Antonio Gordillo Toledo: antonio.gor@berkeley.edu

Quoc Dung Pham: quocdungpham@berkeley.edu

Dataset: Contraceptive

Final Project Write-Up

Abstract

This project involved the use the Contraceptive Method Choice Data Set, which is a subset of a health survey conducted in Indonesia in 1987. Although the dataset is small in features, it encompasses many socioeconomic features. Due to this, many interesting questions can be posed and answered using this dataset. One such question is whether you can accurately predict the use of contraceptives given certain socioeconomic factors. Several data processing methods were used in combination with different models to make an attempt at predicting contraceptive use.

Introduction

The samples in the contraceptive dataset were obtained from married women who were not currently pregnant. Many questions of interest may be posed with the dataset. The dataset itself makes a distinction between no use, short-term use, and long-term use of contraceptives. Before the dataset could be used for predictive modeling, data exploration and processing is required to familiarize oneself with the data and its scope. The data and its features are described in the following section and an overview of the data processing is given after that. All generated tables are presented in the text; figures are presented at the end of the text in the appendix.

Description of Data

The dataset contains nine features that encompass various aspects of life that might affect a woman's use of contraceptives. A description of these features is on Table 1, found below.

Variable	Variable name	Type	Description
Wife’s age	wife_age	Numerical	Age of woman being surveyed
Wife's education	wife_education	Categorical	Level of education of woman being surveyed: 1 = lowest, 4 = highest
Husband’s education	husband_education	Categorical	Level of education of woman’s husband being surveyed: 1 = lowest, 4 = highest
Number of children ever born	num_child	Numerical	Total number of children the wife has had in her lifetime
Wife's religion	wife_religion	Categorical	Religion practiced by woman being surveyed: 0 = Non-Islam, 1 = Islam
Wife’s working status	wife_work	Categorical	Status of work of woman being surveyed: 0 = Working, 1 = Not Working
Husband's occupation	husband_occupation	Categorical	Code of husband’s current occupation: 1, 2, 3, 4
Standard of living index	standard_living	Categorical	Current standard of living the woman being surveyed: 1 = lowest, 4 = highest
Media Exposure	media_exposure	Categorical	Level of media exposure of woman being surveyed: 0 = Good, 1 = Not Good
Contraceptive method used	contraceptive	Categorical	1 = No contraceptive use 2 = Short-term contraceptive use 3 = Long-term contraceptive use

Exploratory Data Analysis

The first method of exploration employed to the dataset was a correlation matrix to present relationships between the different features. An initial correlation matrix was created before realizing that one-hot encoding was necessary to infer any meaningful relationships between multi-class categorical data. After one-hot encoding (described in the following section) was performed, a correlation matrix was created. One interesting thing is that only the largest values are strongly correlated: level of education of 4 among husbands and wives are strongly correlated while this correlation is much lower for other level of education; same thing applies to standard of living level of 4 and level of education of 4. Figure 1 shows the correlation matrix.

After this, distributions were examined within every feature. Figures 2-10 in the appendix show the distributions. The biggest insight drawn from these visualizations is the fact that the dataset is skewed towards more educated women with high standards of living. This is a potential source of bias that is being introduced, as the dataset does not fully represent the experience of woman across all socioeconomic background. As such, any model trained on this dataset is likely to have these biases incorporated into it. When examine the distribution of the number of children, outliers (> 8 children) were found to primarily correspond to woman with higher education and higher standards of living.

One final observation is in the use of contraceptives with respect to standard of living. There appears to be no clear distinction for short-term use of contraceptives; about a third of all woman across all standards of living indicate short-term use of contraceptives. However, long-term use of contraceptives does notably increase as standard of living increases. As such, a predictive model would likely struggle to differentiate between all three categories but might perform better when simply trying to distinguish no-use from use of contraceptives, whether short or long-term use of it. Figures 11 and 12 show this finding.

Description of Methods

Due to the fact that the dataset contains a mixture of numerical and categorical data, some processing was required before proceeding with model creation. Wife's age and Number of Children Ever Born are the only two numerical features. Three possibilities were discussed for numerical features: leave as is, binning, and min-max standardization. Binning would have created many new features because of the integer ranges, so that was discarded. Two DataFrames were created (one with numerical data as is and another with min-max standardization) so that a comparison could be made in how these processing techniques affect model accuracies.

Categorical features were processed via one-hot encoding, except for three features that were already in a binary form: Wife's Religion, Wife Working, and Media Exposure. In the process of performing one-hot encoding, k-1 features were generated for a category with k possibilities. This was done to avoid what is referred to as the Dummy Variable Trap. A total of five DataFrames were generated, described below in Table 2.

Index	Dataset name	Dataset Description
1	data	Original dataset without any modification
2	data_ohe	Dataset with modification of one-hot encoding on categorical variables.
3	data_ohe_s	Dataset with modification of one-hot encoding on categorical variables and min-max standardization on numerical variables.
4	data_s_only	Dataset with modification of min-max standardization on numerical variables and excluding all the categorical variables.
5	data_ohe_only	Dataset with modification of one-hot encoding on categorical variables and excluding all the numerical variables.

Description of Models

For each of the five DataFrames previously described, two models were trained: a Random Forest Classifier (RFC) and a Support Vector Classifier (SVC). Hyperparameter turning was performed to observe how the models performed given different parameters. For the RFC, various number of estimators were use; for the SVC, various values for regularization were used. Table 3 below describes the models accuracy in the multi-class setup. Figures 13-17 show accuracy via hyperparameter tuning.

DataFrame	RFC (Multi-class)		SVC (Multi-class)	
	Max Train Set Accuracy	Max Test Set Accuracy	Max Train Set Accuracy	Max Test Set Accuracy
data	0.958	0.542	0.577	0.606
data_ohe	0.958	0.542	0.582	0.606
data_ohe_s	0.958	0.535	0.754	0.511
data_s_only	0.637	0.532	0.551	0.562
data_ohe_only	0.608	0.410	0.608	0.467

It was of interest to see how the models performed when simply attempting to distinguish between contraceptive use and no use. The training above was repeated on all five DataFrames, but the contraceptive label was reduced to a binary problem. Figures 18-22 show the accuracy during the hyperparameter tuning process.

DataFrame	RFC (Two Class)		SVC (Two Class)	
	Max Train Set Accuracy	Max Test Set Accuracy	Max Train Set Accuracy	Max Test Set Accuracy
data	0.977	0.732	0.739	0.755
data_ohe	0.977	0.698	0.747	0.755
data_ohe_s	0.977	0.691	0.841	0.677
data_s_only	0.769	0.698	0.707	0.715
data_ohe_only	0.733	0.589	0.733	0.627

Summary and Interpretation

Random Forest Classifiers seem to always overfit, regardless of the DataFrame that is passed into the model and regardless of the number of estimators used. SVC performs better than RandomForestClassifiers because it does a better job of generalizing rather than overfitting. Of all five DataFrames used, the best results were obtained using the unprocessed and one-hot encoded DataFrames. This is strange, as it should be expected that one-hot encoding would make a greater difference. Solely using numerical or categorical data as model input drastically decreased the accuracy of the model. The results of the multi-class problem were essentially repeated in the simplified two-class problem, just with higher accuracies.

An interesting thing to note is that using the numerical features to generate standardized features turned out to decrease model performance. This was unexpected, as standardizing tends to help models. The biggest challenge in this project is explaining the low accuracy in the multi class problem. The source is likely a combination of bias in the data, fairly even use of short-term contraceptives across all socioeconomic factors, and the small number of available features. The previously mentioned bias in the dataset (due to the data being skewed towards highly educated woman) imposes a limitation on the scope of the results. A more comprehensive dataset that included a more representative distribution of woman across various socioeconomic factors would strengthen the presented analysis. With more questions added to a future survey, more socioeconomic factors could shed a light on trends that are not immediately obvious in the given dataset. Future work would thus require a more comprehensive dataset. Whether these results would hold in a more general dataset is not clear. Due to the anonymity of the dataset, there were no ethical concerns with disclosing personal information.

Appendix

Figure 1

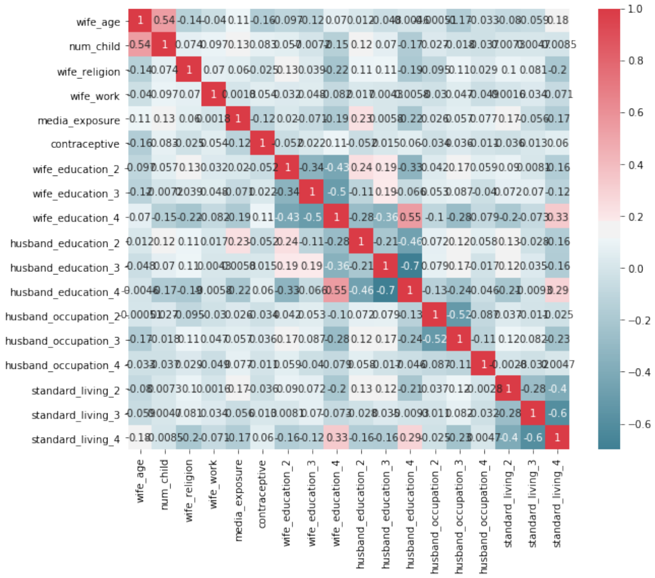


Figure 2

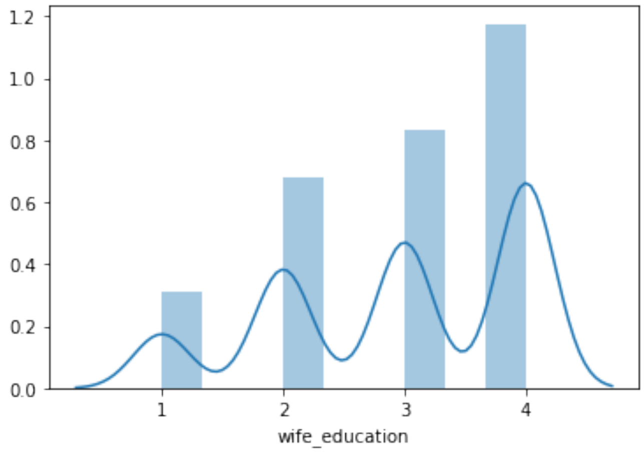


Figure 3

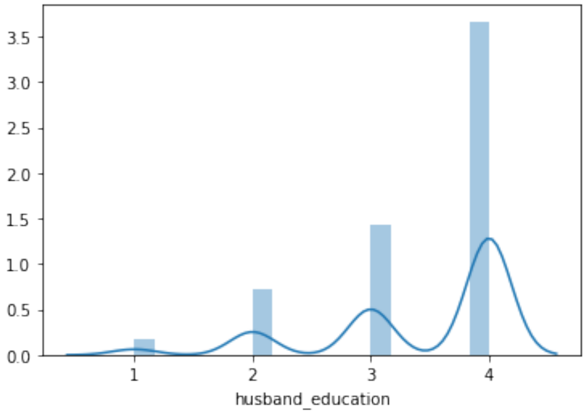


Figure 4

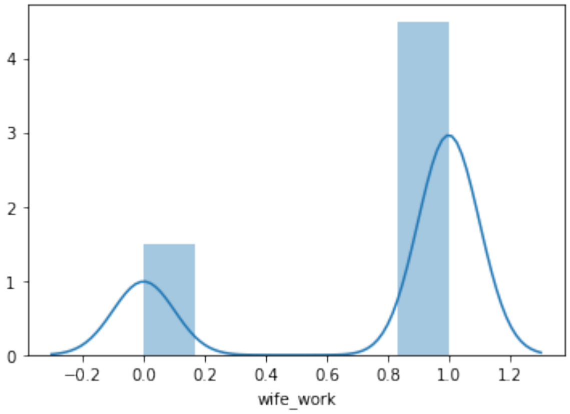


Figure 5

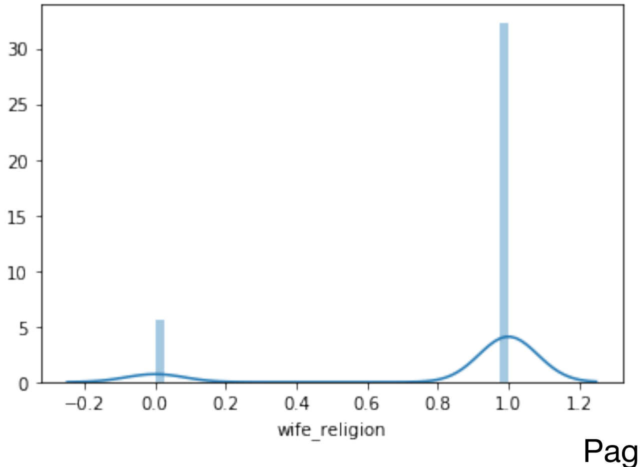


Figure 6

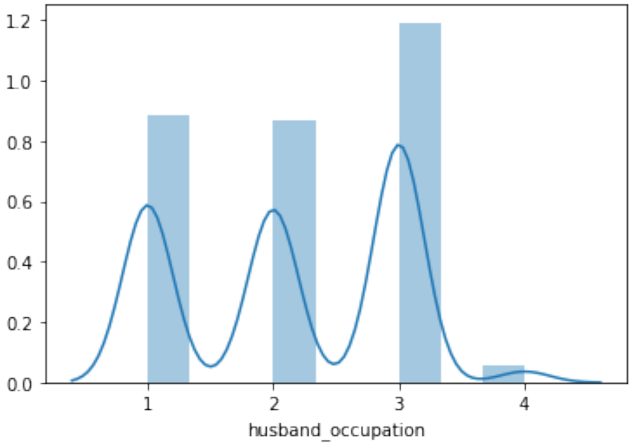


Figure 7

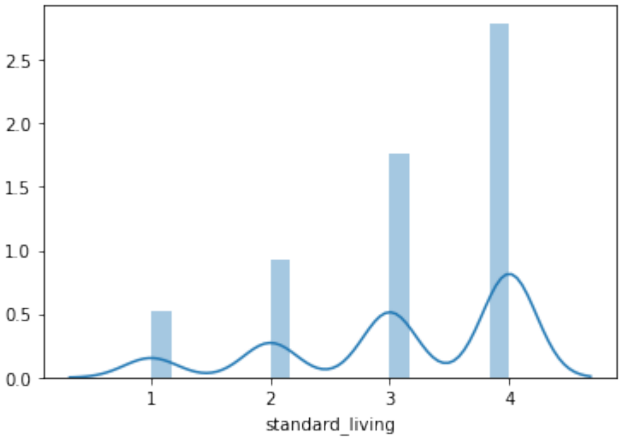


Figure 8

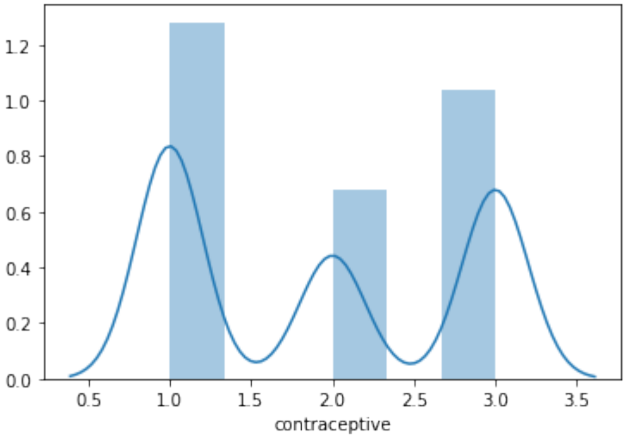


Figure 9

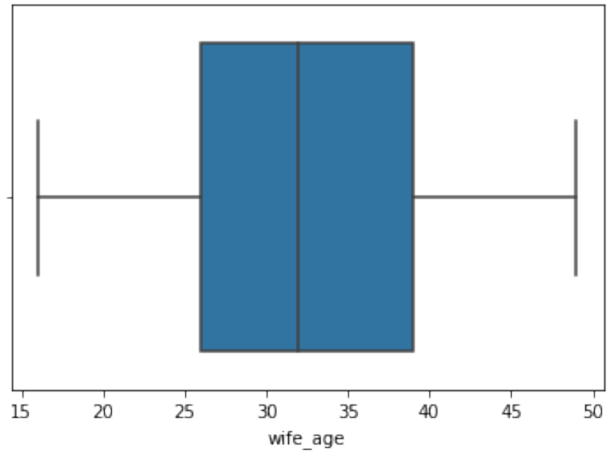


Figure 10

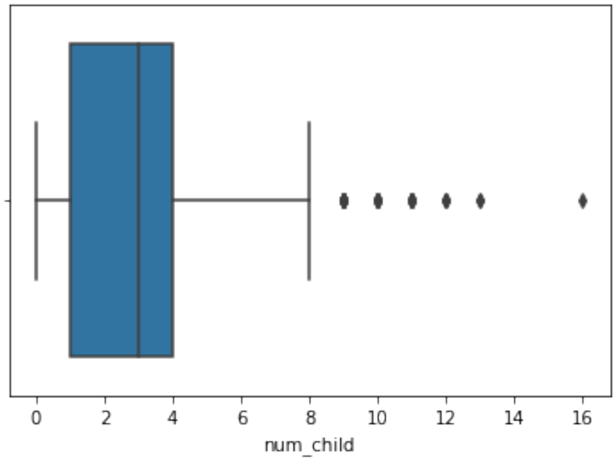


Figure 11

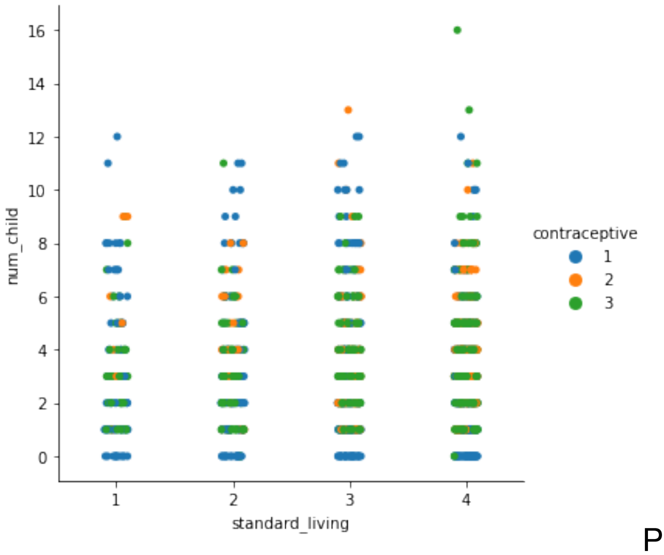
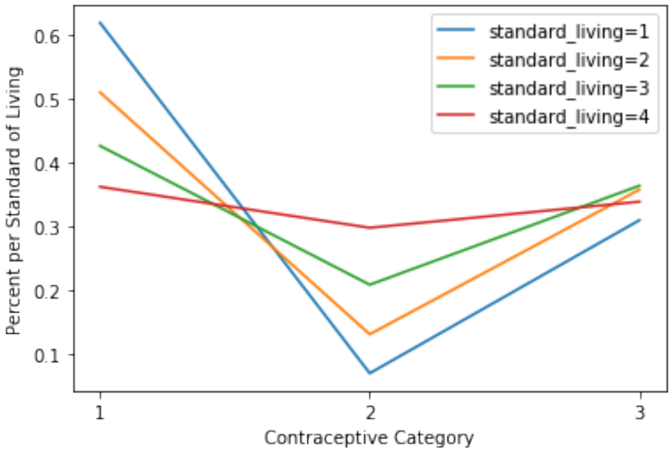


Figure 12



Multi-Class Problem

Figure 13

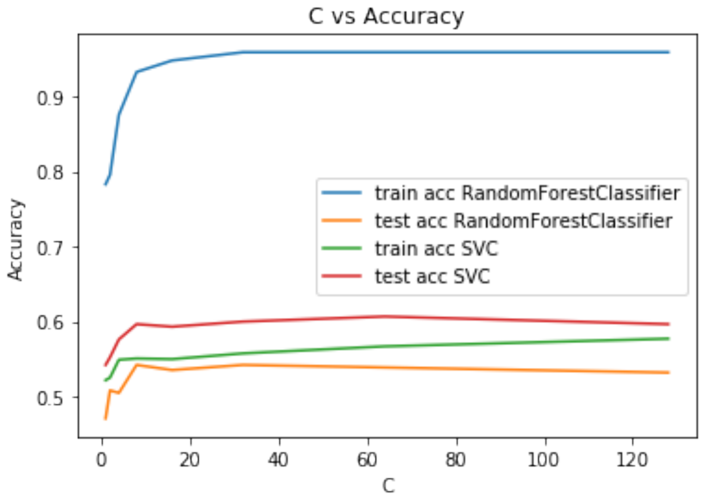


Figure 14

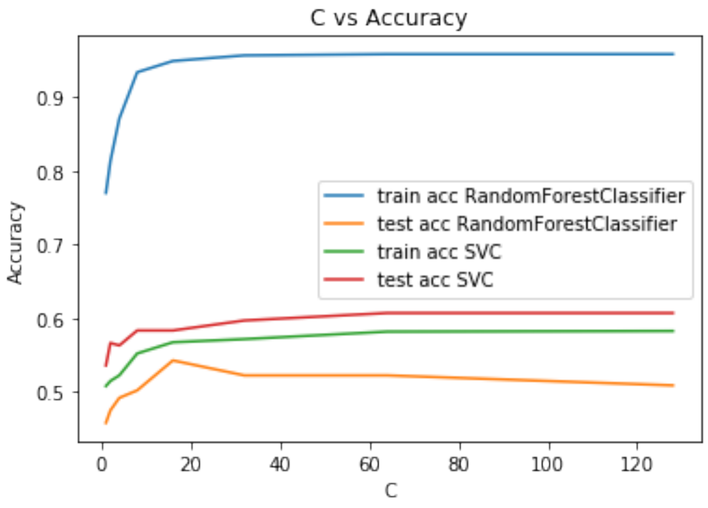


Figure 15

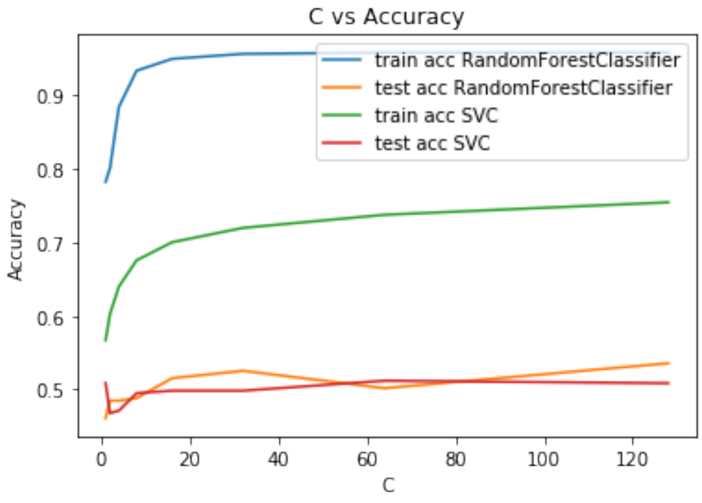


Figure 16

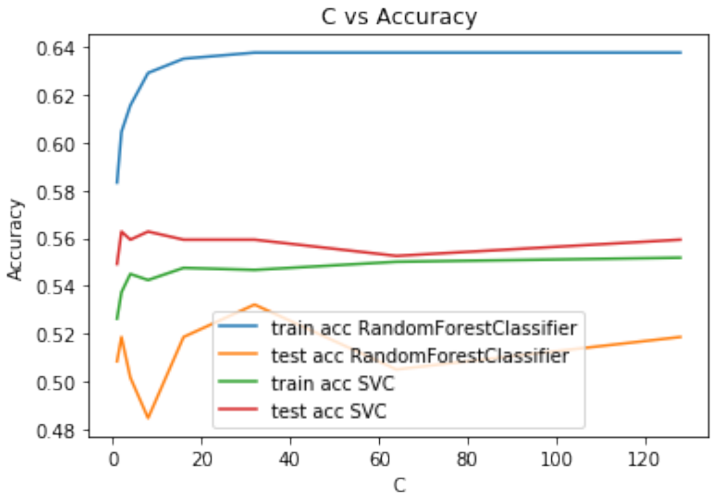
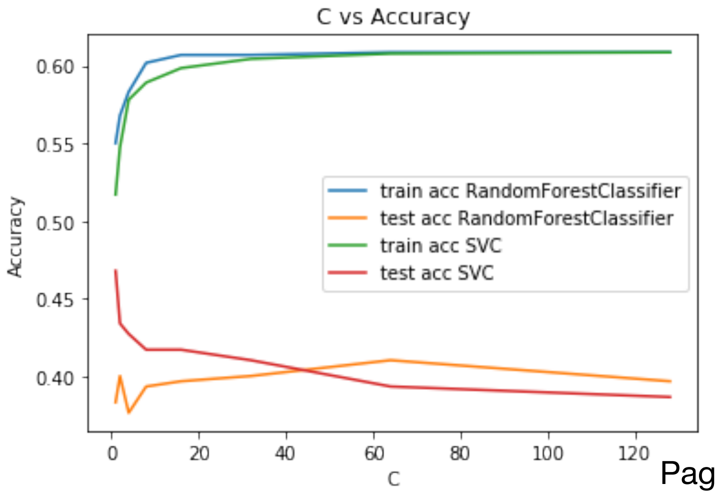


Figure 17



Two-Class Problem

Figure 18

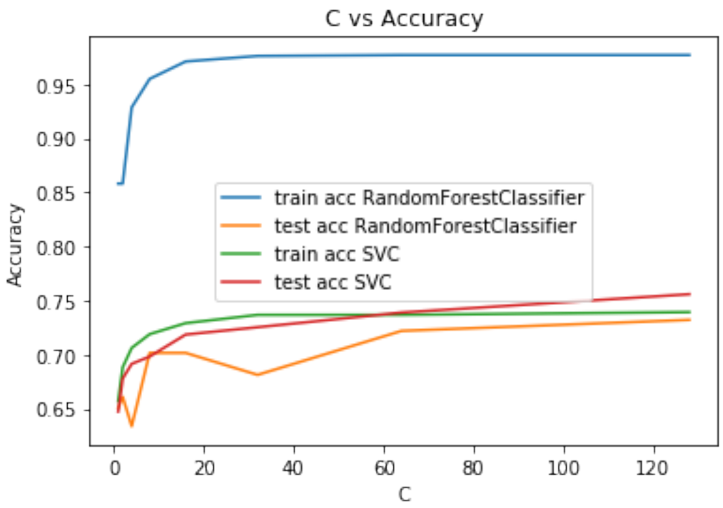


Figure 19

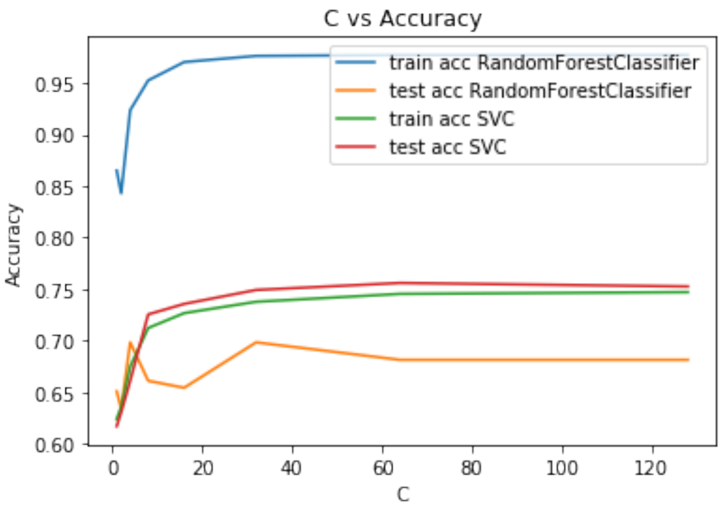


Figure 20

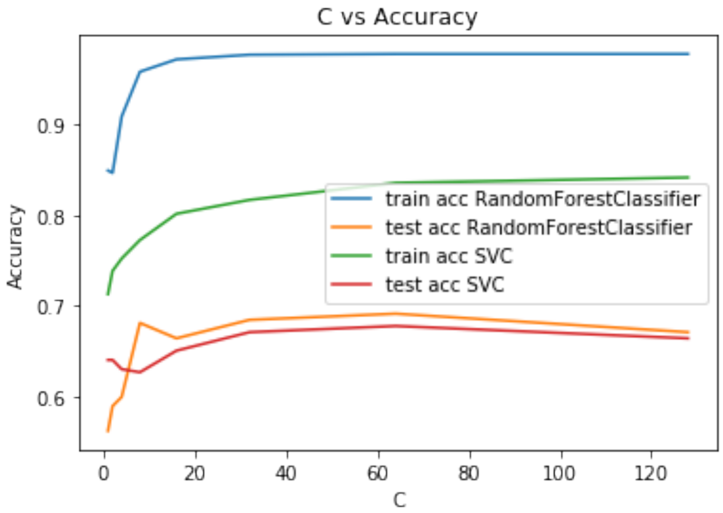


Figure 21

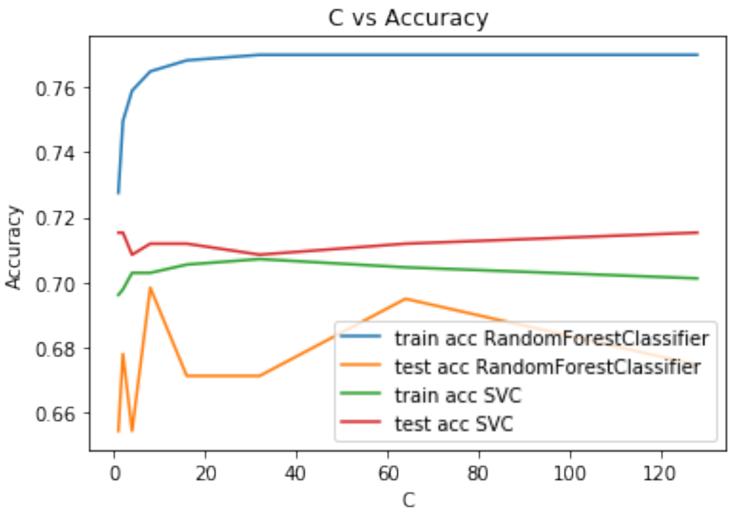
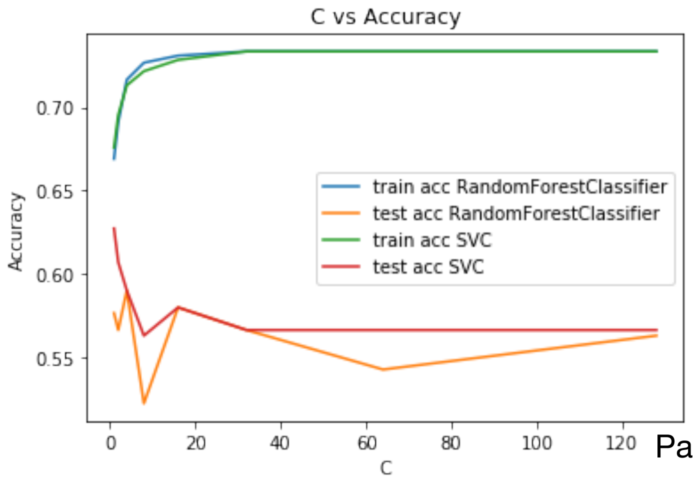


Figure 22


C vs Accuracy
Accuracy
C
train acc RandomForestClassifier
test acc RandomForestClassifier
train acc SVC
test acc SVC