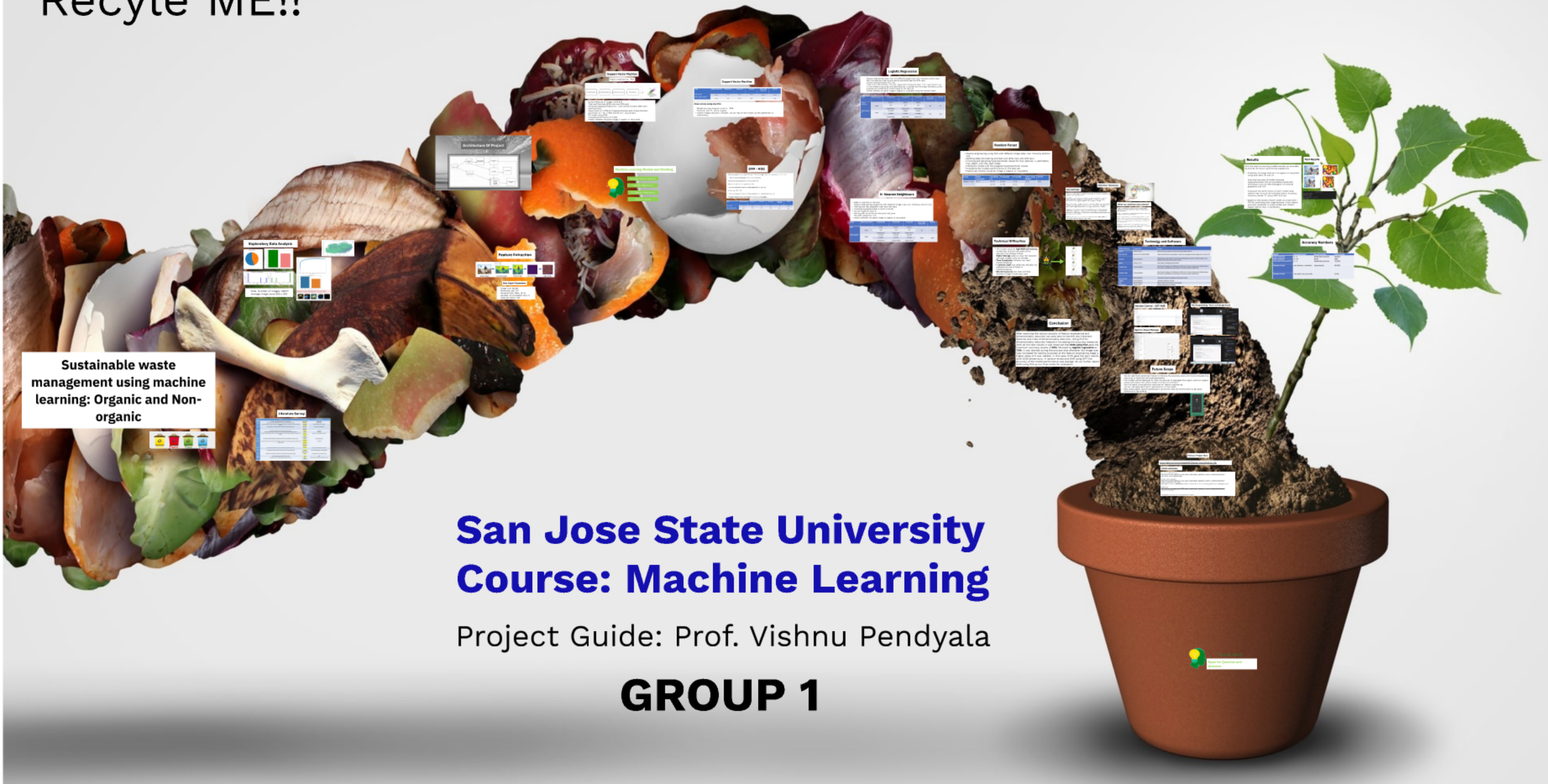
SAVE ME !!!
Recyle ME!!

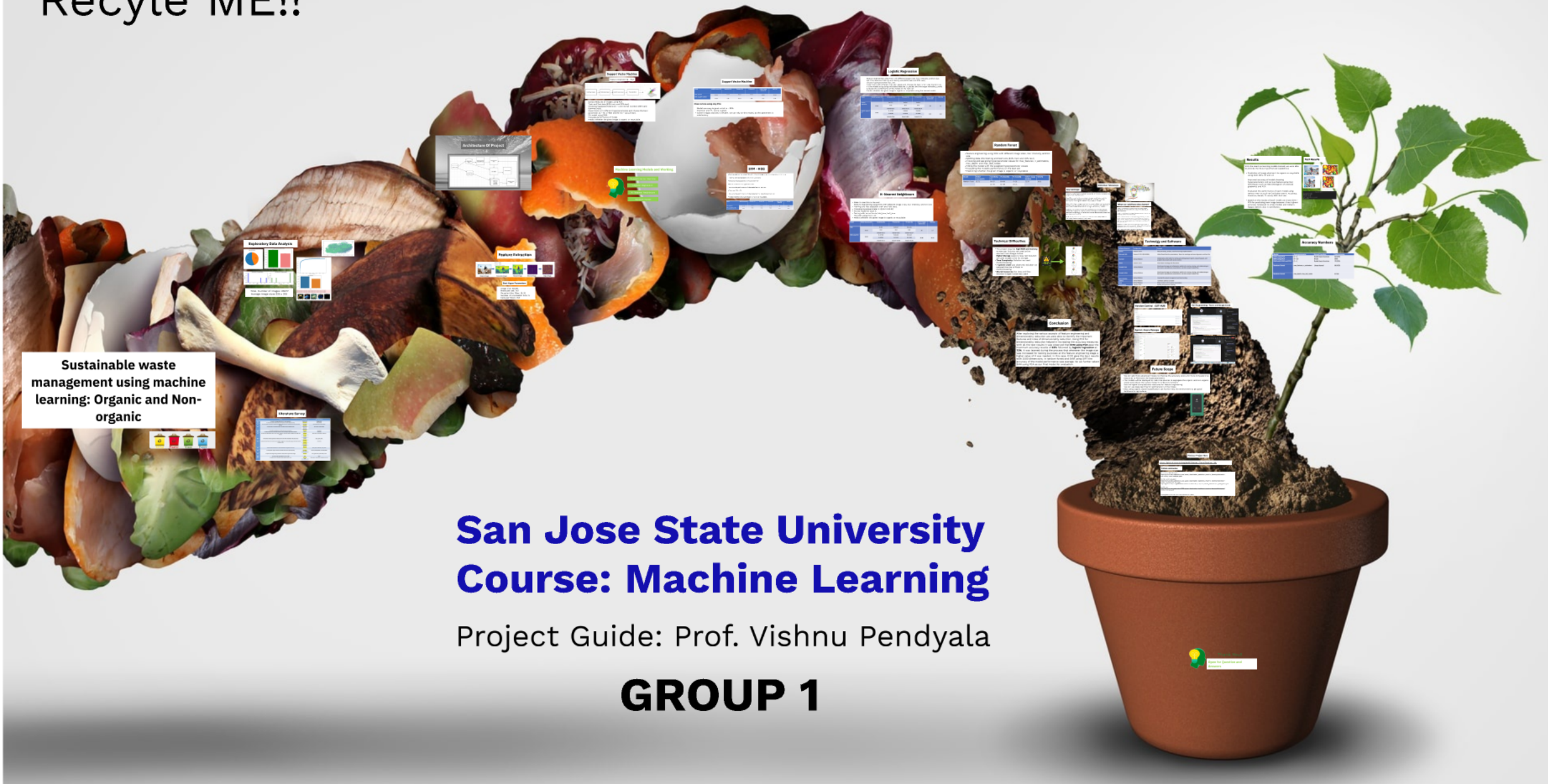


San Jose State University
Course: Machine Learning

Project Guide: Prof. Vishnu Pendyala

GROUP 1

SAVE ME !!!
Recyle ME!!



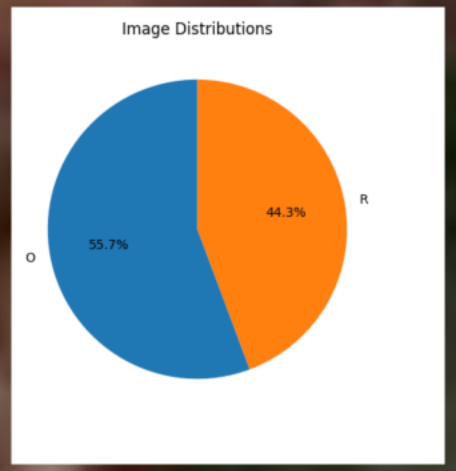
San Jose State University
Course: Machine Learning

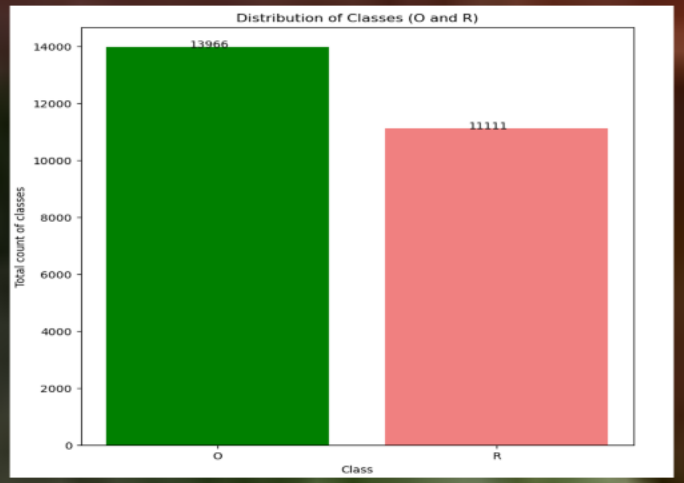
Project Guide: Prof. Vishnu Pendyala

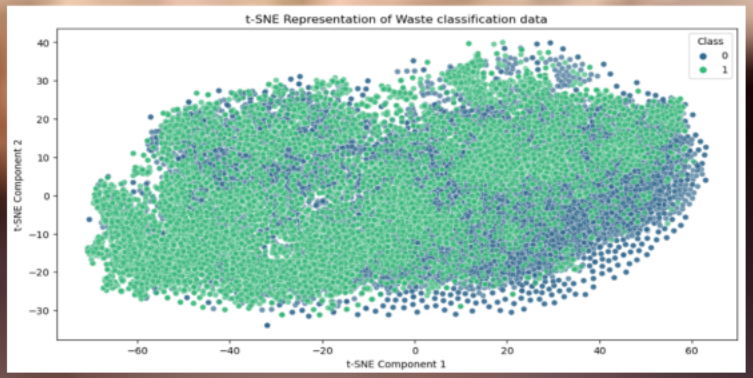
GROUP 1

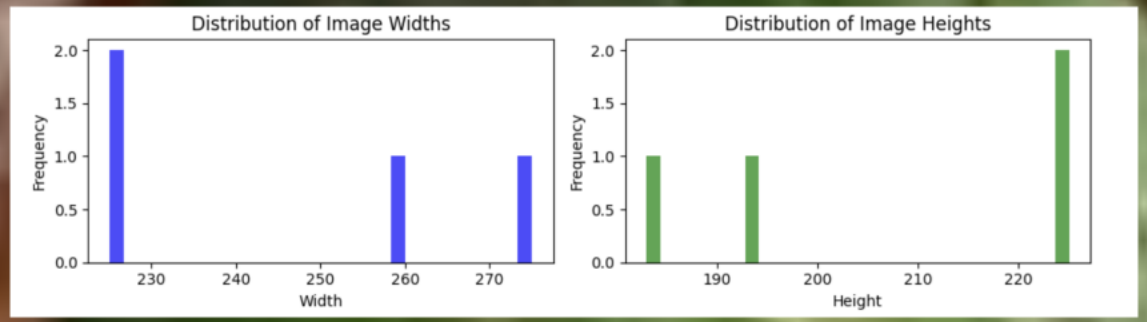
Sustainable waste management using machine learning: Organic and Non-organic

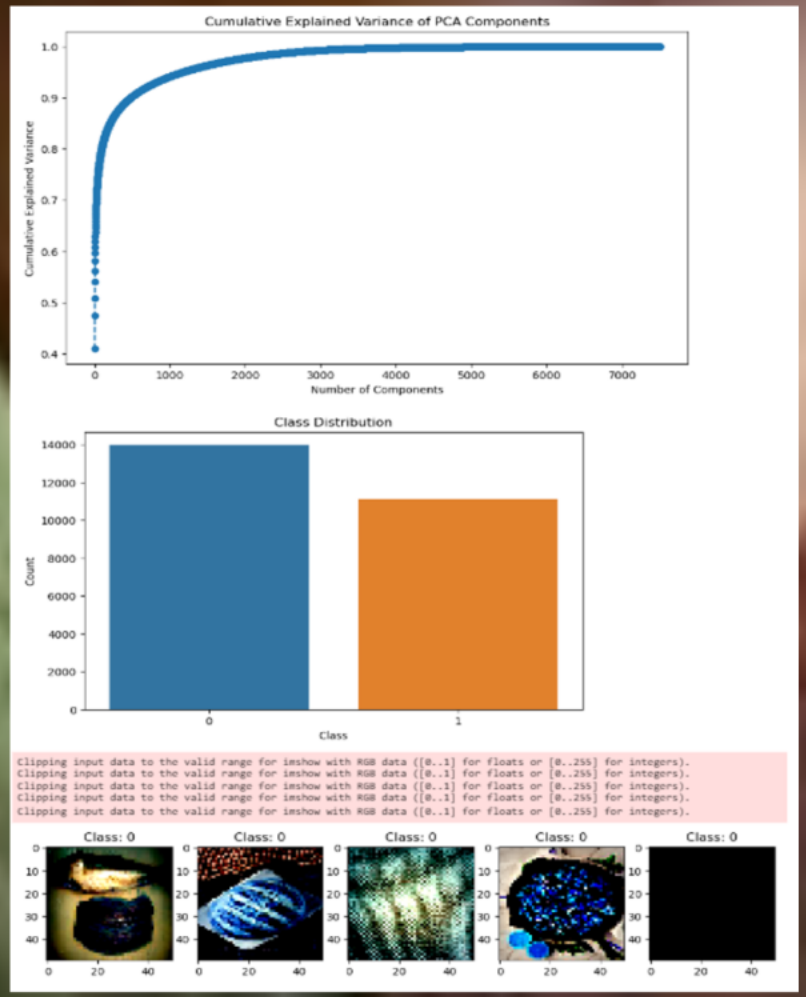
Exploratory Data Analysis









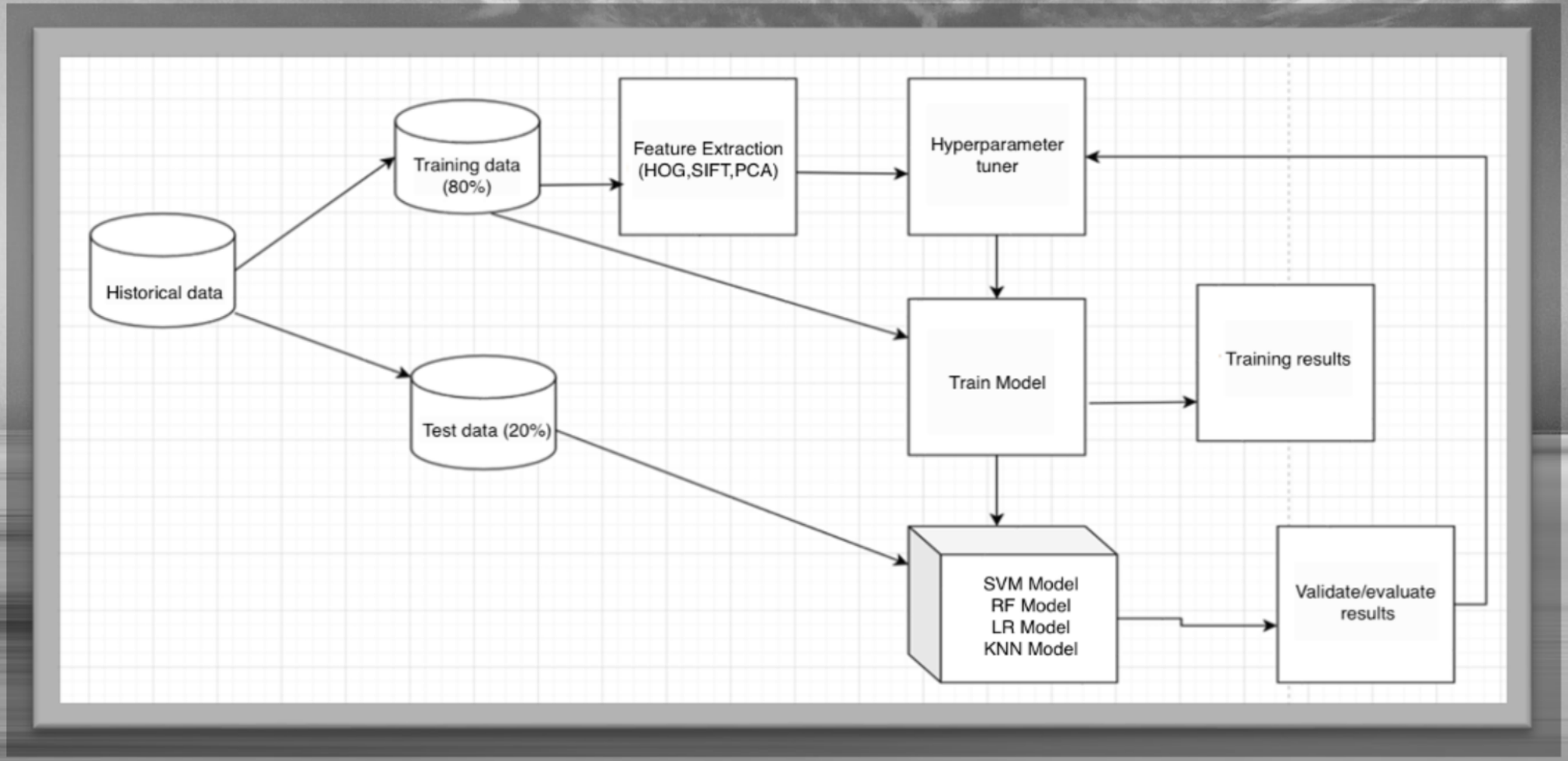


Total Number of Images :25077
Average image sizes 225 x 225

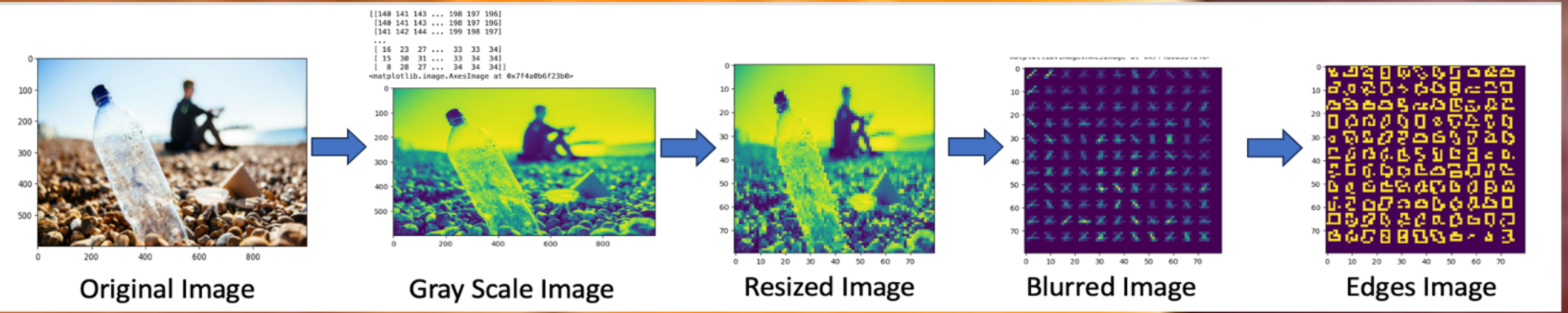
Literature Survey

Sr. No.	Purpose	ML Algorithm	Waste Classification
[1]	To classify the biodegradable waste for reducing pollution	CNN:88.54%	Biodegradable
[2]	To classify the plastic bottles based on their position and color during recycling.	SVM:94.7%	Plastic Bottles
[3]	To use computer vision system using CNN and SVM for classification of handcrafted and non-handcrafted features.	CNN:>90% SVM:>92%	Handcrafted and non-handcrafted
[4]	To extract features using GB and MLP and fragments CDW using RGB images.	MLP:91.3% GB:92.3% CNN:85.9%.	Construction and demolition
[5]	To classify household waste and induvial and community level.	KNN:93%	Household
[6]	To collect waste from the bins when the bins are full by defining the path towards the bins	LR:97%	Smart Bins
[7]	To collect and organize glass, paper, metal, plastic, cardboard, and waste images into RGB format images from datasets.	SVM:85% RF:55% DT:65% CNN:90%	Glass, paper, metal, plastic, cardboard
[8]	To find amount of waste generated on daily basis from the families and between the year 2010-19	LiR:87% SVM:88%	Glass, paper, plastic
[9]	To collect the information from smart bins and notify the magement to empty bins using an automated machine learning process.	KNN:96.3% LR:96.6% SVM:96.8% DT:96.6% RF:97.2%	Smart Bins
[10]	To find out SFIT-PCA features for waste classification using feature extraction	SVM:62%	Glass, paper, cardboard, metal, plastic
[11]	To classify waste using AI based waste classifier with thermo rapid composting	SVM:>85% CNN: >85%	Industrial, biodegradable, nonbiodegradable
[12]	To optimize SVM using boosting methods to classify different types of waste images	SVM:95%	Cans, cigarette butt, plastic bottle, carton
[13]	To classify images based different layers of CNN	CNN	Plastic
[14]	To classify trash for recyclability status using computer vision	SVM:63% CNN	Glass, paper, metal, plastic, cardboard, and trash

Architecture Of Project



Feature Extraction



Best Hyper Parameters

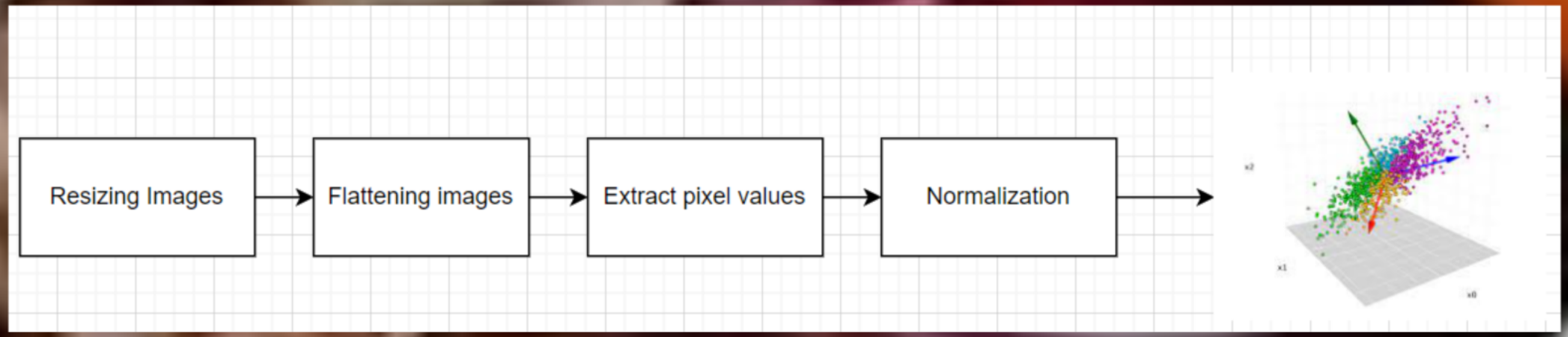
- Image size: 80x80
- Pixels per cell: 7x7
- Gaussian blur filter: (5, 5)
- Number of orientation bins: 5
- Cells per Block: 2x2

Machine Learning Models and Working

- Support Vector Machine
- Logistic Regression
- K- Nearest Neighbours
- Random Forest

Support Vector Machine

Feature Engineering - PCA



- Extract features of images using PCA.
- Train and Test data (80% train and 20% test).
- Choosing hyperparameters (C= 1) and Kernel function (RBF) and Gamma (Auto).
- Experiment with different hyperparameters and choose the best parameter. (C = 10, C=100) and Kernel = (poly,linear).
- Fit model using SVM.
- Evaluate performance of model.
- Predict whether the given image is organic or recyclable.

Support Vector Machine

Model	Model Accuracy (%)	Sensitivity (%)	Precision (%)	F1-Score (%)	Cohen's kappa statistic (%)	MCC (%)
SVM with PCA	82.64	77.47	82.11	79.72	64.56	64.64
SVM using SIFT and PCA	55.80	1.36	46.15	2.64	0.12	0.47

Observations using only PCA

- Model accuracy is good which is ~ 83%.
- Precision and F1- score is good.
- Cohen's kappa statistic is 64.56%, we can rely on this model, as the agreement is satisfactory.

SVM - HOG

- Feature engineering using HOG with different image sizes, blur intensity, and bin size
- Training and Test data(80% train and 20% test).
- Choosing Hyperparameter C=1.0, kernel='rbf'
- Fit the model for the hyperparameter.
- Evaluate the performance of the model on the test set.
- Then use PCA = 2K.
- Evaluate the performance of the model on the transformed test set.
- Predict whether the given image is organic or recyclable.

Model	Model Accuracy (%)	Sensitivity (%)	Precision (%)	F1-Score (%)	Cohen's kappa statistic (%)	MCC (%)
SVM with HOG	67.76	67.76	67.84	67.79	35	35.09
SVM using HOG and PCA	70.95	70.95	71.28	71.01	41.79	41.91

Logistic Regression

- Feature engineering using HOG with different image sizes, blur intensity, and bin size.
- Split the data into training and testing sets (80% train and 20% test).
- Choose hyperparameter max_iter.
- Experiment with different max_iter values and choose the best, which was found to be
- Fit the model using using the scaled features (X_scaled) and the target variable (y_train).
- Evaluate the performance of the model on the test set:
- Predict whether the given image is organic or recyclable using the trained model.

Model	Model Accuracy (%)	Sensitivity (%)	Precision (%)	F1-Score (%)	Cohen's kappa statistic (%)	MCC (%)
Logistic regression	67.86	R= 71.5	R=62.6	R=66.8	18	2.7
		O=0	O=0	O=0		
		Overall=45.21	Overall =0.01	Overall =62.27		
Logistic regressi on	67.86	R= 67.98	R=64.46	R=66.18	35.9	36.2
		O=69.06	O=72.33	O=70.66		
		Overall=67.86	Overall =68.5	Overall =67.9		

K- Nearest Neighbours

- Make it clear this is the end
- Feature engineering using HOG with different image sizes, blur intensity, and bin size
- Training and Test data(80% train and 20% test).
- Choosing Hyperparameter K =[3,5,7,10,15,20]
- Fit the model for each K
- Testing with kernel brute, kdd_tree, ball_tree
- The KNN using PCA = 2K
- Predict whether the given image is organic or recyclable

Model	Model Accuracy (%)	Sensitivity (%)	Precision (%)	F1-Score (%)	Cohen's kappa statistic (%)	MCC (%)
KNN	65.62	R= 99	R=44.57	R=61.65	18	2.7
		O=0	O=0	O=0		
		Overall=45.21	Overall =0.01	Overall =62.27		
KNN using PCA	68.58	R= 67.98	R=64.46	R=66.18	36.88	36.92
		O=69.06	O=72.33	O=70.66		
		Overall=65.11	Overall =65.88	Overall =65.06		

Random Forest

- Feature engineering using HOG with different image sizes, blur intensity, and bin size
- Splitting data into training and test sets (80% train and 20% test)
- Choosing and assigning hyperparameter values for max_features, n_estimators, max_depth, and max_leaf_nodes
- Fitting the model with the assigned hyperparameter values
- Evaluating the model's performance on the test set
- Predicting whether the given image is organic or recyclable

Model	Model Accuracy(%)	Sensitivity(%)	Precision(%)	F1-Score(%)	Cohen's kappa statistics	MCC(%)
Random Forest	66.71	R = 57	R = 64	R = 60	32.93	33.01
		O = 74	O = 69	O = 71		

Results

With the machine learning models trained, we were able to provide the following enhanced capabilities

- Prediction of image whether it is organic or recyclable using SVM, KNN, RF and LR
- Improved accuracy of model choosing hyperparameters and various feature extraction techniques such as HOG (Histogram of oriented gradients) and PCA.
- Evaluated the performance of each model using various metrics such as Confusion Matrix, Accuracy, Precision, Recall, F1-score, ROC-AUC etc.
- Based on the results of each model we chose SVM – PCA for predicting new image because it has highest accuracy compared to other models and Cohen's kappa statistic also is satisfactory

Test Results


```
----------Evaluation metrics----------
Test Accuracy: 82.64%
Sensitivity is 77.47%
Precision is 82.11%
F1-score is 79.72%
Cohen - kappa statistic is 64.56%
MCC is 64.64%
```

Accuracy Numbers

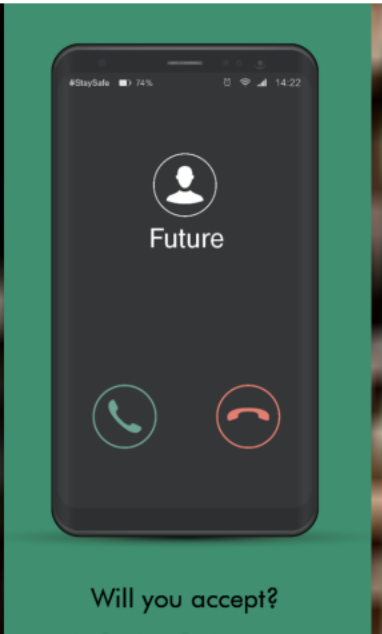
Model	Hyperparameter	Kernel Function	Test Accuracy
SVM using PCA	C = 1	Radial basis function	82.64%
KNN using PCA	K = 20	Brute	68%
SVM with HOG and PCA	C = 1	Radial basis function	70.95%
Random Forest	max_features, n_estimators	Linear kernel	66.42%
Random Forest	max_depth, max_leaf_nodes		67.26

Conclusion

After exploring the various accepts of feature engineering and dimensionality reduction we were able to identify the important features and roles of dimensionality reduction. Using PCA for dimensionality reduction helped in increasing the accuracy measures. With all the test results it was observed that **SVM using PCA** gave the maximum accuracy results of **83%** followed by **logistic regression** at **72%**. It was learned during the process that whenever the image size was increased for testing purposes at the feature engineering stage a higher value of K was needed. In this case, K=20 gave the best results with 2000 dimensions. In random forest and SVM using SIFT the accuracy of the model performance was average. So we further select SVM using PCA as our final model for evaluation.

Future Scope

- We can add more advanced models to improve the accuracy score and more computational resources to fine-tune the hyperparameters.
- The models will be deployed on real-time devices to segregate the organic and non-organic waste and reduce the carbon footprint in the environment.
- Use the higher computational resoucres for feature engineering
- We can use deep learning for optimazation of the model.
- Also using organic waste classification can further help the environment to get good fertilizers for agriculture.

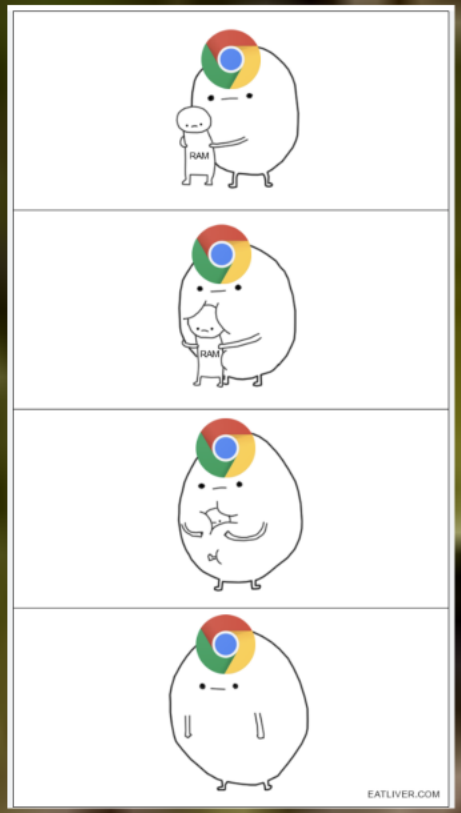


Technolgy and Software

Software Resource	Configuration	Purpose
Python	Version 3.10.0	Data processing, analysis, machine learning implementation
Microsoft 365	Version 16.78.3 (23102801)	Utilize PowerPoint for presentations, Teams for meetings and team alignment, and Excel for
Overleaf	Online Platform	Collaborative LaTeX editor for writing and editing project reports, research papers, and academic documents with version control.
Zoom	Version 5.16.2	virtual team meetings and discussions
Google Drive	Online Platform	Cloud-based storage and collaboration platform for storing, sharing, and collaborating on documents, spreadsheets, presentations, and other project-related files.
Google Collab	Online Platform	Cloud-based storage and collaboration platform for storing, sharing, and collaborating on documents, spreadsheets, presentations, and other project-related files.
Azure DevOps	Online Platform	Essential for project management and task tracking
GitHub	Online Platform	Codebase publicly accessible
Prezi	Online Platform	Make creative and innovative presentation
Grammarly	Version 2.0.0	Spelling and grammar checking

Technical Difficulties

- The project requires **high RAM and memory** to execute: Solution we used online services from Google Collab
- **Higher Storage** space is required: Solution we used Google Drive for storage.
- **Time Complexity**: Solution we used multiprocessing.
- A **system crash** was observed: Solution we reduced the size of Pools in multiprocessing.
- **Shared resources** like data and files:- Solution Google Collab was used



Individual Takeaways

Key learnings:

Yukta :Fixing the hyper parameters especially at the feature engineering stage is most important part.

Maral:This is my first time working on image classification, and this project presents a fantastic opportunity for me to learn various techniques, including the HOG technique, SIFT, and SURF.

Krinal: Got to learn image classification using multiple machine learning models using techniques such as HOG, SIFT and SURF. And when the depth and breadth of the tree is not huge, the accuracy is better

Meghana : To optimize the run time performance of training images (nearly 25k images), we need to reduce the image size and flatten it otherwise the run time will be forever and use dimensionality reduction techniques such as PCA

Sejal: It was nice to explore more on the data. HOG, SIFT, SURF are interesting feature extraction techniques.

What we could have done better?

Yukta:- I could have build functions of KNN from scratch instead of using direct functions. It gives more flexibility to debug the errors.

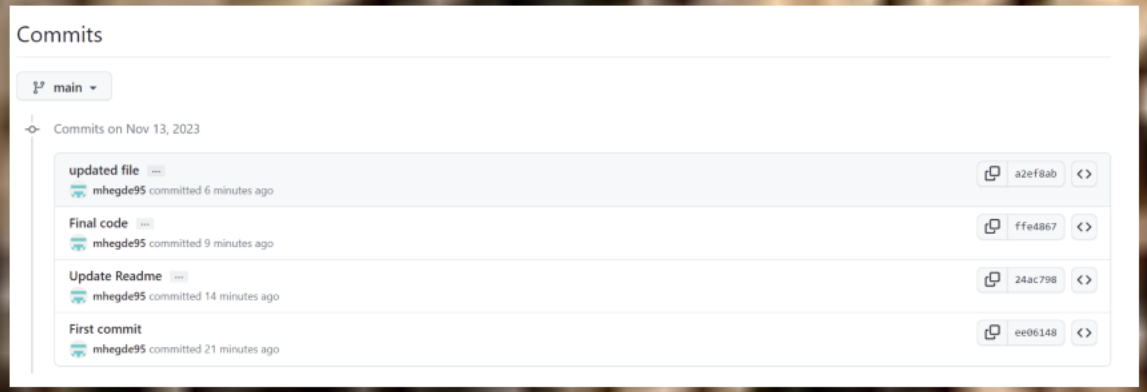
Maral:- I could have done better hyperparameter tuning for my model to achieve improved results.

Krinal:- Random forest is computationally intensive. If I have more computational resources, the performance of random forest model can be further improved by feature engineering/ selection, increasing n_estimators, finding optimum value for max_features, etc.

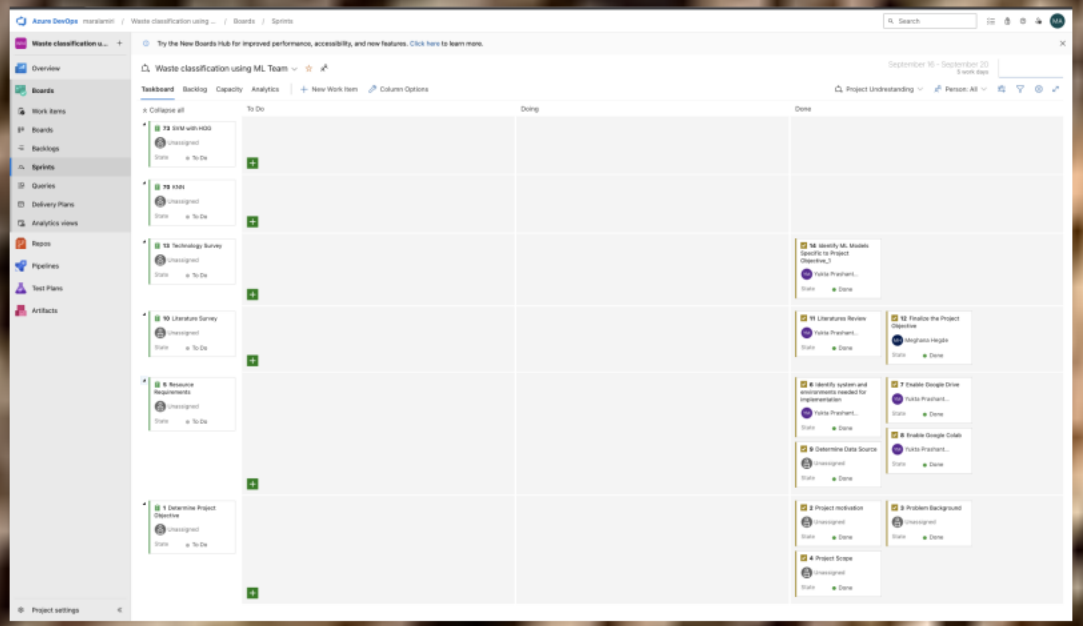
Meghana:- I could have experimented with more number of hyperparameters and make use of GridSearchCV to choose best hyperparameter, to improve accuracy of model.

Sejal : Could have used more optimized features

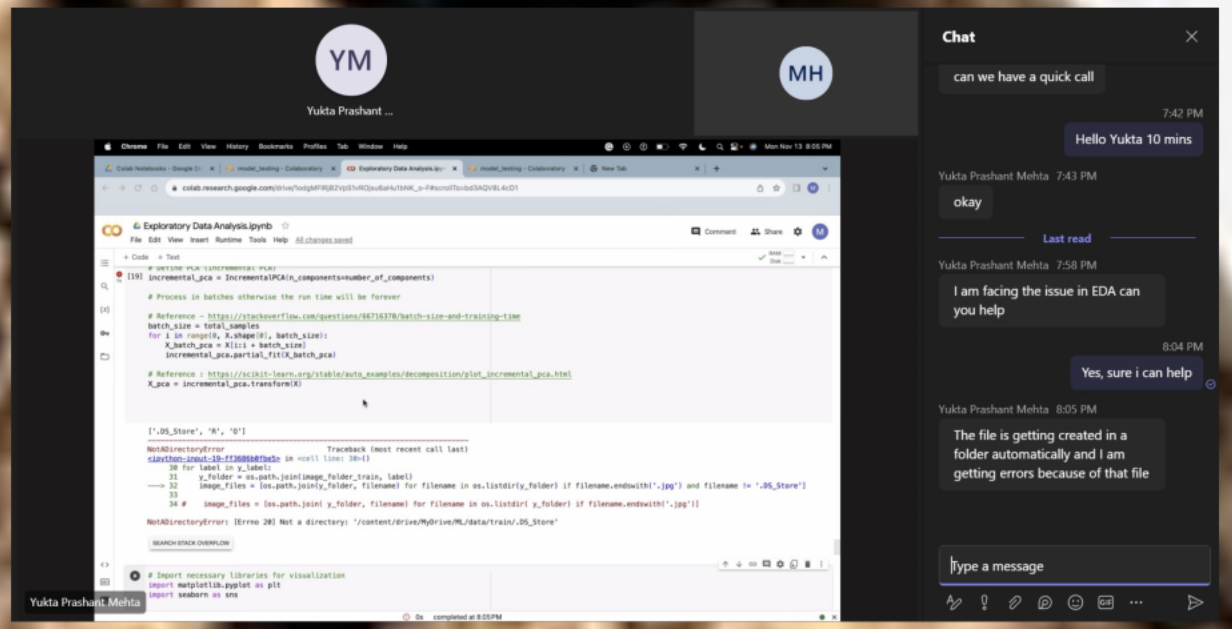
Version Control : GIT HUB

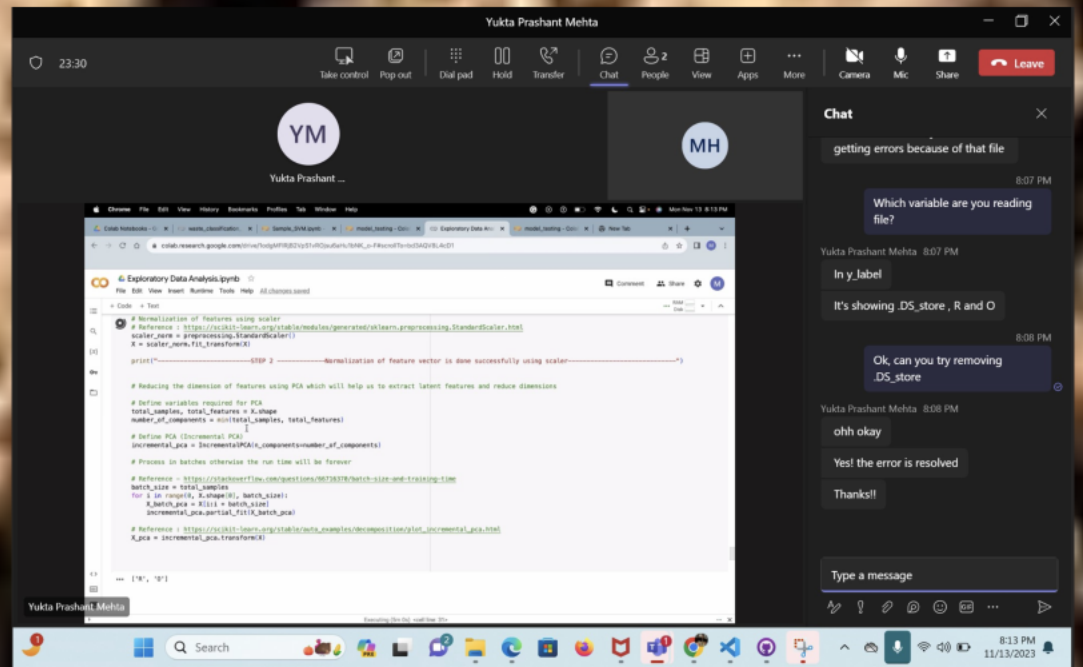


Sprint: Azure Devops



Pair Programming : Teams and Google Colab





Find our Project Here

https://github.com/mhegde95/Waste-Classification-ML

Collab notebooks:

Exploratory Data Analysis:
https://github.com/webrockerz2020/waste_classifcation_traditional_machine_learning/blob/main/Exploratory_Data_Analysis.ipynb

Models and Evaluation:
https://github.com/webrockerz2020/waste_classifcation_traditional_machine_learning/blob/main/waste_classification_ml.ipynb
https://github.com/mhegde95/Waste-Classification-ML/blob/main/Waste_Classification_SVM_PCA.ipynb

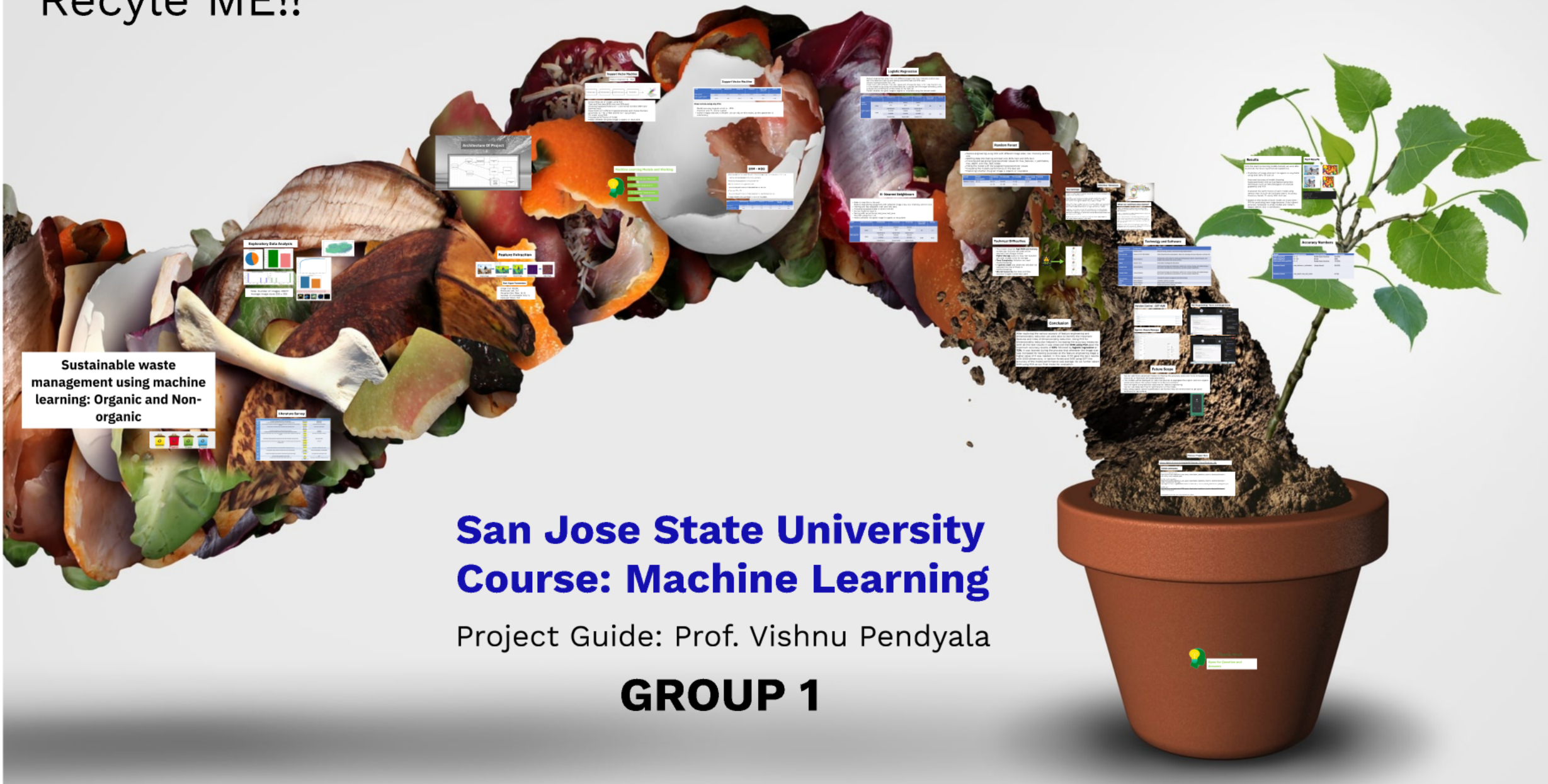
Model Test:
https://github.com/webrockerz2020/waste_classifcation_traditional_machine_learning/blob/main/model_testing.ipynb

Complete code execution take 45mins to 1:30 hr.

Thank You!!

Open for Question and Answers

SAVE ME !!!
Recyle ME!!



San Jose State University
Course: Machine Learning

Project Guide: Prof. Vishnu Pendyala

GROUP 1