# Machine Learning Techniques for Circular Data: A Novel Approach for Decision Trees and Clustering

Snehal Kawale
*Dept.of Statistics, Savitribai Phule Pune University Pune*

***Abstract*— Data can be broadly categorized into linear and circular types, with circular data arising when measurements are directional or angular in nature, such as wind directions or time of day. Unlike linear data, circular data exhibits a cyclical structure—values wrap around at a boundary (e.g., 0° is equivalent to 360°)—posing unique challenges for traditional statistical and machine learning methods. Circular statistics, or directional statistics, specifically address these challenges by accounting for the inherent periodicity of such data. However, existing machine learning techniques have not been adequately adapted to effectively handle circular data. This research aims to bridge this gap by developing novel algorithms and methodologies tailored for circular data analysis. By integrating machine learning with circular statistical principles, our work seeks to enhance predictive modeling and insight extraction from directional datasets, with applications spanning meteorology, biology, and beyond.**

***Index Terms*— Circular Data, Data Visualization, Decision Tree for Circular Data, Machine Learning for Circular Data, Clustering Algorithms, Rainfall Prediction.**

## I. INTRODUCTION

In statistical analysis, data types are generally classified as either linear or circular. While linear data is commonly encountered and extensively studied, circular data—such as angles, time of day, and wind direction—requires specialised treatment due to its inherent periodicity. Unlike linear variables, which have a natural ordering and fixed scale, circular variables wrap around a circle, rendering standard linear methods inappropriate for their analysis.

Circular statistics, also referred to as directional statistics, addresses this challenge by developing tools and methodologies that account for the circular nature of the data. However, despite the increasing availability of circular data in fields such as meteorology, biology, and environmental science, mainstream machine learning techniques remain largely unadapted for such contexts. Algorithms such as decision trees and clustering models, which are fundamental to data-driven discovery, are typically designed for linear spaces and fail to capture the structure of circular data effectively.

This research aims to bridge this methodological gap by developing novel machine learning algorithms tailored specifically for circular data. In particular, we propose a new **decision tree algorithm** and a **clustering approach** that integrate both circular and linear variables. The study also emphasises the importance of appropriate visualisation techniques for circular data to enhance interpretability.

To demonstrate the practical utility of the proposed methods, we apply them to a large-scale meteorological dataset covering 640 districts in India, focusing on the monsoon months. By analysing parameters such as wind direction (circular), time (circular), wind speed, temperature, and precipitation (linear), the study illustrates how incorporating circular statistics into machine learning workflows can significantly improve prediction and classification performance.

## II. RELATED WORK

1. **Mardia and Jupp (2000)** - *Directional Statistics*:
   Mardia and Jupp's book provides the foundational principles for **circular statistics**, including measures like **mean direction**, **circular variance**, and **circular standard deviation**. Their work also introduces visualization techniques such as **Circular Raw Plots** and **Rose Diagrams**, widely used for analyzing angular data.

2. **Al-Daffaie and Khan (2017)** - *Logistic Regression for Circular Data*: Al-Daffaie and Khan extended traditional **logistic regression** to handle **circular predictors** like **wind direction**. Their model adapts cyclic fluctuations to predict binary outcomes (e.g., rainfall prediction), offering a better fit for data where predictors are angular in nature.
3. **Lopez-Cruz and Bielza (2020)** - *Directional Naive Bayes Classifier*: Lopez-Cruz and Bielza introduced the **Naive Bayes classifier** for directional data, using **von Mises** and **von Mises-Fisher** distributions to model circular predictors. They also introduced the **Selective von Mises Naive Bayes**, a variant that selects features based on their mutual information, improving predictive accuracy.
4. **Debnath and Song (2018)** - *Fast Optimal Circular Clustering*: Debnath and Song developed the **FOCC (Fast Optimal Circular Clustering)** algorithm, which improves clustering efficiency for circular data. Their method linearizes circular data to apply traditional clustering techniques and is widely used for clustering **genomic data**, as demonstrated by their **OptCirClust** package.

## III. CLUSTERING ALGORITHM

Traditional clustering algorithms are not well-suited for circular data due to its periodic nature. To address this, we developed a hybrid clustering approach that integrates both circular and linear variables within a unified framework.

### 3.1 Distance Calculation

Linear variables are compared using standard Euclidean distance, while circular variables require angular distance metrics such as chord or cosine distance. Prior to distance computation, all variables are standardised. The combined distance is scaled as:

$$D_{scaled} = D/(1+D)$$

where D is the sum of linear and circular distances. This ensures all pairwise distances lie between 0 and 1.

### 3.2 Clustering Procedure

We apply **agglomerative hierarchical clustering** to the resulting distance matrix. The method begins with each observation as a separate cluster and iteratively merges the closest pairs based on the defined metric. The process is visualised using dendrograms, which support intuitive selection of the number of clusters.

### 3.3 Application Context

Using this approach, we clustered meteorological data across Indian states during monsoon months. Circular variables such as wind direction and time were handled appropriately, and results showed that **mixing height** and **temperature** were key factors driving cluster formation.

## IV. DECISION TREE ALGORITHM

Conventional decision tree algorithms are designed for linear data and do not handle circular variables effectively due to their discontinuous boundaries. To overcome this, we propose a modified decision tree algorithm that accommodates both circular and linear predictors.

### 4.1 Splitting Criteria

For **linear variables**, standard splitting is performed using thresholds that minimise the mean squared error (MSE). For **circular variables**, we apply either a grid search over circular sectors or a 2-means clustering approach to determine optimal split points that minimise **circular variance**.

The best splitting variable at each node is chosen by comparing error metrics: MSE for linear responses, and angular separation or circular variance for circular responses.

### 4.2 Tree Construction and Prediction

The tree is constructed recursively, using the best split at each node until a stopping criterion is met (e.g., maximum depth, minimum samples, or error threshold). Predictions are made by traversing the tree:

- For linear responses: using the mean value at each terminal node.
- For circular responses: using the circular mean.

### 4.3 Application Context

This algorithm was applied to a subset of Indian meteorological data to demonstrate its capability in modelling circular relationships. Results indicate that the method effectively partitions the data and captures the structure of circular predictors such as wind direction and time.

## V. RESULTS

The proposed circular decision tree algorithm was implemented on meteorological data from seven Indian districts. The model successfully identified optimal splits using circular predictors such as wind direction and time, leading to improved prediction accuracy and interpretability. Compared to traditional decision trees, it demonstrated superior handling of circular boundaries and reduced prediction error. The structure of the tree reflected meaningful meteorological relationships, supporting the effectiveness of the proposed approach.

## VI. CONCLUSION

This study presents a novel decision tree algorithm specifically designed for circular data, addressing key limitations of conventional methods when applied to directional variables such as wind direction and time. The proposed model demonstrates improved accuracy, interpretability, and effectiveness in handling circular boundaries, particularly in the context of meteorological analysis.

However, the current implementation is limited by its computational efficiency, particularly when applied to large-scale datasets. Additionally, the splitting strategy, while effective, does not yet incorporate more complex circular clustering techniques that could enhance precision.

Future work will focus on developing scalable implementations, integrating advanced circular clustering methods for node splitting, and extending the algorithm into ensemble frameworks. This will further enhance the robustness and applicability of circular machine learning models across diverse domains.

## ACKNOWLEDGMENT

The author would like to express their sincere gratitude to the Department of Statistics at Savitribai Phule Pune University, Pune for their academic support and guidance throughout this research. We also thank Dr. Akanksha Kashikar for their valuable suggestions and encouragement. Additionally, we acknowledge the Indian Meteorological Department for providing access to the dataset used in this study.

## REFERENCES

[1] Kadhem Al-Daffaie and Shahjahan Khan. “Logistic regression for circular data”.In:*AIP Conference Proceedings*. Vol.1842.1.AIP Publishing.2017.

[2] Tathagata Debnath and Mingzhou Song. “Fast Optimal Circular Clustering and Applications on Round Genomes”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18.6 (2021), pp. 2061–2071. DOI: 10.1109/TCBB.2021.3077573.

[3] Pedro L López-Cruz, Concha Bielza, and Pedro Larrañaga.“Directional naive Bayes classifiers”. In: *Pattern Analysis and Applications*18 (2015), pp. 225–246

[4] K.V.Mardia and P.E.Jupp .*Directional Statistics*.Wiley Series in Probability and Statistics. Wiley, 2009.ISBN: 9780470317815. URL:https://books.google.co.in/books?id=PTNiCm4Q-M0C