Forest Cover Type

Using cartographic variable

Motivation

Global forest cover over the past 60 years has de
has been correlated with many negative impacts
restore forest cover and biodiversity without causi
a region, it is essential that we analyze the type o
that we replicate the same in conservation efforts

One of the ways we can do this is by using data t
are specific to certain kinds of plants, and use it t
given the topological and cartographic features.

Data

The dataset contains **581012** samples and **55** diff
cover found in four different wilderness areas loc
Northern Colorado. Each observation is from a 3

Prediction

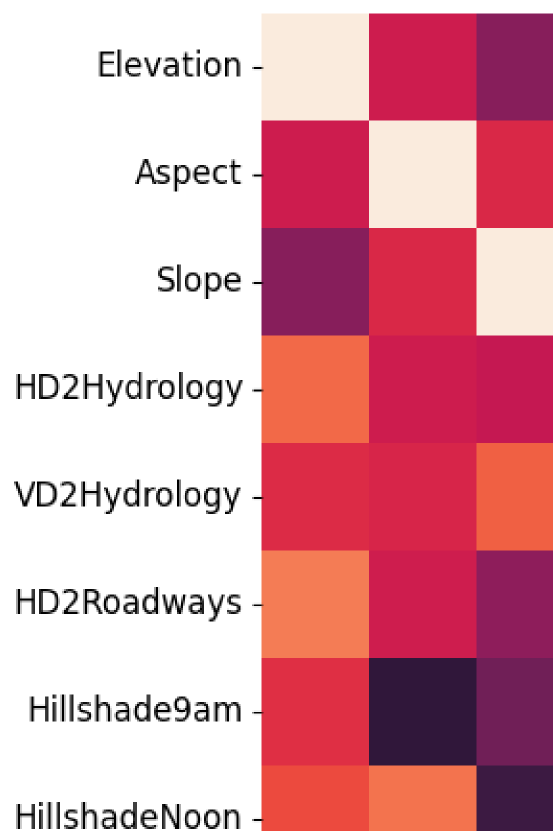
es to classify forest categories

eclined by 81.7 million hectares, which
on the environment. To preserve and
ing a detrimental effect on the ecology of
f flora that is endemic to it, and ensure
.

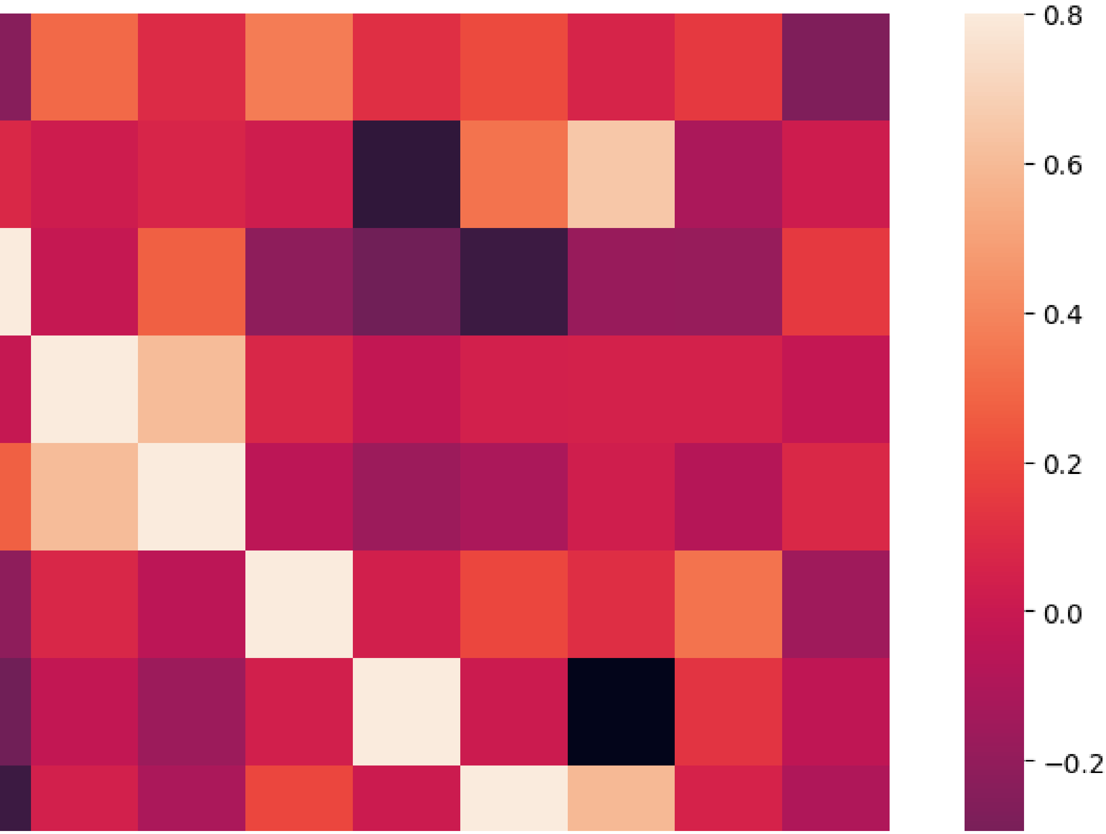
o model the environmental patterns that
o predict which plants would be suitable

ferent features of the types of forest
ated in the Roosevelt National Forest of
0x30m^2 patch.

The heatmap (fig. 2) pre
There is high correlation
Elevation also has high c
certain species of plants
Wilderness_Area_1 and
0, with the absence of bo
pair plot in fig. 3).



sents Pearson's R correlation values for the numerical features.
 between features Aspect, Hillshade_9am and Hillshade_3pm.
correlation with the target variable Cover_Type, suggesting that
 are better adapted to survival at certain heights.
 Wilderness_Area_3 are also highly correlated with Cover_Type
oth of the former signalling the presence of the latter (visible in the



University of Colorado

Feature Engineering

We added a single new feature, Euclidean_Distance_From_Hydrolo
the vertical and horizontal distance from hydrology into a single valu
columns were dropped from the data.

Based on the high correlation between the Aspect and Hillshade9am
column was also dropped.

Modeling and Model Analysis

We ran three different models on our dataset, the the F1 scores for w
The KMeans model was unsupervised, whereas the other two were
supervised fashion. The XGBoost algorithm was set to use 300 estin
subsampling ratio of 0.25. The neural network consisted of ReLU ac
followed by CrossEntropy loss and was trained using SGD (architecu
The models were all evaluated with the F1 score.

o Boulder

gy, which combines
e. The original two

n/3pm, the Aspect

which are in table 3.
trained in a
mators and a
tivated linear layers
ure given in fig. 5).

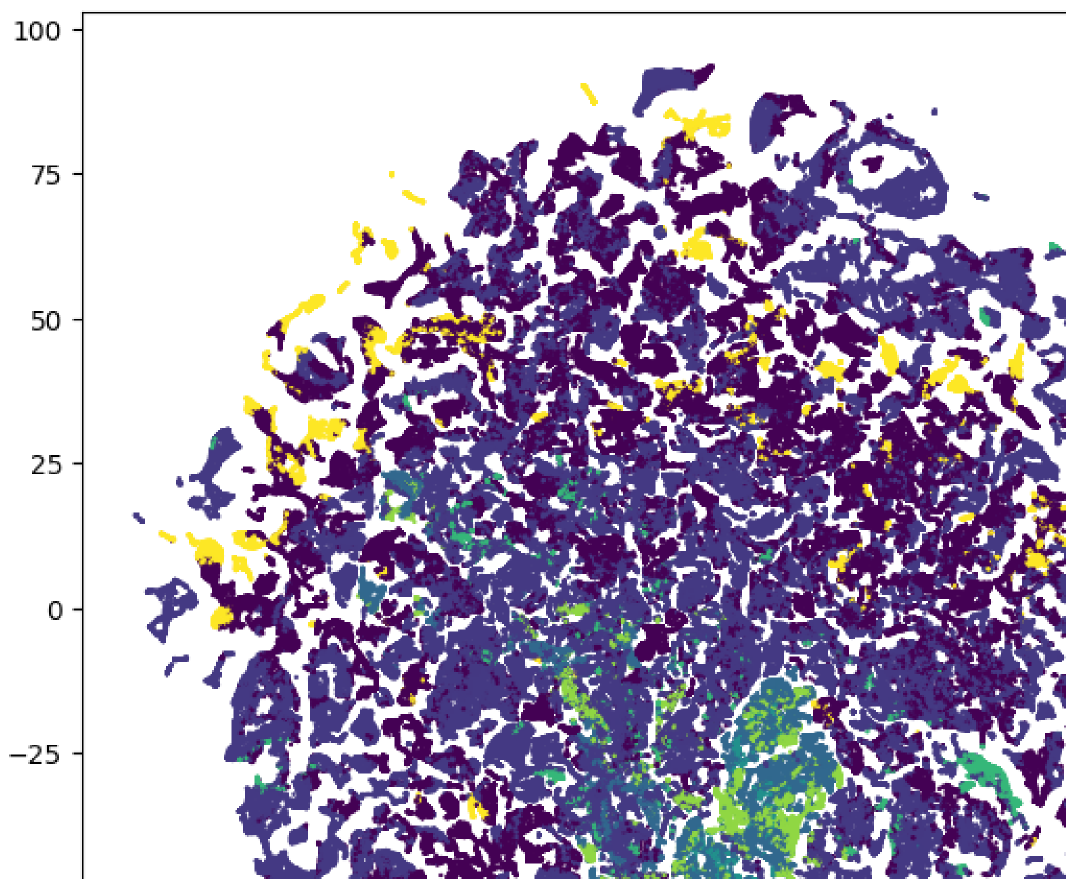
Model	F1 Score

The dataset contains. Some of the critical feature
sources, distance from roadways and soil type.

Numerical Features	**Categorical Featu**
Elevation Aspect Slope Horizontal Distance to Hydrology Vertical Distance to Hydrology Horizontal Distance to Roadways Hillshade at 9 am Hillshade at Noon Hillshade at 3 pm Horizontal Distance to Fire Points	Soil Type Wilderness Area Cover Type *(Prediction T*

Data Analysis

As can be seen in the t-SNE visualization (fig. 1) t
only culsters of cover type 4 and 3 showing any s



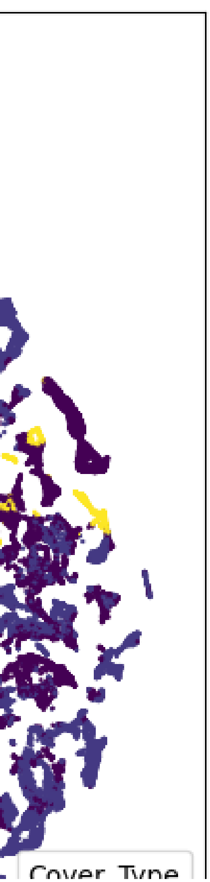
s include elevation, distance from water

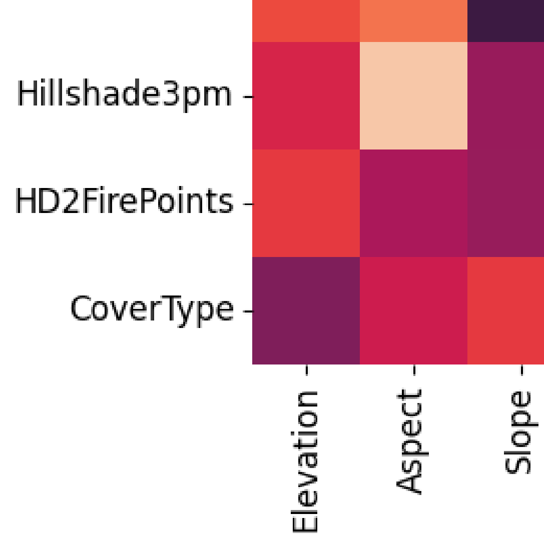
res

Target)

Table 1: Numerical and categorical features in the data.

the dataset is very homogenous with ignificant separation from the rest.





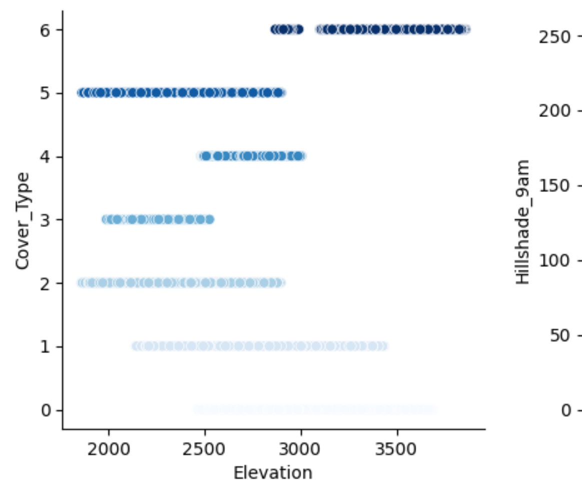
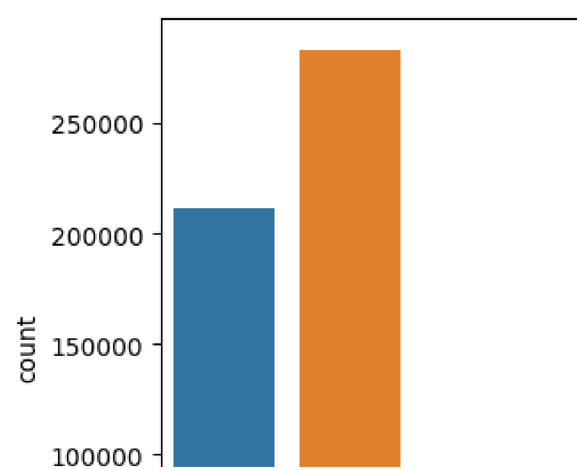


Figure 3: Pa

The dataset demonstrate in fig. 4 we can see that features themselves also scaling/standardization/r given in table 2.



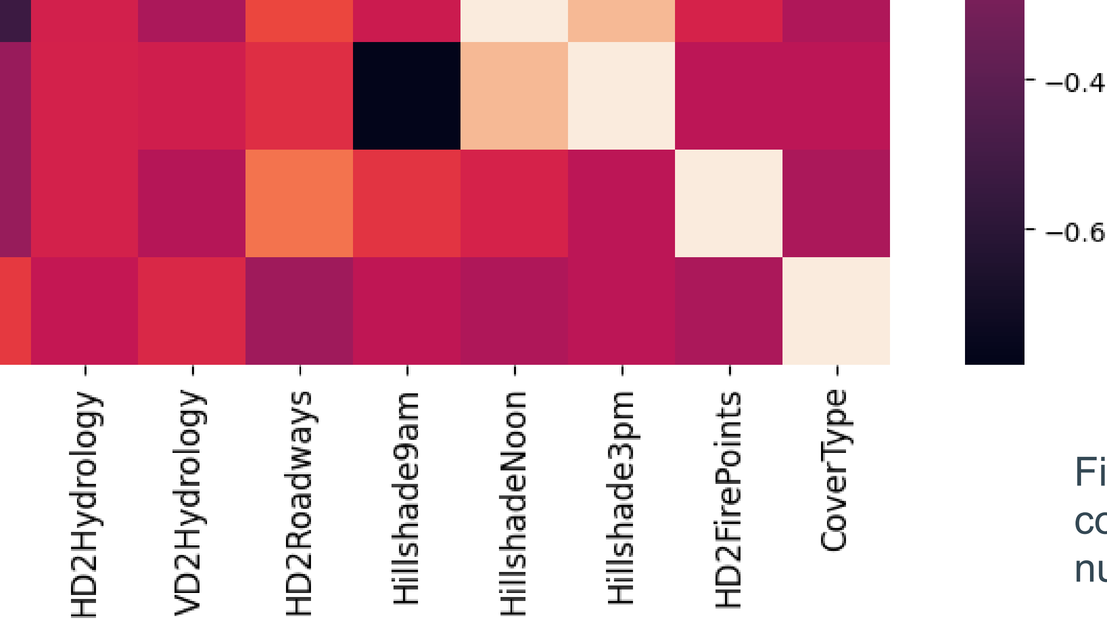
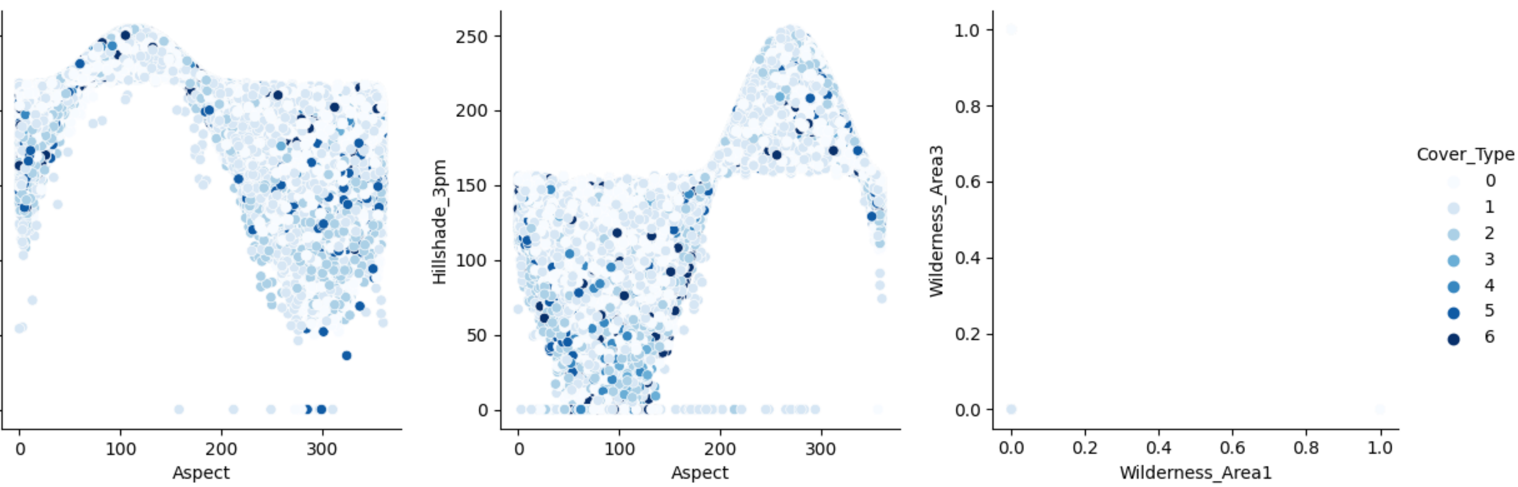


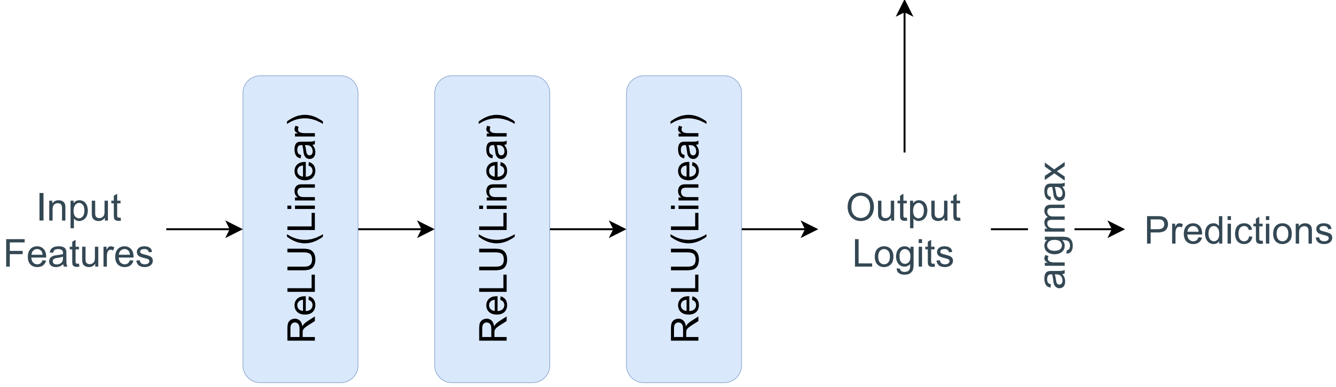
Figure 2: Pearson's R correlation heatmap for numerical features



irplots showing the relationships between some correlated features.

es quite a bit of skew as well. In the classwise sample distribution the dataset is biased heavily towards cover types 0 and 1. The o vary in symmetry and spread, necessitating the need for normalization. The skewed features before and after treatment are

Skew	Before Treatment	After Treatment
High	HDHydrology VDHydrology Hillshade9am HillshadeNoon HDFirePoints	None
Moderate	Elevation Slope HDRoadways	HDHydrology VDHydrology HDRoadways HDFirePoints



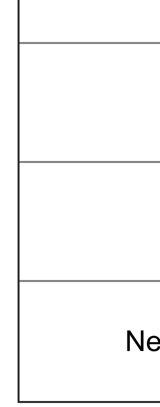


Figure 5: Neural Network Architecture

As expected from the homogeneity of the t-SNE visualization, the KM
unable to find distinct clusters in the data. The XGboost algorithm pe
too was outdone by the neural network which converged to the highe

We also performed ablation tests, removing each feature in turn and
of the final score. Elevation was found to be the most impactful featu
Euclidean_Distance_From_Hydrology and Slope.

Conclusion

In this project we systematically present our methodology to predict t
given cartographic and topological data. We found that there is signi
between such features and the types of plants that will grow in these
demonstrate how we can deal with biases in the data, and model it u
machine learning algorithms. Our experiments result in a neural mod
cover types with a high degree of success.

References

- Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Fore
  Variables (Blackard et al., 1999)
- Accurate Decision Trees for Mining High-speed Data Streams (Gama et al., 2003)
- Round Robin Rule Learning (Furnkranz et al., 2002)

Team

KMeans	0.287
XGBoost	0.686
eural Network	0.915

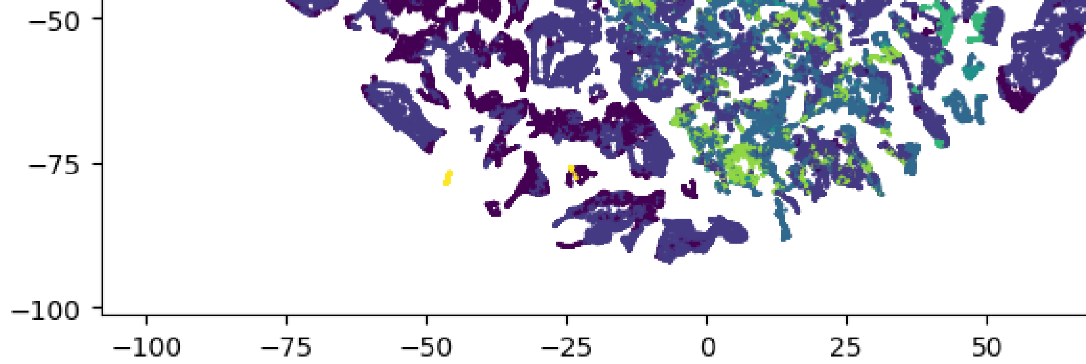
Table 3: Results

Means algorithm was
erformed better, but it
est F1 score of ~91%.

examining its impact
ure, followed by

forest cover types,
ficant correlation
e conditions. We
using three different
del that can predict

est Cover Type from Cartographic


−50
−75
−100
−100
−75
−50
−25
0
25
50

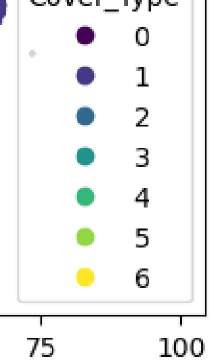


Figure 1: t-SNE visualization of the dataset

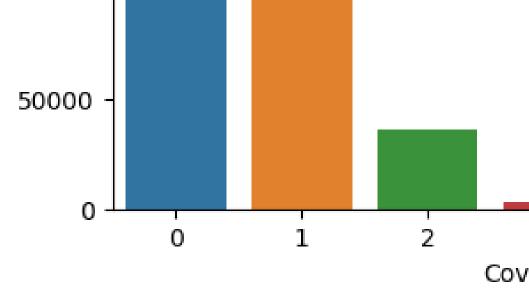
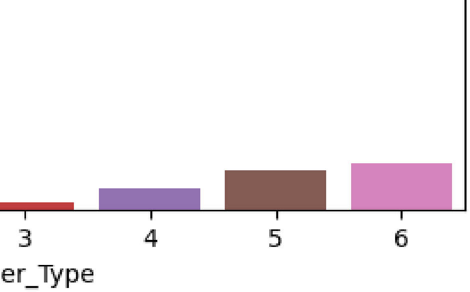


Figure 4: Classwise san



nple distribution

Fair	Aspect Hillshade3pm	Elevation Aspect Slope Hillshade9am HillshadeNoon Hillshade3pm

Table 2: Treating skewed features.

Team

my (ana)conda don't

Aditya Srivastava, Harsh Gupta, Konigari Rachna