Predicting houseplant lifespan on relative indoor air humidity exploiting linear and polynomial regression

Introduction

Houseplants have different needs regarding the relative air humidity. Especially *monstera deliciosa* (monstera) is well-known to thrive in high humidity levels. This indicates that its lifespan might be shorter in spaces where the relative humidity is lower. Young and therefore small monsteras usually cost around 15 € in Finland but a larger 120 cm high monstera costs 100 € in store. [1] Judging by this, the larger the monstera gets, the higher its price becomes. Many people enjoy the look of a larger monstera but are not willing to pay 100 € or more for a plant and therefore the reasonable choice would be to buy a cheap young plant and wait for it to grow.

However, several factors may cause the premature death of a monstera. For this reason, it would be interesting to know what effect the relative air humidity has on the lifespan of a monstera when other care factors, such as light, soil and watering conditions, are comparable. If the lifespan of a monstera could be predicted on relative air humidity, it would help plant-owners to optimise indoor conditions to encourage the growth of the monstera.

This report first discusses the problem formulation where the feature and label are explained and the two separate datasets are described. After the problem formulation, the report explains the methods which were implemented on the data and why were these chosen. In the results chapter it is determined which hypothesis space is the most suitable and accurate and its performance is tested and evaluated using a test set and then the conclusions chapter the methods' performances are evaluated and concluded upon.

Problem formulation

The case presented in the introduction can be defined as a machine-learning problem. The data points in this problem represent different monstera plants that grow otherwise in similar conditions. The varying condition between the data points is the relative indoor air humidity which, in this case, is the feature. This is the only feature which is used to characterise the data point. The label of this machine learning problem is the number of months that the monstera stays alive and grows in the house. These are visualised in table 1.

Datapoint	A monstera plant
Feature	Relative indoor air humidity (percentage)
Label	Lifespan of the monstera (months)

Table 1: Datapoint, feature and label of the machine learning problem.

The problem was chosen due to personal interest on the subject. Optimal humidity conditions for monstera have been observed. However, no dataset could be found for this problem and therefore the data is created synthetically to resemble a real-life dataset. For this dataset, it is assumed that

the higher the humidity levels are, the longer the lifespan of the monstera. The assumption is based on several sources that state that the growth of monstera is optimal in high humidity levels and it originates from tropical conditions. [2,3]

Method

The data used in this project consists of three datasets. The first dataset has 90 datapoints and is used as the training set. The second dataset consists of 50 different datapoints and it is used to validate the hypothesis space. The split between training and validation set is conducted by a single split. The third dataset consists of 50 datapoints and is used to test the chosen hypothesis map and evaluate its performance. All of these are visualised in table 2. These three datasets are fully separate and none of the datapoints used to train the functions were used in validating nor in testing. For each datapoint in all datasets both the feature (relative indoor air humidity) and the label (lifespan of monstera) are known.

	Amount of datapoints	Separation method
Training set	90	Single split
Validation set	50	Single split
Test set	50	Single split

Table 2: Details of the datasets.

Before implementing any machine learning methods on the data, the data was studied manually. This helps with defining the hypothesis space. Based on the scatterplot of the training set – visualised in image 1, next to it in image 2 is the scatterplot of the validation set – the lifespan of the monstera clearly correlates to the relative air humidity. The data seems to be nearly linear. However, the correlation could also be polynomial and therefore both linear regression and polynomial regression up to 4^{th} degree are implemented on this data. These two machine learning methods are then compared and their accuracy in predicting the lifespan is evaluated by exploiting the average squared error loss function.

The hypothesis space h(x) in linear regression follows linear hypothesis map $h(x) = w_0 + w_1x$. The hypothesis space h(x) in polynomial regression, in turn, follows the polynomial hypothesis map $h(x) = w_0 + w_1x + w_2x^2 + \dots + w_rx^r$ where the polynomial degree is r. The squared error loss function $L(x) = (y-h(x))^2$ describes the accuracy of the hypothesis space. In the function y is the actual value of the label and h(x) is the value generated by the hypothesis model. Generally, the smaller the average squared error loss is the better the hypothesis space. However, often functions might overfit on training set and produce a small training error but predict the labels poorly on validation set and produce a large validation error. [4] For this reason, we are looking for the smallest loss function value which is similar in both training and validation sets to avoid over- or underfitting.

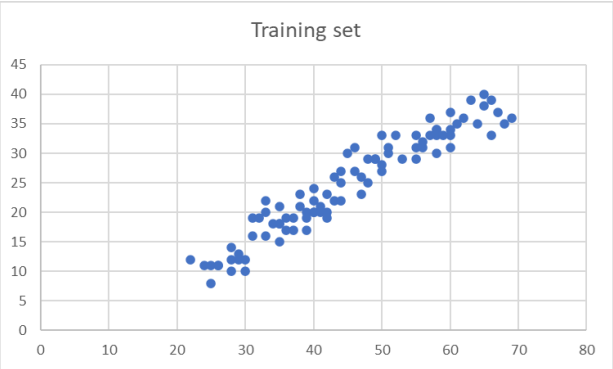

Image 1: Training set scatterplot

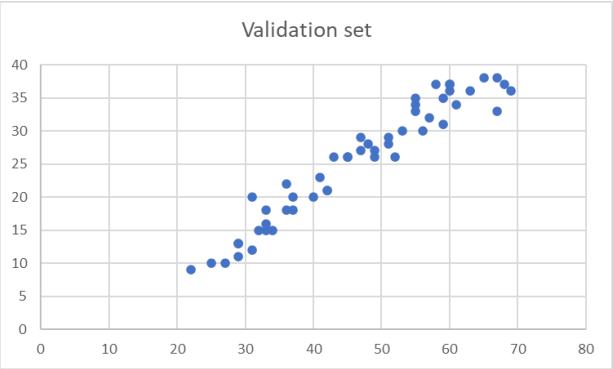

Image 2: Validation set scatterplot

Results

Linear regression and polynomial regression with degree r = 2, 3, 4 were implemented on both training and validation sets and the average squared error loss was calculated for each. All the average squared error losses are presented in the table below (Table 3).

	Linear regression, r = 1	Polynomial regression, r =2	Polynomial regression, r = 3	Polynomial regression, r = 4
Training error	5.868	5.419	27.08	32.08
Validation error	5.590	4.691	29.29	35.23

Table 3: Training and validation errors for each hypothesis map

The smallest training and validation errors are produced by polynomial regression with a maximum degree of two, r = 2. Training, validation, and test set with a fitted second-degree polynomial curve are visualised below (images 3, 4 and 5, respectively).

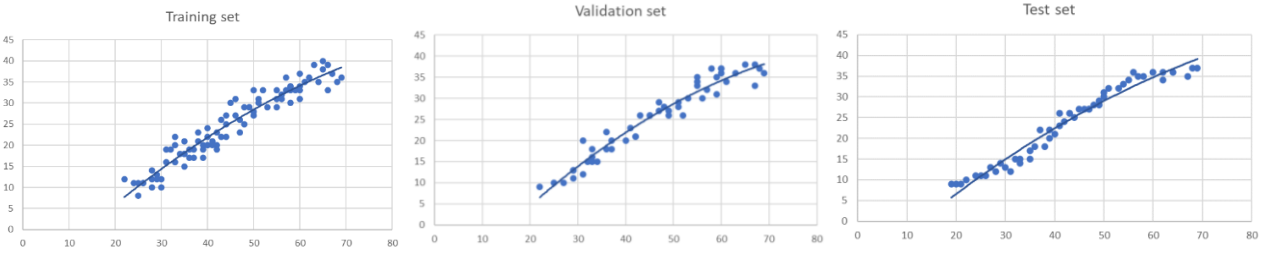

Image 3, 4 and 5: Training, validation, and test set, respectively, scatterplots with the second-degree polynomial curve.

Based on the training and validation errors, the best hypothesis map is the second-degree polynomial regression. The second-degree polynomial regression as the hypothesis map was tested

on a third separate data set called the test set. The test set consisted of 50 datapoints and the average squared error loss for second-degree polynomial regression equals 3.566.

Chosen hypothesis space	Polynomial regression, r = 2
Training error	5.419
Validation error	4.691
Test error	3.566

Table 4: average squared error of each dataset, second degree polynomial regression

Judging by the test error, the second-degree polynomial regression appears to be a reliable predictor for the lifespan of a monstera based on the relative air humidity. The average squared error differs only slightly between the training, validation, and test set.

Conclusions

This report has discussed four different models in predicting the lifespan of a monstera plant based on the relative indoor air humidity. Out of these models, one linear and three polynomials, the second-degree polynomial regression fitted the data best. It yielded both the smallest training error and validation error, 5.4 and 4.6, respectively. The model does not seem to overfit as the validation error and test error (3.6) were smaller than the training error. The data was split into training and validation set by a single split. The training set was larger than the validation set so more outliers occur there, hence the smaller validation error.

Overall, the second-degree polynomial regression predicts the lifespan of a monstera plant fairly accurately, however some improvements could be made. To produce an even more accurate hypothesis for the lifespan, more features such as soil quality, access to sunlight, and frequency of watering could be added. This could possibly yield an even more accurate prediction for the lifespan of a monstera.

References

1. https://www.plantagen.fi/peikonlehti.html
2. https://en.wikipedia.org/wiki/Monstera_deliciosa
3. https://herbaria.plants.ox.ac.uk/bol/plants400/Profiles/MN/Monstera
4. [MLBook] A. Jung, “Machine Learning. The Basics”, 2021, mlbook.cs.aalto.fi