IrisSpeciesPredictor

Tiffany Chu, Gaurang Ahuja, Nguyen Nguyen, Vienne Lee

Summary

This project investigates whether iris species can be predicted using sepal and petal measurements. After loading and validating the dataset, we explore some basic patterns, do some EDA and then train a model. The overall results show that petal measurements provide strong separation between species, allowing the models to achieve high accuracy.

Introduction

The Iris dataset is a well-known benchmark in machine learning, containing measurements of iris flowers collected to study how physical characteristics differ across species.

Feature Summary

There are 4 numerical features in this dataset:

1. `sepal_length`: The length of the sepal (outer part of the flower)
2. `sepal_width`: The width of the sepal
3. `petal_length`: The length of the petal
4. `petal_width`: The width of the petal

There is 1 categorical feature in the dataset: 1. `species`: The target variable (label)

We are using the Iris dataset to answer the question: **"Can we predict the Iris species using petal and sepal measurements?"**

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are not linearly separable from each other (Fisher 1936).

Table 2: Basic data statistics

```
Dataframe has 105 rows
Dataframe has 5 cols
```

Methods & Results

```
File saved to: ./data/iris.csv
```

Table 1: Preview of Iris Data

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Basic Data Stats

In the code below, we check some basic stats for the dataframe such as the number of rows, columns etc. We convert the columns names to a more standard format using underscores and lowercase. We then check how many unique species there are.

Data Wrangling

Since our goal is classification, this section will look at statistics which will help distinguish between the species. Therefore, we will look at things like mean, median, min, max etc for each feature. This will be followed by graphical analysis where we will explore bar plot, pairwise plot, and boxplot.

Table 3: Summary of numeric columns

	sepal_length	sepal_width	petal_length	petal_width
count	105.000000	105.000000	105.000000	105.000000
mean	5.869524	3.050476	3.837143	1.232381
std	0.796039	0.409062	1.714265	0.739217
min	4.400000	2.200000	1.000000	0.100000

Table 3: Summary of numeric columns

	sepal_length	sepal_width	petal_length	petal_width
25%	5.100000	2.800000	1.600000	0.400000
50%	5.800000	3.000000	4.400000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.700000	4.400000	6.900000	2.500000

Table 4: Variation per species

	sepal_length					sepal_width					petal_le
	mean	median	std	min	max	mean	median	std	min	max	mean
species											
setosa	5.000000	5.00	0.336011	4.4	5.8	3.415625	3.4	0.400894	2.3	4.4	1.471875
versicolor	6.000000	5.95	0.505863	4.9	7.0	2.832500	2.9	0.283194	2.2	3.4	4.297500
virginica	6.554545	6.40	0.613948	4.9	7.7	2.960606	3.0	0.298893	2.5	3.6	5.572727

Insights from Data Statistics

From the above tables, we can see: - Most species have approximately the same number of observations - Petal measurements display the strongest separation - Setosa is distinct while versicolor and virginica show some overlap - Petal dimensions have meaningful differences

EDA Plots

Below are the plots visualize the underlying patterns, distributions, and relationships within the three features (sepal_length, sepal_width, petal_length, petal_width) in the Iris dataset.

Pairwise Plot

Pairwise plots show the relationships between feature pair and show any clusters. Figure 1 helps identify which pair of features show the clearest separation and allows us to identify which features are the most informative for prediction.

Boxplot

Figure 2 show how each feature varies across species. The median, spread, outliers, and overlaps helps identify features that are most predictive for classifying the Iris species.

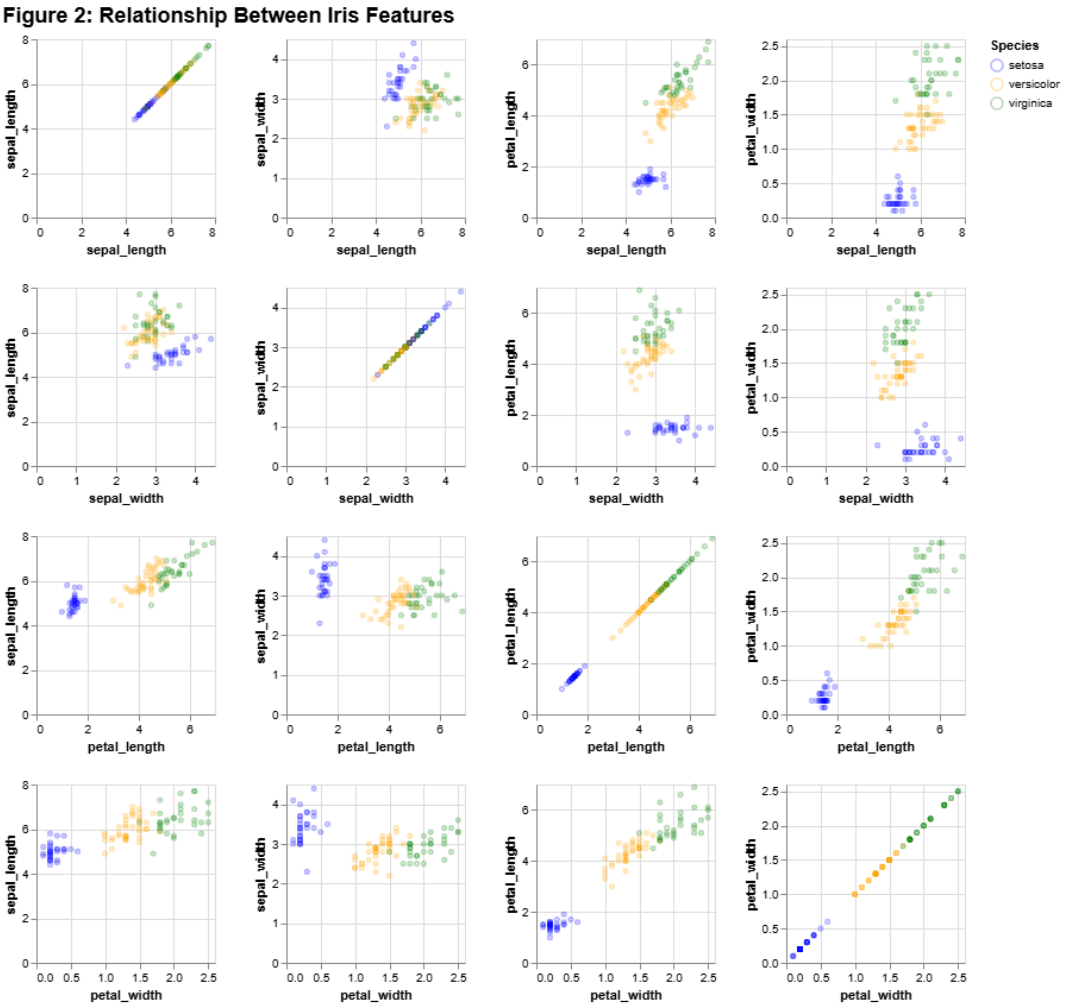

Figure 1: Pairwise Plot of Iris Features

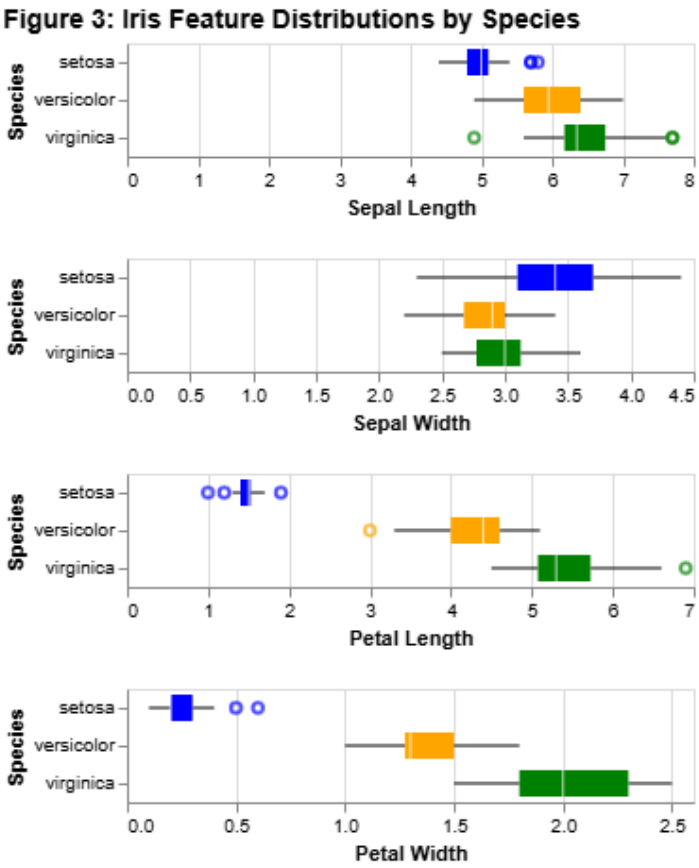

Figure 2: Boxplot of Iris Feature Variation

Model Training

Performs classification or regression analysis and then plot.

estimator	Pipeline(step...lassifier())])
param_distributions	{'decisiontreeclassifier___max_depth': range(1, 20)}
n_iter	50
scoring	None
n_jobs	-1
refit	True
cv	5
verbose	0
pre_dispatch	'2*n_jobs'
random_state	123
error_score	nan
return_train_score	True

copy	True
with_mean	True
with_std	True

criterion	'gini'
splitter	'best'
max_depth	7
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	None
random_state	None
max_leaf_nodes	None
min_impurity_decrease	0.0
class_weight	None
ccp_alpha	0.0
monotonic_cst	None

estimator	Pipeline(step...lassifier())])
param_distributions	{'kneighborsclassifier___n_neighbors': range(1, 20)}
n_iter	50
scoring	None

n_jobs	-1
refit	True
cv	5
verbose	0
pre_dispatch	'2*n_jobs'
random_state	123
error_score	nan
return_train_score	True

copy	True
with_mean	True
with_std	True

n_neighbors	5
weights	'uniform'
algorithm	'auto'
leaf_size	30
p	2
metric	'minkowski'
metric_params	None
n_jobs	None

Table 11: Decision tree model confusion matrix

	setosa	versicolor	virginica
setosa	18	0	0
versicolor	0	10	0
virginica	0	3	14

Table 12: K-NN confusion matrix

	setosa	versicolor	virginica
setosa	18	0	0
versicolor	0	9	1
virginica	0	2	15

Discussion and Analysis

In our investigation of the Iris dataset, we establish whether these features are strong predictors for Iris species classification and rank their predictive power. This discussion begins with the insights derived from our in-depth exploratory data analysis, then the outcomes of our machine learning models, and finally exploring the implications for biological classification and machine learning practices for future research.
To summarize the data, the Iris dataset consists of 150 samples with measurements for four features: sepal length, sepal width, petal length, and petal width, across three evenly distributed species: Setosa, Versicolor, and Virginica.

In our exploratory data analysis (EDA), the summary statistics Table 3 show that petal measurements (length and width) exhibit more obvious differences across species compared to sepal feature measurements. Sepal measurements are not as varied across species and thus provide weaker separation when distinguishing species. Our visualizations of pairwise feature plots Figure 1, and boxplots Figure 2 confirm these patterns, they each reveal the power of petal measurement for distinguishing species during classification. Specifically, the boxplot shows how the species Setosa is easily distinguishable from just its petal measurements.

Both Decision Tree and KNN models perform strongly on this Iris test set (95% vs 93% accuracy, respectively). Petal and sepal feature differences allow decent separation between species, especially between the species Versicolor and Virginica. The misclassifications in confusion matrices show this overlap. The cross-validation results suggest that Decision Trees with max_depth values from 4 to 18 fit the training data perfectly while achieving similarly high validation accuracy (94.3%) on unseen folds. For the KNN hyperparameter tuning, findings show that choosing n_neighbors between 4 and 12 outputs a classification accuracy around 97% on cross-validation test scores.

Examining our confusion matrix for model accuracies: - In the decision tree test performance Table 11, Setosa was classified perfectly (as it is the most distinguishable), while 3 virginica got misclassified as versicolor. - In our KNN model **?@tbl-KNN**, we find that setosa is perfect again, and KNN misclassifies: 1 versicolor as virginica, and 2 virginica as versicolor. - This is expected as the two classes (virginica and versicolor) overlap in features.

Our decision tree model is slightly better than KNN, with better accuracy, fewer total mistakes, perfect versicolor classification, though same performance on setosa and virginica. This suggests high but not perfect predictability of iris species from its measurements, with room for improving classification on overlapping species using more advanced methods or additional features. These results demonstrate a typical well-performing classification pipeline for the Iris dataset, validating the exploratory and modeling insights

The findings that petal measurements strongly separate iris species and have high model accuracy are expected and align with prior knowledge and research on the Iris dataset. The petal length and width consistently show clearer distinctions among the species, particularly

separating Setosa from Versicolor and Virginica, however, Sepal measurements tend to show more overlap and offer weaker discriminatory power.

However, there are some limitations associated with our study, some key limitations of the dataset and the related findings include: The dataset contains only 150 samples with 4 numeric features and 3 species classes, which is considered quite small and simple compared to data on other plant species. This limits the complexity of patterns that can be learned and makes it less applicable to more complex classification tasks. Its size also prevents it from being useful for complex machine learning techniques like deep learning, which require larger datasets. Only sepal and petal measurements are considered, and other potentially informative features such as genetic data, environmental variables, or flower color are not recorded as features in the data. This constrains prediction to specific traits and may limit generalizability to other flowers or datasets
It appears that the species Setosa is clearly separable, while Versicolor and Virginica show overlap in some measurements, leading to some classification errors and ambiguity that models must handle or risk making errors in The dataset is clean with no missing values or measurement noise, which is not typical/ normal in many real datasets. This means models trained here may be optimistic compared to actual real world performance So while appropriate for demonstrating classification approaches and feature importance, the Iris dataset's limitations prevent our model from being broadly applicable and generalizable, which must be considered when interpreting and generalizing findings. We must also note that these methods are derived from the following sources: Pedregosa et al. (2025), UBC Master of Data Science Program (2025a), UBC Master of Data Science Program (2025b), UBC Master of Data Science Program (2025c), and errors from these sources may translate into our research.

The impact of these findings is significant in practical feature selection for machine learning. They show the importance of choosing the most informative features for classification tasks, improving model performance and simplifying models by focusing on fewer but more relevant measurements. This can aid in resource efficiency and interpretability in biological studies, botany, and related scientific work.
This brings up future questions that could use further research: - How do more advanced machine learning algorithms (e.g., SVM) compare in classification accuracy using petal vs sepal features? - Can combining petal and sepal features with additional biological or environmental data improve model robustness or reveal deeper insights? - How well do these findings generalize to other flower species or datasets—can similar feature importance patterns be found? - Could unsupervised learning reveal new subgroups or variations in iris species beyond the three classical ones using these or other features?

In summary, these findings are expected and show the predictive importance of petal dimensions in iris classification. They have implications for feature selection and model design, and point to further possible research on flower classification and its related biological questions

References

Fisher, R. A. 1936. "Iris Dataset." UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/53/iris.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2025. *Scikit-Learn: Machine Learning in Python*. https://scikit-learn.org/stable/.

UBC Master of Data Science Program. 2025a. "DSCI 531: Visualization for Data Science – Course Notes." UBC GitHub Pages. https://pages.github.ubc.ca/mds-2025-26/DSCI_531_viz-1_students/.

———. 2025b. "DSCI 571: Supervised Learning 1 – Course Notes." UBC GitHub Pages. https://pages.github.ubc.ca/mds-2025-26/DSCI_571_sup-learn-1_students/.

———. 2025c. "DSCI 573: Feature and Model Selection – Course Notes." UBC GitHub Pages. https://pages.github.ubc.ca/mds-2025-26/DSCI_573_feat-model-select_students/.