ABSTRACT

Fisheries sector plays significant role in provding the nutritional security of Kerala. It provides livelihood to 2.98 % of total population in the state. Kerala has a prominent place with regard to the marine fish production in the country which contributes to almost 25% of the total marine fish production. The changing climate and oceans have significant impacts on the nation's valuable marine life and ecosystems, and the many communities and economies that depend on them. Forecasting fisheries time series is integral for fisheries management, because it allows policy makers to develop strategy and enact management decisions that can achieve goals in light of uncontrollable events.

Traditionally, time series forecasting has been dominated by linear methods because they are well understood and effective on many simpler forecasting problems. Machine learning brings computer science and statistics together for creating predictive models. ML is a subarea of artificial intelligence (AI), where the main objective of using ML is to practice different algorithms to analyze data, learn from the outcomes, and finally generating prediction accuracy.

Forecasting of fish production is one of the many ways that can contribute to a better decision making for fisheries management. Even though works have been done in this field, majority are confined to traditional or statistical models. Especially in our area of study, Kerala marine environment, no previous work has been done to incorporate machine learning models in the forecast of fish production.

In the present work, four modelling and forecasting techniques were evaluated on the basis of their efficiency to forecast annual marine fish production in Kerala and their ability to utilise auxiliary environmental information: Linear Regression, Multilayer Perceptron, Random Forest Regression and Radial Basis Function. The results reveal that the annual production can be predicted and that the Random Forest Regression model is the best performer overall, characterised by a higher number of stable forecasts, and forecasts with higher precision and accuracy, than the other methods.

ABBREVIATIONS

MAE Mean Absolute Error

ML Machine Learning

MLP Multilayer Perceptron

MLR Multiple Linear Regression

MSE Mean Squared Error

R2 R Squared Score

RBF Radial Basis Function

RFR Random Forest Regression

RMSE Root Mean Squared Error

Chapter 1

Introduction

Fish is a vital source of food for people. It is man's most important single source of high-quality protein : nutritious, cheap and healthy source of protein. The Fisheries sector plays an important role in providing the nutritional health security of a nation.

But more than just a food commodity, the fisheries sector plays a significant role in the socio-economic development of India. Marine fisheries is an important sector in India and contributes 1.2 percent to the nation's Gross Domestic Production. Among the maritime states, Kerala has a prominent place with regard to the marine fish production in the country which contributes to almost 25% of the total marine fish production. Also the Fisheries sector accommodates 2.98% of total population of Kerala, of which 77% are in the marine sector and 23% are in the inland sector.

In the past, the oceans were assumed to be unlimited, with enough fish to sustain an ever-increasing human population. The needs of a burgeoning population, on the other hand, have now far outstripped the seas' sustainable production. And it put natural resources constantly under pressure. The fact that ocean reserves cannot be clearly observed or calculated makes the scenario even worse. Management of marine resources have gradually become more important during these past years because of the increased awareness of these resources becoming limited.

Forecasting of marine fish production is very much essential for proper planning. Fluctuation in marine fish production affect the processing industry, export earnings, employment and income to fishermen community, marketing and cost of marine fish products. Advance information about future production will help in proper planning, storage, distribution and to take necessary measures according to the situation.

Actually we can project into the future using the past patterns. And in the recent decades forecasting has been rapidly developing in the field of fisheries, in describing fishery units and the state of fisheries' resources and management.

The rest of the report is organised as follows: Chapter 2 summarises the existing works which includes the background study and literature review. In chapter 3 we define the problem statement and list out the various objectives of the work. Chapter 4 focuses on the design and implementation where we explain the data-set, the machine learning models, evaluation metrics and the requirements for the work. Chapter 5 analyses the results and interpretation of the results. Chapter 6 concludes the thesis by providing a conclusion and the future scope.

Chapter 2

Existing Methods

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of Machine Learning. With the help of sample historical data, which is known as training data, machine learning algorithms build a mathematical model that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance. Traditionally, time series forecasting has been dominated by linear methods because they are well understood and effective on many simpler forecasting problems. ML is a subarea of artificial intelligence (AI), where the main objective of using ML is to practice different algorithms to analyze data, learn from the outcomes, and finally generating prediction accuracy. ML forecasting algorithms often use techniques that involve more complex features and predictive methods, but the objective of ML forecasting methods is the same as that of traditional methods – to improve the accuracy of forecasts while minimizing a loss function. Generally, it is common to use the linear regression (LR) method to predict time-series datasets. It is obvious that, if the dataset is correlated, then the LR-based ML models produced very good outputs with the more accurate assumption of the predicted values. Nevertheless, if the data are not much

correlated then we need to find different ML models for accurate predictions. Moreover, a single ML model does not perform best in different time-series predictions. In this case, a good choice would be to use ensemble-based ML models for enhancing the prediction accuracy.
ML algorithms normally learn from the data and come up with a prediction accuracy on flight time deviation, weather forecasting, water quality prediction, hydrometeorological forecasting for agricultural decision support, etc. around the world.

2.1 LITERATURE REVIEW

The background study and implementation of the project "Forecasting fish production in Kerala using Machine Learning models" was completed using five base papers as the initial inspiration.

[1] Discussed machine learning techniques and their scope of application in fisheries and aquaculture.Showed that adoption of artificial intelligence tools based on data mining and machine learning algorithms can make significant impact on aquaculture production systems, particularly in the optimisation of feed use, disease prevention, biomass monitoring and market intelligence.

In [2] an attempt has been made to forecast the fish production of India using ARIMA and ANN models. Empirical results clearly revealed that the machine learning technique outperformed the ARIMA model.

[3] made an attempt to evaluate the predictive performances of three modeling techniques, i.e., ARFIMA, ARIMA and NNAR using the total fisheries production (metric tons) data in India. Autoregressive Integrated Moving Average (ARIMA) modeling is a statistical technique used for time series data in order to understand and forecast future trends in a better way. Recently, the ARIMA models have been employed in practice for modeling the data

of total fisheries production in India. In this study, an important family of parametric time series modeling when the order of difference is fractional, called Autoregressive Fractional Integrated Moving Average (ARFIMA), has been proposed for modeling and forecasting the total fisheries production (metric tons) in India. For testing the fundamental assumption of stationarity, Augmented Dickey Fuller (ADF) test was used. We also used a nonparametric model such as Neural Network Autoregressive (NNAR) for investigating the behavior of the data. After the evaluation of different models and perform comparisons based on root mean square error (RMSE) and mean absolute percentage error (MAPE) values, the result indicated that ARFIMA (3, 0.48,0), ARIMA(1,2,1) and NNAR(3,1) were the best models. The current results reflected that ARFIMA model outperformed ARIMA and NNAR models in forecasting the total fisheries prediction.

In [4] tried to forecast the prices in Oman fish markets using data mining algorithms, by means of studying the history of data that will assist to make a proper decision.They considered the fish markets in Sultanate of Oman where it selected 29 markets and 15 fish species in each market. In addition, the data mining algorithms, namely Linear Regression, SMOReg, Multilayer Perceptron, MLP Regressor, and Random Forest has been applied to forecast the prices weekly and monthly. The outcome of this research reveals that Random Forest provides a good performance for the weekly and monthly predictions compared to other algorithms.

[5] presented a novel approach to predict fish production in Malaysia's five major states.They developed an ML based prediction of marine fish and aquaculture production. Based on the feature importance scores, we select the group of climatic variables for three different ML models: linear, gradient boosting, and random forest regression. The past 20 years (2000–2019) of climatic variables and fish production data were used to train and test the ML models. Finally, an ensemble approach named voting regression combines those three ML models. Performance matrices are generated and the results showed that the ensembled ML model obtains R2 values of 0.75, 0.81, and 0.55 for marine water, freshwater, and brackish

water, respectively, which outperforms the single ML model in predicting all three types of fish production (in tons) in Malaysia.

Chapter 3

Problem Statement and Objectives

3.1 Problem Statement

Development and Comparative analysis of Machine Learning models to forecast the total marine fish production in Kerala.

3.2 Objectives

- To forecast the total marine fish production in Kerala.
- To study the effect of environmental factors on these productions.
- Compare the outcome of various models.

In this project, the agenda at hand is to study the trend of the total marine fish production of Kerala and predict the future production figures using the four machine learning models under study : Linear Regression, Multilayer Perceptron, Random Forest Regression and Radial Basis Fuction. Furthermore the comparison of the outcomes of various models is to be carried out and the model with best prediction capability and least error is to be found.

Over the course of this chapter, the initial part of the section covered the five base papers which were studied in order to complete the project at hand. The architecture, implementation, advantages and disadvantages of each paper has been discussed and in the latter part of the chapter, the problem statement and objectives that the project wishes to fulfil has been described in detail. In the next chapter, the design and implementation of the proposed four models is being described.

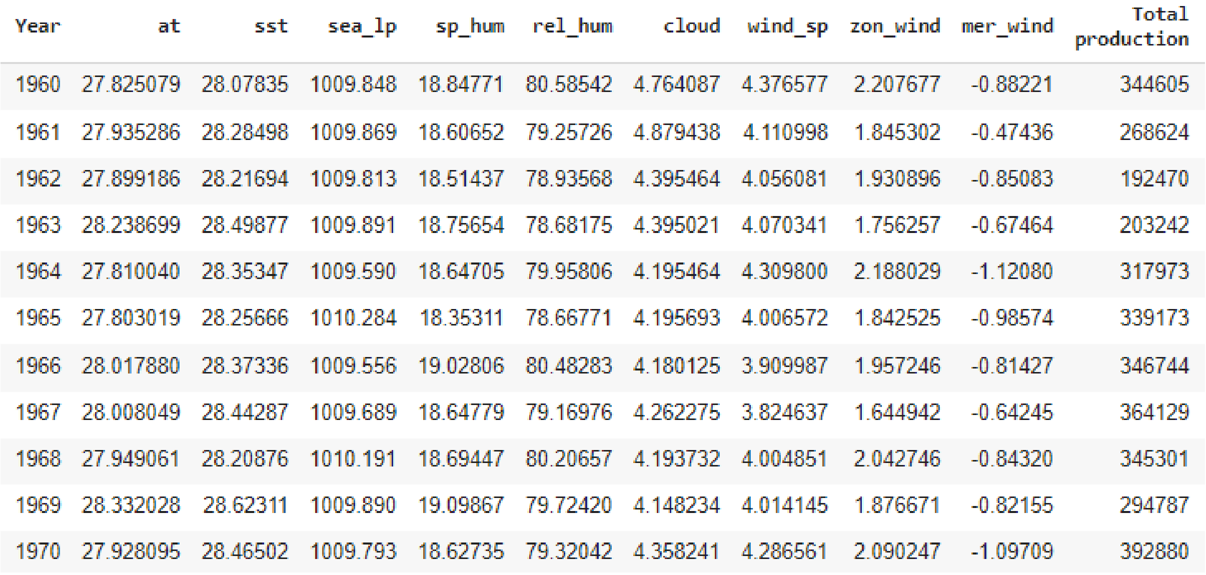
Chapter 4

Design and Implementation

In this chapter, an in-depth analysis of the data-set is being done followed by the data preparation and pre-processing techniques used to prepare the data-set to a manner where it can produce the optimal output from the four given models. This is followed by a discussion on the features extracted from the input data-set. The ML Forecasting Process is also described with focus over the project domain. This is follwed by a detailed description of the various models under study i.e., Linear Regression, Multilayer Perceptron, Random Forest Regression and Radial Basis Function models. And finally the chapter is concluded by stating and defining the various evaluation metrics used to compare the model outcomes.

4.1 DATA-SET

The data-set used in this project is annual the total fish production(in tonnes) figures and values of nine environmental variables related to the domain over the course of 55 years from 1960 - 2014. The various environmental variables colected are Air temperature(degc), Sea Surface temp(degc), Sea level pressure(millibars), Specific humidity(g/kg), Relative humidity(%), Cloudiness(okta), Wind speed(m/s), Zonal wind speed(m/s) and Merid wind speed(m/s). A sample of the data-set is shown in Figure 4.1. This consists of eleven columns, the first one corresponds to the year, the columns two to ten represents the various environment variables and finally the last column corresponds to the total production figures. The environmental variables data is collected from the International Comprehensive Ocean-Atmosphere Data Set(ICOADS). The total marine production data of Kerala is collected from the published reports of Central Marine Fisheries Research Institute(CMFRI), Kochi.



Year	at	sst	sea_lp	sp_hum	rel_hum	cloud	wind_sp	zon_wind	mer_wind	Total production
1960	27.825079	28.07835	1009.848	18.84771	80.58542	4.764087	4.376577	2.207677	-0.88221	344605
1961	27.935286	28.28498	1009.869	18.60652	79.25726	4.879438	4.110998	1.845302	-0.47436	268624
1962	27.899186	28.21694	1009.813	18.51437	78.93568	4.395464	4.056081	1.930896	-0.85083	192470
1963	28.238699	28.49877	1009.891	18.75654	78.68175	4.395021	4.070341	1.756257	-0.67464	203242
1964	27.810040	28.35347	1009.590	18.64705	79.95806	4.195464	4.309800	2.188029	-1.12080	317973
1965	27.803019	28.25666	1010.284	18.35311	78.66771	4.195693	4.006572	1.842525	-0.98574	339173
1966	28.017880	28.37336	1009.556	19.02806	80.48283	4.180125	3.909987	1.957246	-0.81427	346744
1967	28.008049	28.44287	1009.689	18.64779	79.16976	4.262275	3.824637	1.644942	-0.64245	364129
1968	27.949061	28.20876	1010.191	18.69447	80.20657	4.193732	4.004851	2.042746	-0.84320	345301
1969	28.332028	28.62311	1009.890	19.09867	79.72420	4.148234	4.014145	1.876671	-0.82155	294787
1970	27.928095	28.46502	1009.793	18.62735	79.32042	4.358241	4.286561	2.090247	-1.09709	392880

Figure 4.1: Sample Data-Set

4.1.1 Data Preparation

In this study the data preparation steps mainly revolve around data integration and data transformation. The data of the total production figures and the data of various environmental variables were obtained from different sources. These were integrated into a single .csv file inorder to feed as input to the various models under study. The data was further standardised and normalised as per the different model requirements.

4.1.2 Data split

The total data-set as mentioned above contained 55 data points collected annually over the course of the years 1960 - 2014. This data-set is split into training and testing sets. This split is done in a 70%-30% ratio for the Multilayer Perceptron and Radial Basis Function and 80%-20% ratio for the Linear and Random Forest Regression models. The models are then fitted on the training set. Then the testing set is used to evaluate the predictive accuracy of the models.

4.1.3 Variable Selection

Now from the nine different variables available the best combination of variables having highest impact on the total production was selected. Correlation analysis was carried out on the data-set comprising of all nine environmental variables. The study showed that three variables had maximum impact on the dependent or response variable, total production in this case. These are cloud, wind speed and zonal wind. The combination of these three predicate variables had maximum impact on total fish production in Kerala during the period under study.

4.2 ML Forecasting Process

In our project we are using four different models which will be discussed later in this chapter. Each of these models have a basic building block common to any machine learning models as shown in Figure 4.2.

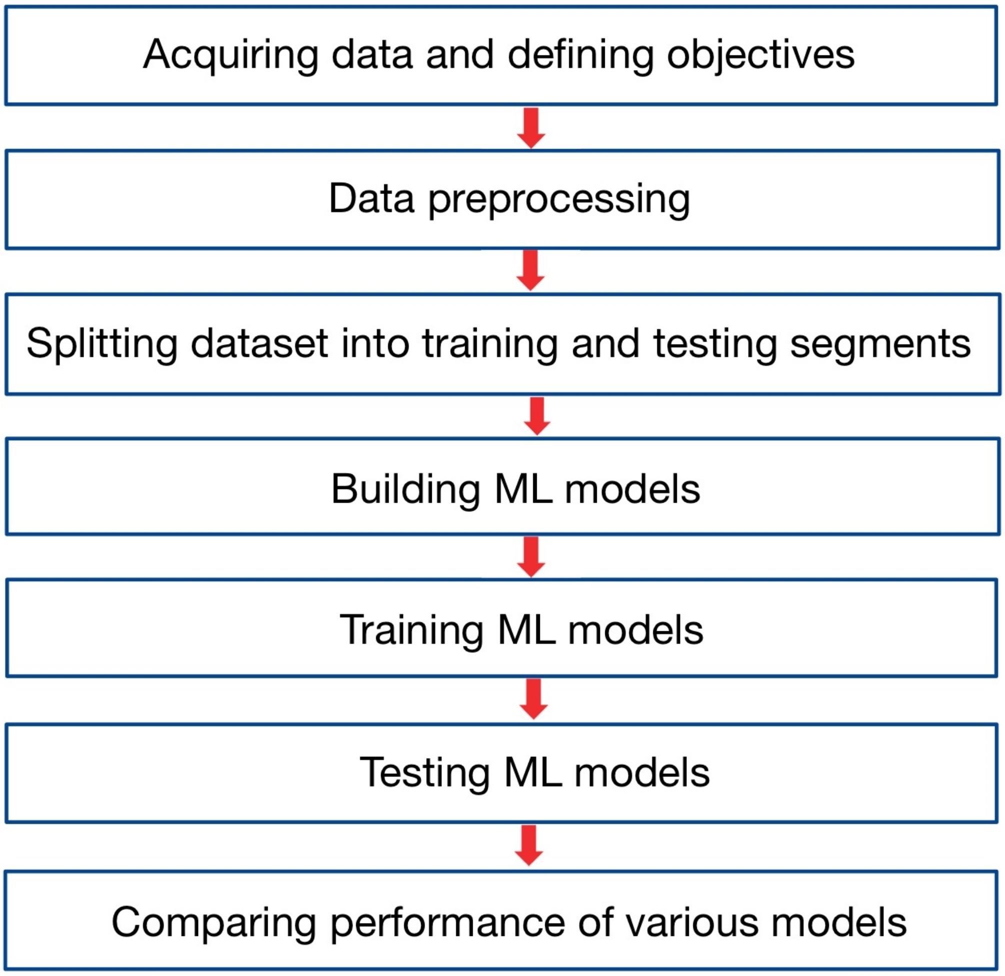

Acquiring data and defining objectives
Data preprocessing
Splitting dataset into training and testing segments
Building ML models
Training ML models
Testing ML models
Comparing performance of various models

Figure 4.2: Design Flow Diagram

First and foremost step is to define the problem domain and acquire sufficient amount of data that has value to our problem statement. So, we must be clear about the objective of the purpose of ML implementation. And these data can be of any format. In this project we are collecting our data mainly in CSV format. Now the collected data is cleaned up and processed as per the model requirements.

The data set is then partitioned into training and testing sets. Now the models are trained and then tested using different cross validation methodologies. The models are fitted using the training data and then the fitted models are evaluated using the testing data.

Each of the models are then evaluated by comparing forecast with actual values. Thus a performance comparison is built up between the models. The best suitable model is then selected based different error factors. Four evaluation metrics are used to compare the model outcomes which will be discussed towards the end of the chapter.

4.3 FORECASTING MODELS

In our project we are mainly trying to forecast the results using four different models, namely

- Linear Regression
- Multilayer Perceptron
- Random Forest Regression
- Radial Basis Function

Linear regression was chosen since it is a simple but powerful algorithm. It is frequently the first algorithm that statisticians evaluate, and the resulting model may be easily examined to reveal information on the linear relationship between input and output attributes. The multilayer perceptron was chosen because it is a general function approximator that can approximate non-linear correlations between input and output, unlike linear regression. The random forest algorithm was chosen since it is also a general function approximator whose model is trained in a profoundly different way than the multilayer perceptron. It is also commonly resistant to overfitting which can be an issue when dealing with user-generated data.

4.3.1 Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

4.3.2 Multilayer Perceptron

A perceptron is a single neuron model that was a precursor to larger neural networks. It is a field that investigates how simple models of biological brains can be used to solve difficult computational tasks like the predictive modeling tasks we see in machine learning. The goal is not to create realistic models of the brain, but instead to develop robust algorithms and data structures that we can use to model difficult problems. The power of neural networks comes from their ability to learn the representation in your training data and how to best relate it to the output variable that you want to predict. In this sense neural networks learn a mapping. Mathematically, they are capable of learning any mapping function and have been proven to be a universal approximation algorithm. The predictive capability of neural networks comes from the hierarchical or multi-layered structure of the networks. The data structure can learn to represent features at different scales and combine them into higher-order features.

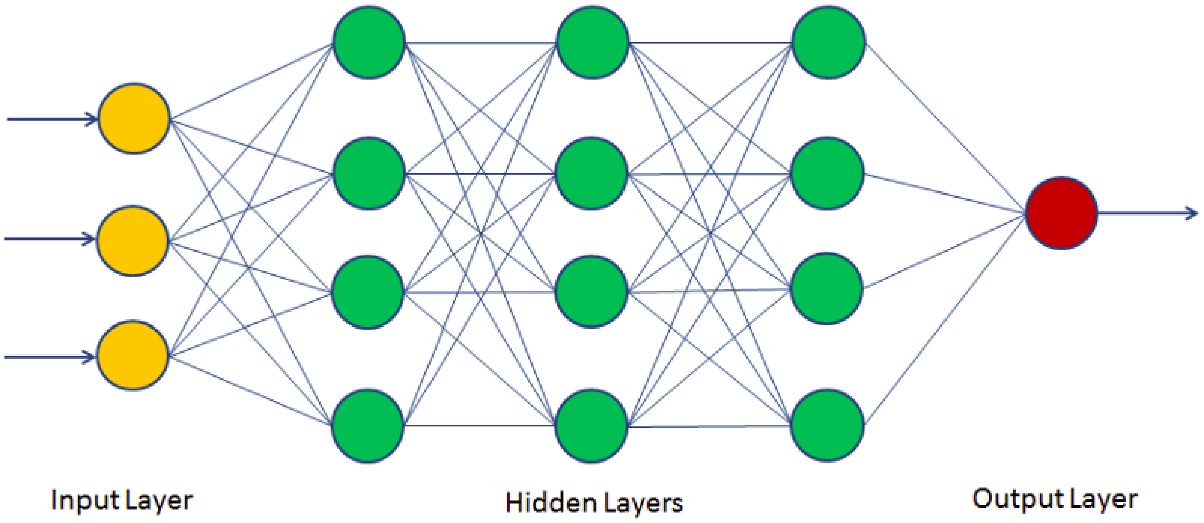


Figure 4.3: Multilayer Perceptron

4.3.3 Random Forest Regression

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models.

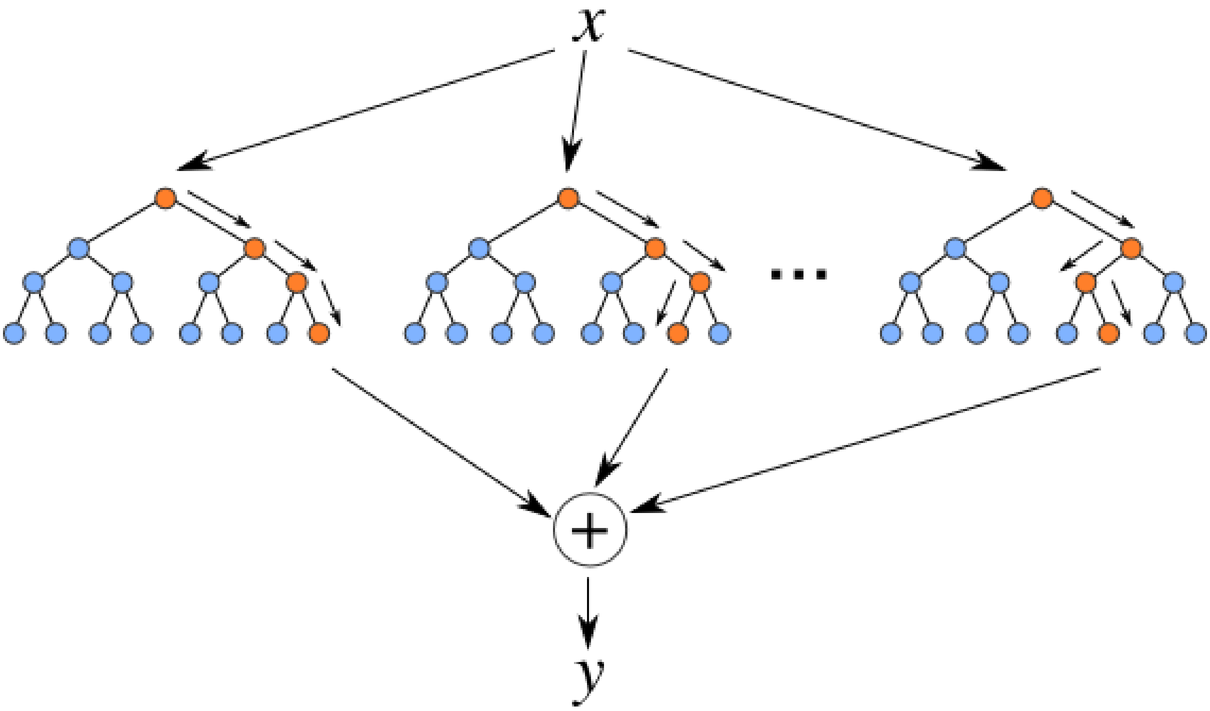


Figure 4.4: Random Forest Regression

4.3.4 Radial Basis Function

RBF networks, a class of feed forward networks have universal approximation capabilities. The design of this network is viewed as a curve fitting approximation problem in a high dimensional space. In its most basic form it involves 3 layers Input layer is made of source nodes that connect the network to its environment. Hidden layer applies a nonlinear transformation from the input space to the hidden space, which is of high dimensionality. Output layer is linear, supplying the response of the network to the activation patterns applied to the input layer. Response of the hidden layer are scaled by the connection weights of the output layer and then combined to produce the network output. In RBF network implementation, the basic functions are usually chosen as Gaussian. The weights connecting the hidden and output units are estimated by linear least squares method.

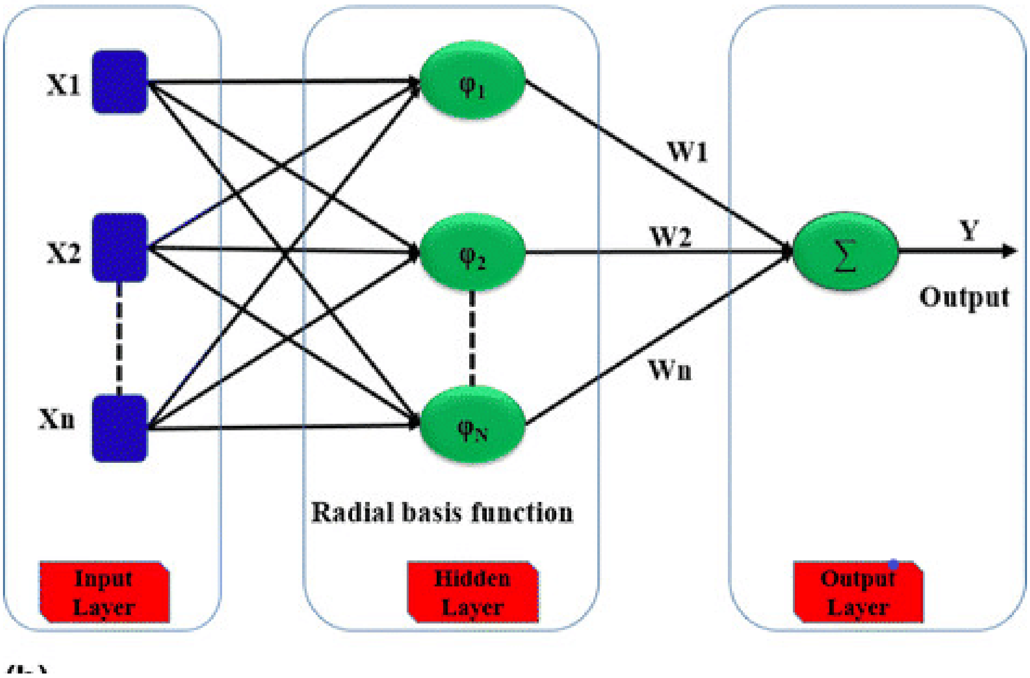


Figure 4.5: Radial Basis Function

4.4 EVALUATION METRICS

Once the model is fitted over the training data-set, the accuracy of the model is to be evaluated by analysing its performance over the test set. This is essential in order to determine the accuracy of performance of the models and thus to finalise if the model can be used feasibly in our specific domain of study. It also makes it possible to compare the performance of the various models and to choose the best one for the particular problem and objectives.

In our study we are mainly using four different evaluation metrics :

- R-squared score
- Mean Absolute Error
- Mean Squared Error
- Root Mean Squared Error

4.4.1 R-squared score

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs. R-squared values range from 0 to 1 and are commonly stated as percentages from 0% to 100%. An R-squared of 100% means that all variations of a dependent variable are completely explained by the independent variable(s).

For practical purposes, the lowest R2 you can get is zero, but only because the assumption is that if your regression line is not better than using the mean, then you will just use the mean value. However if your regression line is worse than using the mean value, the r squared value that you calculate will be negative.

The R2 score can be calculated as follows:

R2 = 1 - $\frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \bar{y}_i)}$

4.4.2 Mean Absolute Error

Mean absolute error (MAE) is a popular metric because, as with Root mean squared error (RMSE), discussed in next subsection, the error value units match the predicted target value units. Unlike RMSE, the changes in MAE are linear and therefore intuitive. MSE and RMSE penalize larger errors more, inflating or increasing the mean error value due to the square of the error value. In MAE, different errors are not weighted more or less, but the scores increase linearly with the increase in errors. The MAE score is measured as the average of the absolute error values. The Absolute is a mathematical function that makes a number positive. Therefore, the difference between an expected value and a predicted value can be positive or negative and will necessarily be positive when calculating the MAE.

The MAE value can be calculated as follows:

MAE = $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

4.4.3 Mean Squared Error

The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points. For every data point, you take the distance vertically from the point to the corresponding y value on the curve fit (the error), and square the value. Then you add up all those values for all data points, and divide by the number of points. The squaring is done so negative values do not cancel positive values. The smaller the Mean Squared Error, the closer the fit is to the data. The MSE has the units squared of whatever is plotted on the vertical axis.

The MSE value can be calculated as follows:

MSE = $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

4.4.4 Root Mean Squared Error

The last metric that we calculate is the Root Mean Squared Error (RMSE). It is just the square root of the mean square error. The RMSE is thus the distance, on average, of a data point from the fitted line, measured along a vertical line. The RMSE is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient.

The RMSE value can be calculated as follows:
RMSE = $\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}$

The meanings of various representations used above are explained below:
n is the number of data points
y_i is the observed value
$\hat{y}_i$ is the predicted value
$\bar{y}_i$ is the mean value

4.5 Requirements

The various hardware and software requirements of the project are as follows. As for the hardware the Colab Workstation was utilized for implementation and analysis of two regression models, Multiple Linear Regression and Random Forest Regression. Other than that a basic computing hardware is only required. The other two Neural Network models, Multilayer Perceptron and Radial Basis Function were implemented on the IBM Statistical Package for Social Sciences(SPSS) software. Other software requirements include the Python libraries like matplotlib, numpy, pandas, sklearn etc.

Chapter 5

Results and Discussion

In this chapter, the results obtained from the four models have been specified in detail.The models have been applied to predict the total marine fish production in Kerala for the period 1960-2014 and to analyse their relationships with environmental factors. The performance of all models was tested by estimating the coefficient of determination R2 between the predicted versus the observed values, by analysing the statistics of their residuals and finally by cross-validating their predictions with unused, during the modelling procedure, test data.

5.1 PRODUCTION TRENDS

The annual Marine fish production in Kerala during 1960-2014 varied between 1.92 lakh tonnes to 8.39 lakh tonnes (Figure 5.1). There has been a spectacular growth in the marine fisheries sector of the state due to fisheries friendly government policies, well developed harvest and post harvest infrastructure and increased demand for sea food both in the domestic and export markets. Kerala has been in the forefront in absorbing innovative and new technologies in fishing practices, which has led marine fisheries to take a complex structure. A growing demand for fish has fuelled a rapid increase of fishing effort in terms of fishing hours through multi-day fishing by the mechanized sector, extension of fishing grounds by the mechanised and motorized fishing crafts.

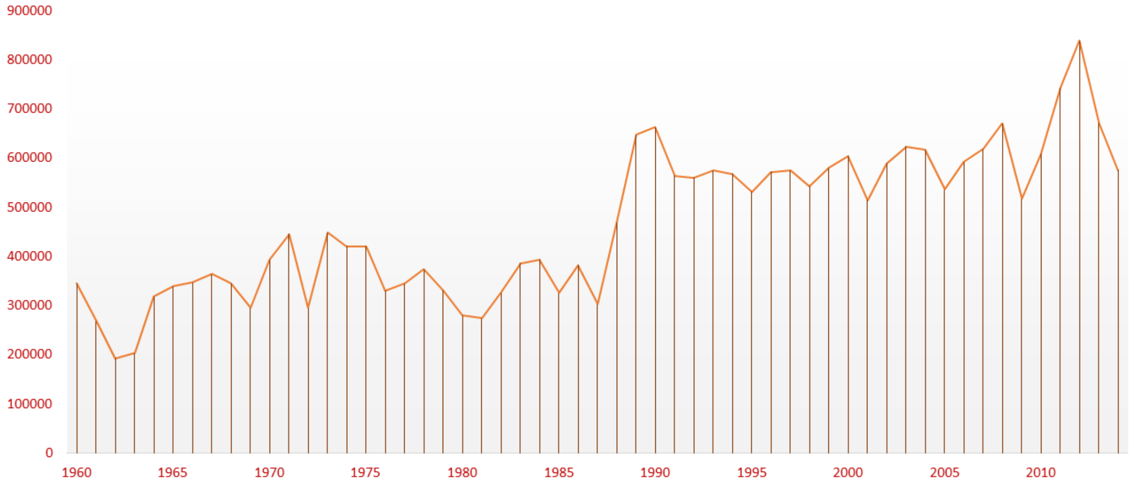


Figure 5.1: Annual Marine Fish Production in Kerala

In Figure 5.1, the X axis corrresponds to the years whose data are collected i.e., 1960 - 2014. Y axis denotes the total production values of the corresponding years.

The environmental variables windspeed, zonal wind and cloud were used as explanatory variables to develop the ML models. Correlation analysis was carried to identify the variables from the nine environmental variables, The average windspeed during 1960-2014 was 4.91m/s and varied between 3.84 and 6.14. The variability and trend in the environmental data are given in Figure 5.2.

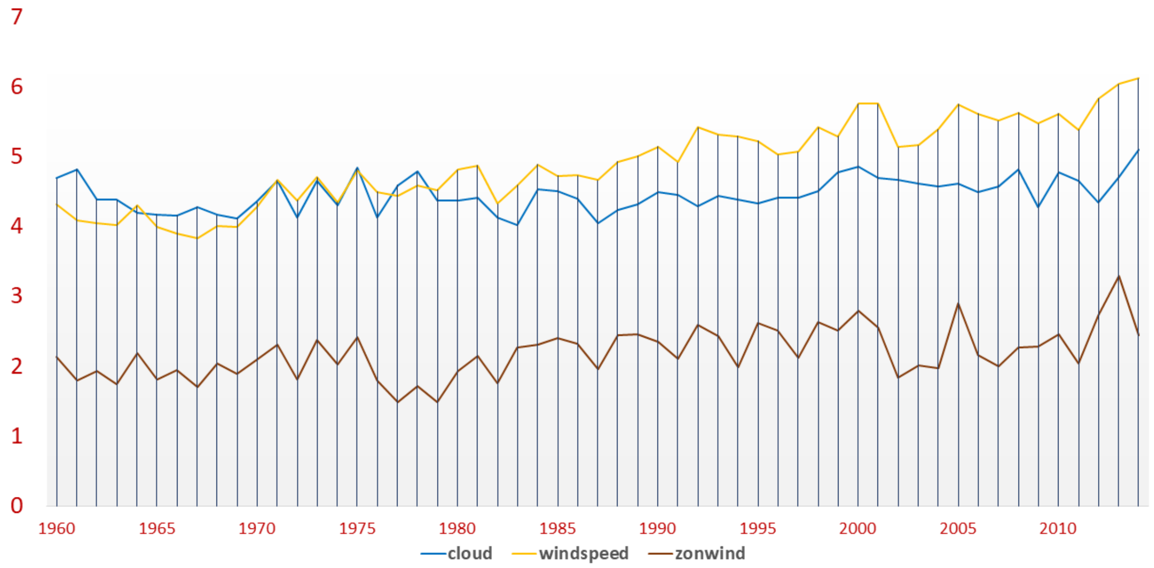
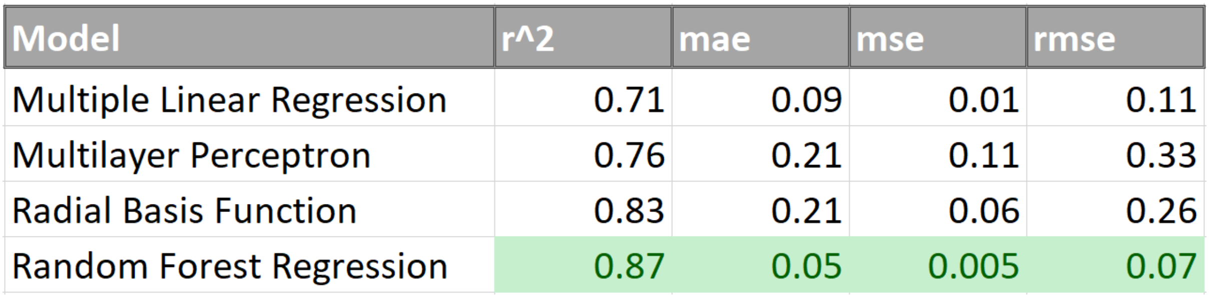


Figure 5.2: Trends in environmental variables

In Figure 5.2, the X axis corrresponds to the years whose data are collected i.e., 1960 - 2014. Y axis denotes the numrical values of the three variables cloud(in okta), wind speed(in m/s) and zonal wind(in m/s) during the corresponding years. The brown plot denotes zonal wind, yellow represents wind speed and the blue plot is of cloud.

5.2 Result Evaluation



Model	r^2	mae	mse	rmse
Multiple Linear Regression	0.71	0.09	0.01	0.11
Multilayer Perceptron	0.76	0.21	0.11	0.33
Radial Basis Function	0.83	0.21	0.06	0.26
Random Forest Regression	0.87	0.05	0.005	0.07

Figure 5.3: Result Evaluation

Figure 5.3 represents the values of the different evaluation metrics used in our study for the four models used. The R2 score is mainly used to analyse the predictive power or accuracy of the models and the different error terms are used to validate the same.

The analysis of the various models, the observations and conclusions are summarised in the following sections.

5.3 Multiple Linear Regression

The Multiple Linear Regression model was used to capture the linear correlations between the predicate variables and the dependent variable i.e., total production. The full data-set was split in a ratio of 80% training data and 20% testing data. The multiple linear regression model from sklearn was then used to fit the training data. The fit model has the intercept value as 0.0968406 and the coefficients corresponding to the three variables in the order cloud, wind speed and zonal wind as -0.06887035, 0.70943185 and 0.06296317 respectively. After the model is fit on the training set, the accuracy of the model is evaluated using the test set. In this study, Multiple Linear Regression model produced the least R2 score among the four models used. It generated a R2 score of .71. This means that model could explain 71% of the variations in the total fish production in Kerala.

5.4 MULTILAYER PERCEPTRON

The Multilayer Perceptron (MLP) neural network model was trained with a back-propagation learning algorithm which uses the gradient descent to update the weights towards minimizing the error function. The data were randomly assigned to training 70% and testing 30% subsets. The training dataset is used to find the weights and build the model. The testing data is used to find errors and prevent overtraining during the training mode. Before training, all covariates were normalized which returns values between 0 and 1.

The network diagram that used to predict the production using windspeed, cloud and zonal wind, is shown in Figure 5.4. The diagram shows the 3 input nodes, one hidden nodes and the output node representing production.

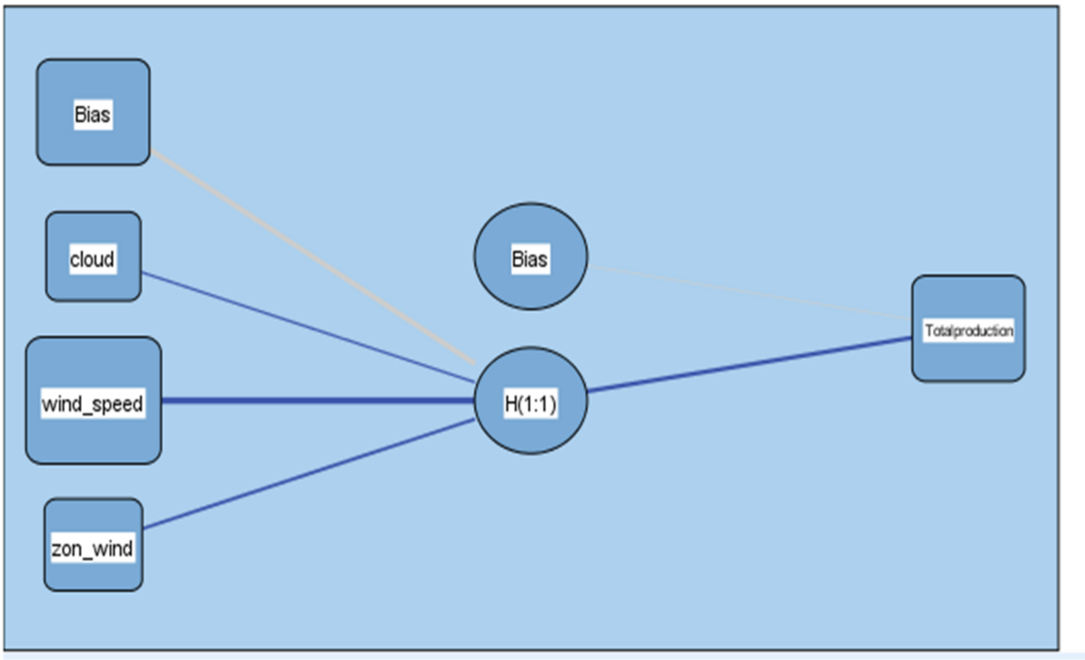


Figure 5.4: Network diagram of Multilayer perceptron

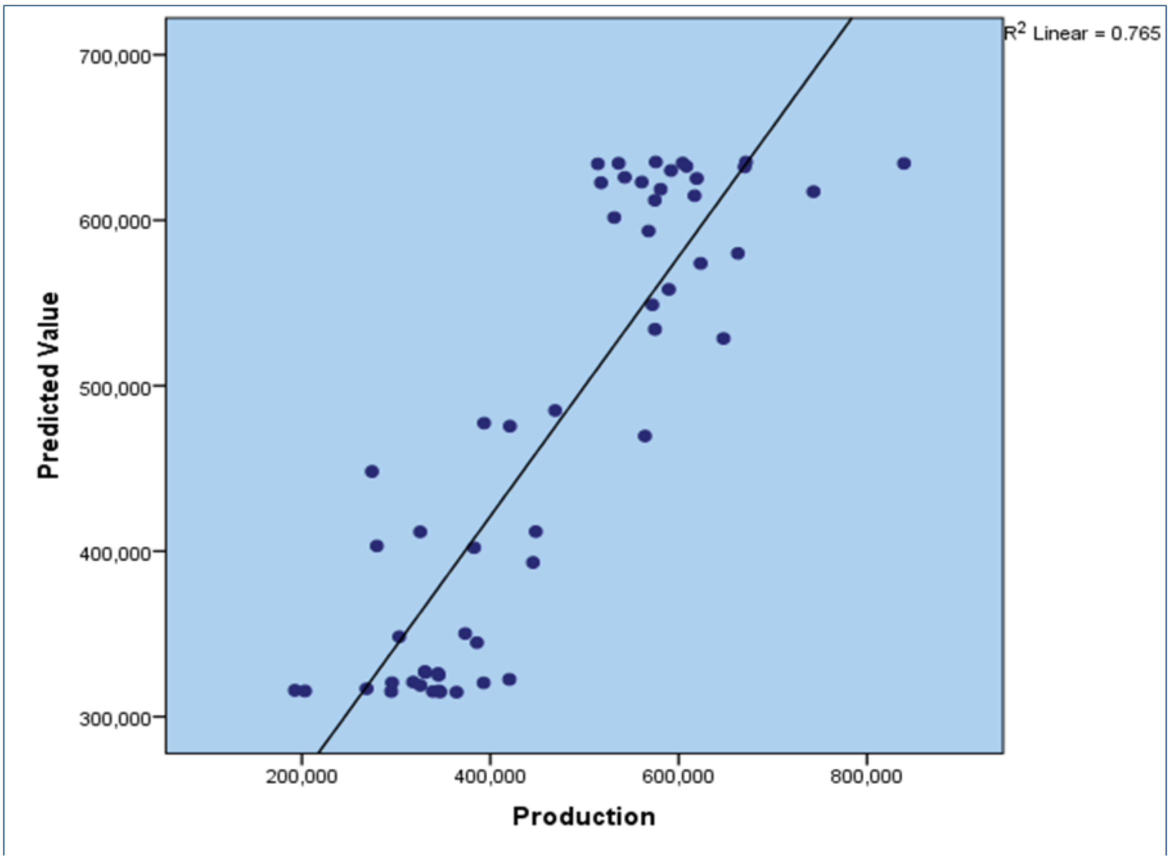


Figure 5.5: Predicted values of production using multilayer perceptron

Figure 5.5 displays the predicted values of production using the model. The X axis denotes the actual production values and the Y axis denotes the production values predicted by the model. These values are given in tonnes.

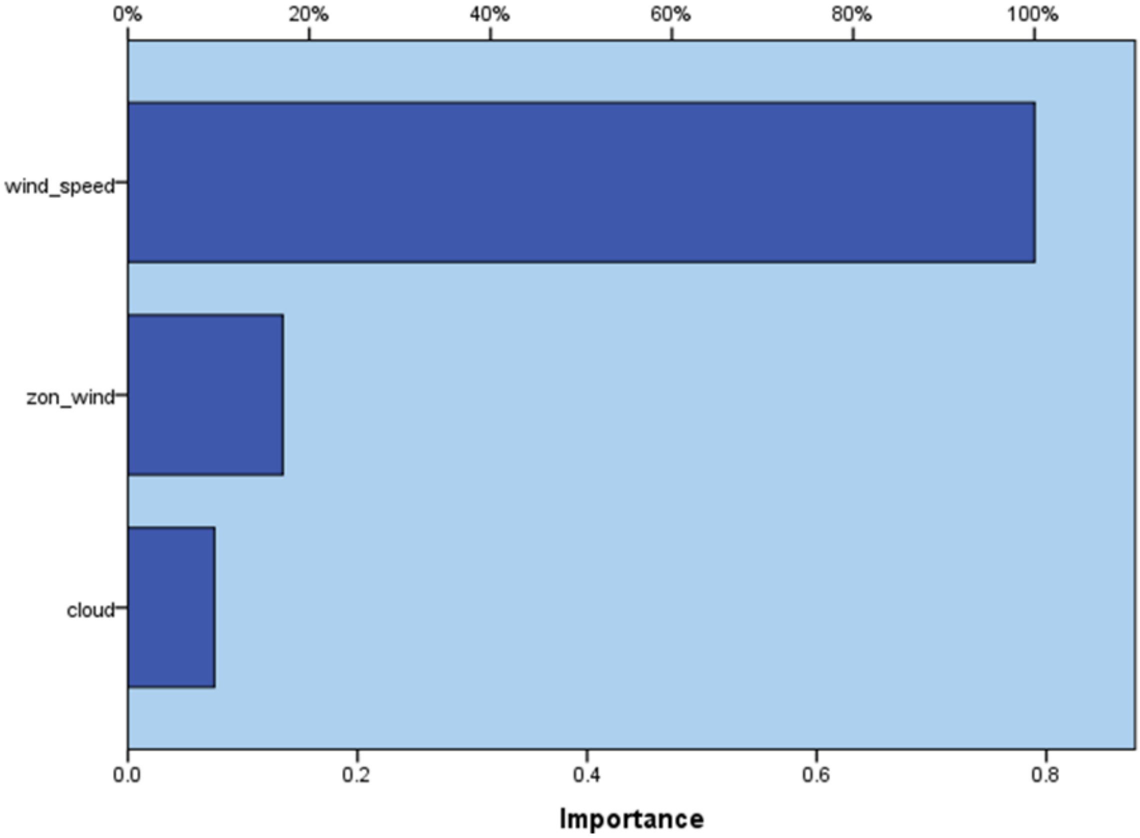


Figure 5.6: Normalised importance of the variables

Figure 5.6 depicts the importance of the variables, i.e how sensitive is the model is the change of each input variable.From the chart is apparent that windspeed have the greatest effect on production and cloud is having the least predictive power.

5.5 RANDOM FOREST REGRESSION

The Random Forest Regression turned out to best model in terms of prediction accuracy among all the four models used in the study. This model produced an R2 score of .87 and also had the least values for all the error terms used. Here also the full data-set was split in a ratio of 80% training data and 20% testing data. Hyperparameter tuning was performed to find out the best set of parameters. Only number of decision trees varied from the default parameter values of the Random Forest Regression model of sklearn. The number of trees was determined by gradually incrementing the number of trees till a point after which the R2 score started decreasing. The number of trees that gave best result in our case is 142.

5.6 RADIAL BASIS FUNCTION

For Radial Basis Function, the target variable was the total marine fish production from Kerala for the period 1960-2014 and the same set of environmental variables were used for prediction. For radial basis function analysis, 80% cases were assigned to the training sample, 20% to the testing sample. Network information is given in Figure 5.7.

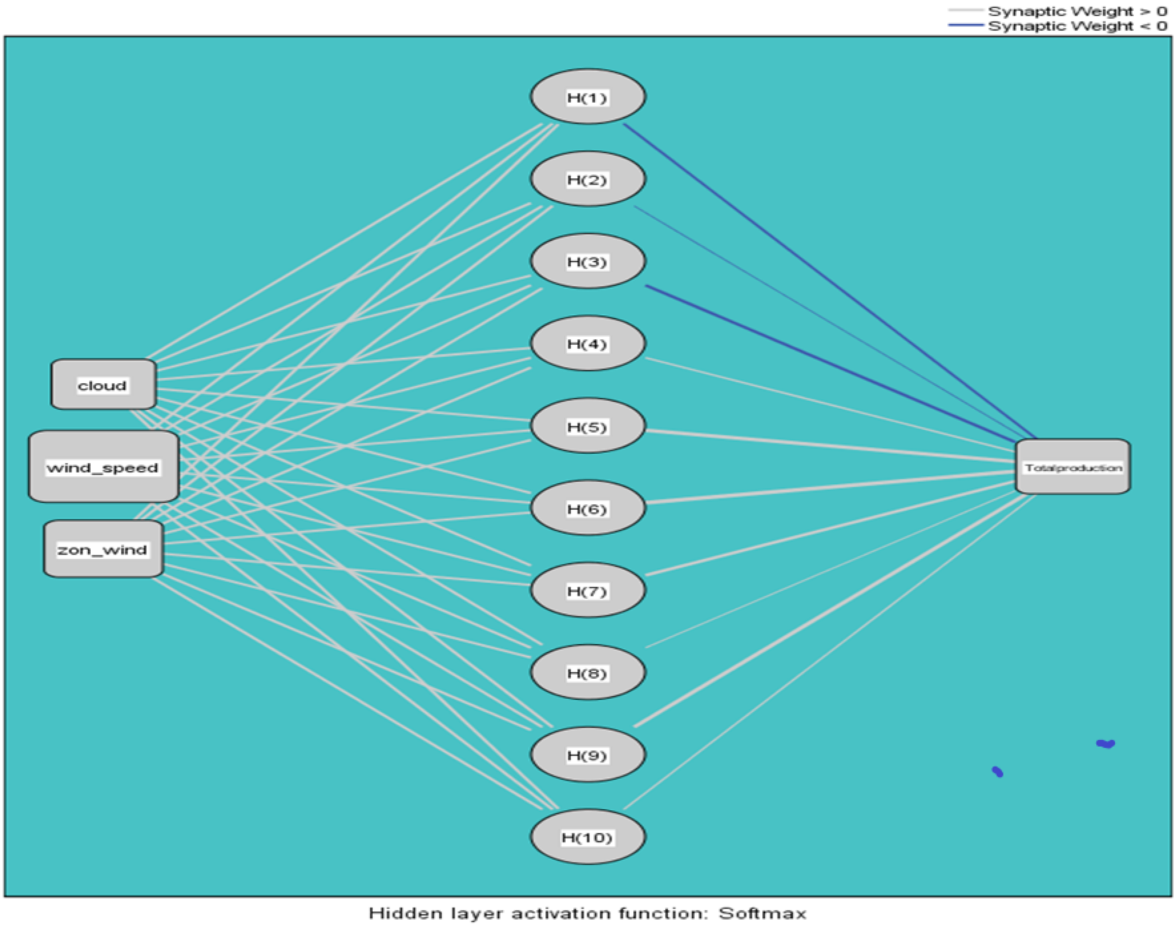


Figure 5.7: Network Architecture for Radial basis function

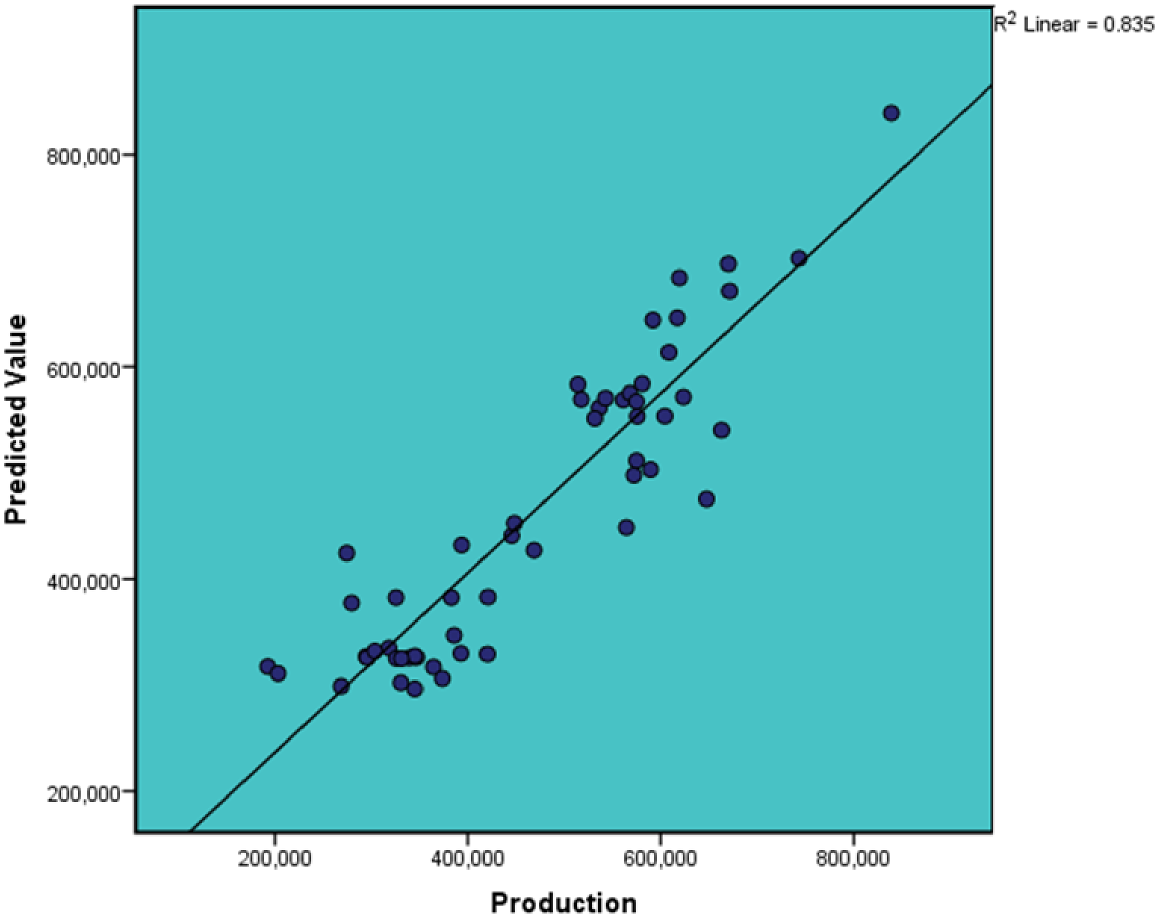


Figure 5.8: Oserved and predicted values using Radial Basis Function

Figure 5.8 shows the comparison between real and estimated production for RBF, which gives an R2 value of 83.5% .he X axis denotes the actual production values and the Y axis denotes the production values predicted by the model. These values are given in tonnes.

5.7 Prediction Graph

The graph between actual and predicted values is one of the simplest ways to analyse the performance of a model. These graphs are generated for each of the four models.

5.7.1 Multiple Linear Regression

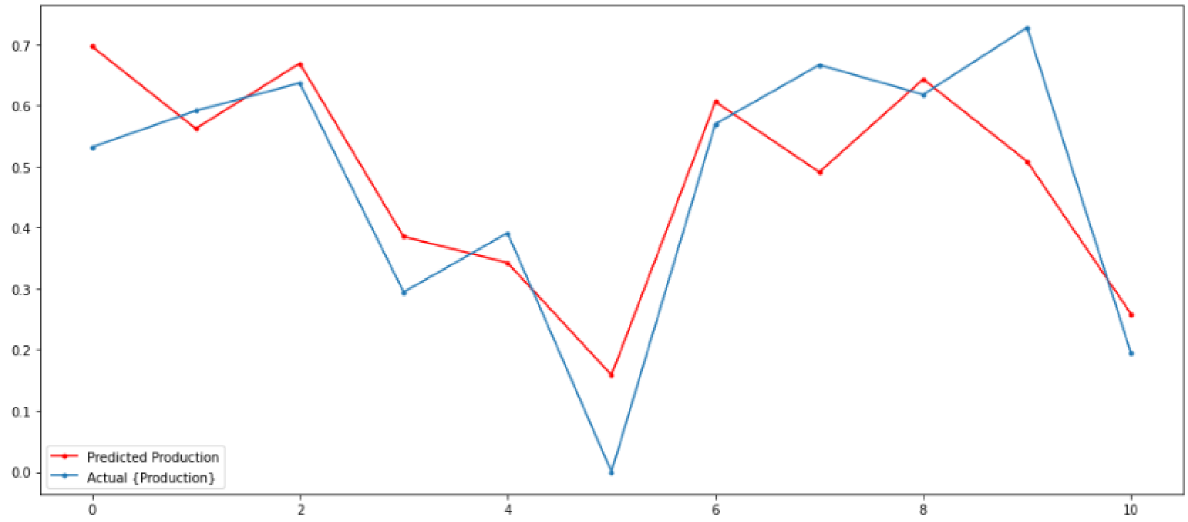


Figure 5.9: Prediction Plot 1

Figure 5.9 represents the prediction plot of Multiple Linear Regression. X axis denotes the years in the test set. Y axis denotes the normalised production value.The blue plot denotes the Actual Production and the red plot represents the Predicted Production.

5.7.2 Multilayer Perceptron

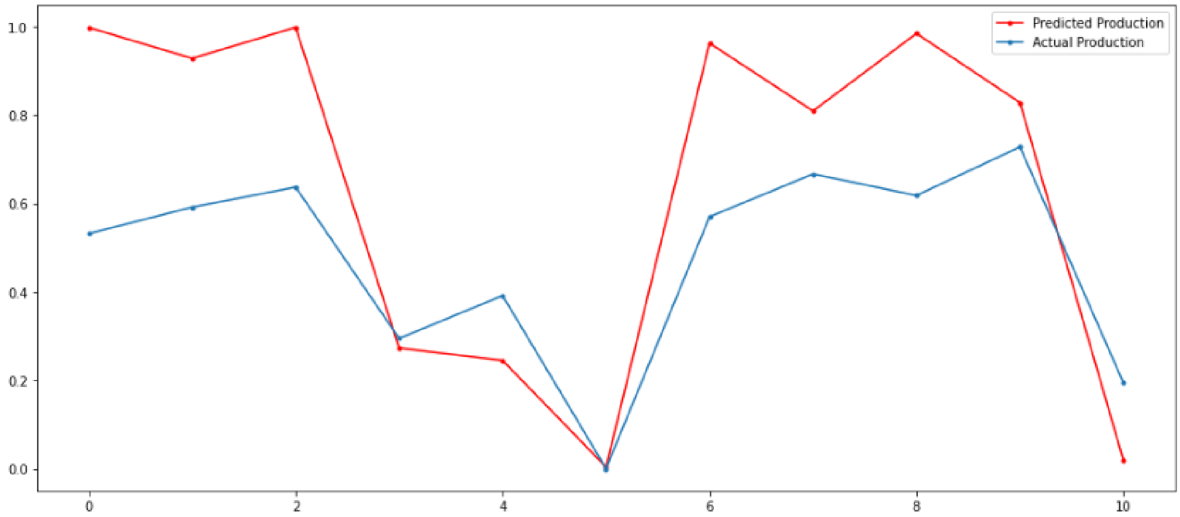


Figure 5.10: Prediction Plot 2

Figure 5.10 represents the prediction plot of Multilayer Perceptron. X axis denotes the years in the test set. Y axis denotes the normalised production value.The blue plot denotes the Actual Production and the red plot represents the Predicted Production.

5.7.3 Radial Basis Function

Figure 5.11 represents the prediction plot of Radial Basis Function. X axis denotes the years in the test set. Y axis denotes the normalised production value.The blue plot denotes the Actual Production and the red plot represents the Predicted Production.

5.7.4 Random Forest Regression

Figure 5.12 represents the prediction plot of Random Forest Regression. X axis denotes the years in the test set. Y axis denotes the normalised production value.The blue plot denotes the Actual Production and the red plot represents the Predicted Production.

From the graphs also it is clearly evident that the Actual and Prediction graphs are very

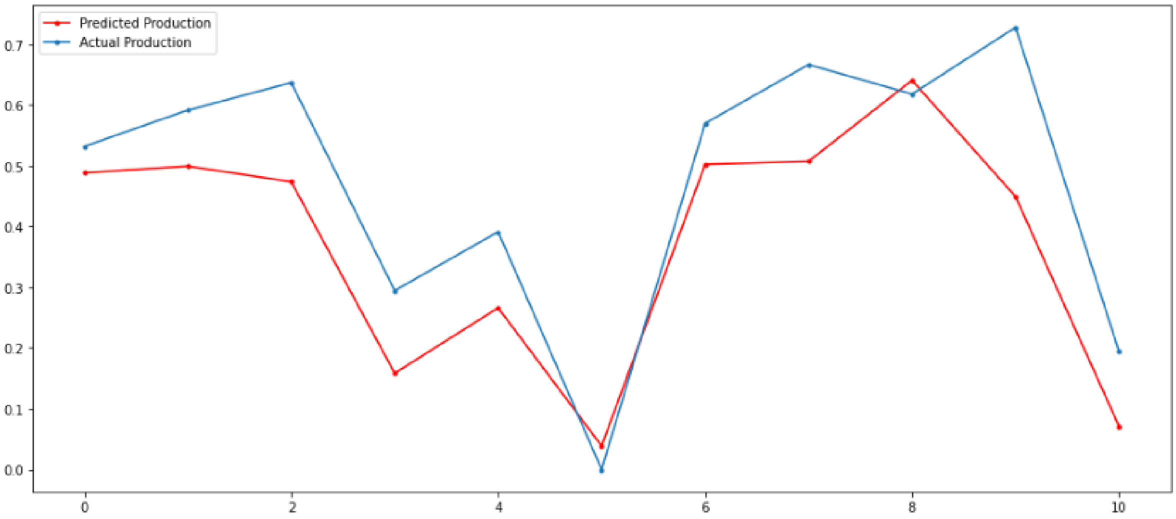


Figure 5.11: Prediction Plot 3

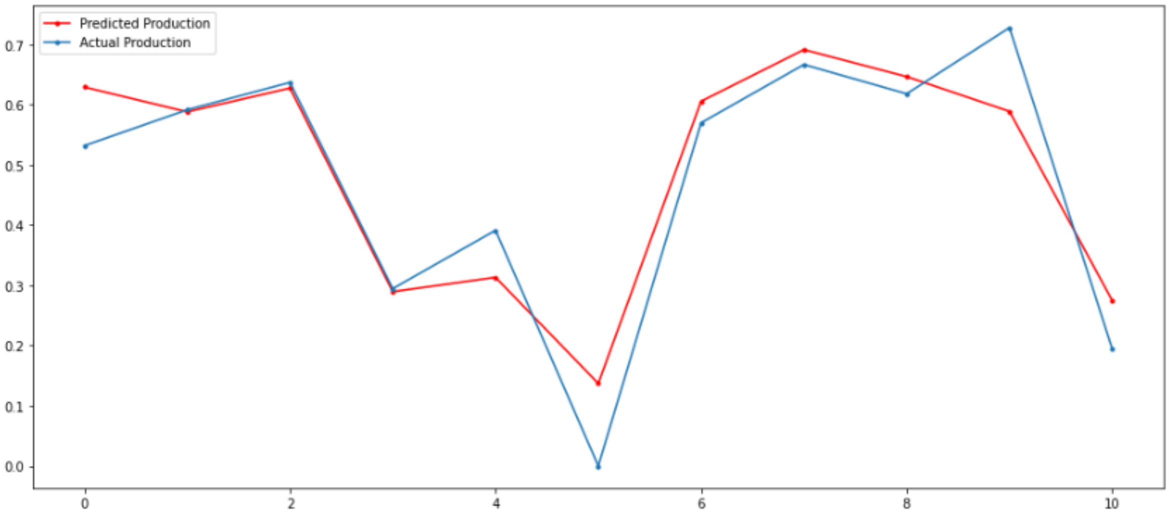


Figure 5.12: Prediction Plot 4

much closer in the Random Forest Regression. The R2 score and other error measures also suggest that the model with highest predictive power is the Random Forest Regression model.

5.8 Discussion

Marine fisheries resources are invisible, frequently migrating and easily affected by the changes in the sea. Moreover, fish is a natural resources that are renewable - they are capable of growth but are not inexhaustible. These characteristics make it unique and complex and hence difficult to monitor, manage and intervene. Also the populations of aquatic organisms are generated and maintained through complex biotic and abiotic interactions within the marine ecosystem. Apart from environmental factors, fisheries is affected by the productivity of the seas, the ecosystem and species interactions, pollution, prey-predator relationships etc. Fishing is also an important economic activity and the catches landed are highly dependent on the number of fishing boats operated, the cost of fishing and the new methods of fishing. Thus, the marine ecosystems remain in a constant imbalance, and they are characterized by the presence of high stochastic levels and non-linear relations. Sometimes, the non-linear relations are not evident and present difficulties for modeling. However, during recent years, artificial intelligence has become a technique supporting both the management of large databases and the use of algorithms, which although complex in structure, give results quite easy to interpret. The purpose of the use of forecasting models in fisheries is to provide to those responsible for the management of resources and users, with the information on the biological and/or environmental effects of fisheries on stocks from a future perspective.

Chapter 6

Conclusion and Future Scope

Forecasting fish production is a critical element tool for fisheries managers and policymakers to make short-term quantitative recommendations for fisheries management. In this study, it is proposed to develop models for forecasting of fish production in Kerala and the catch of major fish species using different machine learning tools. Accurate forecasts of 1–2 years may provide useful information to fishery resource managers, fishermen, market managers and the fishing industry. The performance of the models could be further enhanced by the inclusion of additional information on rainfall, sea-surface temperature and other environmental variables. This study is expected to provide an ML model-based framework that delivers a reliable and accurate estimation of fish production in Kerala.

As a future scope the specialisation of this production for selected species can also have greater impact on the fishing domain in Kerala. Our project initially aimed to conduct the prediction of certain major species like Oil Sardine, Mackerel, Perches, Prawns and Anchovies. But we were not able to conduct the same due to the unavailability of proper data-sets corresponding to the same. But these few species contribute a larger portion of the total fish production in Kerala. So analysis of these species can help a lot in making better decisions related to this domain.