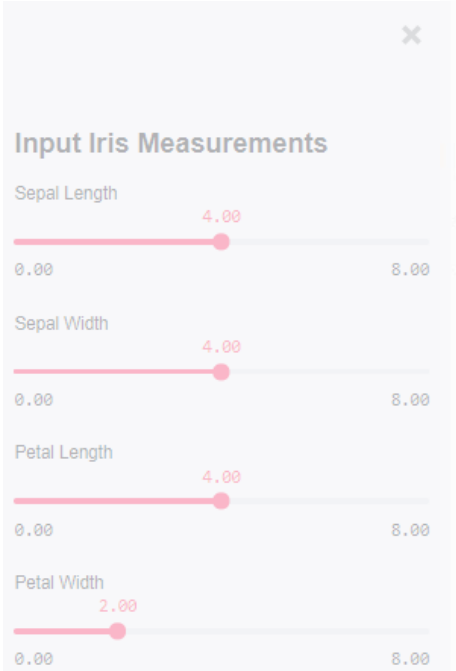
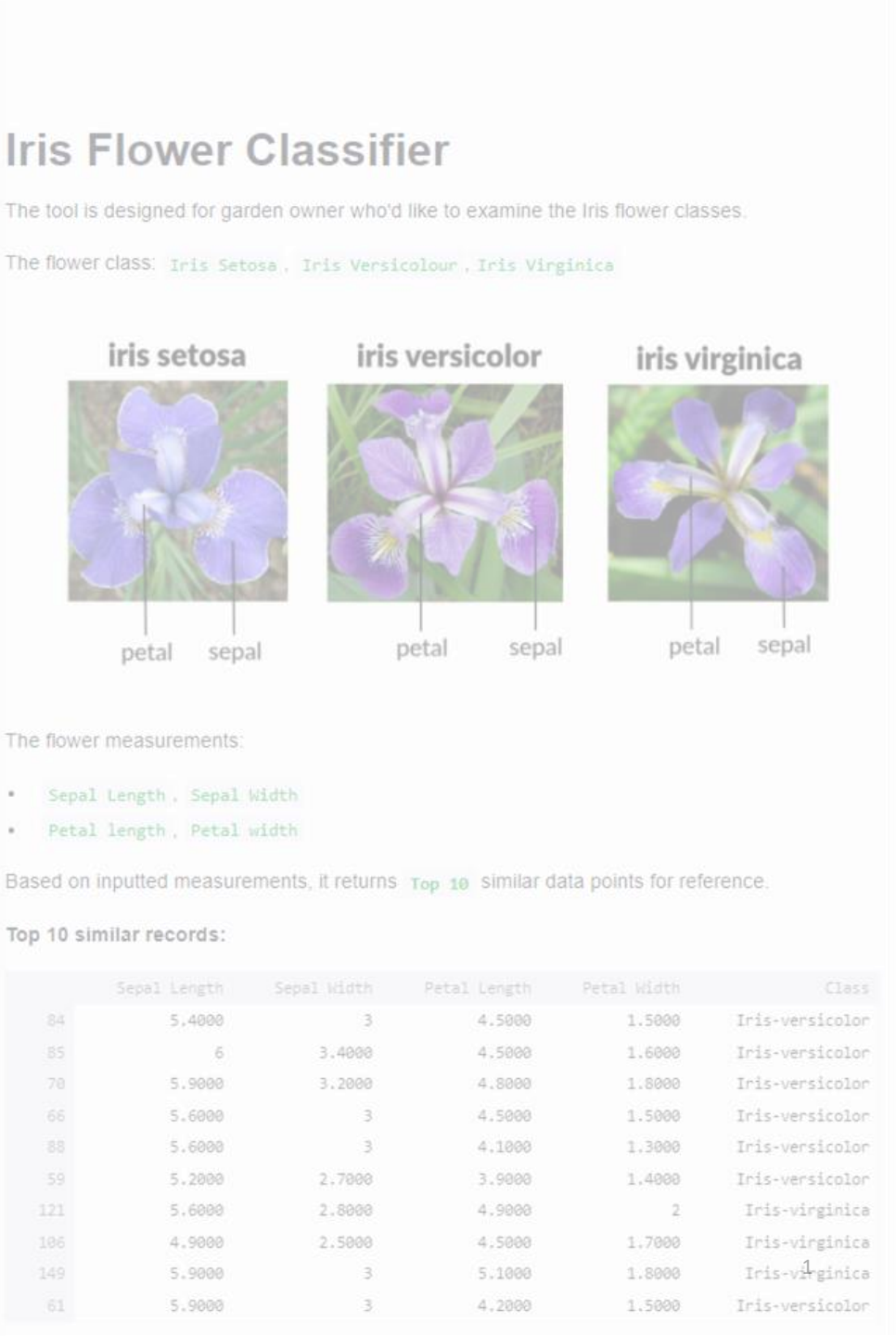
Iris Flower Classifier App

The app allows garden owner examine Iris flower classes

@ Jason Zhang

- **Dataset**
- **Methodology**
- **Results**
- **Next Steps**





	Sepal Length	Sepal Width	Petal Length	Petal Width	Class
84	5.4000	3	4.5000	1.5000	Iris-versicolor
85	6	3.4000	4.5000	1.6000	Iris-versicolor
70	5.9000	3.2000	4.8000	1.8000	Iris-versicolor
66	5.6000	3	4.5000	1.5000	Iris-versicolor
88	5.6000	3	4.1000	1.3000	Iris-versicolor
59	5.2000	2.7000	3.9000	1.4000	Iris-versicolor
121	5.6000	2.8000	4.9000	2	Iris-virginica
106	4.9000	2.5000	4.5000	1.7000	Iris-virginica
149	5.9000	3	5.1000	1.8000	Iris-virginica
61	5.9000	3	4.2000	1.5000	Iris-versicolor

Click Here or Scan Me

SCAN ME

Building a better working world

DATASET

Iris Dataset Assessment: good quality dataset with well-balanced labels

```
from urllib.request import urlretrieve
import pandas as pd
import numpy as np

iris = 'http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'
urlretrieve(iris)

df = pd.read_csv(iris, sep=',', header = None, names=['Sepal Length', 'Sepal Width', 'Petal Length', 'Petal Width', 'Class'])
```

```
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Sepal Length  150 non-null    float64
 1   Sepal Width   150 non-null    float64
 2   Petal Length  150 non-null    float64
 3   Petal Width   150 non-null    float64
 4   Class         150 non-null    object
dtypes: float64(4), object(1)
```

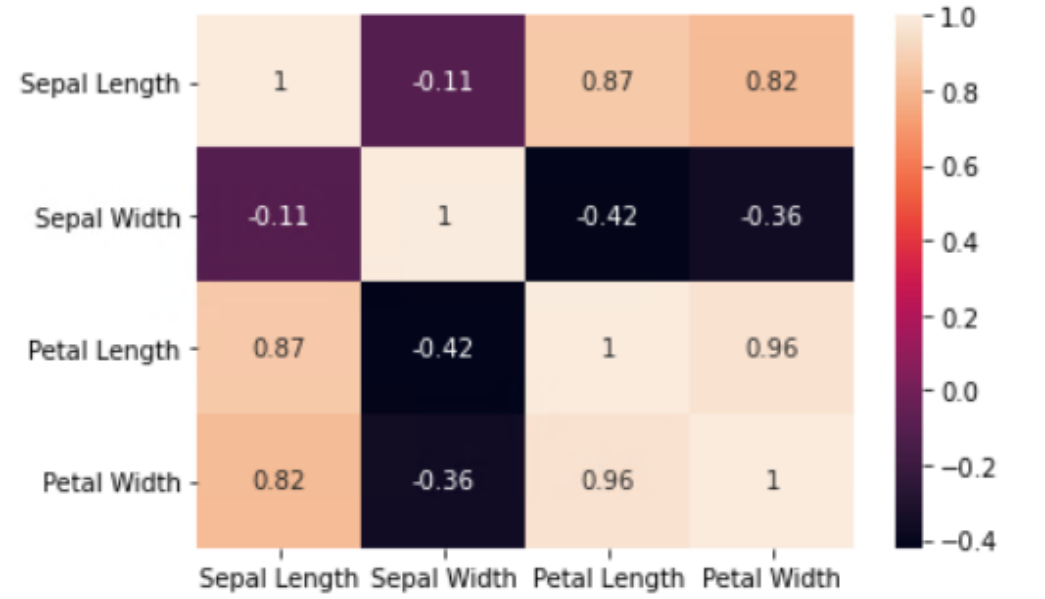
	Sepal Length	Sepal Width	Petal Length	Petal Width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

1. The dataset shape is 150(rows) *5(columns). No null values in the dataset.
2. Class Type is well balanced. Each has 50 samples.
3. Different Measurement has different scales. Sepal length ranges from 4.3 to 7.9, while Petal Width ranges from 0.1 to 2.5.

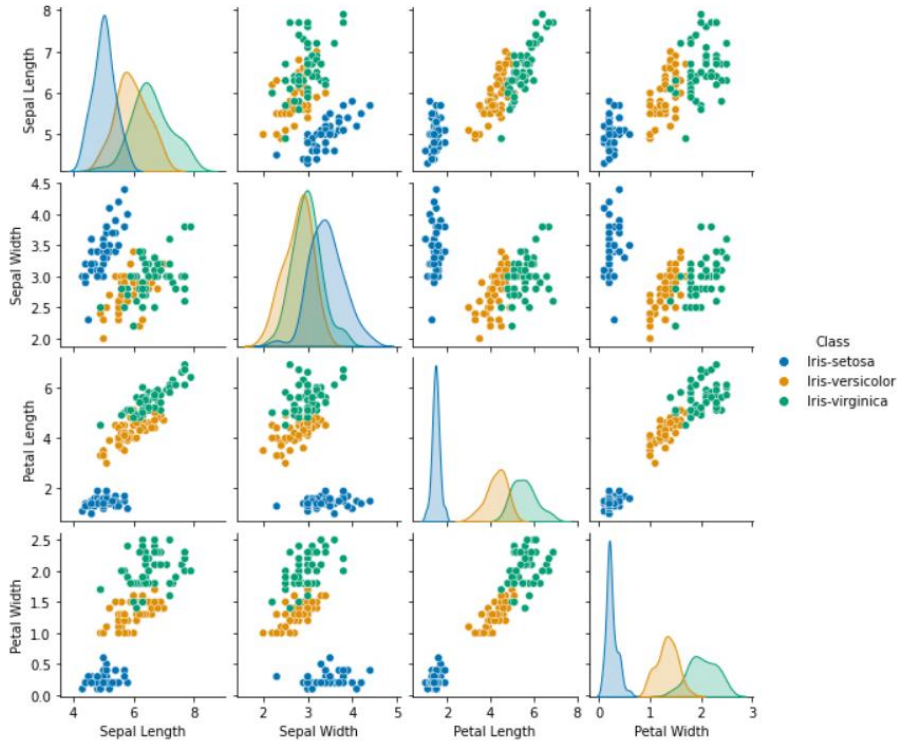
DATASET

EDA: Relationship Between Sepal and Petal Measurements

Correlation Heatmap



Pair Plot

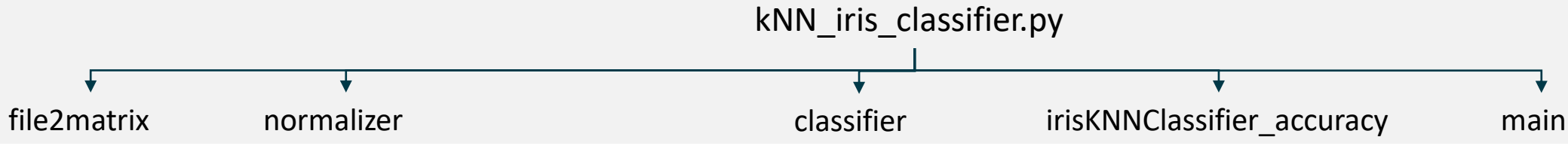


1. Sepal width has negative correlation with petal length and width, while sepal length shows positive relationship with petal measures.
2. Iris-Setosa has much smaller petal length/width, which means petal length/width can be a good indicator of Iris-Setosa.
3. Iris-Versicolor tends to have smaller septal length/width and petal length/width compared to Iris-virginica.

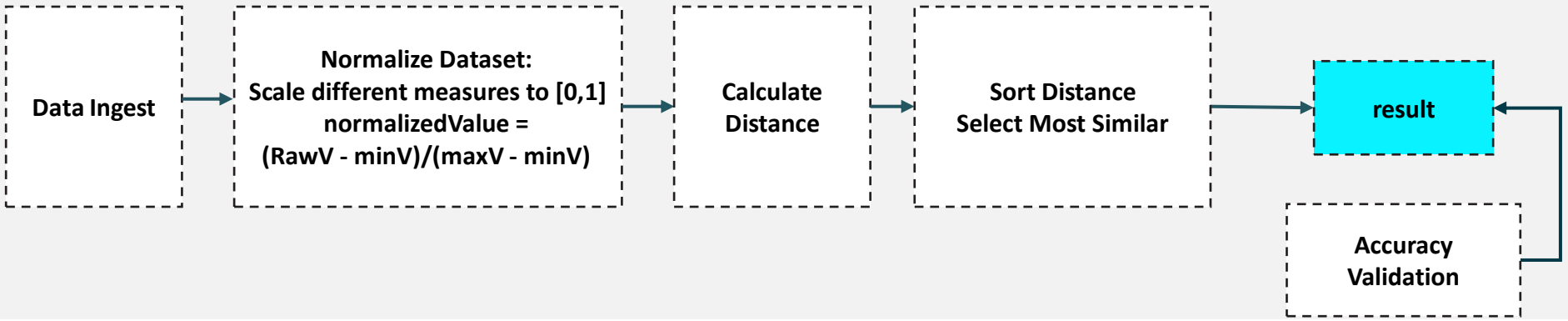
METHODOLOGY

KNN with Normalized Euclidean Distance to classify the Iris flower class

Functions



Process Flow



Viz

Sepal Length	Sepal Width	Petal Length	Petal Width	Class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa

0.222222	0.625	0.0677966	0.0416667
0.166667	0.416667	0.0677966	0.0416667
0.111111	0.5	0.0508475	0.0416667
0.0833333	0.458333	0.0847458	0.0416667
0.194444	0.666667	0.0677966	0.0416667
0.305556	0.791667	0.118644	0.125

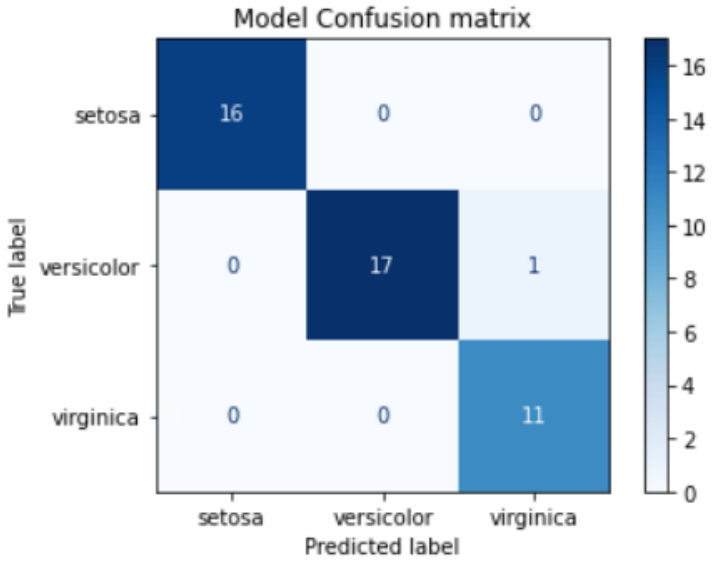
New Data Points

								Distance	k=2	
				0.242	0.614	0.071	0.042			
0.222	0.625	0.068	0.042					0.02286	1	Iris-setosa
0.167	0.417	0.068	0.042					0.21125	0	
0.111	0.500	0.051	0.042					0.17474	0	
0.083	0.458	0.085	0.042					0.22270	0	
0.194	0.667	0.068	0.042					0.07103	1	Iris-setosa
0.306	0.792	0.119	0.125					0.21157	0	
...	...	...	...					...	...	...

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

RESULTS

Overall Accuracy : 97.8% Accuracy (70% Train / 30% Test, KNN n_neighors = 10)



Based on inputted measurements, it returns `Top 10` similar data points for reference.

Top 10 similar records:

	Sepal Length	Sepal Width	Petal Length	Petal Width	Class
56	6.3000	3.3000	4.7000	1.6000	Iris-versicolor
91	6.1000	3	4.6000	1.4000	Iris-versicolor
51	6.4000	3.2000	4.5000	1.5000	Iris-versicolor
85	6	3.4000	4.5000	1.6000	Iris-versicolor
70	5.9000	3.2000	4.8000	1.8000	Iris-versicolor
63	6.1000	2.9000	4.7000	1.4000	Iris-versicolor
127	6.1000	3	4.9000	1.8000	Iris-virginica
78	6	2.9000	4.5000	1.5000	Iris-versicolor
138	6	3	4.8000	1.8000	Iris-virginica
149	5.9000	3	5.1000	1.8000	Iris-virginica

1. 1/150 incorrect prediction (versicolor being predicted as virginica).
2. User can input the iris measurements by sliding the bar. It returns top 10 most similar records based on normalized Euclidean distance.

Next Steps

Suggestions to improve the classification accuracy

1. Normalize the dataset based on StandardScaler, instead of MinMaxScaler.
- Standard scaler removes the mean and scaling to unit variance.
- Min/Max scaler scales each feature to range (0,1)

2. Classify classes based on other values, such as 5 or 7, instead of using 10 nearest neighbors.

3. Collect more data samples to increase accuracy. Current dataset contains only 150 samples.

4. Explore other models, such as decision tree classifier / random forest / support vector classifier etc.