Iris Flower Classification and Predictions

I. INTRODUCTION

This report will perform supervised learning operations using Orange on a data set of Iris flowers provided by the Orange tool. Classification methods will be used to predict the genus of the Iris flower as the target and metric based measurements will be used as the features. Several models will be employed in this effort to predict the genus type and each model's accuracy as well as implementation details will be recorded and compared to the other models to determine which model yields the best results. In each case there will be a common training and testing set used. The goal is to understand which genus types are the easiest and hardest to classify based on the metrics provided from the data set.

Future sections of this report describe the dataset, the methodology, results accompanied by a discussion, and concluding thoughts. Section II contains a description of the dataset used for this analysis as well as a detailed table of all the attributes along with stacked histograms. The methodology and model choices for classification is presented in section III which will reply heavily on scatter plots. Section IV contains a report and discuss of the results based on metrics such as F1, precision, and recall. Finally, section V will provide the conclusion as to which classification model provides the best predictions and which genus is the easiest and hardest to predict from using classification models.

II. DATA DESCRIPTION

There are five columns in total contained in the data set as depicted in Table I. The first column identifies the name of the flower and will be the target of prediction models. The next two columns measure the flowers SEPAL LENGTH and SEPAL WIDTH respectively as nonnegative numeric values in units of centimeters. The last two columns of Table I measure the flowers PETAL LENGTH and PETAL WIDTH respectively as nonnegative values, also in units of centimeters. SEPAL LENGTH, SEPAL WIDTH, PETAL LENGTH, and PETAL WIDTH will be used to create meaningful classifications in sections II and III.

TABLE I.

Attribute	Type	Example Value	Description
NAME	Nominal (string)	"Iris-setosa"	Iris Genus Name
SEPAL LENGTH	Numeric (real)	5.0	Sepal Length (cm)
SEPAL WIDTH	Numeric (real)	3.8	Sepal Length (cm)
PETAL LENGTH	Numeric (real)	1.4	Sepal Length (cm)
PETAL WIDTH	Numeric (real)	0.2	Sepal Length (cm)

The attributes recorded in Table I can be better understood when their frequencies are analyzed. Using histograms with stacked columns, the distribution of data collected for SEPAL LENGTH, SEPAL WIDTH, PETAL LENGTH, and PETAL WIDTH are easily visualized in Figures I, II, III, and IV of page two.

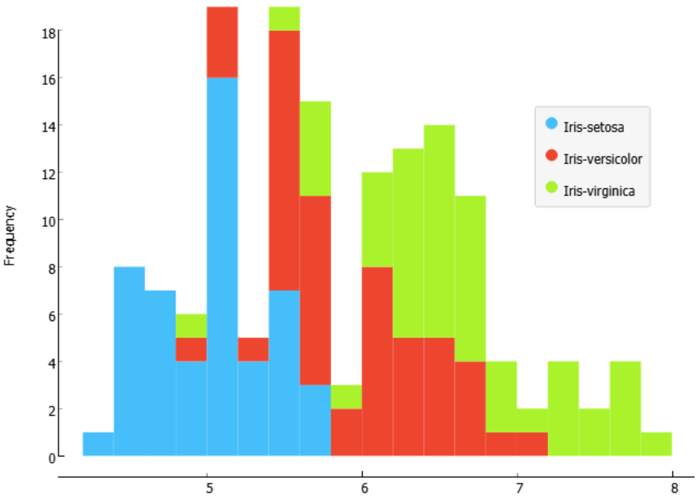


Fig. 1. Frequency distribution of
SEPAL LENGTH

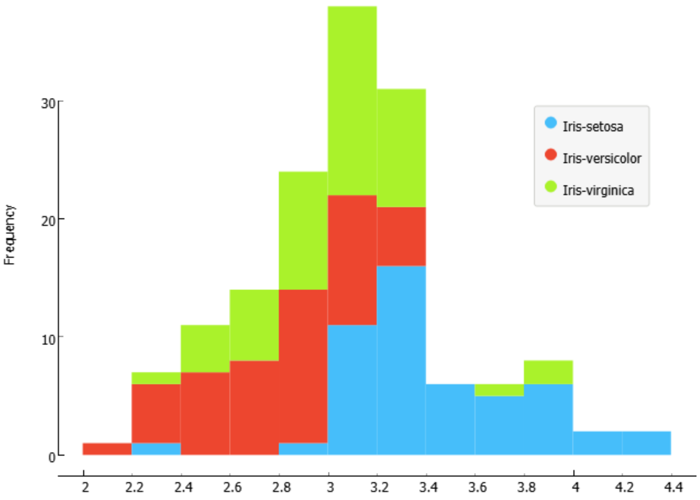


Fig. 2. Frequency distribution of
SEPAL WIDTH

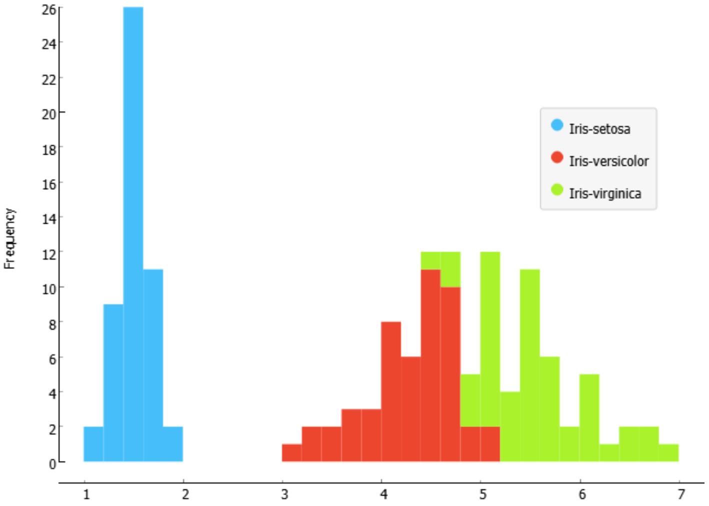


Fig. 3. Frequency distribution of
PETAL LENGTH

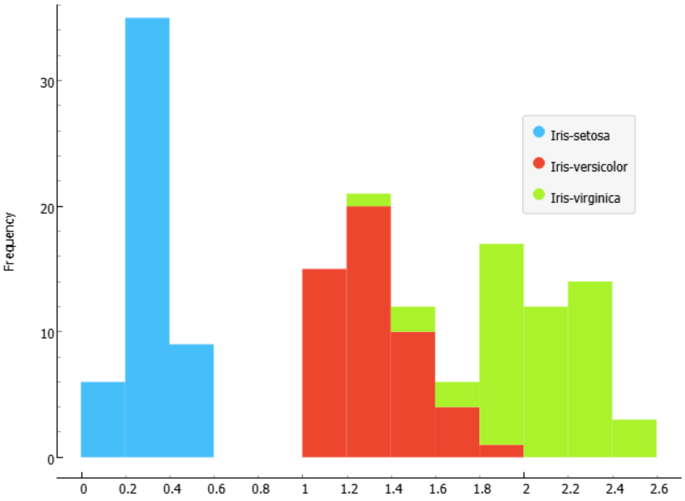


Fig. 4. Frequency distribution of
PETAL WIDTH

III. Methodology

The data provided by the Orange Tool for Iris flowers was analyzed in the following ways. First, the two most meaningful attributes was determined by using a tree model to create a decision tree of the four features in the data set. The decision tree created by this method contained a maximum depth of five layers as shown in Figure 5. The first layer of the decision tree used the PETAL LENGTH as the first split decision. The second layer contained only one other split decision based on PETAL WIDTH. From these two decision splits it was determined that the PETAL LENGTH and PETAL WIDTH were the two most meaningful features and would be used as the axis for creating scatter plots.

After the axis for the scatter plots was determined, three other models were used classify the data set. In total the four models used to classify the data set are: kNN (k – Nearest Neighbors), logistic regression, random forest, and tree (decision tree). Each of these four models was applied to classify the data set and the results of each are recorded in the form of confusion matrices. The training and testing of the data in all cases used a cross validation method where the data set was randomly divided into five sets. In this way the training and testing data was always split along eighty and twenty percent lines. Confusion matrices were used in this instance to measure the number of correct predictions as well as the number of false positives and false negatives. These values are non-trivial because they will in Section IV to measure the F1, Precision, and Recall values.

Finally, the number of classifications and misclassifications are visualized using scatter plots. The data set contains three unique target values for predicting the flower’s genus. Each predicted genus data point is represented in its respective scatter plot using the same color patterns as Figures I – IV. Scatter plots were chosen as the chief visualizer to showcase the Euclidean distance between data points and as binary representations for misclassification where an unfilled circle represents a misclassification and a filled circle represents a correct prediction.

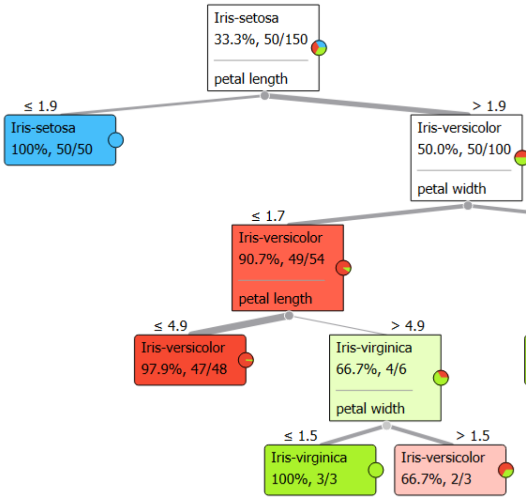
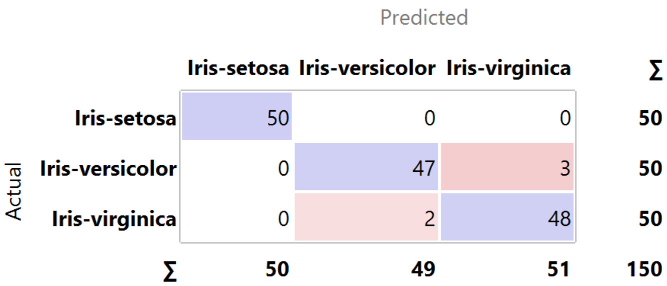


Fig. 5. Decision Tree of Iris Flower data set

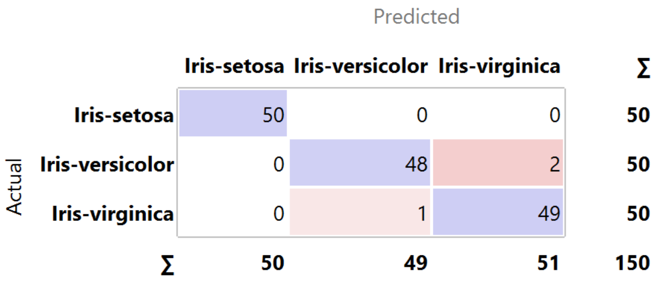
IV. Results and Discussion

Figures VI, VII, VIII, and IX show the confusion matrices that were created with their respective prediction models are input. The values that form a diagonal line from the top left corner to the bottom right corner count the number of correct predictions while the numbers outside these boxes count the number of misclassifications and what they were misclassified as.



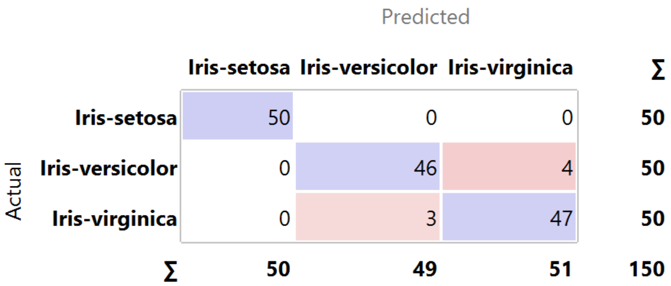
		Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Actual	Iris-setosa	50	0	0	50
	Iris-versicolor	0	47	3	50
	Iris-virginica	0	2	48	50
	Σ	50	49	51	150

Fig. 6. Confusion Matrix of kNN



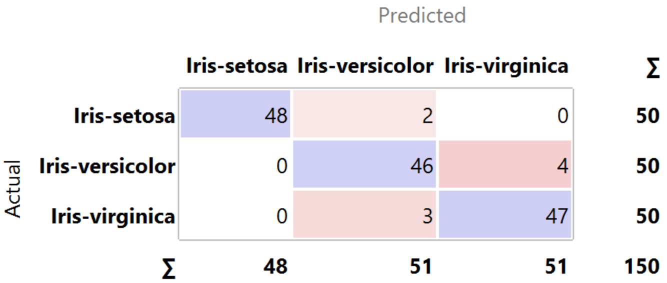
		Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Actual	Iris-setosa	50	0	0	50
	Iris-versicolor	0	48	2	50
	Iris-virginica	0	1	49	50
	Σ	50	49	51	150

Fig. 7. Confusion Matrix of Logistic Regression



		Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Actual	Iris-setosa	50	0	0	50
	Iris-versicolor	0	46	4	50
	Iris-virginica	0	3	47	50
	Σ	50	49	51	150

Fig. 8. Confusion Matrix of Random Forest



		Predicted			
		Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Actual	Iris-setosa	48	2	0	50
	Iris-versicolor	0	46	4	50
	Iris-virginica	0	3	47	50
	Σ	48	51	51	150

Fig. 9. Confusion Matrix of Tree

These misclassification and correct prediction counts are then used to determine the F1, Precision, and Recall values. The Precision value is a measure of how accurate our model is with respect to how high the false positive count is. The Recall value is a measure of how accurate our model is with respect to how high the false negative count is. F1 combines Precision and Recall into a single measure that controls the weight of both and commonly reduces both to a harmonic mean. In this report, F_β is replaced with F1 because our $\beta = 1$.

$$Precision = \frac{\text{\# True Positive}}{\text{\# True Positive +\# False Positive}} \tag{1}$$

$$Recall = \frac{\text{\# True Positive}}{\text{\# True Positive +\# False Negative}} \tag{2}$$

$$F1 = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{3}$$

The confusion matrices are then mapped onto a scatter plot which visualizes the classifications of each individual point. As mentioned in the methodology section of this report, filled circle represent correct predictions and unfilled circles represent misclassified or incorrect predictions.

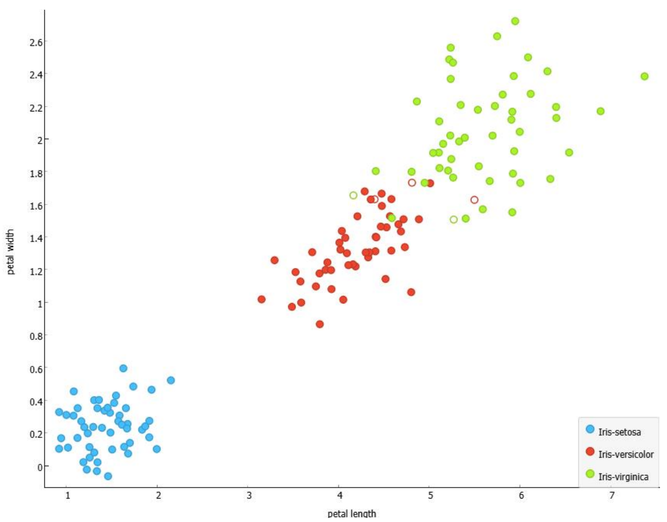


Fig. 10. Scatter Plot of kNN

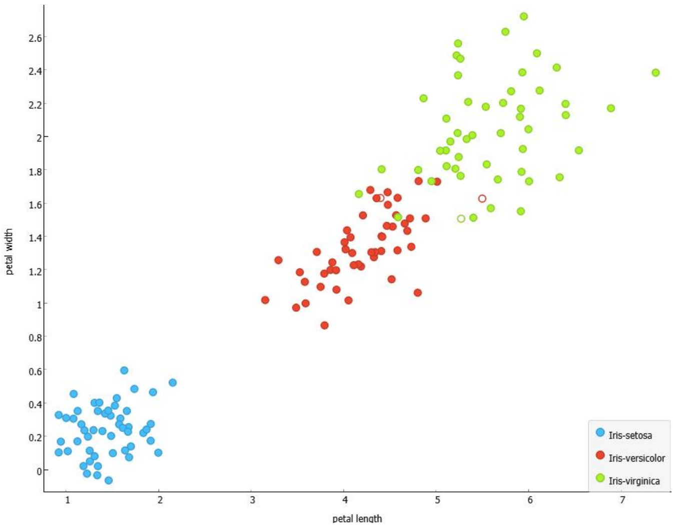


Fig. 11. Scatter Plot of Logistic Regression

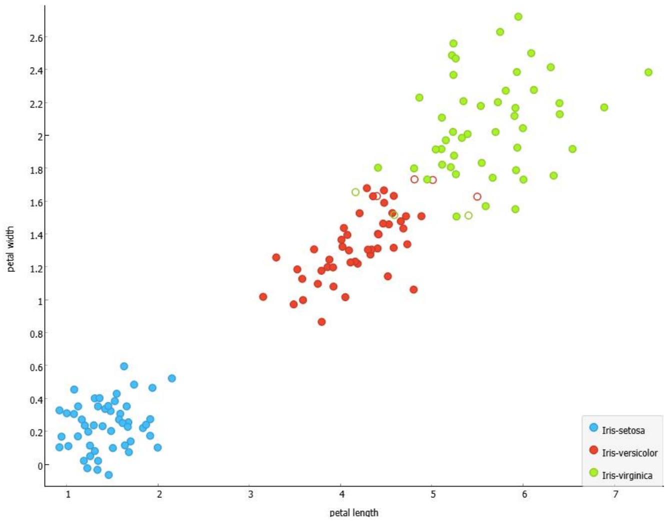


Fig. 12. Scatter Plot of Random Forest

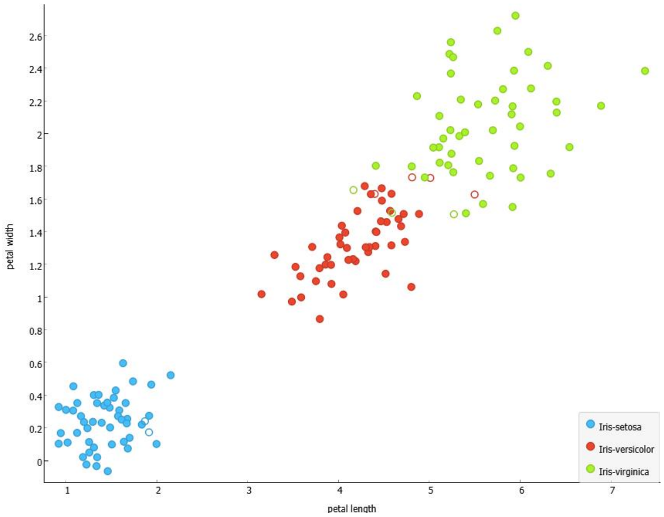


Fig. 13. Scatter Plot of Tree

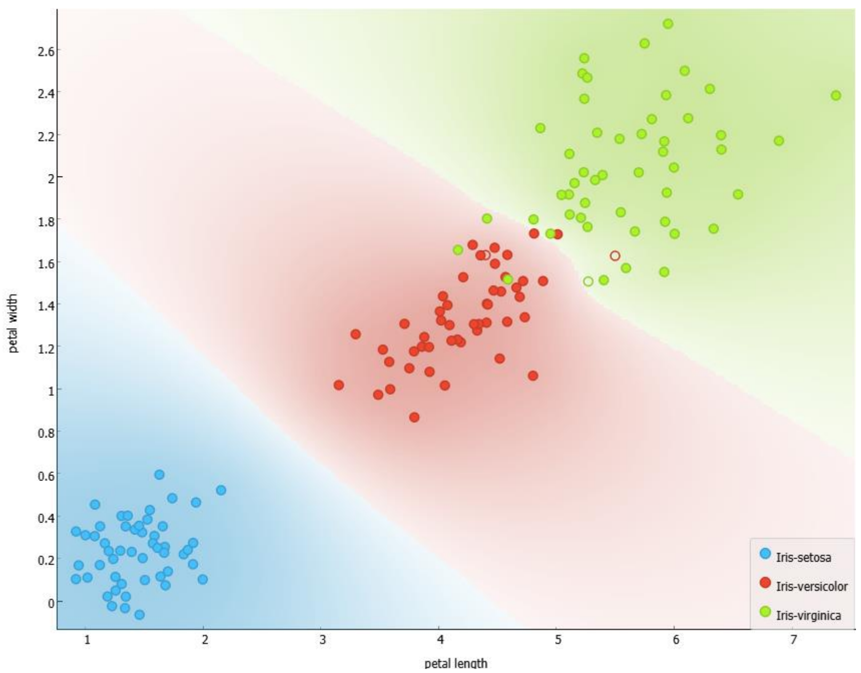


Fig. 14. Scatter Plot of Logistic Regression with Color Map

V. Conclusions

The results from section IV can all be recorded in Table II with F1, Precision, and Recall values.

TABLE II.

Table Evaluation Results			
Model	**F1**	**Precision**	**Recall**
Logistic Regression	0.980	0.980	0.980
kNN	0.967	0.967	0.967
Random Forest	0.953	0.953	0.953
Tree	0.940	0.941	0.940

From Table II it can be concluded that the logistic regression model yielded the best results and is the best classification for Iris flowers. After applying a color map to Figure XI a clear distinction can be made between the three target values of Iris Flowers. The clearest of which is for Iris-setosa while the clusters of data points for Iris-versicolor and Iris-virginica. From this it can be concluded that Iris-setosa is the easiest to classify while Iris-versicolor and Iris-virginica are more difficult.