PLANTIS EX MACHINA: PREDICTING PLANT STRESS WITH MACHINE LEARNING

FALL 2025
DATA 602
Maxfield Raynolds

Abstract

This study investigates environmental and nutrient factors that contribute to plant stress and develops predictive models to classify stress level based on these conditions. Effective management of plant stress is essential for growers, and identifying which factors most strongly influence stress can support more consistent plant health.

Ten plants were monitored every six hours. At each interval, eleven environmental and nutrient features were recorded, and a qualitative stress rating was assigned as Healthy, Moderate Stress, or High Stress. Analysis revealed an association between stress level and two key features: soil moisture and nitrogen level.

Several machine learning models were evaluated for their ability to predict stress level from these features. Logistic Regression achieved 76% accuracy, Support Vector Machines achieved 98% accuracy, and K-Nearest Neighbors reached 97%.

While these results demonstrate that plant stress can be predicted reliably within this controlled dataset, their generalizability is limited to this specific plant variety and environmental context.

Agenda

- Overview
- Data Source
- Data Wrangling
- Exploratory Data Analysis
- Data Analysis
  - *Logistic Regression with Recursive Feature Elimination*
  - *Support Vector Machines*
  - *K-Nearest Neighbors*
- Conclusions

Overview

- Modern Agriculture
  - *Data intensive growing of plants*
    - Environmental sensors and testing
      - *Temperature, humidity, nutrient levels, etc.*
- How to make that data useful?
  - *Identify critical metrics*
  - *Predict plant health status based on measured metrics*
  - *Anticipate what will cause plant stress*

Data Source

- Dataset of Plant-Health-Data from:
  - *https://www.kaggle.com/datasets/ziya07/plant-health-data*
- 10 plants
  - *Monitored every 6 hours for 30 days*
  - *11 features (environmental and plant nutrition based)*
  - *At every recorded event a qualitative stress level was assigned:*
    - Healthy
    - Moderate Stress
    - High Stress

Data Wrangling

A preview of the data:

- A timestamp
- A plant id
- 11 numeric features/measurements
- A qualitative plant health status

	timestamp	plant_id	soil_moisture	ambient_temperature	soil_temperature	humidity	light_intensity	soil_ph	nitrogen_level	phosphorus_level	potassium_level	chlorophyll_content	electrochemical_signal	plant_health_status
0	2024-10-03 10:54:53.407995	1	27.521109	22.240245	21.900435	55.291904	556.172805	5.581955	10.003650	45.806852	39.076199	35.703006	0.941402	High Stress
1	2024-10-03 16:54:53.407995	1	14.835566	21.706763	18.680892	63.949181	596.136721	7.135705	30.712562	25.394393	17.944826	27.993296	0.164899	High Stress
2	2024-10-03 22:54:53.407995	1	17.086362	21.180946	15.392939	67.837956	591.124627	5.656852	29.337002	27.573892	35.706530	43.646308	1.081728	High Stress
3	2024-10-04 04:54:53.407995	1	15.336156	22.593302	22.778394	58.190811	241.412476	5.584523	16.966621	26.180705	26.257746	37.838095	1.186088	High Stress
4	2024-10-04 10:54:53.407995	1	39.822216	28.929001	18.100937	63.772036	444.493830	5.919707	10.944961	37.898907	37.654483	48.265812	1.609805	High Stress

```
# View the first five lines of the dataset
df.columns = df.columns.str.lower()
df.head()
[134]
```

Data Wrangling

- 10 plants
  - *120 entries for each plant*

```
df.groupby('plant_id').size()
✓ [12] 28ms

plant_id
1     120
2     120
3     120
4     120
5     120
6     120
7     120
8     120
9     120
10    120
dtype: int64
```

Data Wrangling

- Summary statistics of the numeric features
  - *Very different ranges*
  - *Important to scale data when applying certain models*

```
df.drop(columns=['plant_id']).describe()
[137]
```

	soil_moisture	ambient_temperature	soil_temperature	humidity	light_intensity	soil_ph	nitrogen_level	phosphorus_level	potassium_level	chlorophyll_content	electrochemical_signal
count	1200.000000	1200.000000	1200.000000	1200.000000	1200.000000	1200.000000	1200.000000	1200.000000	1200.000000	1200.000000	1200.000000
mean	25.106918	23.999130	19.957794	54.853165	612.637265	6.524102	30.106751	30.264484	30.112088	34.749591	0.987764
std	8.677725	3.441561	2.932073	8.784916	228.318853	0.581755	11.514396	11.466846	11.668085	8.766995	0.575116
min	10.000724	18.001993	15.003710	40.028758	200.615482	5.507392	10.003650	10.017690	10.000606	20.025511	0.002376
25%	17.131893	21.101766	17.353027	47.019694	416.878983	6.026042	20.249774	20.894445	19.585561	27.463350	0.487982
50%	25.168333	23.889044	19.911473	54.692069	617.240221	6.540524	30.138590	30.019385	30.495054	34.433427	0.981647
75%	32.370231	27.042634	22.596851	62.451053	811.474690	7.030039	40.184737	40.131459	40.108296	42.232637	1.473142
max	39.993164	29.990886	24.995929	69.968871	999.856262	7.497823	49.951136	49.980700	49.981945	49.990811	1.996116

Data Wrangling

- No missing data
- Data types are appropriate for analysis

```
# Inspect dataset for missing values
df.isna().any()
[149]
```

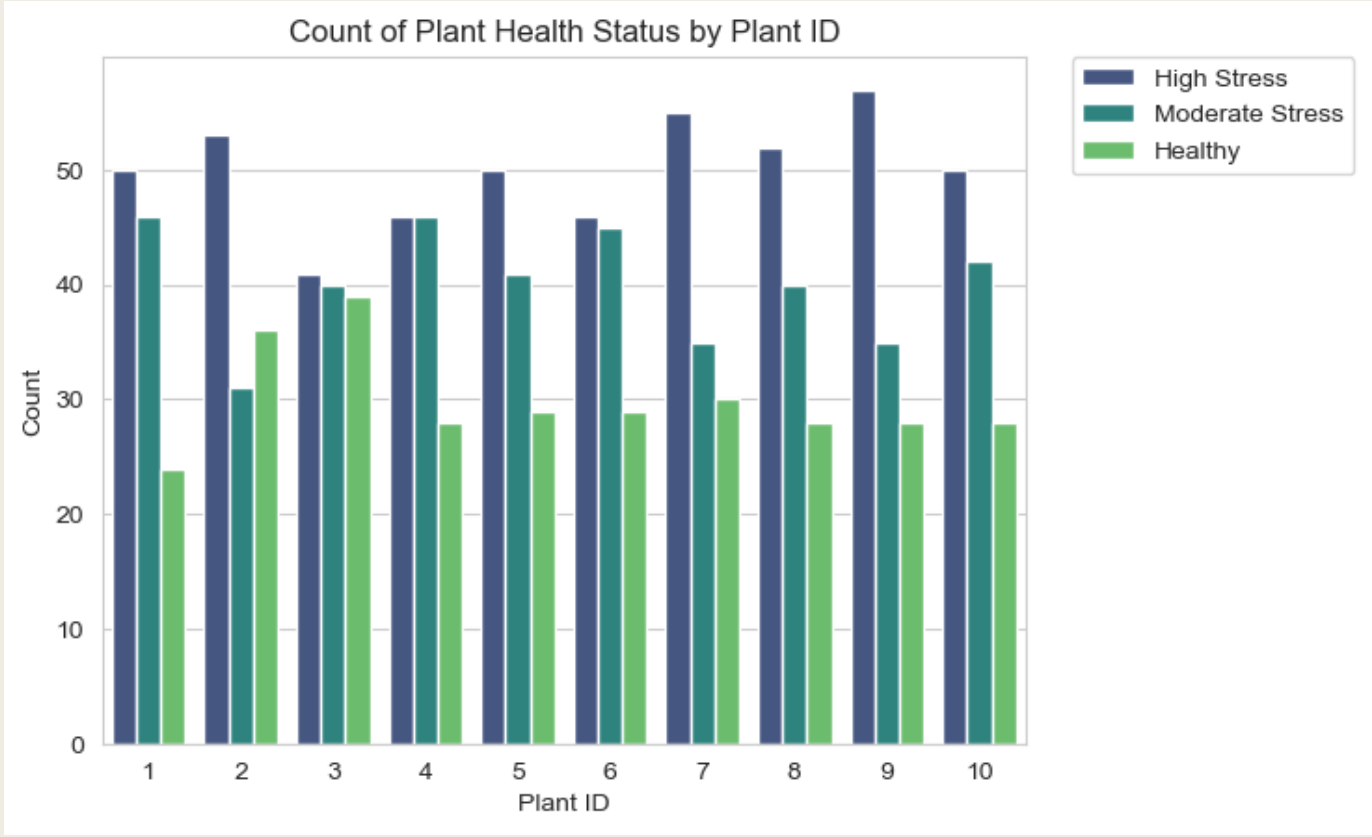
```
timestamp                False
plant_id                 False
soil_moisture            False
ambient_temperature      False
soil_temperature         False
humidity                 False
light_intensity          False
soil_ph                  False
nitrogen_level           False
phosphorus_level         False
potassium_level          False
chlorophyll_content      False
electrochemical_signal   False
plant_health_status      False
stress_encoded           False
dtype: bool
```

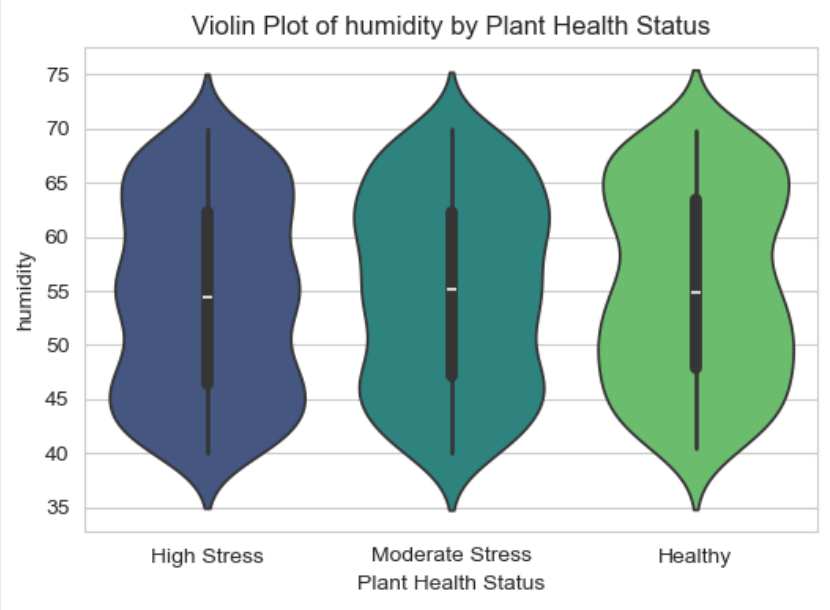
```
# Inspect data types
df.dtypes
[150]
```

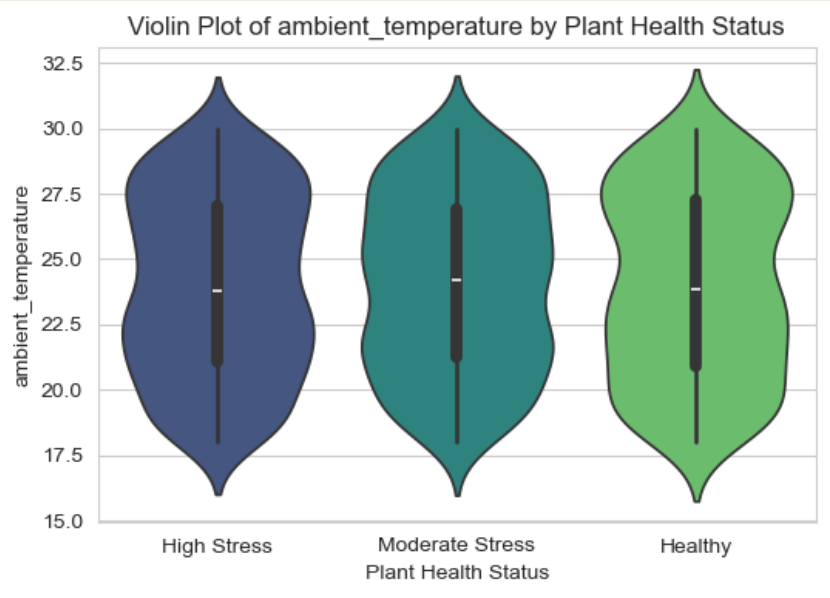
```
timestamp                datetime64[ns]
plant_id                         int64
soil_moisture                  float64
ambient_temperature            float64
soil_temperature               float64
humidity                       float64
light_intensity                float64
soil_ph                        float64
nitrogen_level                 float64
phosphorus_level               float64
potassium_level                float64
chlorophyll_content            float64
electrochemical_signal         float64
plant_health_status             object
stress_encoded                   int64
dtype: object
```

Exploratory Data Analysis

- Relatively even distributions of status assignments by plant
- More moderate and high stress statuses than healthy statuses

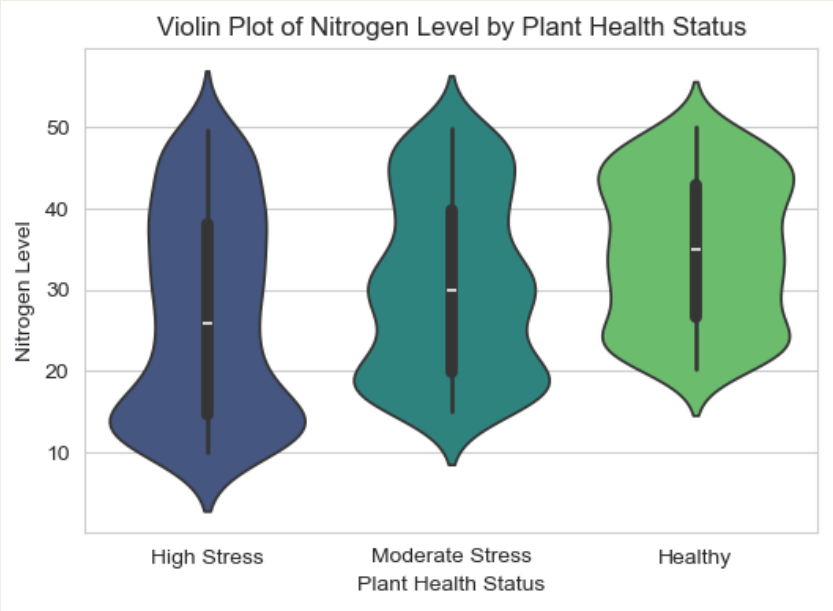


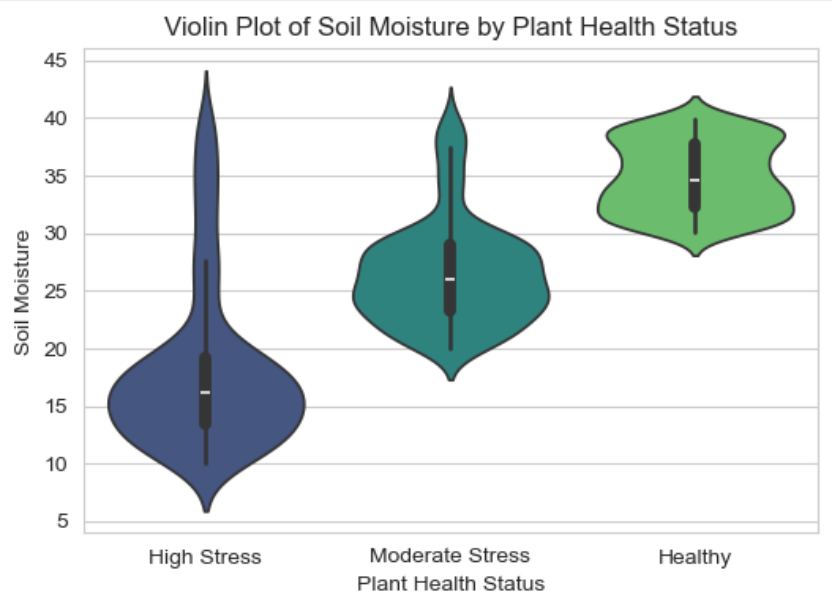




Exploratory Data Analysis

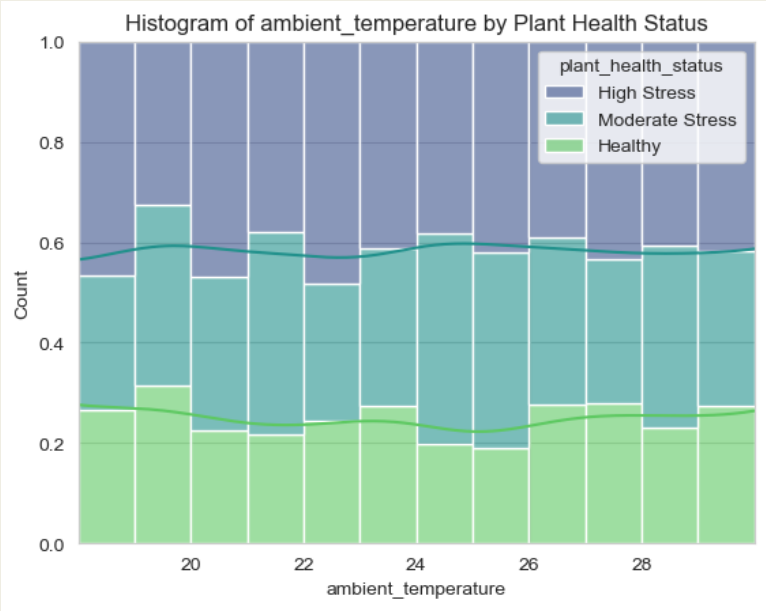
- Violin plots of features by plant health status
- Majority were non-distinct
- Shown are two examples of non-distinct plots, the violin plots look relatively event at all the plant health statuses

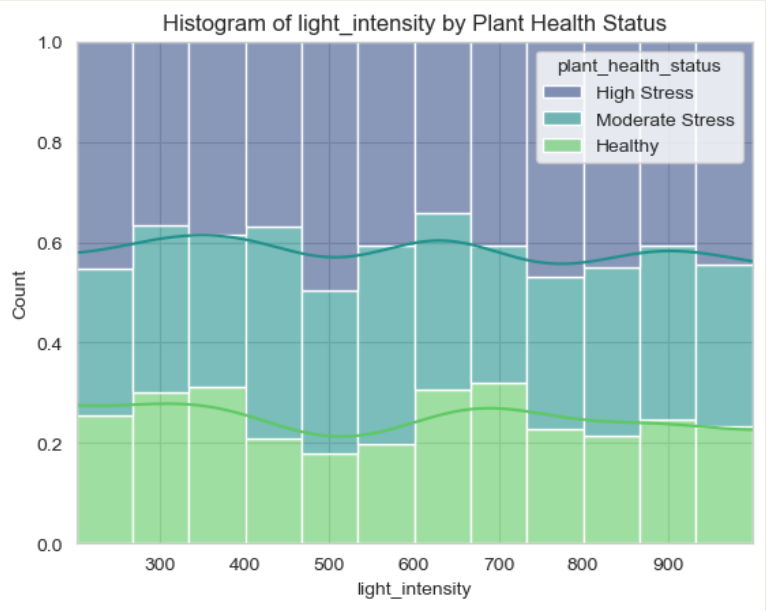




Exploratory Data Analysis

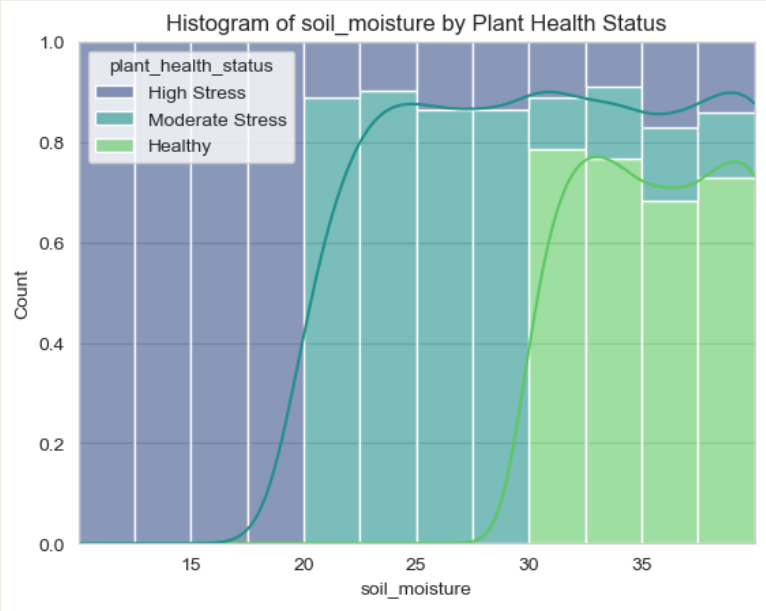
- Violin plots of features by plant health status
- Majority were non-distinct
- Two were variable:
  - *Nitrogen level*
  - *Soil moisture*

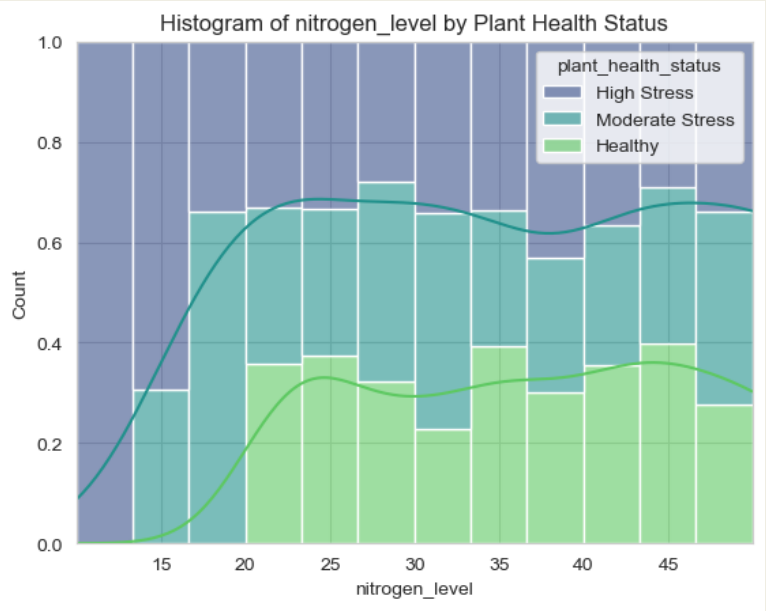




Exploratory Data Analysis

- Histogram plots of features by plant health status
- Similar to violin plots
- Majority were non-distinct
- Shown are two examples of non-distinct plots, the violin plots look relatively event at all the plant health statuses



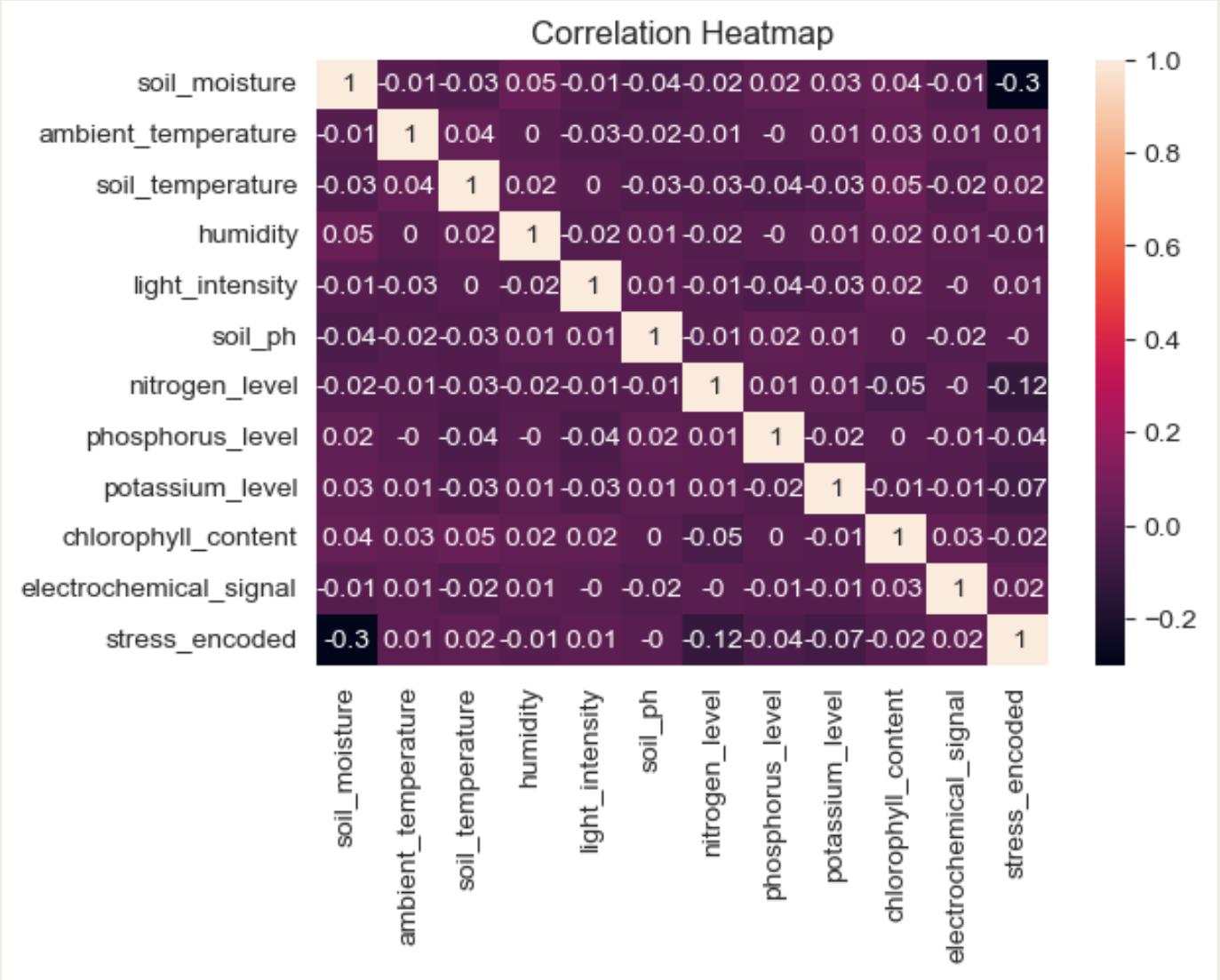


Exploratory Data Analysis

- Histogram plots of features by plant health status
- Similar to violin plots
- Majority were non-distinct
- Two were variable:
  - *Nitrogen level*
  - *Soil moisture*

Data Analysis

- Correlation heatmap confirms violin plots and histogram indications
- Largest correlation values with health status:
  - *Soil moisture -0.3*
  - *Nitrogen level -0.12*
- Both negatively correlated



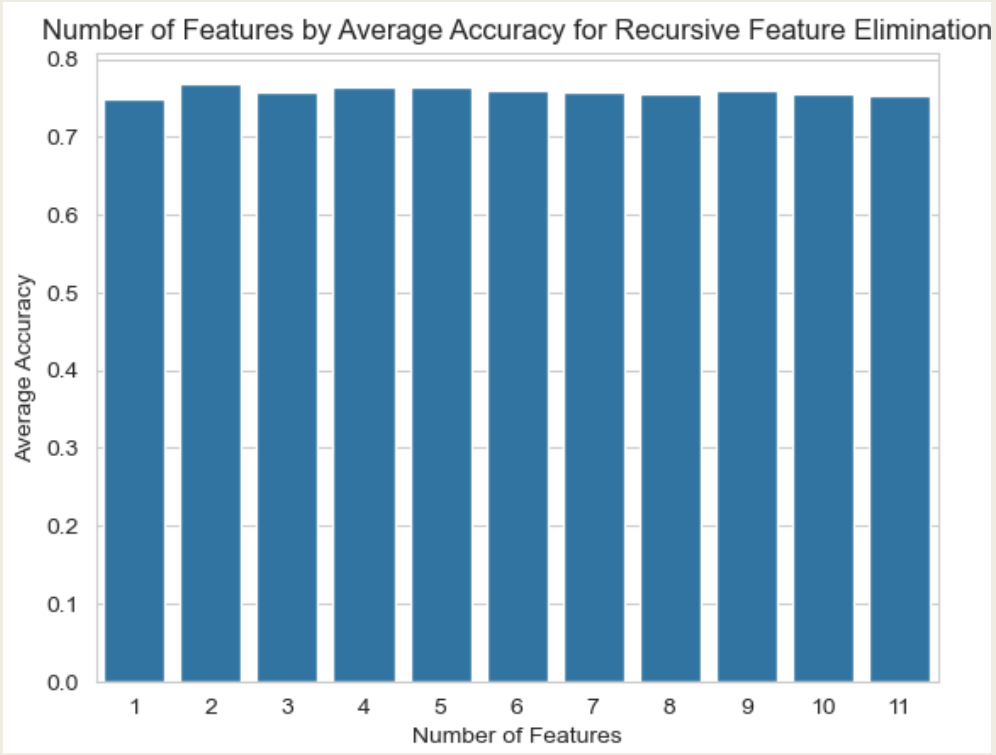
Data Analysis

- Confirm correlations are statistically significant by checking the p-values
- p-values below 0.05 are statistically significant
- Three below this threshold:
  - *Soil moisture*
    - p-value = 0.000
    - Correlation = -0.3
  - *Nitrogen level*
    - p-value = 0.000
    - Correlation = -0.12
  - *Phosphorus level*
    - p-value = 0.24
    - Correlation = -0.04

```
Optimization terminated successfully.
         Current function value: 0.532071
         Iterations: 44
         Function evaluations: 51
         Gradient evaluations: 51
                             OrderedModel Results
==============================================================================
Dep. Variable:     plant_health_status   Log-Likelihood:                -478.86
Model:                    OrderedModel   AIC:                             983.7
Method:            Maximum Likelihood   BIC:                             1046.
Date:                Wed, 10 Dec 2025
Time:                        21:28:17
No. Observations:                 900
Df Residuals:                     887
Df Model:                          11
==============================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
soil_moisture                -0.3781      0.019    -19.988      0.000      -0.415      -0.341
ambient_temperature          -0.0359      0.025     -1.431      0.152      -0.085       0.013
soil_temperature             -0.0439      0.028     -1.540      0.124      -0.100       0.012
humidity                     -0.0044      0.010     -0.461      0.645      -0.023       0.014
light_intensity           3.795e-05      0.000      0.102      0.919      -0.001       0.001
soil_ph                      0.0427      0.145      0.295      0.768      -0.241       0.327
nitrogen_level               -0.1103      0.009    -12.165      0.000      -0.128      -0.093
phosphorus_level             -0.0168      0.007     -2.260      0.024      -0.031      -0.002
potassium_level              0.0052      0.007      0.720      0.471      -0.009       0.019
chlorophyll_content          0.0048      0.010      0.486      0.627      -0.014       0.024
electrochemical_signal       -0.0325      0.145     -0.224      0.823      -0.317       0.252
Healthy/Moderate Stress     -17.5903      1.646    -10.684      0.000     -20.817     -14.363
Moderate Stress/High Stress   1.3619      0.056     24.218      0.000       1.252       1.472
==============================================================================
```

Data Analysis

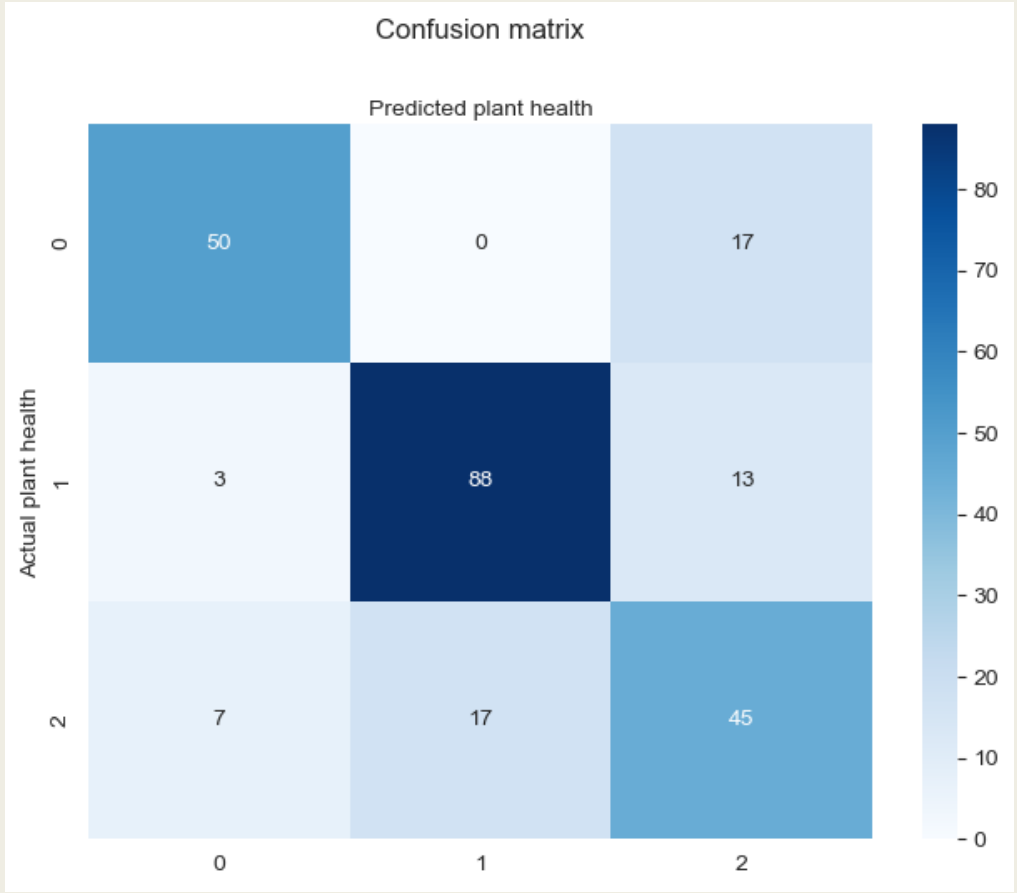
- Recursive Feature Elimination
  - *Used logistic regression as model*
  - *A "for loop" to determine the best number of features to use*
- Selected two features
  - *Soil moisture*
  - *Nitrogen level*
- Accuracy = 76.9%
- Same features identified during exploratory data analysis



	n	mean	selected_features
1	2	0.769167	[soil_moisture, nitrogen_level]
3	4	0.764167	[soil_moisture, soil_temperature, nitrogen_lev...
4	5	0.763333	[soil_moisture, ambient_temperature, soil_temp...
5	6	0.760000	[soil_moisture, ambient_temperature, soil_temp...
8	9	0.759167	[soil_moisture, ambient_temperature, soil_temp...
2	3	0.758333	[soil_moisture, soil_temperature, nitrogen_level]
6	7	0.756667	[soil_moisture, ambient_temperature, soil_temp...
9	10	0.755000	[soil_moisture, ambient_temperature, soil_temp...
7	8	0.755000	[soil_moisture, ambient_temperature, soil_temp...
10	11	0.753333	[soil_moisture, ambient_temperature, soil_temp...
0	1	0.748333	[soil_moisture]

Logistic Regression

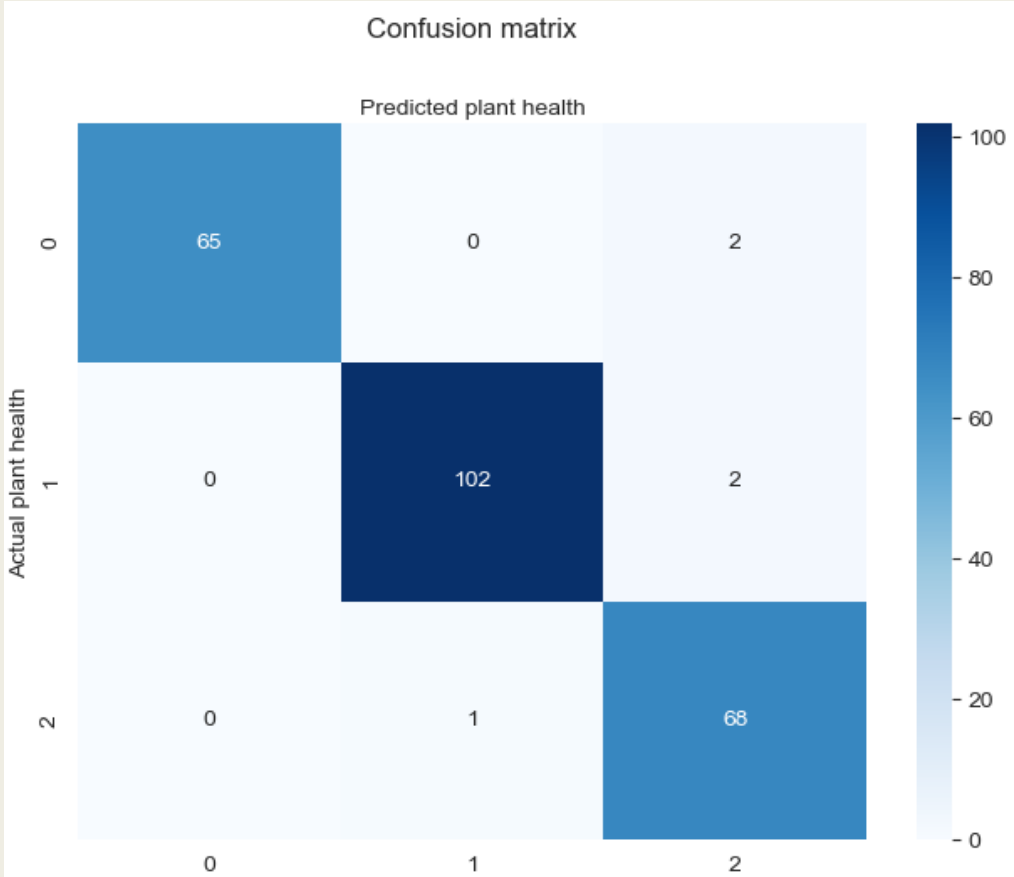
- Logistic Regression
  - *Cross validation*
  - *Scaled features*
  - *Using two selected features*
    - Soil moisture
    - Nitrogen level
  - *To confirm the results of the RFE*
- Accuracy = 76.9%
  - *Confirms the RFE*
- Model performance is not exceptional
  - *Struggles with High Stress plants*
- Confusion Matrix:
  - *0 = Healthy*
  - *1 = Moderate Stress*
  - *2 = High Stress*



	precision	recall	f1-score	support
Healthy	0.83	0.75	0.79	67
Moderate_stress	0.84	0.85	0.84	104
High_stress	0.60	0.65	0.62	69
accuracy			0.76	240
macro avg	0.76	0.75	0.75	240
weighted avg	0.77	0.76	0.76	240

Support Vector Machines (SVM)

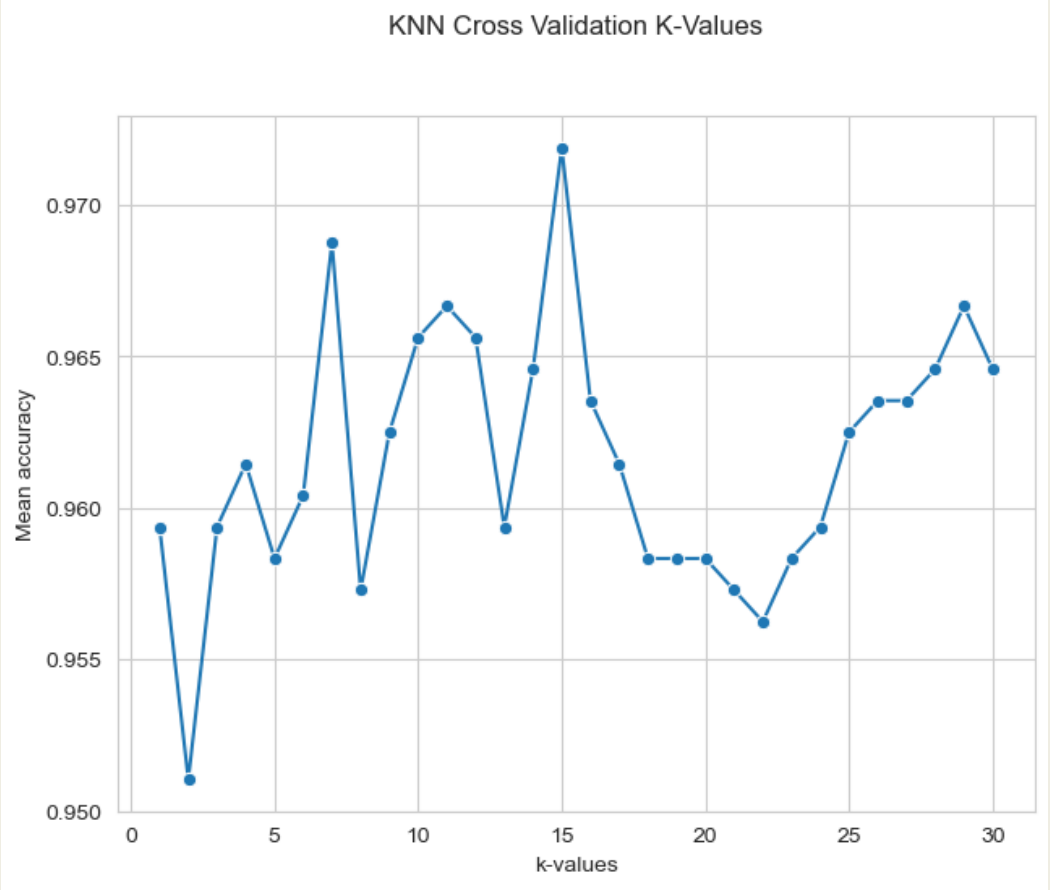
- SVM
  - *Scaled features*
  - *Using two selected features*
    - Soil moisture
    - Nitrogen level
- Accuracy = 98%
- Model performance is exceptional
  - *Biggest error was false positives of High Stress*
- Confusion Matrix:
  - *0 = Healthy*
  - *1 = Moderate Stress*
  - *2 = High Stress*



	precision	recall	f1-score	support
Healthy	1.00	0.97	0.98	67
Moderate_stress	0.99	0.98	0.99	104
High_stress	0.94	0.99	0.96	69
accuracy			0.98	240
macro avg	0.98	0.98	0.98	240
weighted avg	0.98	0.98	0.98	240

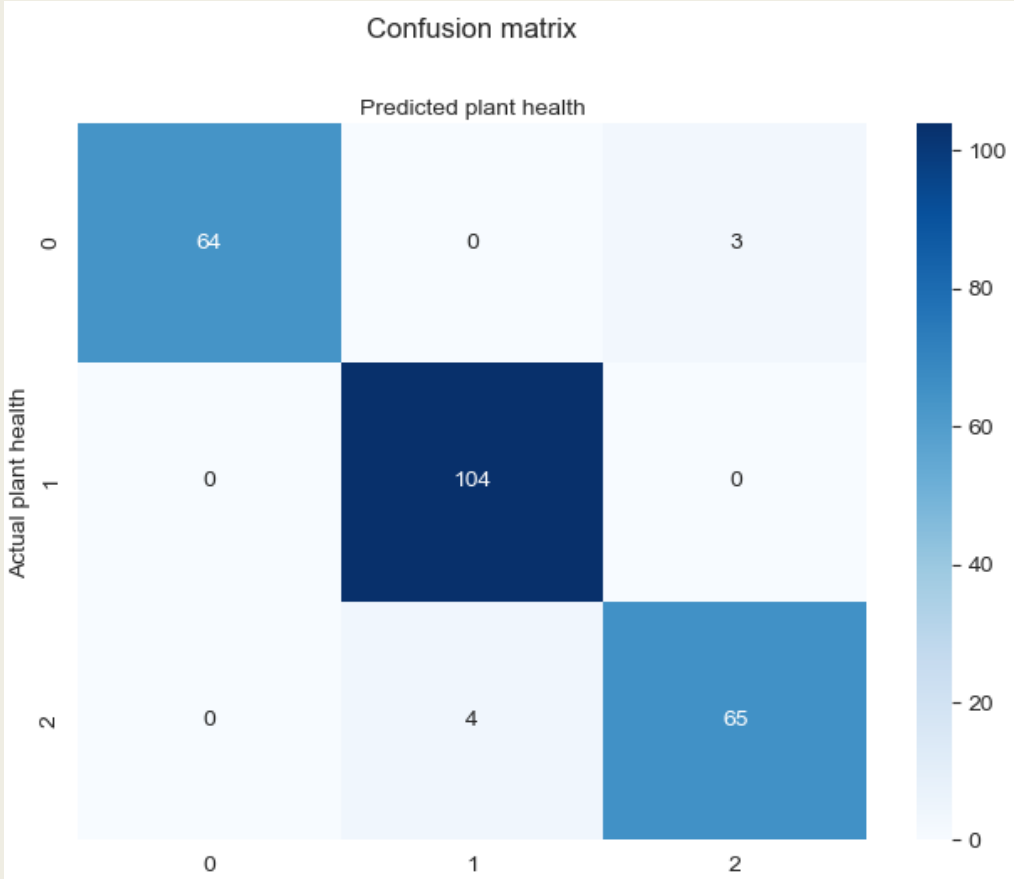
K-Nearest Neighbors (KNN)

- KNN
  - *Scaled features*
  - *Cross validation*
  - *Selected k-value:*
    - k-value = 15
  - *Using two selected features*
    - Soil moisture
    - Nitrogen level



K-Nearest Neighbors (KNN)

- KNN
  - *Scaled features*
  - *Cross validation*
  - *For loop to select k-value:*
    - Selected k-value = 15
  - *Using two selected features*
    - Soil moisture
    - Nitrogen level
- Accuracy = 97%
- Model performance is exceptional
  - *Biggest errors:*
    - False positives of High Stress when actually Healthy (3)
    - False positive of Moderate Stress when actually High Stress (4)
- Confusion Matrix:
  - *0 = Healthy*
  - *1 = Moderate Stress*
  - *2 = High Stress*



	precision	recall	f1-score	support
Healthy	1.00	0.96	0.98	67
Moderate_stress	0.96	1.00	0.98	104
High_stress	0.96	0.94	0.95	69
accuracy			0.97	240
macro avg	0.97	0.97	0.97	240
weighted avg	0.97	0.97	0.97	240

Conclusion

- Within this dataset:
  - *From a range of environmental measurements -> identified two critical features*
  - *Confirmed the two features were critical using Recursive Factor Elimination*
  - *Instantiated three models to predict health status*
    - Logistic Regression -> 76% accurate
    - Support Vector Machines -> 98% accurate
    - K-Nearest Neighbors -> 97% accurate
  - *Excellent model performances*
  - *These results are not likely universally applicable, useful for:*
    - This specific data
    - This specific plant
    - This specific growing context
  - *Potentially this process would be useful to apply to other contexts*