Experiments over Apparent Forecast models

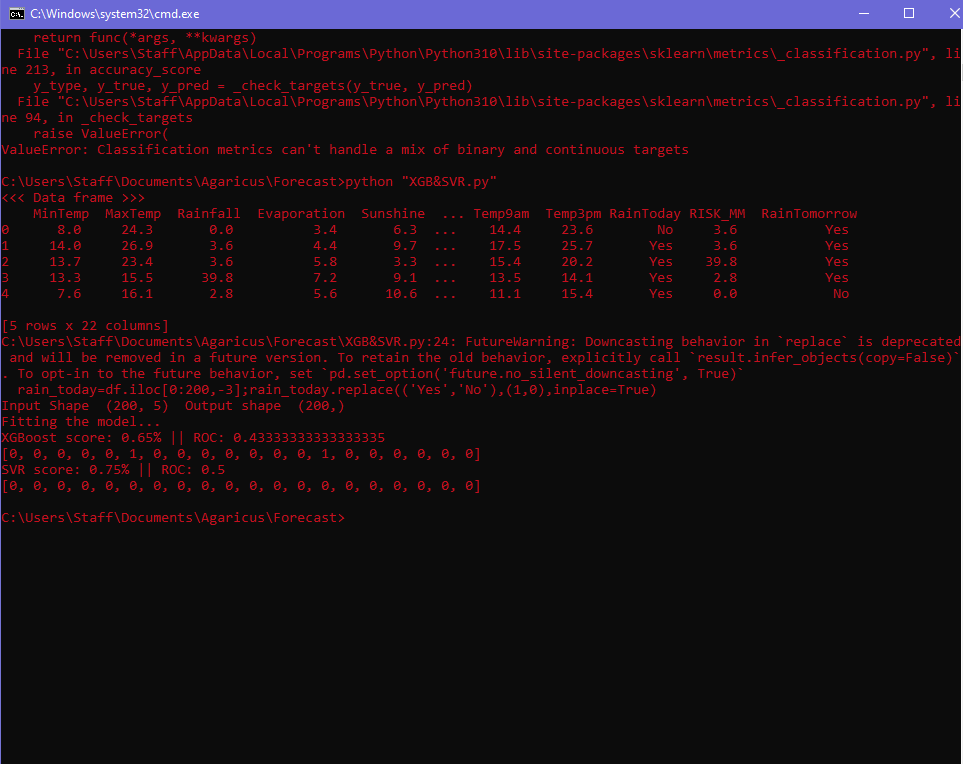
Forecasting mushroom production or growth is tricky considering that fact that space available to each mushroom to grow alongside its myco and surrounding environment play a momentous role in it. Generally traditional machine learning models work decent in time series data and for précised forecast more powerful algorithms like gaussian, differential propagation and deep learning are used. In this case, experiments started with simple traditional algorithms which are described and induced below.

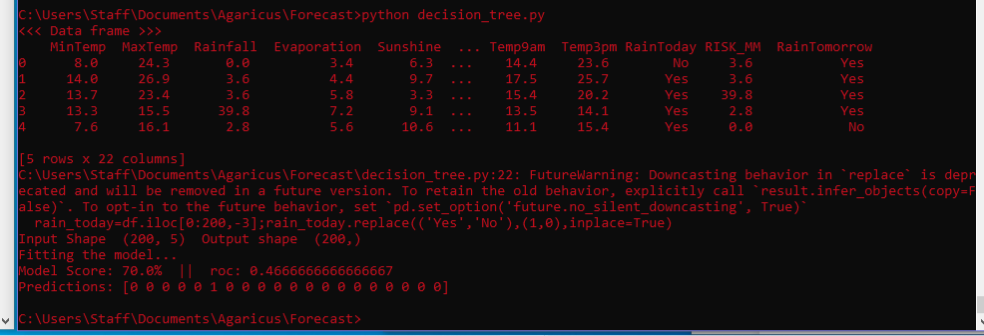
Some Traditional machine learning models previously tried:

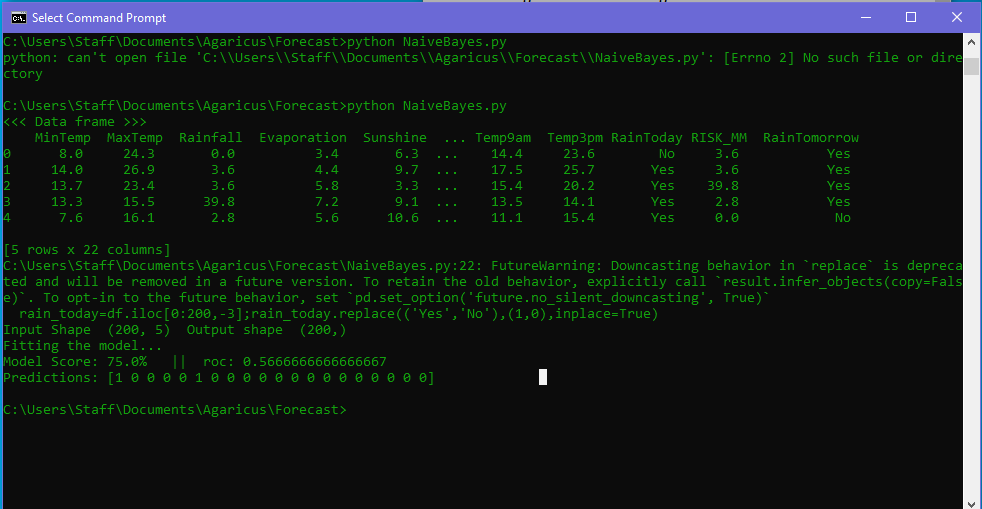
To find a suitable model first similar dataset was searched. A weather forecast dataset matched the requirement. The environment inside a mushroom tunnel has similar factors that resembles with weather, moreover the dataset is timeseries data and variance seems like that of the dataset can be fetched in mushroom tunnels.

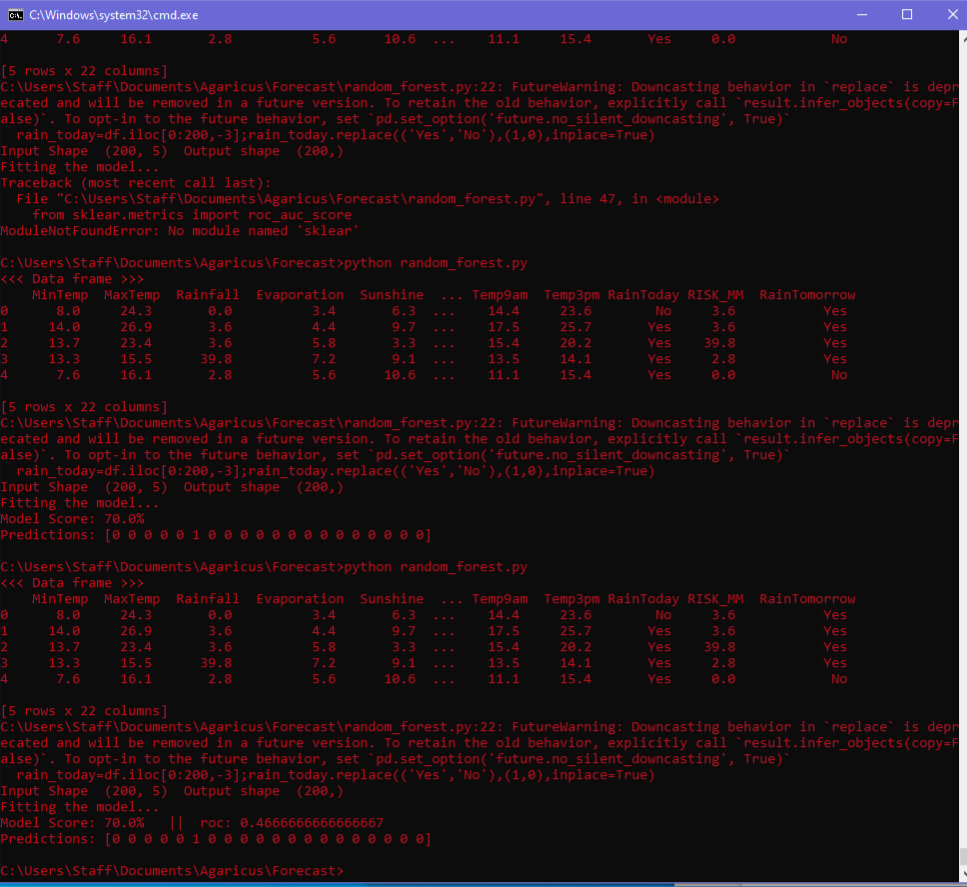
Several models were trained and evaluated over that dataset to see how better the models were performing. It seemed conventional models like Random Forest, Decision tree, Naive bayes, even boosting algorithms could not go beyond 75% accuracy. Output valuation in terms of model score, ROC and

Experiments over models with weather dataset

Img.1 XGBoost

Img.2 Decision Tree

Img.3 Naive Bayes

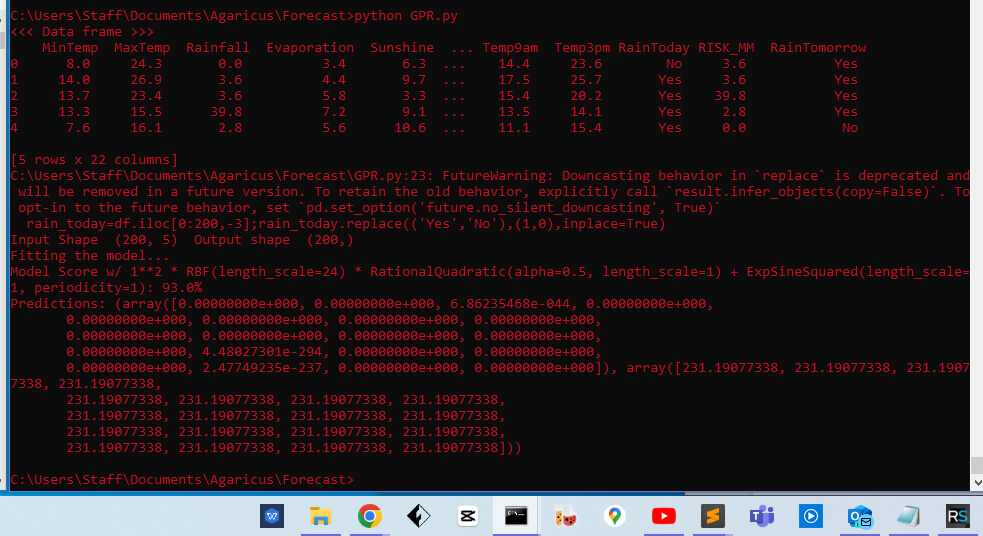


Img.4 Random forest

Moving to Regression:

As it can be interpreted that the models are not fit for the data giving not more than 70% accuracy. With having no satisfactory results found, Gaussian process regression (GPR) was tried what gave 93% accuracy. Img.5 shows the performance of GPR of the dataset.

Model result with accuracy

Img.5 Gaussian Process Regression(GPR)

Inference:-

It seemed regression might be a good choice to tackle this problem. ARIMA is wellknown for regression based powerful forecasting. In stock markets ARIMA is reliable in use. Therefore, ARIMA was taken into consideration to forecast over the weather dataset.

**Conclusion:**

RF,DT, NB,GPR need input features like growth from a certain time to another, for example 7 am to 10 am as input and as output the growth in 6 hours ahead for different sizes. This is the scenario of training data.

The drawback is these models make forecast at a specific time window, let us say in 6 hours and 12 hours. If the model is trained over 6 hours ahead data as output, it can only forecast 6 hours ahead not in 3 hours or 9 hours.

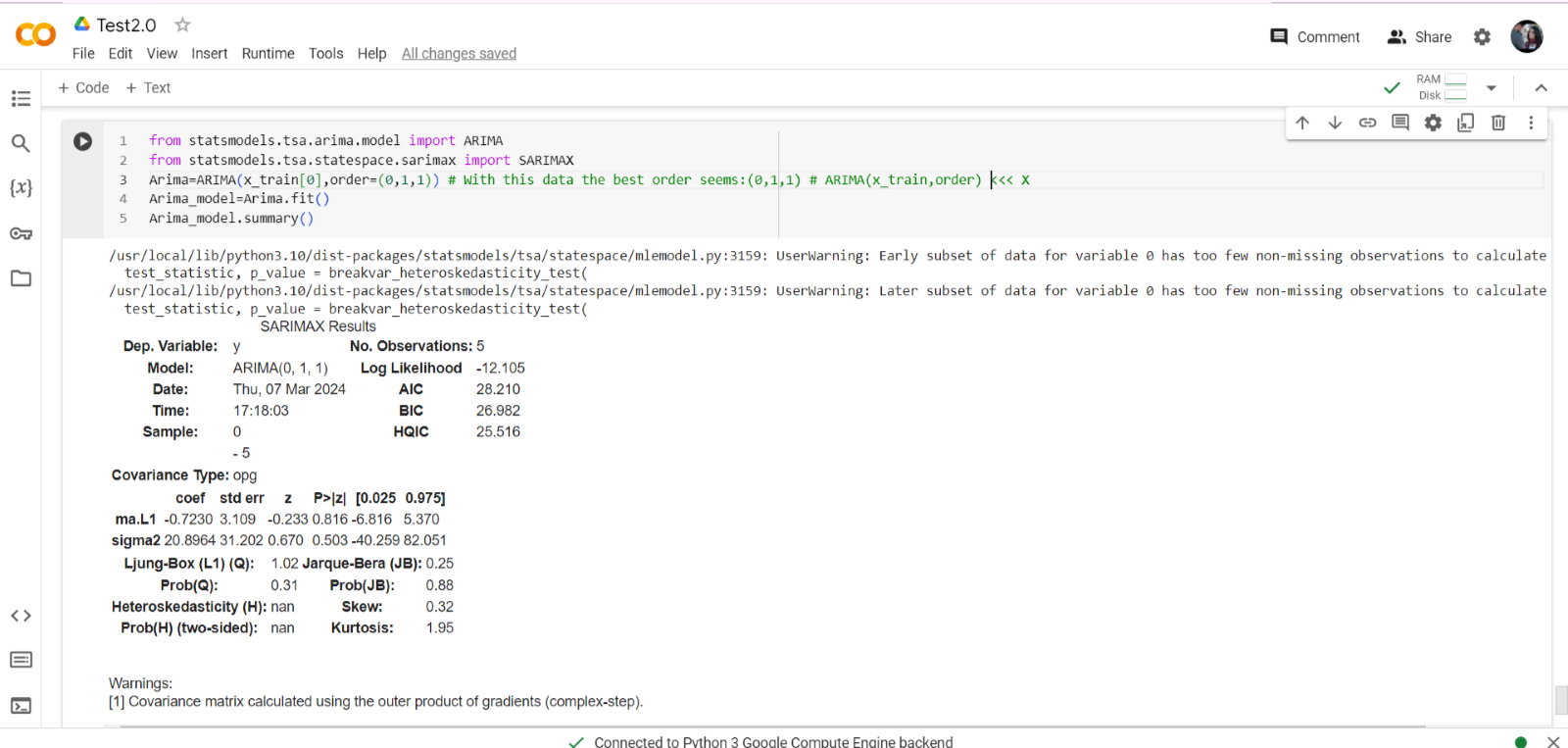
Another drawback is that, the input feature needs to be huge for example every hour scan data starting from 7 am to 7pm is not sufficient to train these models.

One thing can be done, a full day scan data of every hour can be fed to the model to forecast 6 hours ahead which is next day. Which means Every day scan is required for making forecast for the next day

Experimental results over ARIMA:

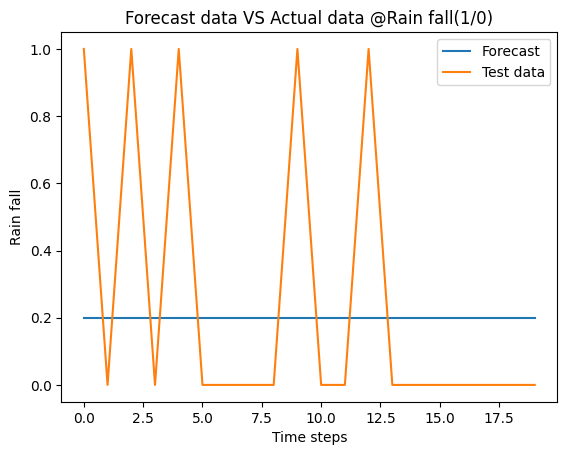
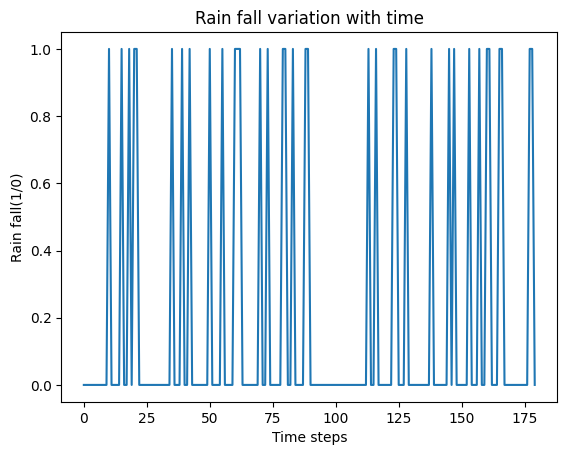
In the beginning ARIMA (Autoregressive Integrated Moving Average) was tested and it forecast on 6 and 12 hours ahead of time. Img.5 describes the insight of the data, if the data is stationary (P-value <0.05) and what are the statistical data (mean, median, std etc.)can come out of.

Statistical Insight of the data



Img.5

ARIMA forecast vs Test data comparison



Graph.1 ARIMA Train data Graph.2 ARIMA Forecast vs Actual data

After performing the test over weather dataset, the following inference can be reached. As mentioned before weather dataset is analogous to the dataset can be prepared by scanning the mushroom beds, hence with the results got after running the model on weather dataset, it can be said that the inference drawn will not be that divergent to that of the actual mushroom growth dataset.

Inference:

1. ARIMA needs to be applied to individual sizes and it will forecast based on that input. What means It cannot take multiple features one at a time e.g. 20mm time-series and 21mm time-series cannot be fed altogether but one by one after it does forecast one after another.

2. The feature input needs to be large, here hourly growth data was fed and that stood out *not to be sufficient.*

4. Data under test needs to be stationary and the data came out to be non-stationary so first the dataset was made differenced to make it stationary to feed into ARIMA

Restrictions:-

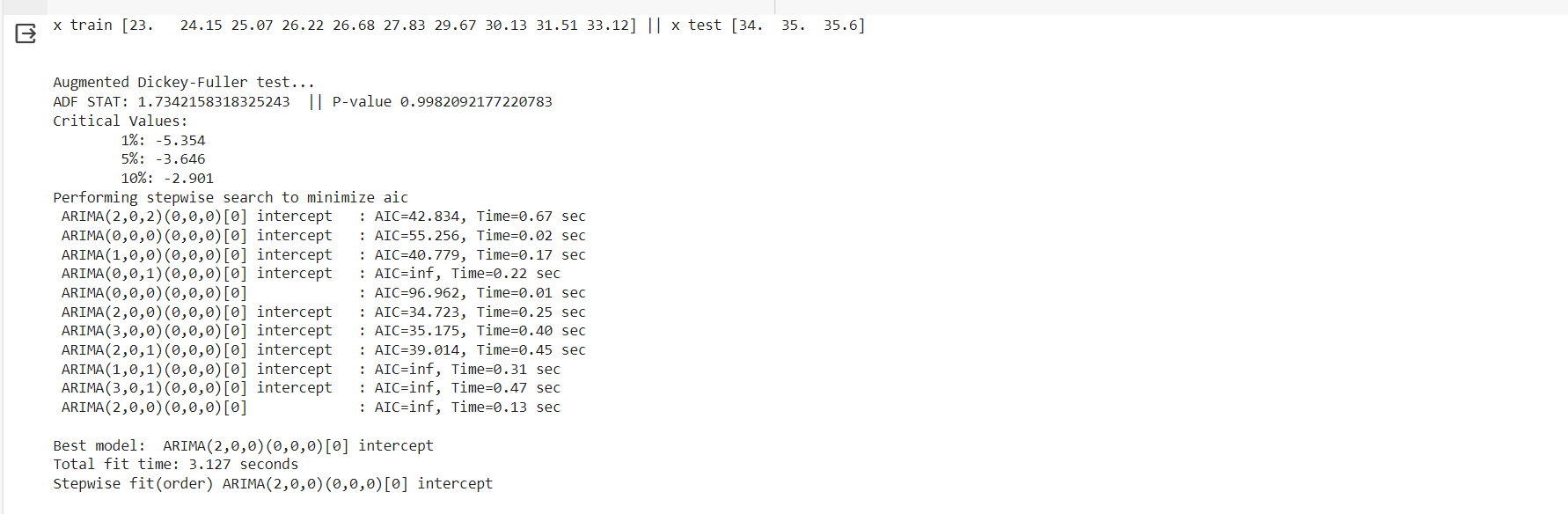
The dataset used was medium in size. Bigger dataset could have given more inside of the suitability of the regression model. Actual mushroom dataset may differ in properties and inner statistics w.r.t this dataset

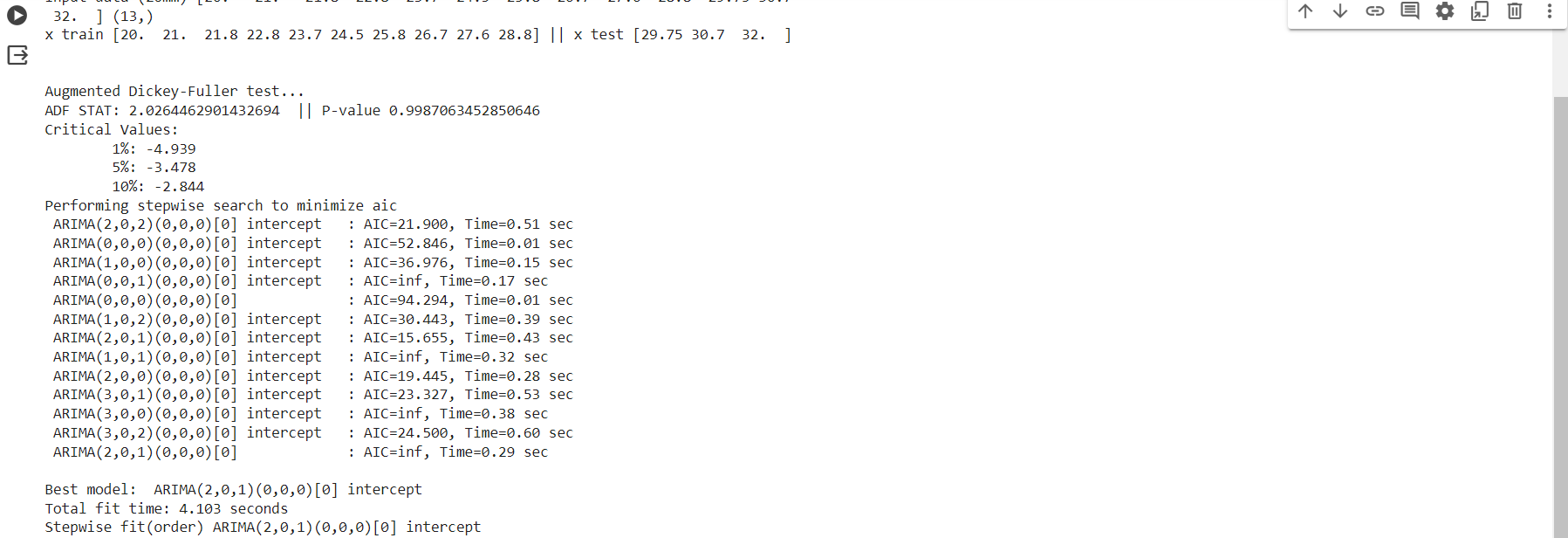
ARIMA on Mushroom Growth dataset:-

As mentioned above fake dataset consists of mushroom growth of each size starting from 20 mm to 65mm over every hour.

Add Fuller (ADF) test result:

It seems the dataset is non-stationary with p-value for 20mm being 0.99 which is >0.05. Other soizes e.g. 21mm, 22mm,23mm and so on gives same ADF test result.

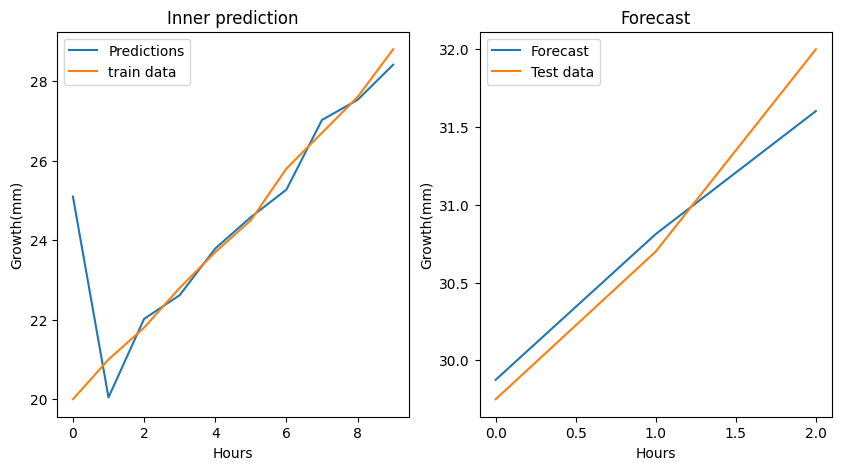
Img. 6 ADF test over 20mm size Growth data series

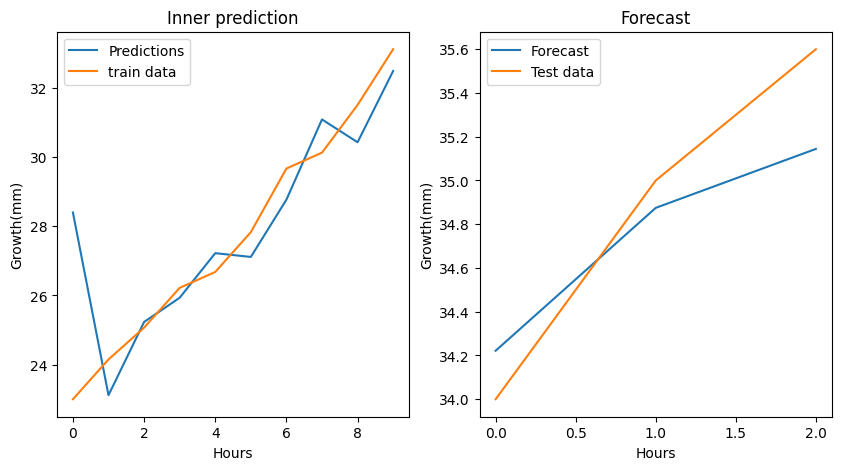
Img. 6 ADF test over 23mm size Growth data series

ARIMA forecast comparison:-

Prediction is result within the training time frame whereas forecast is outbound prediction.

If dataset has 7 am to 10 am data, prediction will give data with more details and finer insight between 7am to 10am while forecast will give data after 10 am.

Graph.3 ARIMA @ 20mm

Graph.4 ARIMA @ 23mm

Inference:-

1. With Hourly data, ARIMA is not able to understand data variations with much accuracy giving slightly different result than actual data.
2. To make forecast for a whole bed, the model needs to run for the number of different sizes available on bed as this forecast happens for a single size once at a time
3. Dataset needs to be huge. Every hour scan data for 12 hours does not seem to be sufficient. Forecast is solely dependent on the dataset prepared earlier.

How it might work:

1. At the morning a whole bed scan needs to be performed. All different sizes needs to be listed.
2. ARIMA would be run for individual sizes one by one giving 6/12 hours of forecast
3. For each sizes a forecast in 6/12 hours can be made which will give net weight of mushroom from a single shelf and eventually from a tunnel.

Step taken:-

A test was performed whether a different dataset can do better forecast. A time series dataset with hourly growth was created and it acted as a fake dataset for the test performed afterwards.

Growth rate analysis has been done-[Growth Rate.docx](https://universityoflincoln-my.sharepoint.com/:w:/g/personal/akundu_lincoln_ac_uk/EWg5-b7ECjtMofShcVPrHl0Bx_5IrTB-ydjor9nQsjg0zA?e=TDEgUe) -and better models than ARIMA is being searched.

Things to keep in mind:

The model which is going to be trained on the growth dataset will be used to make forecast. This forecast is supposed to be made at the beginning of the day which is 7 am in the morning which means, the forecast model is supposed to take one time scan data and forecast in 6 hours ahead.

At moment ARIMA can be used for forecasting specific size of mushroom. Once the whole dataset is ready and ARIMA is trained on the same, it can forecast next time steps e.g. for 6 hours/12hours and 24 hours ahead in case in need.

Requirement:

ARIMA needs a full 12 or 24 hours scan prior to the day of forecast. For example, if farm needs forecast for tomorrow, a full scan needs to be performed today.

Cons:

ARIMA requires more than 12 hours of regular scan happening every hour