International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 01 Jan. 2025, pp: 373-379 www.ijaem.net ISSN: 2395-5252

Comparative Study between AR-MEM Model and Decision Tree

Aswini Shaji Archa, Libya Thomas
Department of Electronics and Communication College of Engineering Trivandrum, Affiliated to APJ Abdul Kalam Technological University
Department of Electronics and Communication College of Engineering Trivandrum, Affiliated to APJ Abdul Kalam Technological University

Date of Submission: 15-01-2025 Date of Acceptance: 25-01-2025

ABSTRACT—Predictive modeling is understanding complex sys- tems and decision-making informed by the parameter space of various domains. The aim of this study is to compare two predictive models, the Auto-Regressive Maximum Entropy Model (ARMEM) and the Decision Tree model, for predicting the performance of a specific variable, Surface Ocean Direction, from a dataset. The dataset, obtained through High-Frequency Radar (HFR) measurements around Koko Head, was taken as a case study to test these models.
ARMEM is a time-series model that merges the autoregressive methods with the principle of maximum entropy, thus being highly appropriate for high-resolution spectral analysis and noisy or incomplete data. On the other hand, the Decision Tree model works through recursive partitioning of data and thereby provides intuitive, interpretable predictions through capturing the underlying linear and nonlinear relationships.
Keywords-: AR-MEM, Decision Tree, HFR Ocean Current Data

I. INTRODUCTION

This project aims to do a comparative study between two models - Auto Regressive Maximum Entropy Model and Decision Tree, by predicting the variable ‘Surface Current Direction’ in the HFR Oceanic Dataset.

Using the dataset collected from the University of Hawaii’s Global High-Frequency Data Repository, the study aims to provide an analysis between the models.

The ability to understand and predict ocean surface currents is very important for many civil, environmental, and scientific applications, including maritime navigation, pollution control, and ecological studies. High-frequency radar (HFR) systems have become a well-established and reliable tool for mapping ocean surface currents.

This particular comparative study is done between the two models by comparing the Mean Squared Error (MSE), Mean Average Error (MAE), and R-squared (R^2); As well as by plotting graphs between the actual value and predicted value, the variance plot and the residual plot. Ultimately this study aims to contribute to a better understanding of machine learning and statistical modeling in various scientific and practical contexts.

This work introduces the contributions of:

A. Data Source

The dataset was obtained from National Centres for Envi- ronmental Information. Specifically, from the Surface Ocean velocities obtained by HF radar from stations located along coastal waters of Hawaii, North Slope Alaska, Puerto Rico/Virgin Islands, eastern US/Gulf of Mexico and western US.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 01 Jan. 2025, pp: 373-379 www.ijaem.net ISSN: 2395-5252

Fig. 1. Dataset in its Orginal Form.

The orginal dataset was obtained in .rdl format, and for ease of implementation, it was converted into .csv format.

Details of the dataset:-

- %FileType: LLUV rdls : The radial component data of surface currents collected by High-Frequency Radars.
- Dataset originates from the University of Hawaii’s Global High-Frequency Data Repository, ie this dataset is focused in Hawaii; Specifically, the radar site is named ”Koko Head.”
- The geographical origin of the radar site in decimal degrees is latitude: 21.2610° and longitude: -157.7030°.
- The bandwidth of the transmitted radar signal is 100 kHz.
- The dataset focuses on the time 2023-12-01 00:00:00 (December 1, 2023, at midnight UTC)

B. Features and Variables

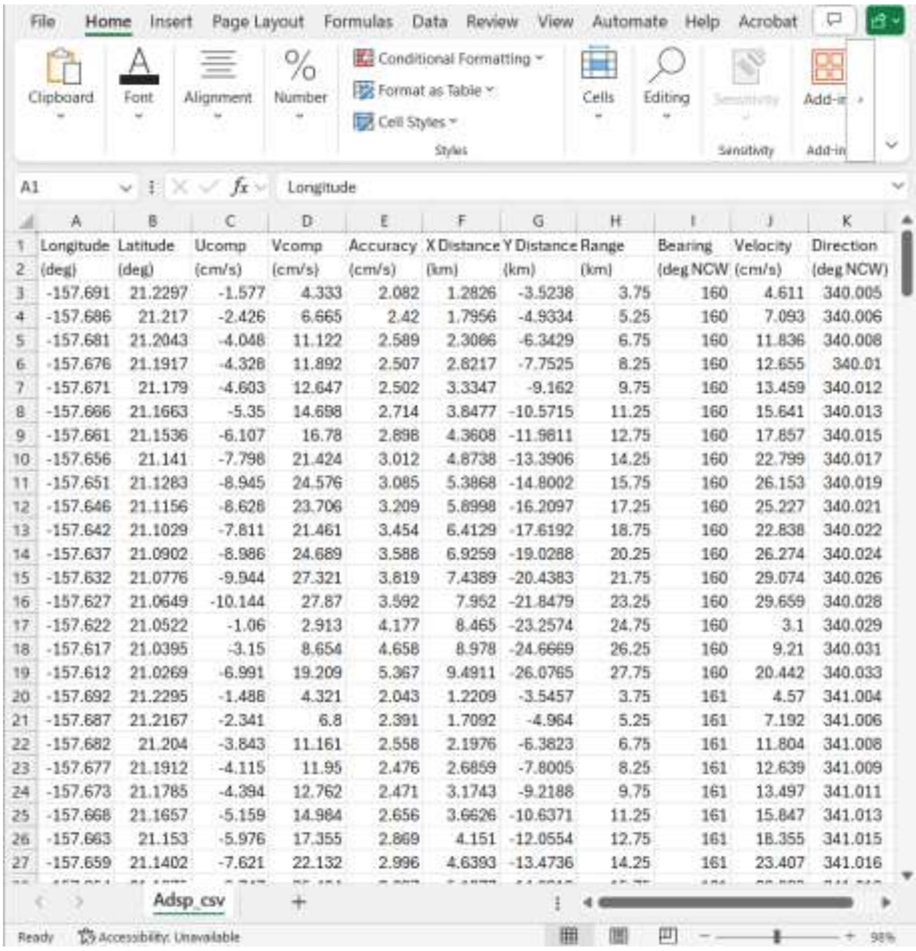
The above mentioned dataset contains the following vari- ables:

- Longitude and latitude coordinates of measurements.

-Components of velocity in the U (eastward) and V (north- ward) directions.

- Error accuracy associated with the measurement.
- Distance in the X (eastward) and Y (northward) directions.
- Radial distance from the radar in kilometers.
- Bearing (direction) from the radar to the measurement point.
- Heading direction.
- Total velocity magnitude.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 01 Jan. 2025, pp: 373-379 www.ijaem.net ISSN: 2395-5252



	A	B	C	D	E	F	G	H	I	J	K
1	Longitude	Latitude	Ucomp	Vcomp	Accuracy	X Distance	Y Distance	Range	Bearing	Velocity	Direction
2	(deg)	(deg)	(cm/s)	(cm/s)	(cm/s)	(km)	(km)	(km)	(deg NCW	(cm/s)	(deg NCW)
3	-157.691	21.2297	-1.577	4.333	2.082	1.2826	-3.5238	3.75	160	4.611	340.005
4	-157.686	21.217	-2.426	6.665	2.42	1.7956	-4.9334	5.25	160	7.093	340.006
5	-157.681	21.2043	-4.048	11.122	2.589	2.3086	-6.3429	6.75	160	11.836	340.008
6	-157.676	21.1917	-4.328	11.892	2.507	2.8217	-7.7525	8.25	160	12.655	340.01
7	-157.671	21.179	-4.603	12.647	2.502	3.3347	-9.162	9.75	160	13.459	340.012
8	-157.666	21.1663	-5.35	14.698	2.714	3.8477	-10.5715	11.25	160	15.641	340.013
9	-157.661	21.1536	-6.107	16.78	2.898	4.3608	-11.9811	12.75	160	17.857	340.015
10	-157.656	21.141	-7.798	21.424	3.012	4.8738	-13.3906	14.25	160	22.799	340.017
11	-157.651	21.1283	-8.945	24.576	3.085	5.3868	-14.8002	15.75	160	26.153	340.019
12	-157.646	21.1156	-8.628	23.706	3.209	5.8998	-16.2097	17.25	160	25.227	340.021
13	-157.642	21.1029	-7.811	21.461	3.454	6.4129	-17.6192	18.75	160	22.838	340.022
14	-157.637	21.0902	-8.986	24.689	3.588	6.9259	-19.0288	20.25	160	26.274	340.024
15	-157.632	21.0776	-9.944	27.321	3.819	7.4389	-20.4383	21.75	160	29.074	340.026
16	-157.627	21.0649	-10.144	27.87	3.592	7.952	-21.8479	23.25	160	29.659	340.028
17	-157.622	21.0522	-1.06	2.913	4.177	8.465	-23.2574	24.75	160	3.1	340.029
18	-157.617	21.0395	-3.15	8.654	4.658	8.978	-24.6669	26.25	160	9.21	340.031
19	-157.612	21.0269	-6.991	19.209	5.367	9.4911	-26.0765	27.75	160	20.442	340.033
20	-157.692	21.2295	-1.488	4.321	2.043	1.2209	-3.5457	3.75	161	4.57	341.004
21	-157.687	21.2167	-2.341	6.8	2.391	1.7092	-4.964	5.25	161	7.192	341.006
22	-157.682	21.204	-3.843	11.161	2.558	2.1976	-6.3823	6.75	161	11.804	341.008
23	-157.677	21.1912	-4.115	11.95	2.476	2.6859	-7.8005	8.25	161	12.639	341.009
24	-157.673	21.1785	-4.394	12.762	2.471	3.1743	-9.2188	9.75	161	13.497	341.011
25	-157.668	21.1657	-5.159	14.984	2.656	3.6626	-10.6371	11.25	161	15.847	341.013
26	-157.663	21.153	-5.976	17.355	2.869	4.151	-12.0554	12.75	161	18.355	341.015
27	-157.659	21.1402	-7.621	22.132	2.996	4.6393	-13.4736	14.25	161	23.407	341.016

Fig. 2. Dataset in .csv format.

In this project, Direction is the Dependent Variable, and the others are the Independent Variables.

C. Models

The aim of this project is to do an analysis using two mod- els: Auto-Regressive Maximum Entropy Model and Decision Tree.

- **AR-MEM Model**

The Auto-Regressive Maximum Entropy Model (ARMEM) is a powerful tool for statistical analyses, combining elements of autoregressive modeling and the maximum entropy principle to apply to time series. It uses the data in a linear form of its preceding values, and it can grasp temporal dependencies very well, so it can capture underlying patterns.

A salient feature of ARMEM is the application of the maxi- mum entropy principle, which makes spectral estimates the least biased and most uniform. This approach minimizes assumptions about the underlying data distribution, making the model particularly effective in handling noisy or incomplete datasets.

ARMEM is applied in signal processing, geophysics, and oceanography, where high-resolution spectral analysis is criti- cal. By combining autoregressive modeling with the maximum entropy principle, ARMEM provides a robust and computa- tionally efficient framework for analyzing complex time-series data.

- **Decision Tree**

A decision tree is a supervised learning algorithm that can be used for classification and regression tasks. It is structured like a flowchart, where each internal node represents a decision or split based on the value of a specific feature, and each leaf node corresponds to a final outcome or prediction. The splits are made recursively, dividing the data into smaller subsets until a stopping criterion.

The working mechanism of a decision tree is akin to answering a series of yes/no or binary questions. At each step, the algorithm evaluates the feature that best separates the data into homogenous groups.

Decision Trees are applied in many fields. For instance, in marketing, they are widely used for customer segmentation. Based on the analysis of customer attributes such as age, purchasing behavior, and income, decision trees can segment customers and provide the best strategies tailored to each group. This makes them very useful in making data-driven decisions in different industries.

D. Evaluation Metrics

These above mentioned models are applied to analyze the relationship between various input parameters (such as velocity, range, bearing, etc.) and the oceanic current prediction, with the objective of comparing their performance in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE), R-Squared (R^2).

- **Mean Absolute Error (MAE)**

Measures the magnitude of the errors in a set of average predictions. It is the average of the absolute differences between predicted and actual values.

- **Mean Squared Error (MSE)**

Measures the average of the squared differences between predicted and actual values, giving more weight to large errors.

- **R-squared (R^2)**

Measures how well the models predictions match the actual data. if R^2 is close to 1, it means the model is doing a great job at predicting, while a value closer to 0 means its not doing well.

E. Plots

The models were also analyzed using 3 different plots. With the help of these plots, its easier to understand which model outperforms the other.

- **Actual v/s Prediction Plot**

This plot compares the predicted values to the actual values for both the models. This shows how well the models are performing.

- **Residual Plot**

This plot visualizes the residuals (differences between the predicted and actual values). This helps assess how evenly errors are distributed.

- **Variance Plot**

A variance plot visualizes the distribution of residuals (errors) from a model's predictions, showing how often each error value appears.

II. LITERATURE SURVEY

In paper [5] describes basic decision tree issues and current research points. Decision tree techniques have been widely used to build classification models as such models closely resemble human reasoning and are easy to understand. In this, the study reviewed existing literature on decision tree algorithms, focusing on classification and regression trees. It further explored various algorithms such as CART, SPRINT, and SLIQ, and discussed their strengths and weaknesses. The paper also describes the strengths and weaknesses of decision trees and further indicates potential avenues for future research. It suggests that model complexity should be balanced with interpretability and generalization performance. It also suggests bias combinations and new algorithms to enhance the performance of decision trees.

In paper [6] presents an optimum combination of two sta- tistical techniques to improve the skill of long-range weather forecasts in sub-Carpathian zones compared to plane zones.

The paper focuses on using Extended Empirical Orthogonal Functions (EEOF) decomposition with a 3-month data window for temperature and precipitation fields in Romania. The paper also applied an auto-regressive model with parameters determined using the maximum entropy method (AR-MEM) to forecast time series of the EEOF components. In the paper the AR-MEM model showed improved forecast skill for temperature fields in the central part of Romania. However, the forecast based on the EEOF component for precipitation was less skillful. The paper also highlights the potential of combining EEOF decomposition with AR-MEM for long-range weather forecasting.

III. METHODOLOGY

The study was followed up by the following method

- The data was initially loaded from the dataset in csv file format.
- Then it was made sure that the column names matched the feature names mentioned.
- The feature columns were then converted into numeric values so that any non-numeric values can be handled by coercing them to NaN.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 01 Jan. 2025, pp: 373-379 www.ijaem.net ISSN: 2395-5252

- Also to preprocess the dataset, any rows with missing values in the selected feature of target columns were dropped.
- Eventually, the dataset was split into training and test data, where 80% was the training data and 20% the test data
- For training the model under AR-MEM, lags were adjusted to fit the training data, and predictions were made using this fitted model.
- For training the model under the Decision Tree model, GridSearchCV was used to find the best hyperparameters, and the decision tree was fit using these hyperparameters; and finally, predictions were made using this model.
- For evaluation of these two models, MSE, MAE and R^2 metrics were calculated.
- Also to perform a visual analysis the prediction v/s actual plot, residual plot, and variance plot are added.

IV. RESULTS AND INFERENCES

Below are the results obtained through this study.

	Model	MSE	MAE	R2
0	AR-MEM	23.572304	4.315809	-0.043792
1	Decision Tree	0.268503	0.506034	0.988111

Fig. 3. MSE MAE AND R^2 values

Based on the above shown Model Metrics it can be under- stood that:

- Since the MSE and MAE values are on the lower side for the decision tree, it indicates that the the model shows a better performance. Whereas, for the AR-MEM model, the value lies on the higher end, therefore the model doesn't show a good performance
- For the case of R^2, usually values closer to 1 indicate better model performance. Therefore, since the decision tree shows a value very close to 1, it can be concluded as the better model when compared to AR-MEM. Also the negative value indicates poor model performance.

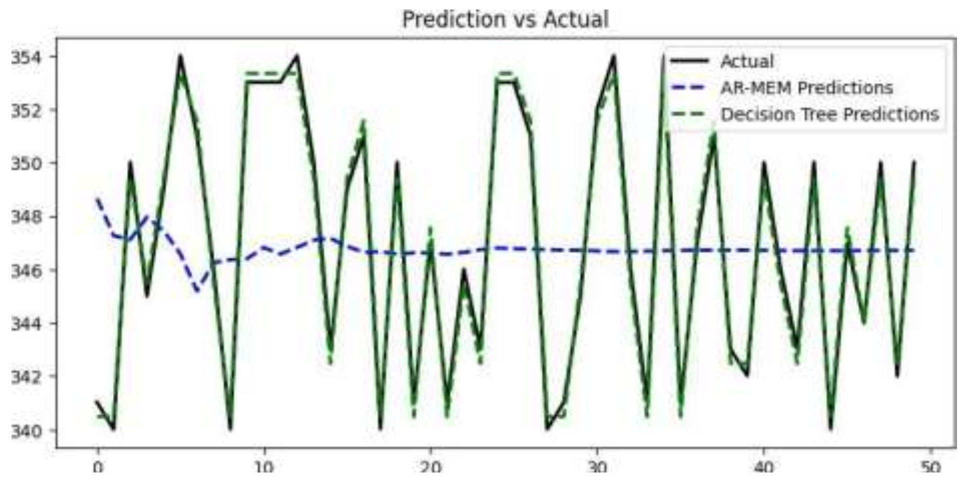


Fig. 4. Prediction v/s Actual Values.

In the above plot, the X-axis shows the data points, and the Y-axis the values of target variables.
From the plot, it can be interpreted that

- The Decision Tree predictions closely follow the actual values, indicating higher accuracy.
- The AR-MEM predictions appear more smoothed and less accurate, deviating from the actual values.

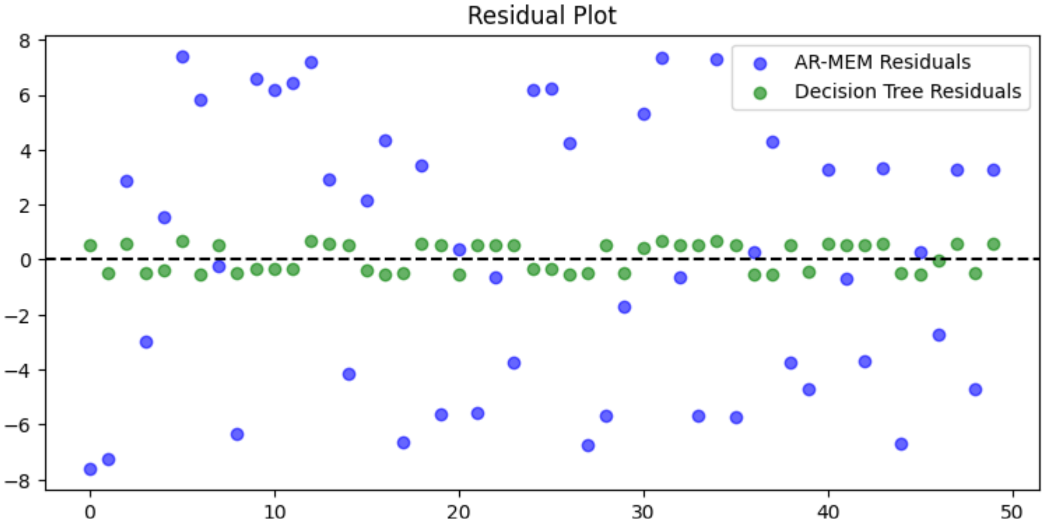


Fig. 5. Residual Plot.

In the above plot, the X-axis shows the data points, and the Y-axis the Residuals (errors) of the predictions.
From the plot, it can be interpreted that

- The Decision Tree residuals are more tightly clustered around zero, indicating better accuracy.
- The AR-MEM residuals are more scattered, indicating larger prediction errors.

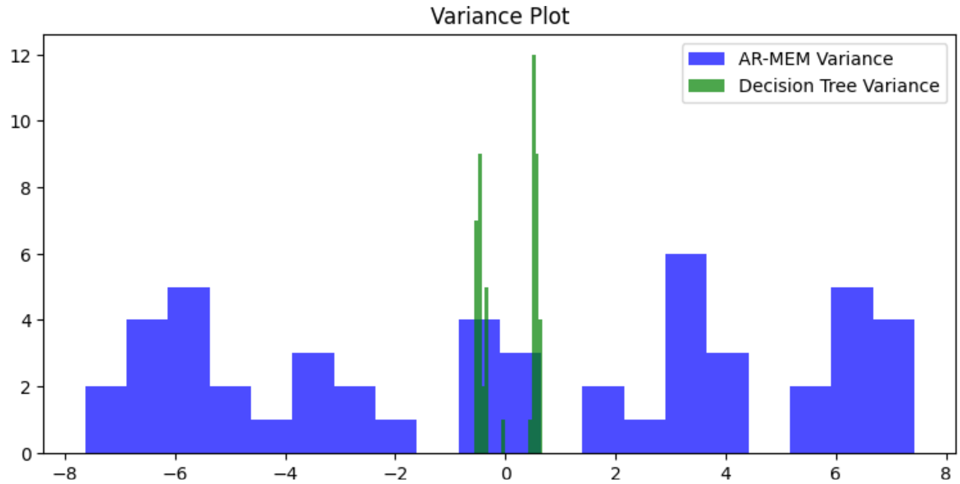


Fig. 6. Variance.

In the above plot, the X-axis shows the Variance of AR- MEM model, and the Y-axis the Variance of Decision Tree. From the plot, it can be interpreted that

- The Decision Tree model has a higher range of residuals close to zero, indicating lower variance and better perfor- mance.
- The AR-MEM model has more varied residuals, indicationg higher variance and less reliable predictions.

V. CONCLUSION

Therefore, from this study, it can be summarized that, the Decision Tree model has a significantly better performance in predicting the target variable (Surface Current Direction) than the AR-MEM model, based on lower errors (MSE and MAE) and a higher R^2 value.

Further visual analysis of the plots confirms that the Decision Tree model predictions are really very close to the real values of the target variable (Surface Current Direction), with the residuals being compactly clustered around zero, which means minimal bias and good performance.

Hence it can be concluded that that the Decision Tree model is much better at capturing intricate and nonlinear patterns within the dataset.

REFERENCES

[1]. Domps, Baptiste & Dumas, Dylan & Gue´rin, C-A & Marmain, Julien. (2021). High-Frequency Radar Ocean Current Mapping at Rapid Scale With Autoregressive Modeling. IEEE Journal of Oceanic Engineering.
[2]. 46. 891-899.
[3]. Ulrych, T. J., and T. N. Bishop (1975), Maximum entropy spectral anal- ysis and autoregressive decomposition, Rev. Geophys., 13(1), 183–200.
[4]. K. G, Y. Sushmitha, K. L. Saranya, P. Naga Ramya Sri and P. Amulya, ”Rain fall Prediction Using Deep Learning and Machine Learning Techniques,” 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023.
[5]. Lee, S., Park, S. (2013). Maximum Entropy Test for Autoregressive Models. In: Huynh, VN., Kreinovich, V., Sriboonchitta, S., Suriya, K. (eds) Uncertainty Analysis in Econometrics with Applications. Advances in Intelligent Systems and Computing, vol 200. Springer, Berlin, Hei- delberg.
[6]. Kotsiantis, S.B. Decision trees: a recent overview. Artif Intell Rev 39, 261–283 (2013).
[7]. I. Mares, ”Long Range Forecasting in the Mountain and Hill Zones in Romania by Means of an AR-MEM Model,”