ML Project Proposal

TITLE: Forest cover type prediction.

TEAM MEMBER NAMES:

- Siddhi Degaonkar
- Sanika Pol
- Vaidehi Sonar
- Tanvi Rasam

INTRODUCTION:

Predicting the forest cover type using various types of models i.e. linear Classifier(Naive Bayes Classifier), tree based Classifier(Random Forest Classifier), clustering(K-Means Classifier), etc. Also, we will combine feature selection methods such as Lasso, Random Forest, Principal Component Analysis[1] with each of the classifier models. The idea is to compare the accuracy of each model and how the feature selection method affects each of the classifier. Using this analysis, we will ultimately try to improve the accuracy for forest cover type prediction and give the best suitable combination of feature selection method and a classifier with the highest accuracy[1].

DATASET:

Our main aim is to predict the forest cover type (the predominant kind of tree cover) from cartographic variables. The actual forest cover type for a given 30 x 30 meter cell was determined from US Forest Service (USFS) Region 2 Resource Information System data. Independent variables were then derived from data obtained from the US Geological Survey and USFS. The data is in raw form and contains binary columns of data for qualitative independent variables such as wilderness areas and soil type.

This dataset includes information on tree type, shadow coverage, distance to nearby landmarks (roads etcetera), soil type, and local topography.

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.

To predict an integer classification for the forest cover type. The seven types are:

Integer value between 1 and 7, with the following key:

1 - Spruce/Fir

2 - Lodgepole Pine

3 - Ponderosa Pine

4 - Cottonwood/Willow

5 - Aspen

6 - Douglas-fir

7 - Krummholz

Data Fields:

1.Elevation - Elevation in meters
2.Aspect - Aspect in degrees azimuth
3.Slope - Slope in degrees
4.Horizontal_Distance_To_Hydrology - Horz Dist to nearest surface water features
5.Vertical_Distance_To_Hydrology - Vert Dist to nearest surface water features
6.Horizontal_Distance_To_Roadways - Horz Dist to nearest roadway
7.Hillshade_9am (0 to 255 index) - Hillshade index at 9am, summer solstice
8.Hillshade_Noon (0 to 255 index) - Hillshade index at noon, summer solstice
9.Hillshade_3pm (0 to 255 index) - Hillshade index at 3pm, summer solstice
10.Horizontal_Distance_To_Fire_Points - Horz Dist to nearest wildfire ignition points
11.Wilderness_Area (4 binary columns, 0 = absence or 1 = presence) - Wilderness area designation
12.Soil_Type (40 binary columns, 0 = absence or 1 = presence) - Soil Type designation
13.Cover_Type (7 types, integers 1 to 7) - Forest Cover Type designation

Soil_Type1 to Soil_Type40 [Total 40 Types]

There are 4 different wilderness areas and 40 different soil types. These are already converted to machine-learning readable binary formats(One Hot encoding) .

LITERATURE SURVEY:

- **Classifying Forest Cover type with cartographic variables via the Support Vector Machine, Naive Bayes and Random Forest classifiers[1] :**

In the above approach, the forest type classes is predicted using various classification methods such as the Support Vector Machine, Decision Trees' Random Forest and the Naive Bayes classifiers. The focus is on overall accuracy and accuracy with respect to each class. For dimensionality reduction methods like Lasso, Random forest and PCA are used and then models such as Naive Bayes, SVM, Random forest Classifiers are used to predict the cover types. According to this approach, Random Forest worked best with PCA and gave the highest accuracy possible and Naive Bayes performed worst with PCA. The work can be extended to Independent Component Analysis to improve the accuracy of Naive Bayes. SVM required highest computational time where as Naive Bayes required the least.

- **Comparative Accuracy of Different Classification Algorithms for Forest Cover Type Prediction[2] :**

The forest cover dataset which consist of predominant kind of tree cover and various attributes which are **cartographic** uses different algorithms for predicting the cover type using cartographic variables. The

algorithms used are firstly **Regression**, **Decision tree**, **Random forest**, **Gradient boosting machines.** Using these machine learning methods models have been developed and tested for accuracy ranging from 19.4% to 74.8%. The result of this paper is the forest cover type prediction based on cartographic variables. An important aspect of the study is the use of different performance measures to evaluate the learning methods. It is obtained that Random Forest gives better prediction with 74.8% accuracy.

- **Classifying Forest Cover Type using Cartographic Features[3] :**

This paper aims to predict forest cover type by incorporating cartographic data and a variety of classification algorithms. To avoid overfitting the models employed 10-fold cross validation. For some 44 binary attributes(4 for wilderness area and 40 for y soil type) out of 54 were omitted to reduce overfitting and increase testing performance. Principal component Analysis is used for dimensionality reduction. For 10 dimensional PCA-transformed data (capturing 99.997% of the variance) performed only marginally worse than using all 54 dimensions.Computational time of Multi-class SVM was reduced with PCA. This SVM uses the "one vs. one" approach to multi-class classification. When SVM was trained on the entire dataset accuracies obtained were 81.35% for training and 78.24% testing, while without booleans were 75.21% and 72.75%, respectively. K-Means Clustering was also developed by running the algorithm 10 times to increase its accuracy. For K-means when k=7 ie. number of forest cover types, the model performed poorly. Thus, no. of clusters were increased ie. allowing each type to have more than one cluster which decreases the error. SVM classifier outperformed models used in studies by Blackard and Dean [4] which reported 70.58% accuracy when classifying forest cover type using an artificial neural network and 58.38% classification accuracy using discriminant analysis

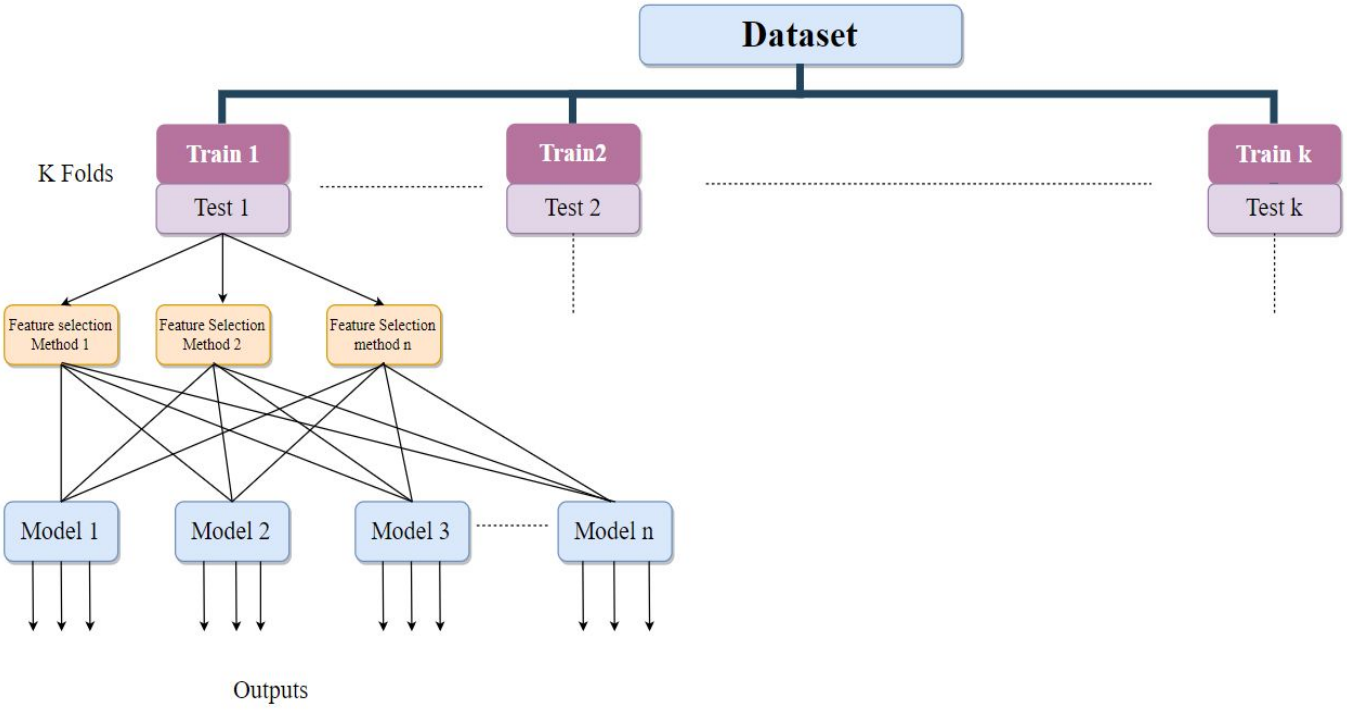
METHOD:

A general overview of our project is to apply different Feature selection approaches like Random forest, Lasso regression and PCA on different Machine Learning models such as Discriminant Analysis, K-Means clustering, Random Forest and Naive Bayesian and study which approach is best for each model. [1][3]

The model's performance highly depends on the data feed. Thus, the data preprocessing is critical. Since our data is already encoded, in preprocessing, we would perform further step of normalization as most of the attributes are quantitative with different units. Also, for methods like Lasso and PCA, normalization is essential for its performance. For getting basic insights from the dataset, we will visualize, analyze and understand the data.

Next we will split our data into training and testing using k-fold cross validation to avoid overfitting our models.[3] On each of the k-folds we perform different feature selection methods on the training data. Feature Selection greatly impacts the performance of the model as irrelevant or partially relevant features can have a negative impact. Every model is then trained on the selected features of the different methods individually. This will be done by calling the train_step of each model.

Once our training is complete, we proceed for testing and comparison. Here, we use our testing data to evaluate each model's performance. This step is meant to be representative of how the model will perform on unseen data. Further, we would analyze the results based on parameters like accuracy, computation time, F-score and AUC. Depending on these parameters, which model performed the best with which feature selection method can be concluded. Fig 1 shows the our system architecture



PLAN

PROJECT TIMELINE:

TEAM MEMBER ROLES

Task	Siddhi	Sanika	Vaidehi	Tanvi
Project group and topic finalization	✔	✔	✔	✔
Data exploration and analysis	✔	✔	✔	✔
Planning	✔	✔	✔	✔
Project proposal	✔	✔	✔	✔
Changes as per feedback in the proposal	✔	✔	✔	✔
Data preprocessing	✔	✔	✔	✔
Exploration of models and algorithms	✔	✔	✔	✔
Model finalization	✔	✔	✔	✔
Mid report	✔	✔	✔	✔
Model training	✔	✔	✔	✔
Model evaluation and tuning	✔	✔	✔	✔
Final report	✔	✔	✔	✔

AIM OF THE PROJECT:

The Forest Cover type prediction Project is going to answer the following questions:

Q1. What is the forest cover type?

Q2. What is the accuracy of different models without feature selection?

Q3. What is the accuracy of different models with feature selection methods?

Q4. Which is the best and worst pair of feature selection method and a particular type of model?

DIFFERENCE (IS OUR IDEA NOVEL?)

As per our knowledge, various approaches have been used in the past to explore this dataset and predict the cover type. Considering [1], we intend to adopt a similar approach, compare and simultaneously improve the accuracy of different models. But rather than picking up classifiers randomly we will compare one of each type of classifier (eg. a linear classifier, a tree based classifier,etc). Further, the novel approach we will use is, we will compare a linear classifier and a tree based classifier with different feature selection methods(lasso, Random Forest, PCA, etc) and try to get the best feature selection method suited for a particular type of model while giving a higher accuracy as possible.

WHAT WILL WE LEARN ?

We will learn to work with real cartographic data. We will learn how to prepare the dataset so that it is usable to our model and how can we fit our models properly. We will be ables to understand what features are relevant and irrelevant to our target. We will be able to compare and analyze what factors, what feature selection methods are appropriate and works better for a certain type of model and which are not suitable. We will be able to understand which models have greater accuracy and how the accuracy can be improved.

REFERENCES

[1] Hugo Sjoqvist. "Classifying Forest Cover type with cartographic variables via the Support Vector Machine, Naive Bayes and Random Forest classifiers". Orebro University - School of Business and Economics, Spring 2017.
https://pdfs.semanticscholar.org/fc79/be46c8c9499df564bdcfd670fb28b0909b35.pdf

[2] "Forest Cover Type Prediction using Cartographic Variables",
https://www.ijcaonline.org/archives/volume182/number30/wagh-2018-ijca-918191.pdf

[3] Kevin Crain, Kevin Crain. "Classifying Forest Cover Type using Cartographic Features". Stanford University - CS 229: Machine Learning - December 2014
http://cs229.stanford.edu/proj2014/Kevin%20Crain,%20Graham%20Davis,%20Classifying%20Forest%20Cover%20Type%20using%20Cartographic%20Features.pdf

[4] Jock A. Blackard, Dr. Denis J. Dean "Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables". Computers and Electronics in Agriculture (1999)
https://www.sciencedirect.com/science/article/pii/S0168169999000460

[5]Jock A. Blackard, Dr. Denis J. Dean, Dr. Charles W. Anderson ." UCI Cover Type Dataset"
https://archive.ics.uci.edu/ml/datasets/covertype