COVER TYPE PREDICTION USING RANDOM FOREST ALGORITHM

K. Manikanta and K. Yaswanth and G. Anil Kumar
Department of Artificial Intelligence

ABSTRAT:

Machine Learning bring drastic changes in technologies .This machine learning algorithms are used to train the machines, This algorithms can classify the data and also predict the output of the data. Machine learning is a subset of artificial intelligence (ai) that focuses on developing algorithms and statistical models that allow computers to learn and improve from experience without being explicitly programmed. The primary goal of machine learning is to enable computers to identify patterns in data and make predictions or decisions based on those patterns. Machine learning has applications in various domains, including image and speech recognition, natural language processing, recommendation systems, medical diagnosis, financial modeling, and autonomous vehicles, among others.

Key Words:

Introduction:

Predicting forest cover type from cartographic variables only (no remotely sensed data). The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types). This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado.

These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices. As for primary major tree species in these areas, Neota would have spruce/fir (type 1), while Rawah and Comanche Peak would probably have lodgepole pine (type 2) as their primary species, followed by spruce/fir and aspen (type 5). Cache la Poudre would tend to have Ponderosa pine (type 3), Douglas-fir (type 6), and cottonwood/willow (type 4).

1)Data set used:

Predicting forest cover type from cartographic variables only (no remotely sensed data). The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types). This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices

2)Additional Variables:

Given is the attribute name, attribute type, the measurement unit and a brief description. The forest cover type is the classification problem. The order of this listing corresponds to the order of numerals along the rows of the database.

Name / Data Type / Measurement / Description

Elevation / quantitative /meters / Elevation in meters

Aspect / quantitative / azimuth / Aspect in degrees azimuth

Slope / quantitative / degrees / Slope in degrees

Horizontal_Distance_To_Hydrology / quantitative / meters / Horz Dist to nearest surface water features

Vertical_Distance_To_Hydrology / quantitative / meters / Vert Dist to nearest surface water features

Horizontal_Distance_To_Roadways / quantitative / meters / Horz Dist to nearest roadway

Hillshade_9am / quantitative / 0 to 255 index / Hillshade index at 9am, summer solstice

Hillshade_9am / quantitative / 0 to 255 index / Hillshade index at 9am, summer solstice

Hillshade_Noon / quantitative / 0 to 255 index / Hillshade index at noon, summer soltice

Hillshade_3pm / quantitative / 0 to 255 index / Hillshade index at 3pm, summer solstice

Horizontal_Distance_To_Fire_Points / quantitative / meters / Horz Dist to nearest wildfire ignition points Wilderness_Area (4 binary columns) / qualitative / 0 (absence) or 1 (presence) / Wilderness area designation Soil_Type (40 binary columns) / qualitative / 0 (absence) or 1 (presence) / Soil Type designation

Cover_Type (7 types) / integer / 1 to 7 / Forest Cover Type designation.

3)Algorithms used:

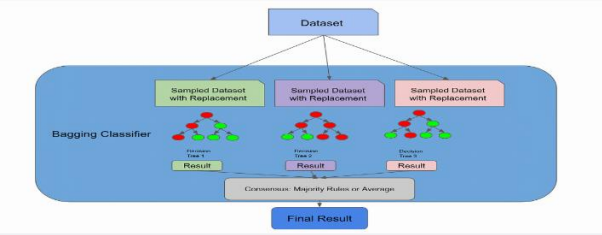
There are various classification algorithms that can be used based on the nature of the data and the problem at hand. Some common algorithms include*:*

4)Decision Tree Algorithm:

Decision tree algorithm is a classification algorithm used for Machine

learning. It classify the data set into n decision trees.Random Forest is built upon the concept of decision trees, which are tree-like structures where each internal node represents a decision based on a feature attribute, and each leaf node represents the class label (in classification) or the predicted value (in regression).

5)Architecture of Decision Tree:

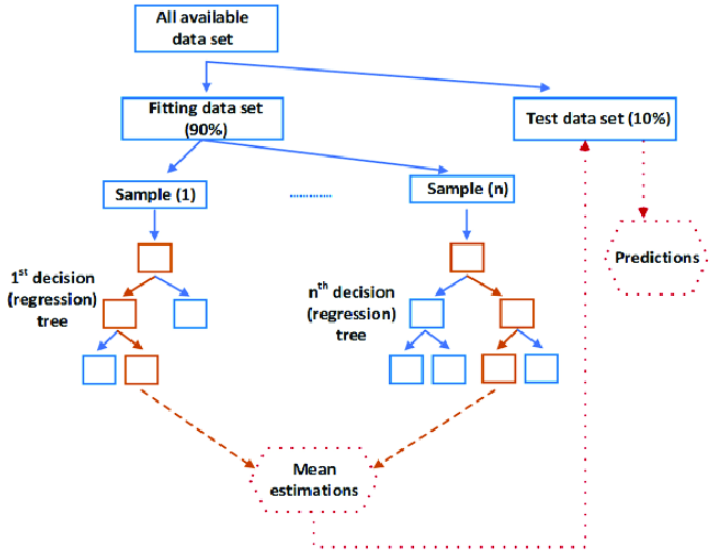


6)Random Forest Algorithm:

Random Forest is a powerful ensemble learning technique used for both classification and regression tasks in machine learning. It is based on the concept of decision trees and combines the predictions of multiple individual decision trees to improve overall accuracy and reduce overfitting. Random Forest is built upon the concept of decision trees, which are tree-like structures where each internal node represents a decision based on a feature attribute, and each leaf node represents the class label (in classification) or the predicted value (in regression).

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. Each decision tree is trained independently on a random subset of the training data and features**.**

7)Architecture of Random Forest Algorithm:



8)Training Analysis:

In every machine learning algorithms involve training and testing of the data set and predicting the output. For these we have used python packages like pandas, numpy, sklearn, matplot etc. These packages is the collection of modules .These packages are used to data processing and data visualization.

9)Testing and Result Analysis:

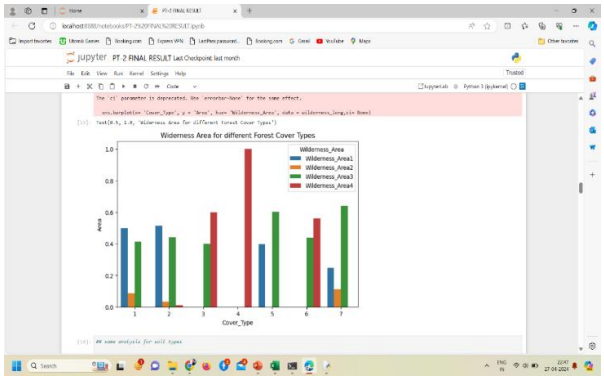
Tasting is a preprocessing of the given data set using machine learning techniques. These testing is also used for the visualization of the data using some python packages like pandas, numpy, sklearn, matplotlib etc.These testing is used to predict the output using some

machine learning algorithms like Random Forest, Decision Tree etc.

1)Bar plot:

These bar maps shows us the Aspect and wilderness_ness_Areas of each cover type these bar map is very easy to understand and in the below image shows that aspect of wilder_ness areas in seven cover types

The blue one represents the wilder_ness_area1 and the yellow one represents the wilder_ness_area 2 and green represents the wilder_ness_Area 3 and the last one represents the wilder_ness_area 4 and the highest percentage is in the graph is wilder_ness_area 4.



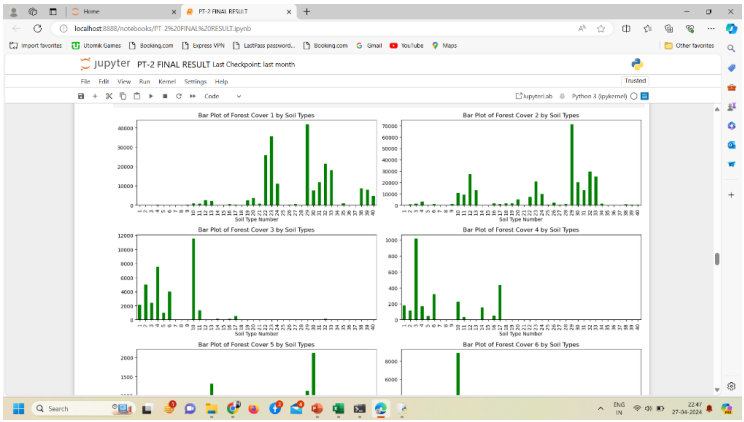
Explanation:

This image is used to Areas in different cover types and this is bar plot in these bar plot we taken x-axis as cover types and y-axis as areas.

These bar plot is plotted by using seaborn module in python it is used to visualize the data in understanding farmet.

2)Subplot:

These barplot shows the soil_types in each cover_type in these the barplot has the soil_types in different cover_type and which soil has the highest value among them.

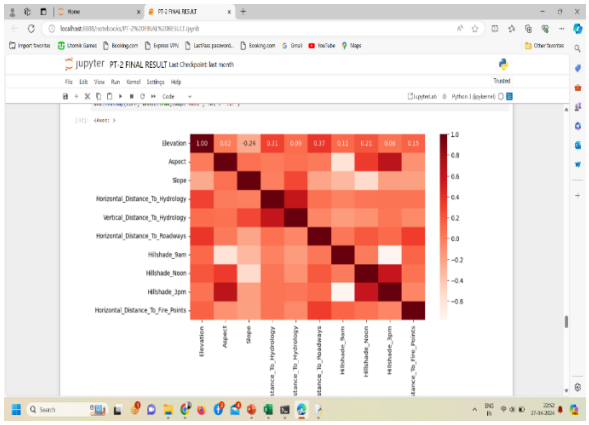


3)Box Plot:

These box_plot shows the values of Entities in cover_types . These like Evalution, Aspect, Slope, Horizontal_Distance_to_Hydrology, Vertical Distance_to_Hydrology, etc.

4)Heat Map:

These heat_map represents the Negative and Positive co-relation between the Entities and also represents the highest value among the Entities and also the bar beside of the heat_map has the different colors which represents the negative and paositive value of the entities.



Conclusion:

In conclusion, machine learning is a powerful field within artificial intelligence that focuses on developing algorithms and models to enable computers to learn from data and make predictions or decisions without explicit programming. Key concepts in machine learning include supervised learning, unsupervised learning, reinforcement learning, and various algorithms tailored to specific tasks.

One such important algorithm is Random Forest, which is a popular ensemble learning technique used for both classification and regression tasks. Random Forest combines the predictions of multiple decision trees trained on random subsets of data and features to improve accuracy and reduce overfitting. Its ability to handle large datasets, high dimensionality, and provide insights into feature importance makes it widely applicable in diverse domains.

References:

https://link.springer.com/article/10.1007/s00521-020-05139-4

https://www.sciencedirect.com/science/article/pii/S003442570700449X

https://www.mdpi.com/2072-4292/12/7/1135

https://books.google.co.in/books?hl=en&lr=&id=vLiTXDHr_sYC&oi=fnd&pg=PA3&dq=references+for+cover+type+prediction+using+machine+learning&ots=C_lvwA_Ffn&sig=F4D0YBp9reTQi79pnFznqJgObQs