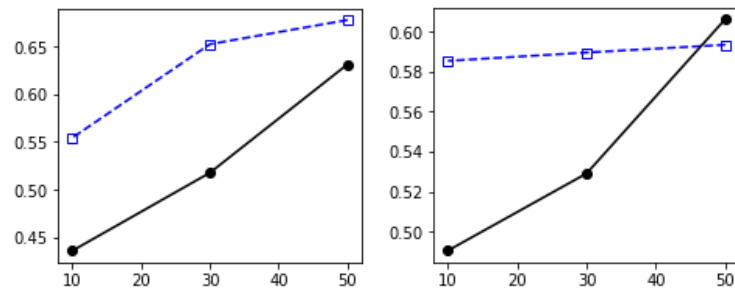

Precision and Recall:

For Test Size 0.5 Not scaled
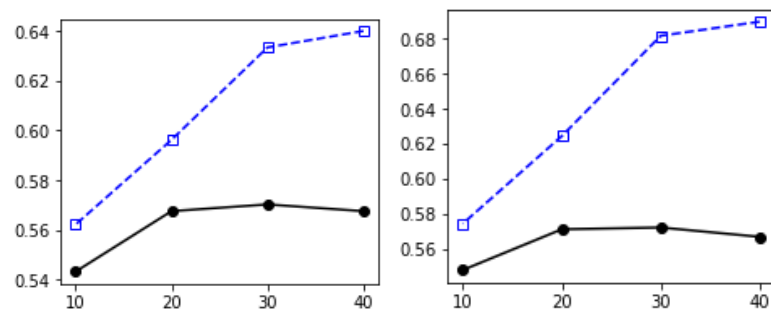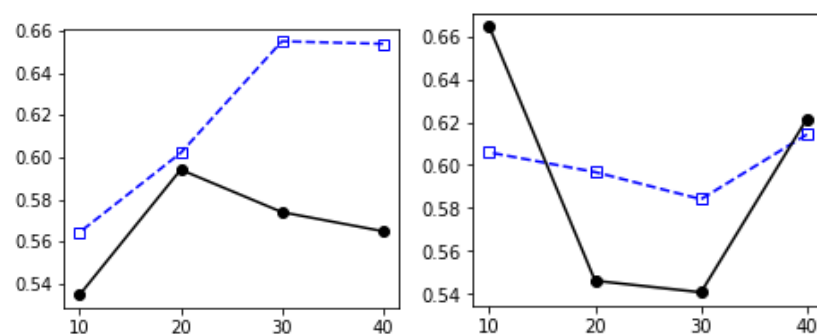
Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

Accuracy and Area Under the Curve:

Precision and Recall:

Heart Disease Dataset 22
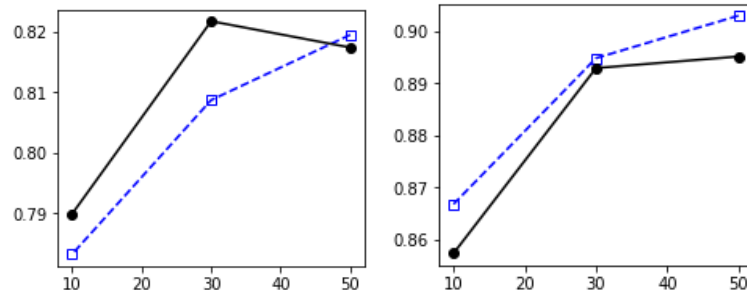Source: https://www.kaggle.com/ronitf/heart-disease-uci
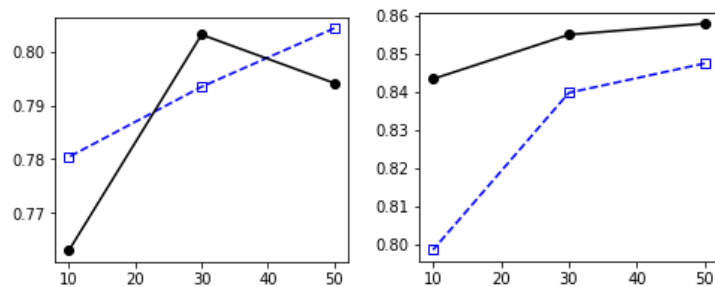-Sample Size: 303
-Attributes: 12
-138:165 of 0:1
-Information: This dataset contains multiple individuals all of whom have different symptoms, which could contribute to the binary dependent variable, whether or not they have heart disease.
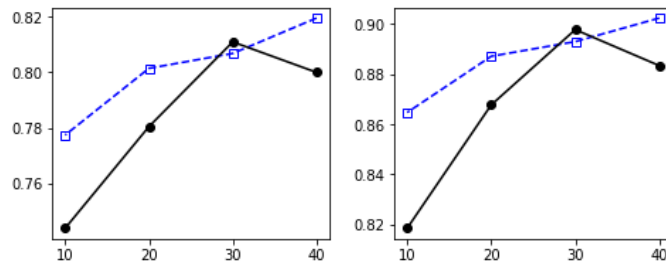Accuracy and Area Under the Curve:
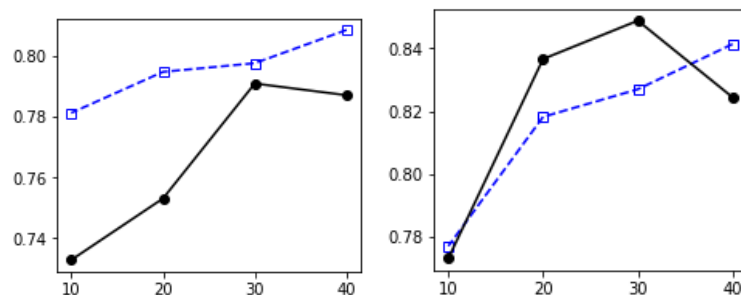

Precision and Recall:

For Test Size 0.6 Rescaled
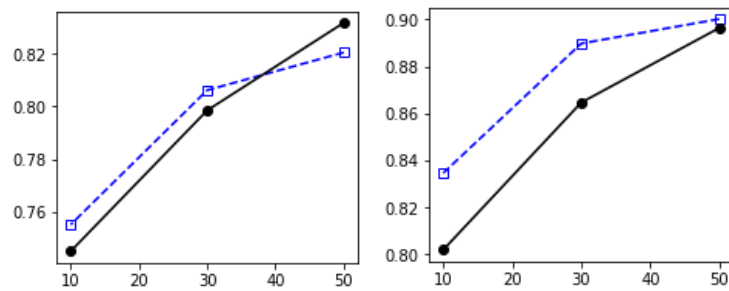Accuracy and Area Under the Curve:
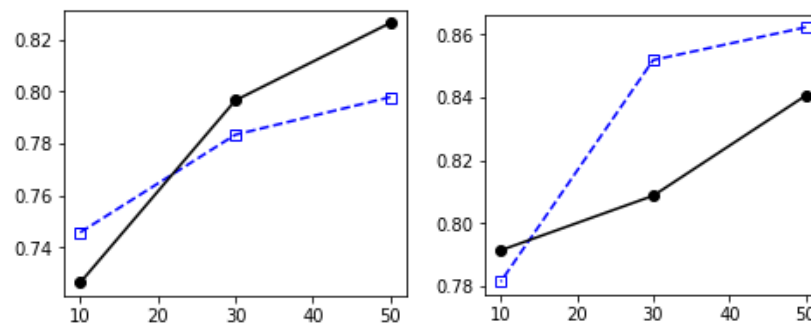

Precision and Recall:

For Test Size 0.5 Not scaled
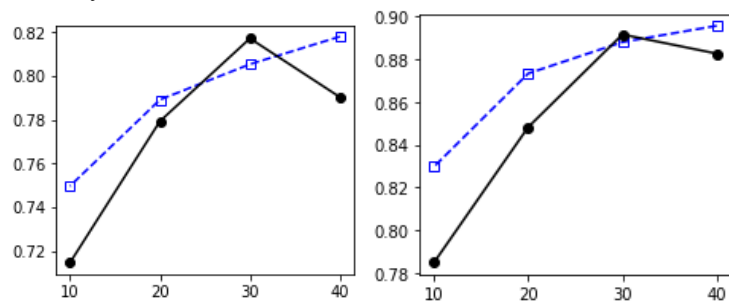Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

Accuracy and Area Under the Curve:

Precision and Recall:
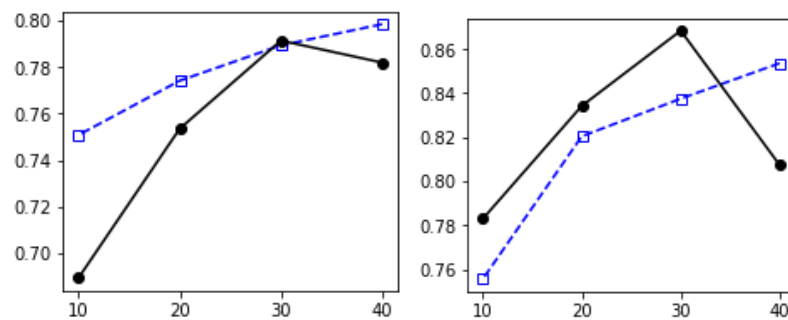
Source: https://www.kaggle.com/jeevannagaraj/indian-liver-patient-dataset
-Sample Size: 584   -Attributes: 8 -167:416 of 0:1[2]
-Information: This dataset contains information on multiple people who do or don't have liver disease, and their lab samples comparing distinct factors such as albumin and bilirubin levels, and the binary variable, which was presented in terms of 1's and 2's, as to whether or not they are a liver disease patient.
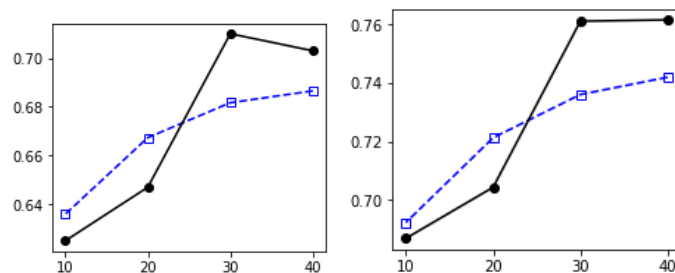Accuracy and Area Under the Curve:
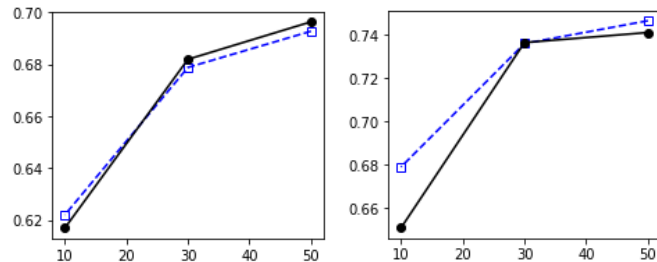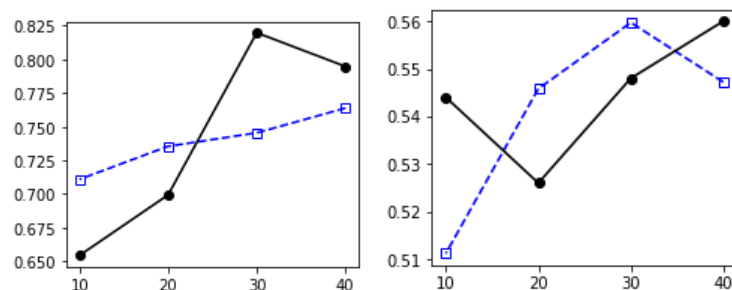

Precision and Recall:
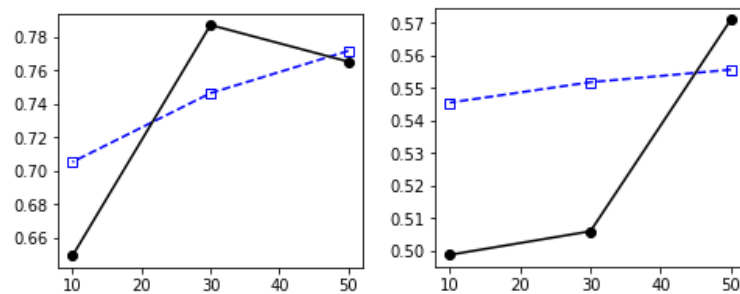
For Test Size 0.6 Rescaled
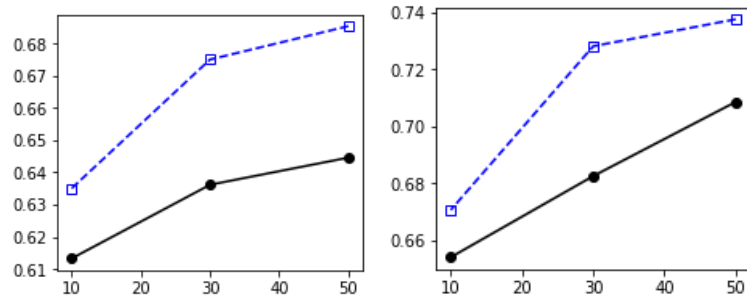Accuracy and Area Under the Curve:
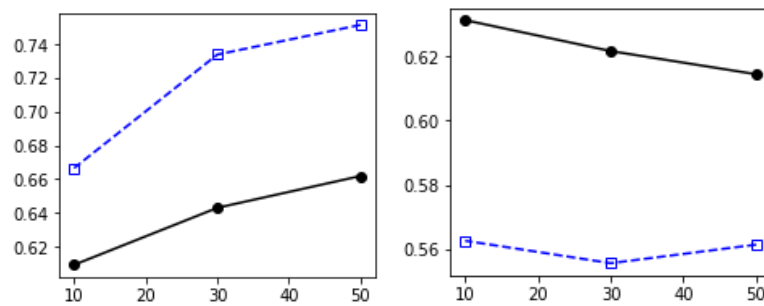

Precision and Recall:

---

2 Was represented as 2:1 in dataset, so all of the 2's were replaced with 0.

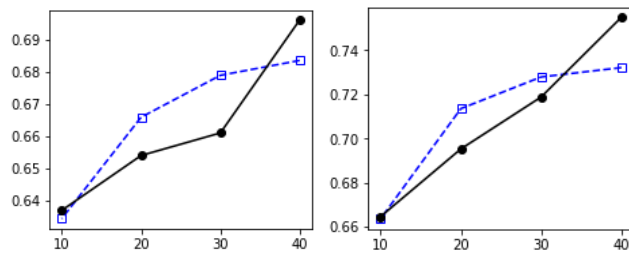**For Test Size 0.5 Not scaled**

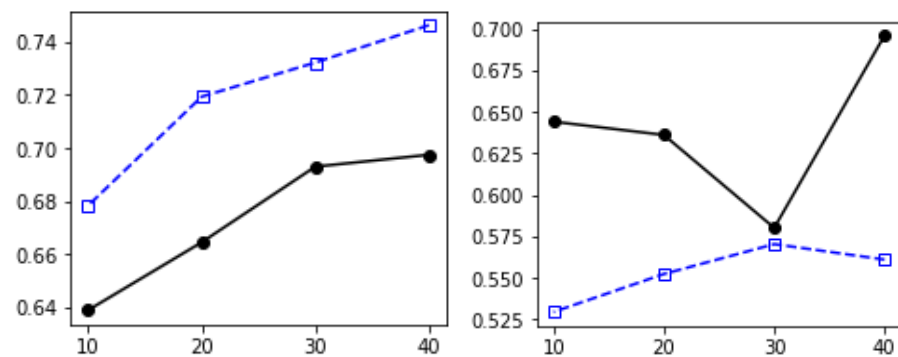Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.6 Not scaled

Accuracy and Area Under the Curve:

Precision and Recall:
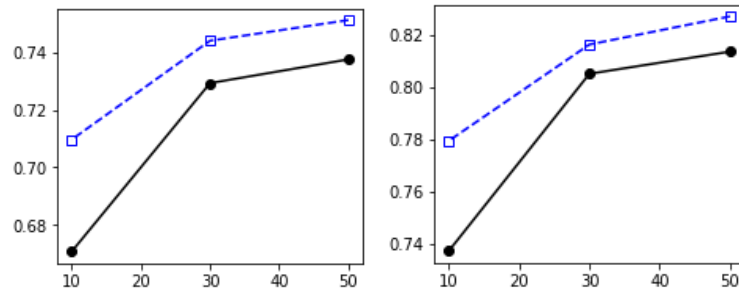
Language Impairment Dataset 24
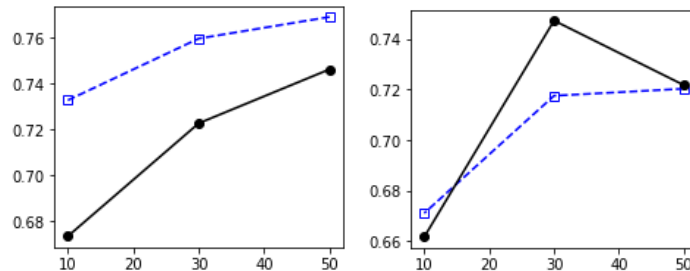Source: https://www.kaggle.com/dgokeeffe/specific-language-impairment
-Sample Size: 1164   -Attributes: 61   -896:268 of 0:1
-Information: This dataset provides developmental information on different children studied, and if given different factors studied, the child has a developmental disorder tied to language impairment. The binary dependent variable is represented by the presence of language impairment or not.
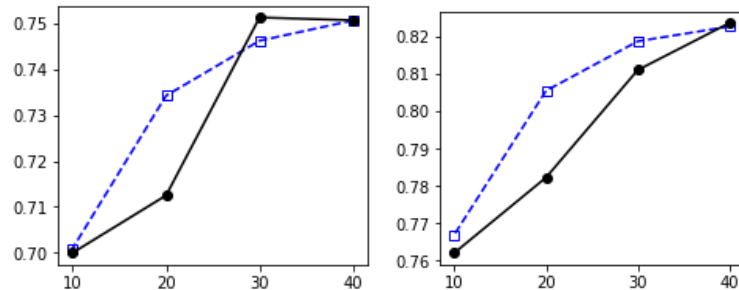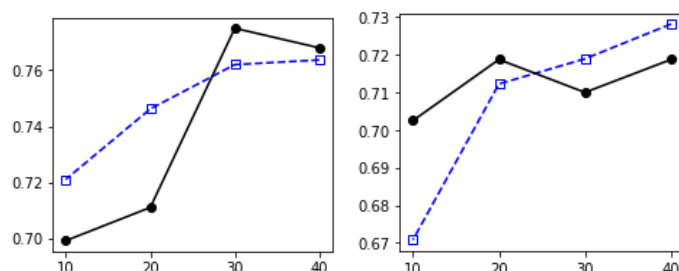Accuracy and Area Under the Curve:

Precision and Recall:

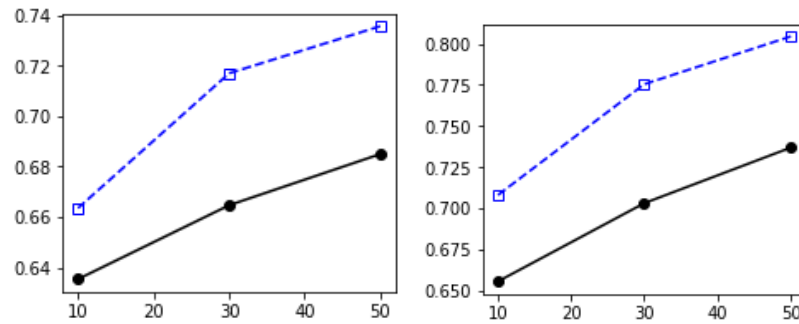**For Test Size 0.6 Rescaled**
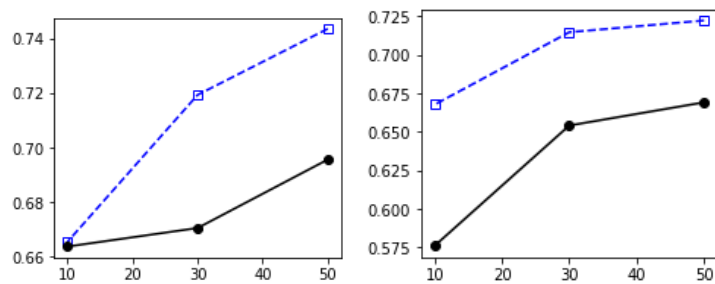Accuracy and Area Under the Curve:

Precision and Recall:

For Test Size 0.5 Not scaled
Accuracy and Area Under the Curve:

Precision and Recall:

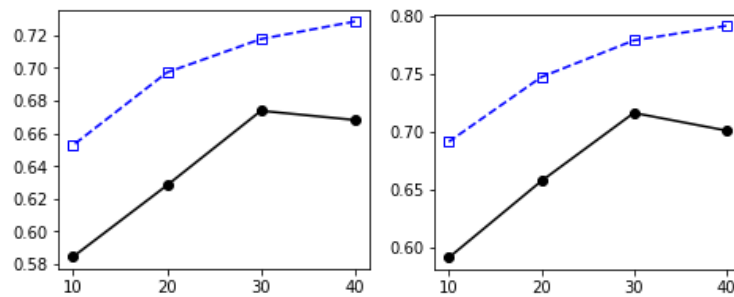**For Test Size 0.6 Not scaled**
Accuracy and Area Under the Curve:

Precision and Recall: