Project Report

Subject

Industrial Analytics

Topic of the Project

Product Demand Forecasting

Honey Production

Group Members

Hanish Raval

Dharmikkumar Anaghan

Introduction :

Before moving to our topic we would like to answer to the question of how important honeybees are for human nutrition. Then we will explain the importance of honeybees to our diet using a variety of facts. These include information on the percentage of crops they pollinate, the types of crops they pollinate, and the value of the crops they pollinate.

According to the U.S. Department of Agriculture, these underappreciated workers pollinate 80 percent of our flowering plants, which account for one-third of our total food supply. Their loss could not only impact staple crops such as apples, broccoli, strawberries, nuts, asparagus, blueberries and cucumbers, but could also threaten our cattle and dairy industries if alfalfa is no longer available as feed. A Cornell University study estimates that honeybees pollinate $14 billion worth of seeds and crops in the U.S. each year. If honeybees disappear, they could take most of the insect-pollinated crops with them, potentially reducing humanity to a water-only diet.

The loss of wild pollinators in Europe is attributed to a number of factors, including climate change, invasive alien species, habitat and land use changes.

Pollinators can come into contact with toxins through contact with spray residues on plants or ingestion of contaminated pollen

and nectar. Contamination of nest sites and nesting material also poses a risk.

According to the European Red List of Threatened Species the population of about one in three bee and butterfly species is declining, and about one in ten species is threatened with extinction.

After considering All these points, the decline of honey is dependent upon the following reasons.

- Habitat loss
- Climate change
- Pesticides
- GM (Genetically Modified) crops
- Diseases
- Invasive species

From the aforementioned reasons mainly two reasons are considered as most for decreasing the production of honey first is climate change and second is pesticide.

In this Project we are forecasting the honey production of the USA of 2012 and we have data of 1998 to 2011. The number of variables we have to predict the total production and The data is taken from the Kaggle platform named as Honey production of USA.

All the variables of this data is shown below:

Numcol: Number of honey producing colonies. Honey producing colonies are the maximum number of colonies from which honey was taken during the year. It is possible to take honey from colonies which did not survive the entire year.

Yieldcolon: Honey yield per colony. Unit is pounds.

Totalprod: Total production (numcol x yieldpercol). Unit is pounds.

Stocks: Refers to stocks held by producers. Unit is pounds.

Priceprlb: Refers to average price per pound based on expanded sales. Unit is dollars.

Prodvalue: Value of production (totalprod x priceperlb). Unit is dollars.

Additional data

Consumption: (Totalprod - Stocks) unit is pounds.

After reading some articles we get to know the production of honey is higher in summer as compared to winter and monsoon so as per this basis we put some estimated relation variable of summer and winter for better prediction purposes.

Summer_Prod: Production of summer (Production * 0.6).

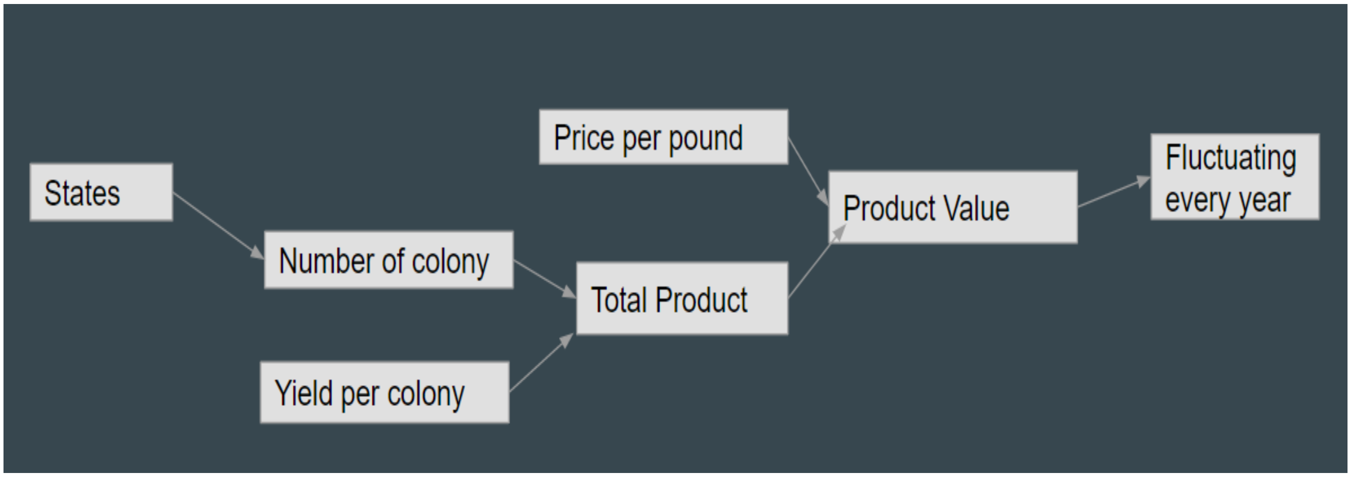
Winter_Prod: Production of winter(Production * 0.4).

The following steps are done after preparing the data for making a better prediction model.

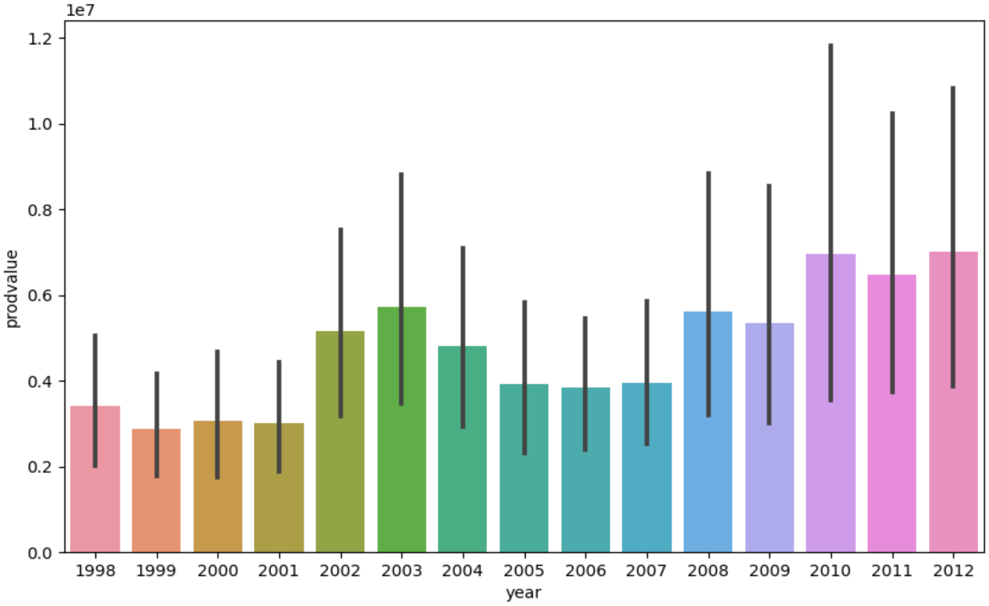
- Data Schema
- Data Visualization
- Feature Selection
- Tree Model
- Neural network Model Building
- Comparison of Tree and Neural network
- Evaluation

❖ Data Schema

The above Figure shows the dependency of variables. The number of colony is



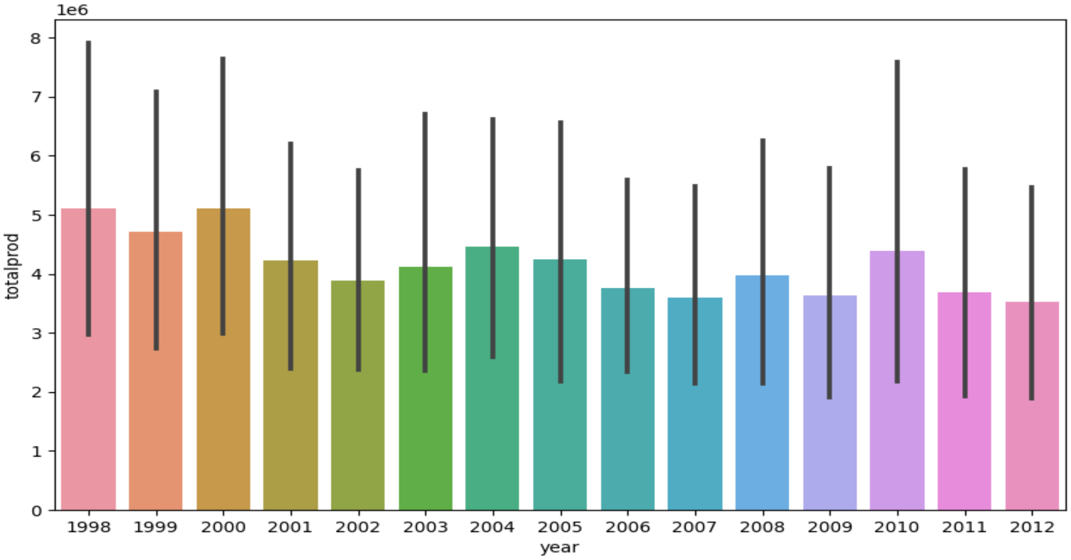
❖ Data Visualization

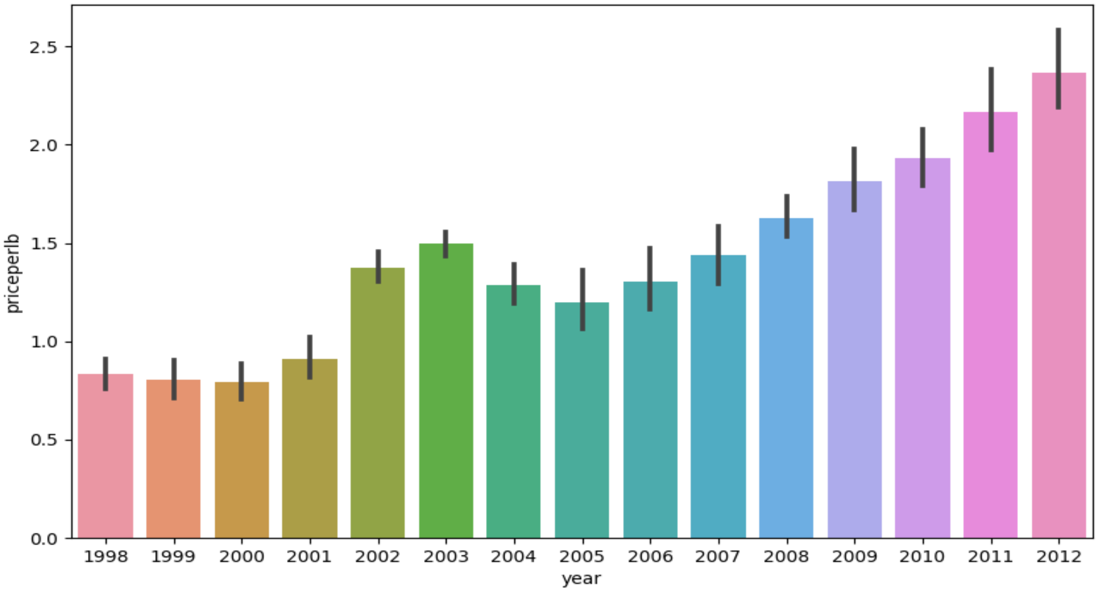


The above bar plot of Production value vs year in the span of year 1998 to 2012. The production value is the product of total production and price per pound.

The below two bar plots of total production and price per pound show's variation in each year. The price per pound is increasing every year because of its demand and change in currency but total production seems to be decreasing as compared between the year 1998 to 2012.

The black line in every bar plot represents the standard deviation or in short error to find the average mean of every year. The big and small line depicts the large and minimum error respectively.





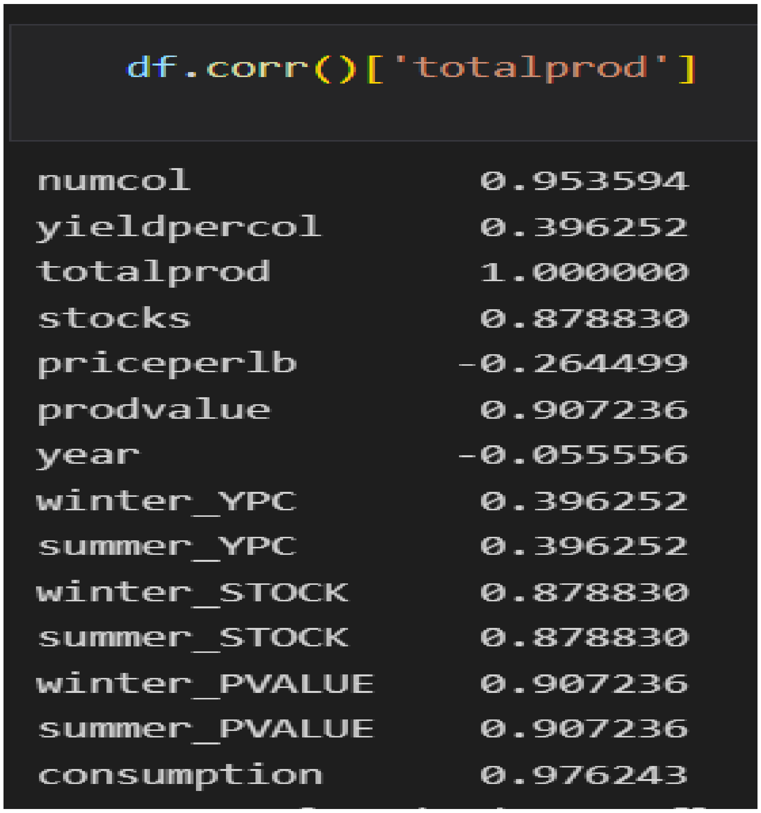
After considering all three variables we can conclude that the increasing trend of production is mainly because of increasing the price per pound of honey. However, The total production is decreasing.

Every year all the variables show extremely different performance and by selecting all the variables as an input variable may decrease the prediction accuracy. So the next step is to do the Feature selection of input variables for better prediction.

❖ Feature Selection

It is the major task while making any machine learning model. The selection of input variables can be done by checking the correlation between the response variable.

In this project we considered total production as a response or output variable. Below we can see the percentage of correlation of each variable to the total production.

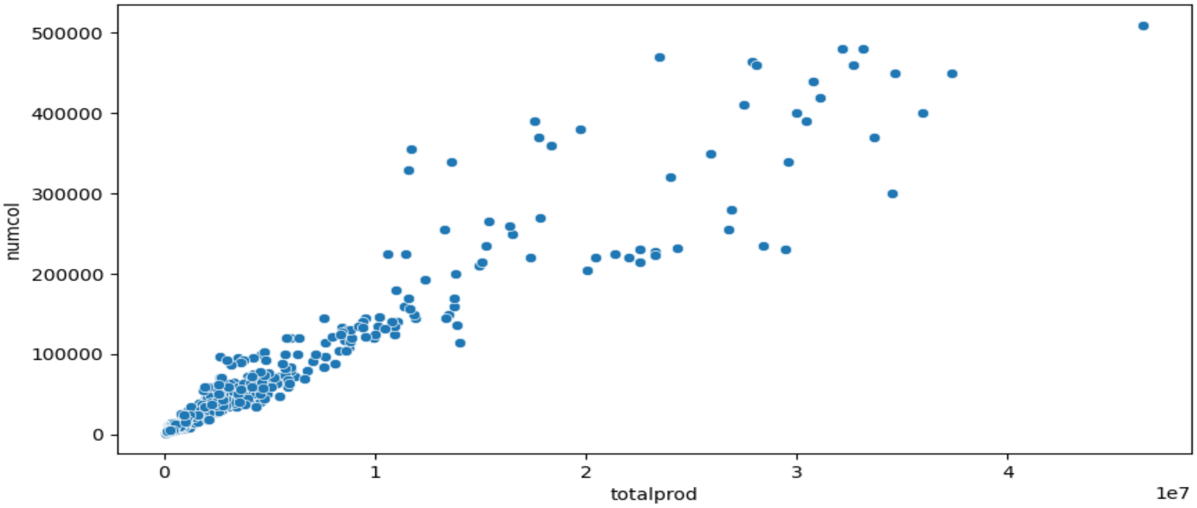


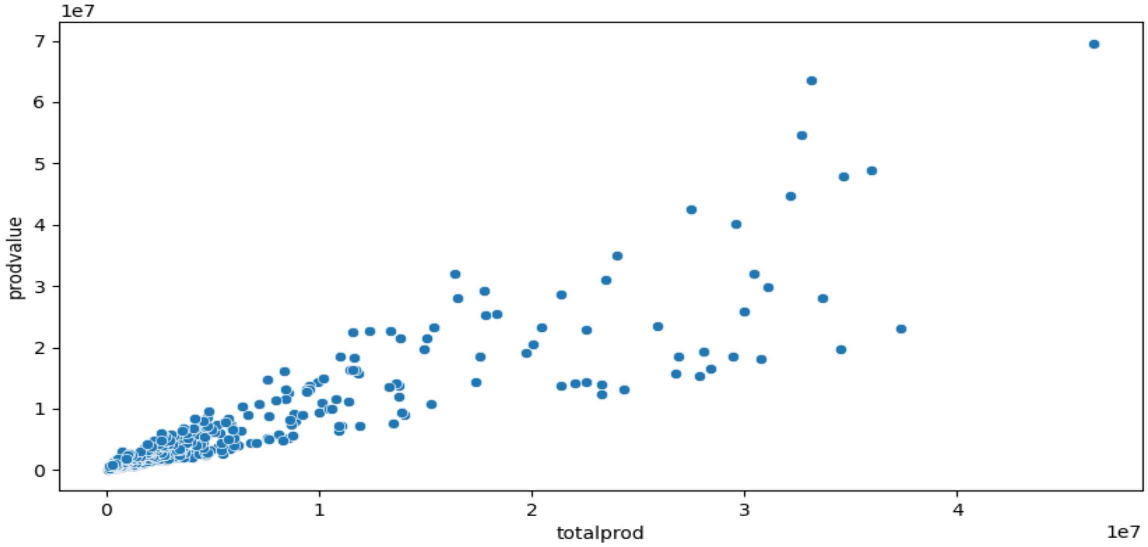
```
df.corr()['totalprod']
```

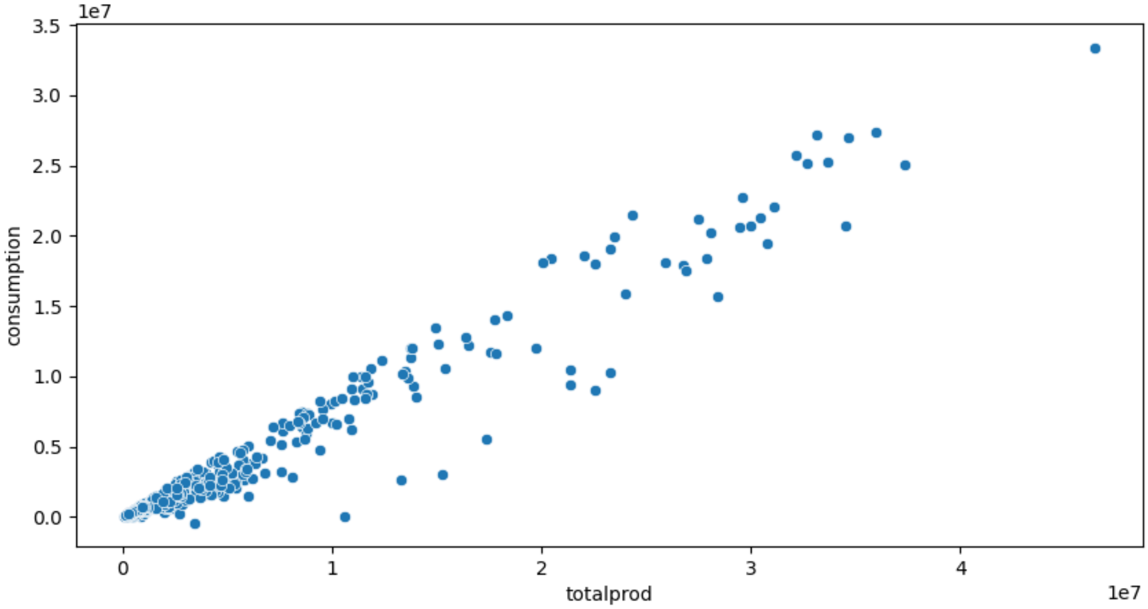
```
numcol             0.953594
yieldpercol        0.396252
totalprod          1.000000
stocks             0.878830
priceperlb        -0.264499
prodvalue          0.907236
year              -0.055556
winter_YPC         0.396252
summer_YPC         0.396252
winter_STOCK       0.878830
summer_STOCK       0.878830
winter_PVALUE      0.907236
summer_PVALUE      0.907236
consumption        0.976243
```

From the above result of correlation we can say that the numcol (number of colonies), stocks, production value and consumption gives the highly correlated results to total production.

For better understanding let's take a look at the linear relation of these variables to the total production.







In all the three scatter plots we can see the good linear relation with total production hence this is the reason to select these feature as Input variables for making a good model and giving a better forecasting results of honey production.

The next step is to select the method of building the model. It is a problem of forecasting , Hence it is said to be a regression problem. Here we have noisy data so we could not get a good result by linear regression so the next step is to make a model by regression tree and neural network and compare the results.

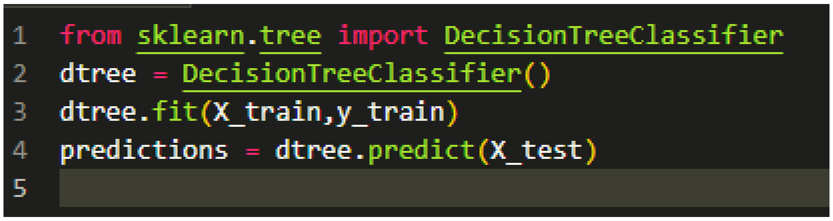
After discussing the result , select an appropriate method and make a final model.

❖ Decision Tree

A decision tree is a supervised machine learning tool that may be used to classify or forecast data based on how queries from the past have been answered. The model is supervised learning in nature, which means that it is trained and tested using data sets that contain the required categorisation.

First data is splited using test train split in python. X_data is feature data and Y_data is actual production value.

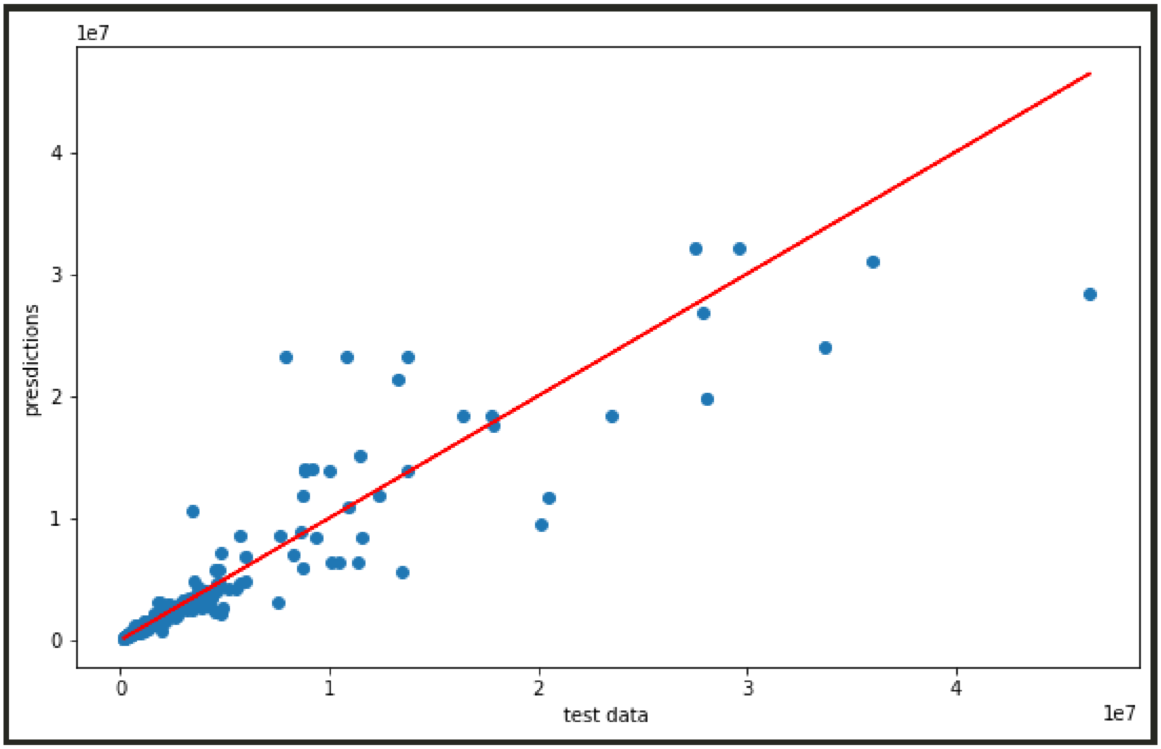
Below is code that we implemented in jupyter file.



```python
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier()
dtree.fit(X_train,y_train)
predictions = dtree.predict(X_test)

```

Then decision tree classifier is imported and fit the features and actual production values.

From the prediction,variance score is calculated which is 0.82. Hence it is clearly proved that accuracy of decision tree is not very accurate compared to neural network.



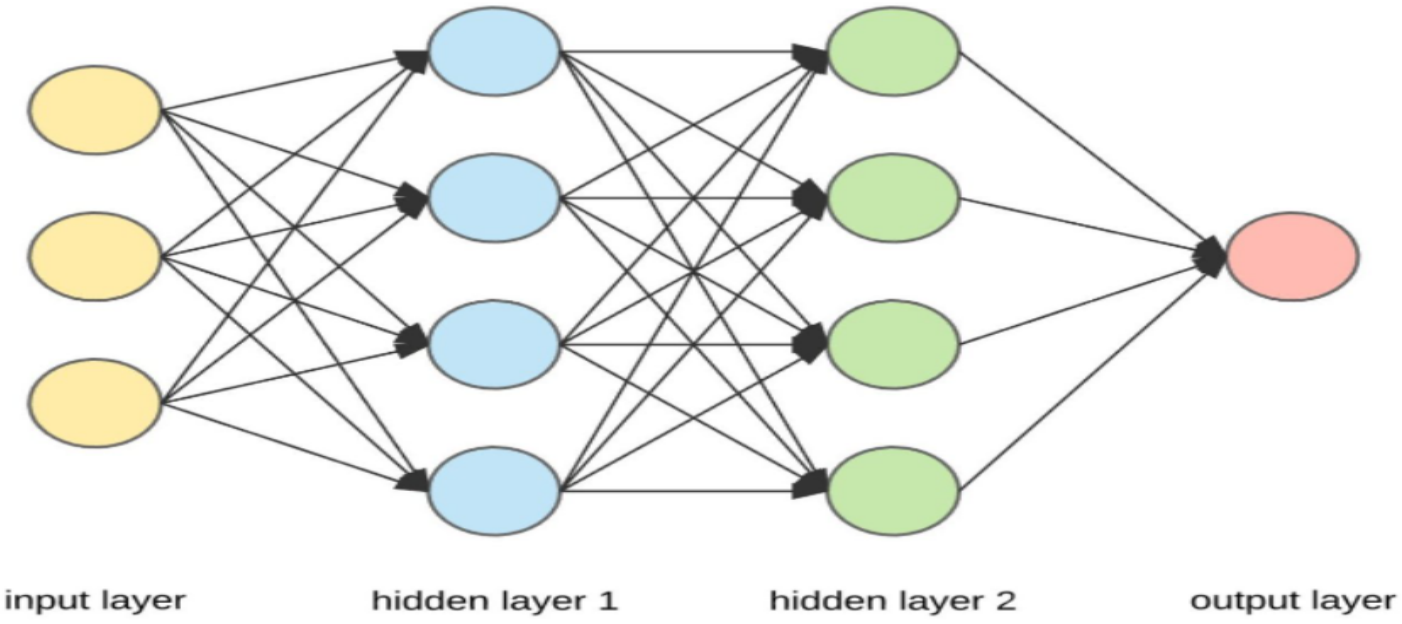
From the graph of test data and prediction resulted from decision tree we can conclude that as blue dots are too far from red actual line it has higher residual value which makes poor accuracy in predictions.

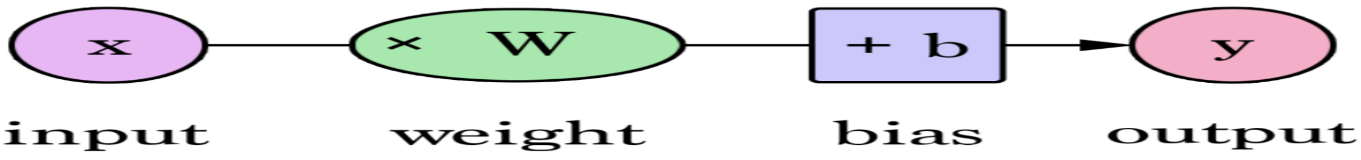
Now below we are going to implement neural network as tree model does not predicted accurate production value.

❖ Neural Network

A neural network can be understood as a network of hidden layers, an input layer and an output layer, that attempts to mimic the functioning of a human brain.

The hidden layers can be visualized as an abstract representation of the input data itself. These layers help the neural network understand various features of the data using its own internal logic.





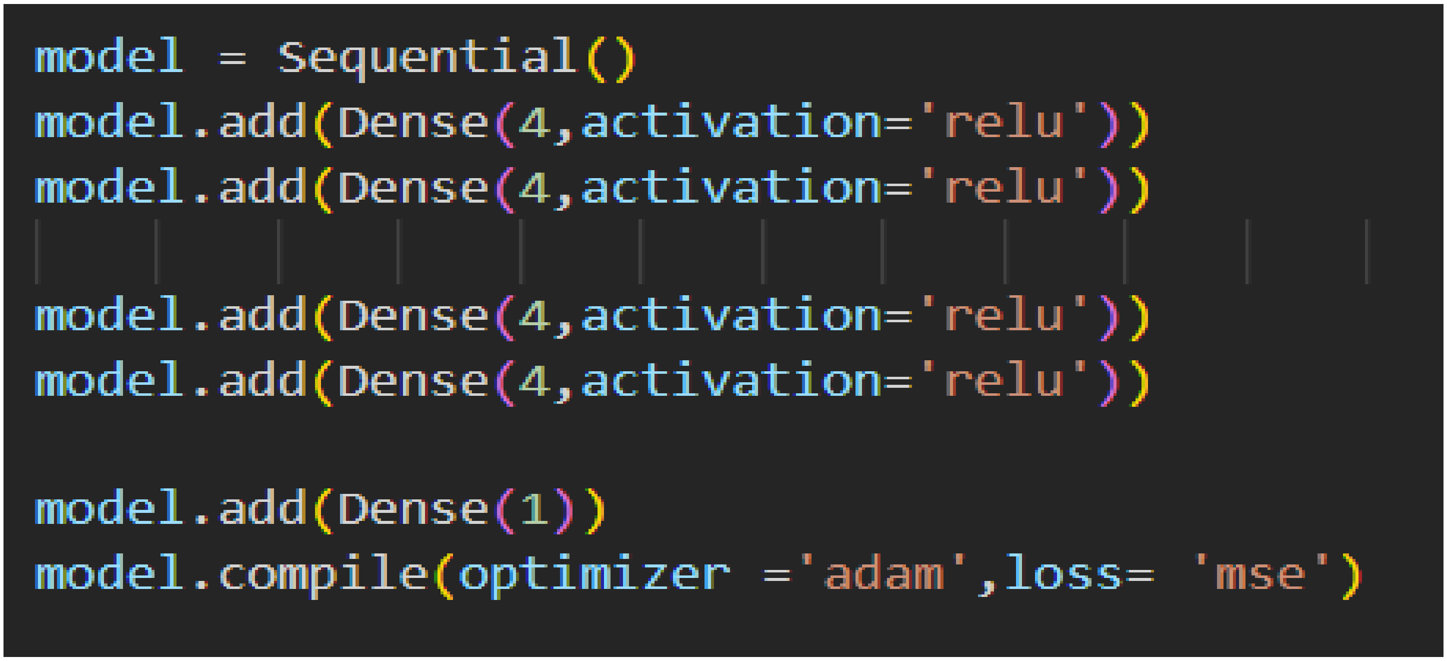
These neural networks are non-interpretable models. Non-interpretable models are those that cannot be interpreted or understood even if we observe the hidden layers.

Working of Neural network: In our project we took 4 input layers and 4 hidden layers. At each hidden layer the output is updated by adding weight and biases to obtain approximate result compared to the test data.

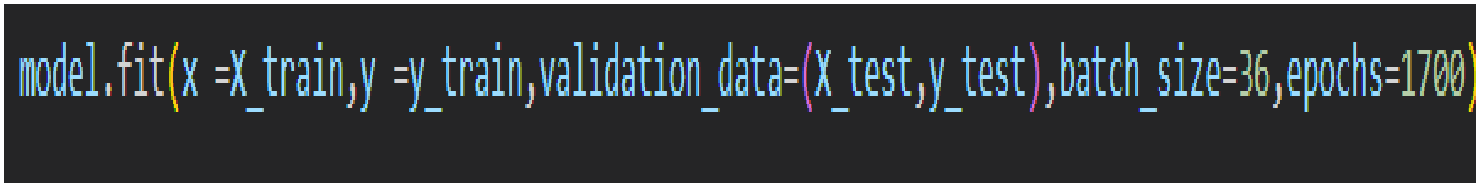
Number of epochs is also selected to adjust the weight and biases here we used 1700 epochs that means this model runs upto 1700 times to obtain the similar result as we have in test data and try to keep the same losses of actual data and test data as a result we can get a better prediction.

Gradient descent optimizer is used as Adam and Activation is set to be a relu(rectifier linear unit) which means the output generated by each layer should be in the form of positive number all negative values are considered as 0 and positive value remains same in each layer by using this activation.

The size of the validation is 0.30 percent of the training data. The batch size is 32 means at every iteration we consider 32 rows of selected features that can be changed at every iteration. The model can be seen below.



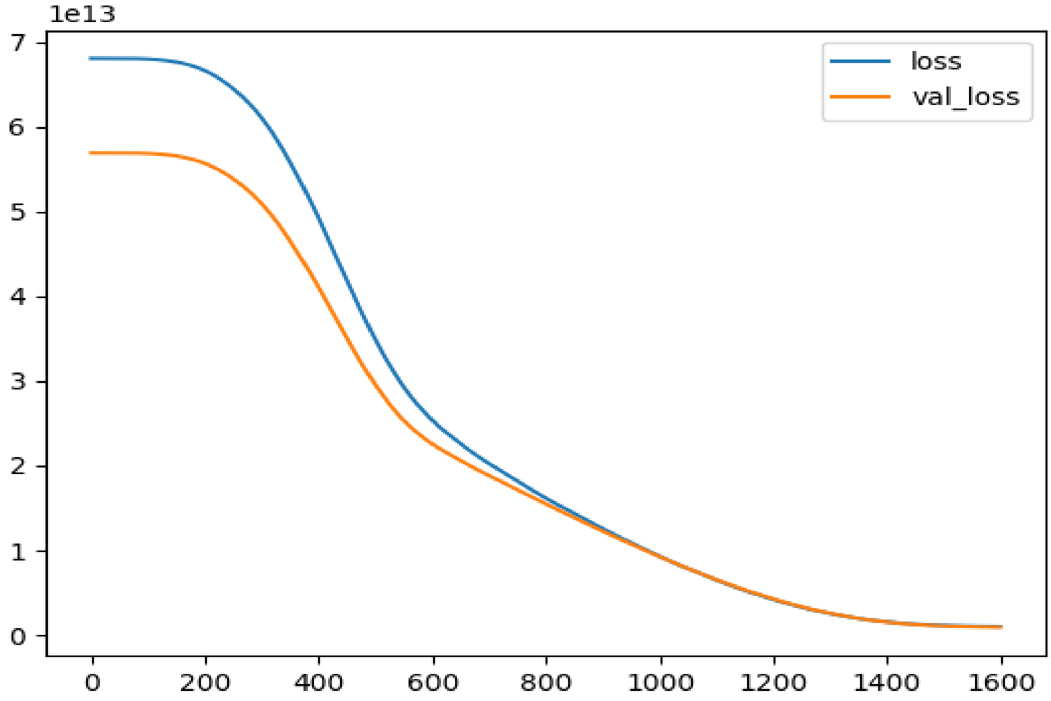
```
model = Sequential()
model.add(Dense(4,activation='relu'))
model.add(Dense(4,activation='relu'))

model.add(Dense(4,activation='relu'))
model.add(Dense(4,activation='relu'))

model.add(Dense(1))
model.compile(optimizer ='adam',loss= 'mse')
```



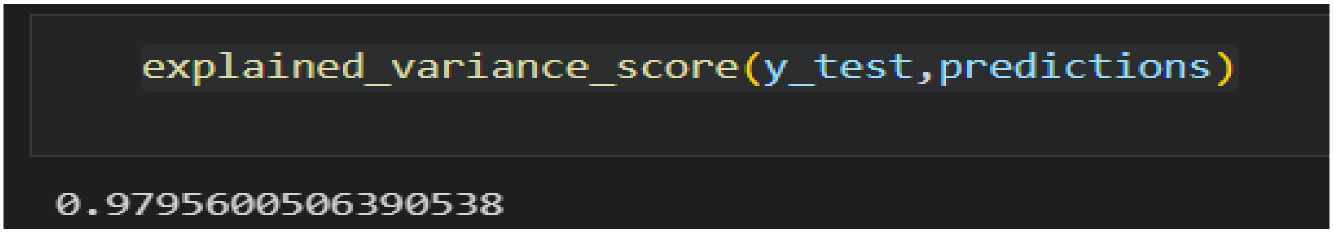
```
model.fit(x =X_train,y =y_train,validation_data=(X_test,y_test),batch_size=36,epochs=1700)
```

After fitting the model, we take a look into the losses of validation and test losses.

As shown in below fig around 1600 epochs the validation and test losses stop to overfitting which means the required weight and biases are used to update to get a similar result of test data and the losses are equal.



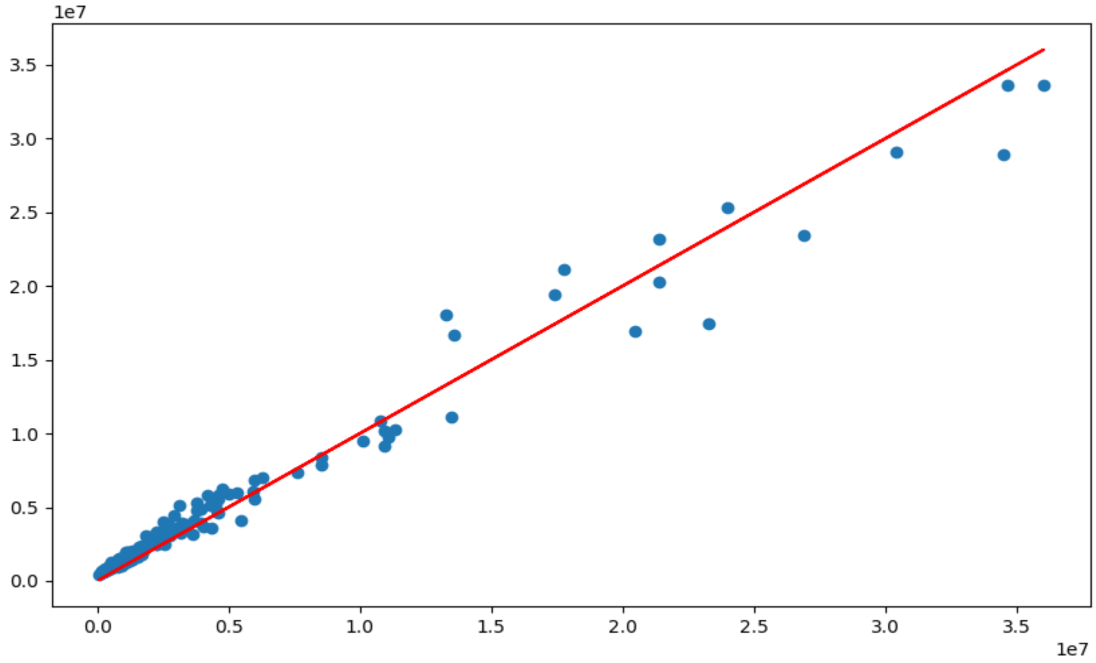
Another way to see how the model performs is to check the variance score of test data and prediction by which we can easily identify the performance of the model.



```
explained_variance_score(y_test,predictions)

0.9795600506390538
```

Varianced score is 0.97 that means the 97% of prediction is similar to the test data, where 1 is said to be a 100% match.

For better understanding it is good to look at the plot of real and prediction data as in the scatter form

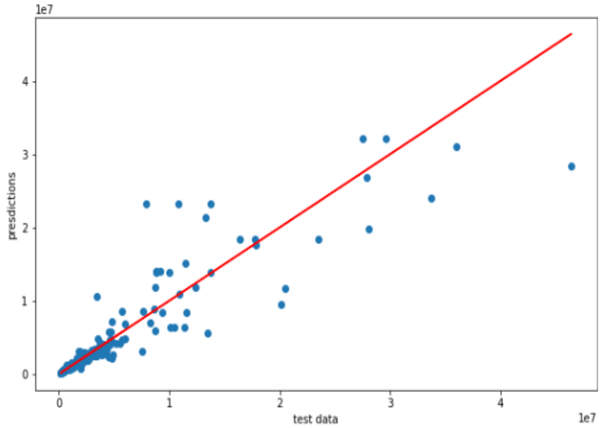


As shown in the above fig red line represent test data and blue dots represent the prediction with a variance score of 0.97. We can also depict that the residuals are also less and prediction is very close to the actual data.

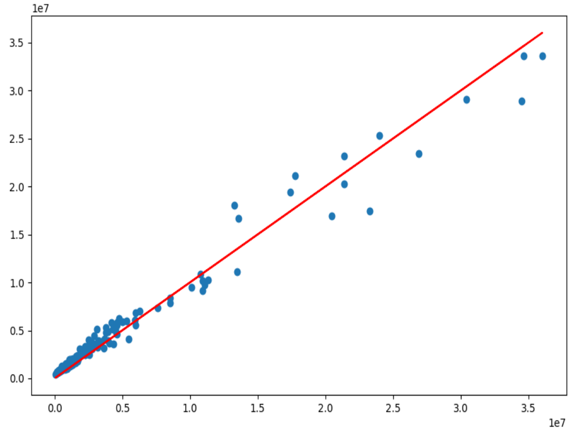
Now the question arises which model we should use: decision tree or neural network.

By comparing both methods we have a good variance score in neural networks .

Compared plots of prediction in both methods



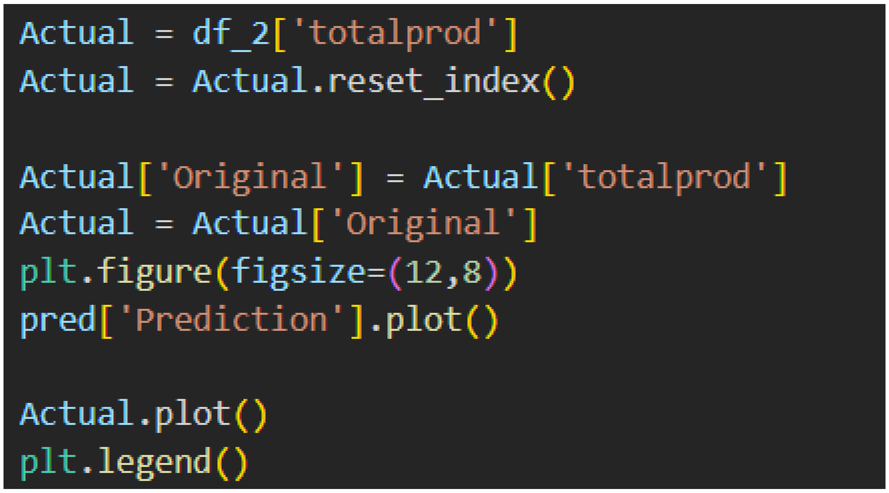
Decision Tree



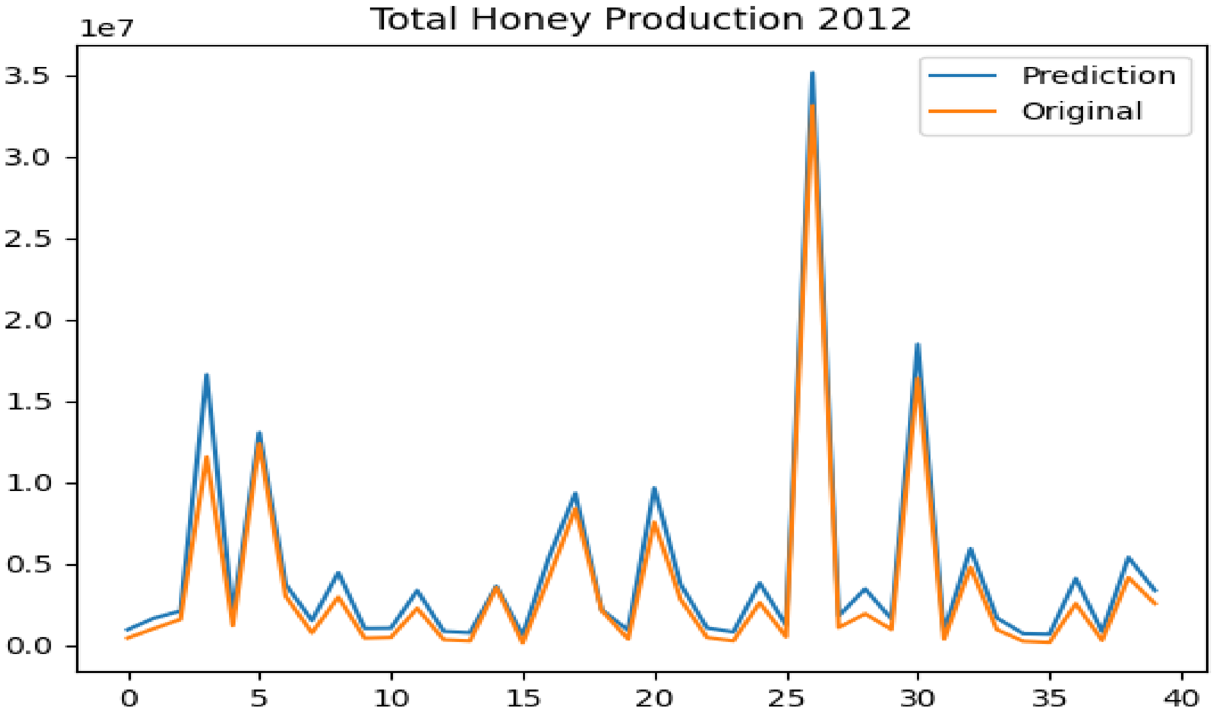
Neural networks

In the plot of a neural network the prediction is close to the real data so it is better to use this model for prediction. There are some drawbacks, like time consuming and more complex but it gives better accuracy than regression trees so we make a model of neural network.

The next step is to save the model and the main task is to check the model with other new data. We have data from 1998 to 2012 so we extract the 2012 data for forecasting purposes. We did not use 2012 data for testing or training. After the model is built we put 2012 data for forecasting and check the result by comparing predicted forecasting total production result and real 2012 total production. By putting in new data we can see how the model performs for the new one because the new data is not available in the test and validation data. As shown below the plotting code in python.



```
Actual = df_2['totalprod']
Actual = Actual.reset_index()

Actual['Original'] = Actual['totalprod']
Actual = Actual['Original']
plt.figure(figsize=(12,8))
pred['Prediction'].plot()

Actual.plot()
plt.legend()
```



The above figure shows the result of forecasting of honey production 2012 and the original total production and the model gives a good prediction.

❖ Tools and libraries used :

- Python
- Jupyter
- Scikit Learn
- Pandas
- Matplotlib
- Pandas
- Numpy
- Tensorflow
- Seaborn

❖ Summary

- In this project we did data featuring , extracting and preparing.

- Feature selection is done with the help of correlation.

- It is the problem of forecasting hence we must go through regression model building.

- Linear regression is not suitable due to the noisy data then we first try to regression tree and then neural network (sequential).

- After comparing the results of both methods. Neural network gives the best prediction so we decided to choose the neural network for forecasting.

- Finally , we put the 2012 data in a model for forecasting the total production of honey and the result is quite similar to the original one. As a result we can say that our model gives a better prediction

References

- Matplotlib : https://miro.medium.com/max/805/1*aUSZsGFCMPNYCkQygs4aGQ.jpeg
- Numpy:https://upload.wikimedia.org/wikipedia/commons/thumb/3/31/NumPy_logo_2020svg/2560px-NumPy_logo_2020.svg.png
- Python: https://docs.servicestack.net/assets/jupyter-python.6188762b.png
- Jupyter :https://docs.servicestack.net/assets/jupyter-python.6188762b.png
- Matlab: https://logos-world.net/wp-content/uploads/2020/12/MATLAB-Logo-700x394.png
- Pandas: https://upload.wikimedia.org/wikipedia/commons/thumb/e/ed/Pandas_logo.svg/1280px-Pandas_logo.svg.png
- Excel: https://upload.wikimedia.org/wikipedia/commons/thumb/3/34/Microsoft_Office_Excel_%282019%E2%80%93present%29.svg/2203px-Microsoft_Office_Excel_%282019%E2%80%93present%29.svg.png
- Figure: Created by Author
- Data : Honey Production in the USA (1998-2012) | Kaggle
- https://friendsoftheearth.uk/nature/what-are-causes-bee-decline
- https://busybees99.wordpress.com/2018/05/03/the-importance-of-honeybees/
- https://enrd.ec.europa.eu/sites/default/files/enrd_publications/enrd_protecting_pollinators_from_pesticides_policy_insight.pdf
- https://uniaobahia.org/why-are-honey-bees-declining/
- https://justbeehoney.co.uk/blogs/just-bee-honey-blog/why-are-bees-dying