Training Day 14 Report

Date: 18 July 2025
Topic: Decision Tree, Random Forest, and Support Vector Machine (SVM)

Summary of the Day:
Day 14 of training focused on advanced supervised machine learning algorithms—Decision Trees, Random Forests, and Support Vector Machines (SVM). The session built upon previous discussions of regression and classification models, extending into tree-based and hyperplane-based methods for both classification and regression tasks.

Morning Session:
The trainer began by revisiting **classification and regression** and introduced **Decision Trees** as a model that mimics human decision-making. The concept of nodes, branches, and leaf nodes was explained using flowchart-like structures. The **root node** represents the first decision point, **branches** represent decision paths, and **leaf nodes** represent final outcomes or classifications.

Key highlights included:
Understanding the limitations of linear models (like Linear Regression and KNN) in handling non-linear and high-dimensional data. Decision Trees provide a structured way of taking conditional decisions through "if-else" logic. Concept of **root node, decision node, branch, and leaf node** in tree-based models. Practical example: predicting the **risk of heart attack** based on features like age, weight, and smoking habits. The instructor explained how decision trees select features using mathematical metrics to create optimal splits. Two major parameters were introduced:
Gini Impurity: Measures the impurity or uncertainty at a node. Formula: $G = 1 - \Sigma(p■^2)$ **Entropy and Information Gain:** Entropy measures randomness (ranges 0–1), and Information Gain helps determine the best feature for a split. It was demonstrated how Gini Impurity = 0 indicates a **pure node** (all samples belong to one class), while higher values represent mixed or impure nodes.

Afternoon Session:
The session continued with the **Random Forest algorithm**, which builds upon the Decision Tree concept. The trainer explained:
Random Forest is an **ensemble method** combining multiple Decision Trees to improve accuracy and reduce overfitting. Each Decision Tree is trained on a random subset of data and features (a concept called **bootstrapping or bagging**). During prediction, results from all trees are combined using: **Majority voting** for classification problems. **Average value** for regression problems. This aggregation helps in minimizing bias and variance compared to a single Decision Tree. An intuitive analogy was provided—taking advice from multiple people (trees) instead of a single friend (tree) for better, unbiased decisions. The importance of parameter tuning, tree depth, and number of estimators was also discussed, emphasizing how the Random Forest prevents overfitting while maintaining interpretability.

Evening Session:
The final part of the day introduced **Support Vector Machines (SVM)**. This model is widely used for classification and regression but excels in high-dimensional data. The instructor discussed:
SVM works by finding the **optimal hyperplane** that separates different classes in a feature space. The **support vectors** are the data points that lie closest to the separating boundary. The **margin** between classes should be maximized for better generalization. SVMs can handle non-linear separations by transforming data into higher dimensions using the **kernel trick**. Example: For

circularly distributed data (non-linear), a linear boundary cannot classify effectively. SVM projects data into higher dimensions, allowing for separation using a hyperplane.

Key Learnings: Decision Tree helps visualize and make logical decisions based on conditional features. Random Forest improves upon Decision Trees using ensemble learning to boost performance. Understanding **Gini Impurity**, **Entropy**, and **Information Gain** for feature selection.

SVM introduces the concept of hyperplanes and margins for classification. Real-world applicability of these algorithms in domains like healthcare, finance, and marketing analytics. **Conclusion:**
Day 14 concluded with a strong understanding of non-linear models and ensemble techniques. These methods are crucial for handling real-world datasets that are complex and multidimensional. The session also highlighted the importance of documentation, particularly the **scikit-learn library**, as an essential resource for implementing machine learning algorithms. The next sessions will focus on **project implementation and model deployment**.

Overall Experience:
Day 14 was an in-depth exploration of tree-based and kernel-based machine learning algorithms. The hands-on explanation using real-life analogies made the session highly interactive and valuable. The knowledge of impurity measures, ensemble methods, and SVM theory added substantial depth to understanding supervised learning models.