A Dynamic Two-stage Machine Learning Approach for the Selection of a UE-VBS in a 5G Network

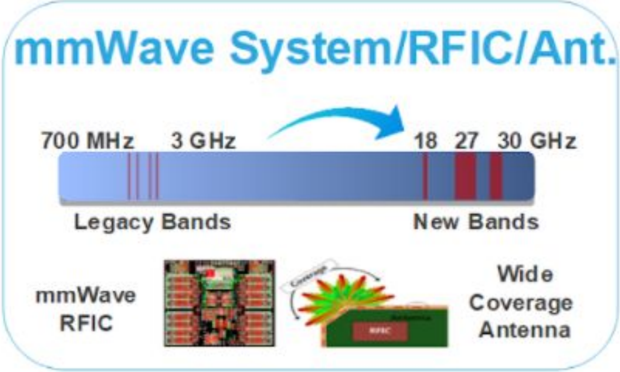
Nivedhitha D
Sriram M

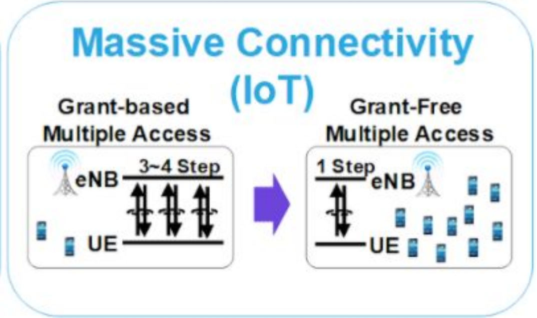
Department of Computer Science & Engineering
SSN College of Engineering

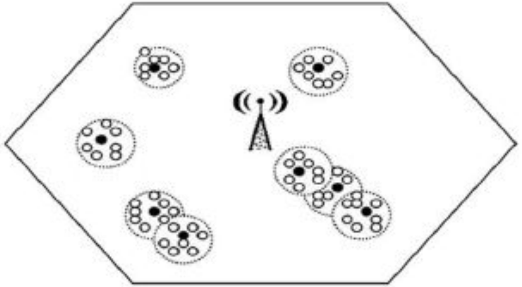
26 October 2021

Agenda

1. Introduction
2. Problem Description
3. Literature Survey
4. Experimental Setup
5. Proposed Methodology
    a. Clustering
    b. Classification
6. Results
    a. Model Evaluation Metrics
    b. Statistical Analysis
7. Conclusions
8. Further Improvements
9. Acknowledgements

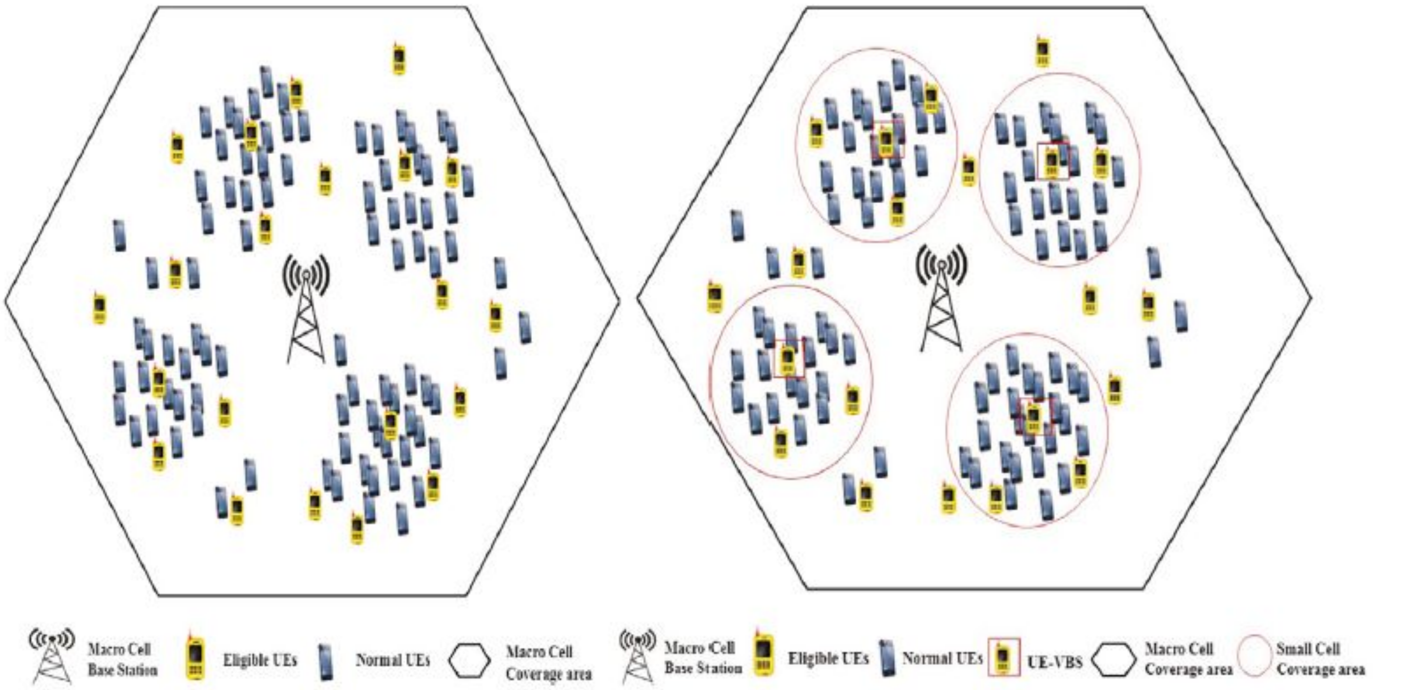



Massive Connectivity (IoT)
Grant-based Multiple Access
Grant-Free Multiple Access
eNB
3~4 Step
UE
1 Step
eNB
UE



Introduction

The 5G cellular network is a new generation of mobile networks that will be critical in meeting current and future needs for wireless access by heterogeneous devices in a dense network. This study builds on the **User Equipment-based Virtual Base Station (UE-VBS)** concept, which allows smartphones to **offer the services of a Base Station to UEs** in the perimeter by capitalizing on the massive connectivity and the enhanced resources of the new UE generation in terms of connectivity, computing, and battery/power. The selection of a qualified UE to become a VBS is difficult due to the **asymmetric dynamics of human movement in the temporal and spatial domains, UE hardware configurations**, and **network infrastructures**. In a 5G dynamic network architecture, a Machine Learning (ML) model can be used to automate the dynamic selection of eligible UEs that can be activated on the fly into a UE-VBS to **support data rate expansion** and **improve QoS** in locations where infrastructure is lacking and a more agile network operation is required.



(a) No UE-VBS activated: All connections are handled by the BS

(b) UE-VBS activated: VSCs are deployed to complement and support the network infrastructure

UE-VBS Concept: **Formation of VSC among UEs.**

Problem Description

- The main objective of this paper is to develop a **Two-stage Machine Learning Engine** that can be used to activate UE-VBS at places adjacent to where data consumption occurs dynamically, resulting in enhanced service quality across the cluster
- The proposed solution divides the process of selecting a UE-VBS into two stages: (i) **clustering** UEs based on their **distance from the BS** and (ii) **binary classification** of UEs in each cluster into **eligible and ineligible UEs** based on the device's **QoS, QoE, data rate, distance from the BS, power/battery, and processing power**
- Finally, a UE is activated to become the UE-VBS (cluster head)

Literature Survey

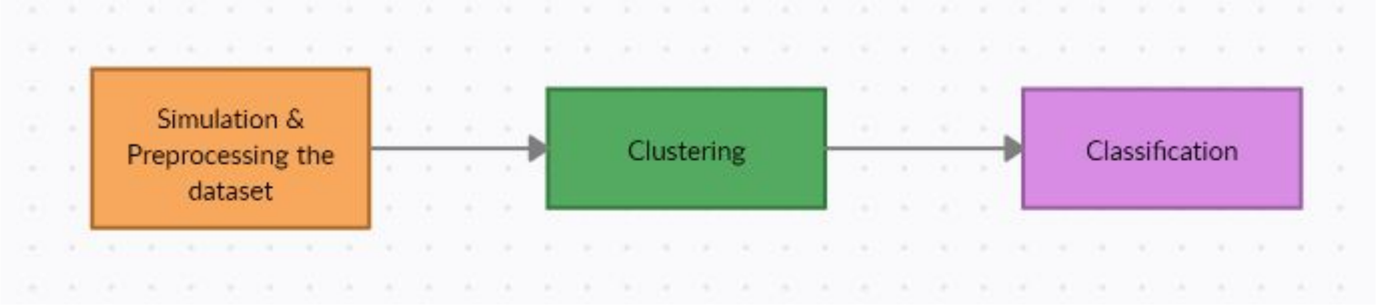
Name	Authors	Methodology Used
Selection of UE-based Virtual Small Cell Base Stations using Affinity Propagation Clustering,(2018)	P. Swain, C. Christophorou, U. Bhattacharjee, C. M. Silva and A. Pitsillides	Affinity Propagation Clustering
Dynamic selection of Virtual Small Base Station in 5G Ultra-Dense Network using Initializing Matching Connection Algorithm(2019)	K. Venkateswararao, P. Swain, C. Christophorou and A. Pitsillides,	User Equipment-based Virtual Base Station (UE-VBS) concepts that enhances the Smartphones of the general population into Virtual Small Base Stations, IMCA and VSC activation algorithms for UE selection

Inference from Literature Study

- To understand the background of the project
- To acquire the domain knowledge to select the required parameters to generate the dataset for training the machine learning model
- To evaluate the machine learning model and improve interpretability

Proposed Methodology

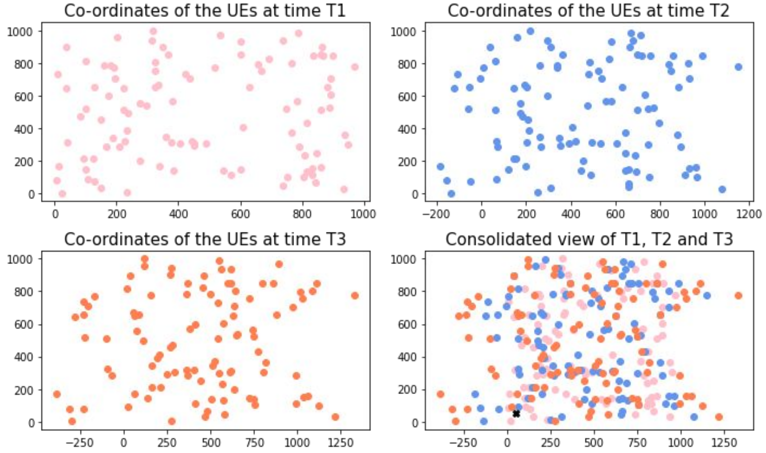
- Simulation of the UE geographical location using a non-uniform distribution in space & time domains
- Clustering the UEs based on the distance from the macro BS
- Classification based on the eligibility to become a UE-VBS



Experimental Setup

- The position of UEs were initialized as Cartesian pairs i.e., (x, y) coordinates in space. The angle, speed and time intervals were initialized randomly to demonstrate the movement of UEs in a **non-uniform distribution** that facilitates **dynamic selection**
- The experiments were also conducted for **various population densities** to further improve the reliability of the model in real world scenarios
- Following this, the coordinates were sent for clustering

Time Snapshots



Co-ordinates of the UEs at time T1

Co-ordinates of the UEs at time T2

Co-ordinates of the UEs at time T3

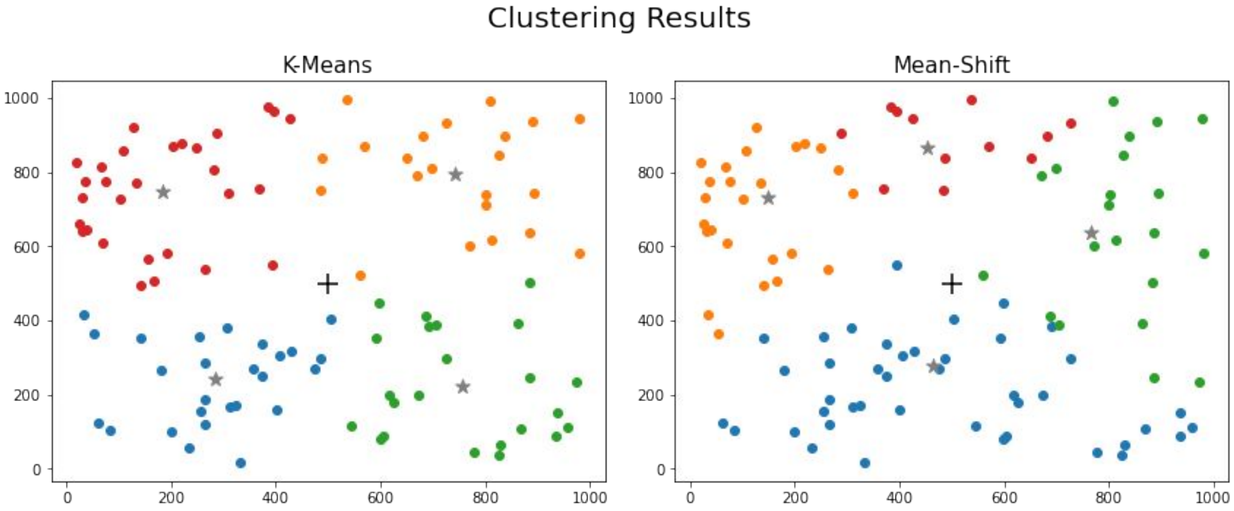
Consolidated view of T1, T2 and T3

Clustering

The nodes were clustered based on the distance from the base station. Two algorithms were considered to be effective while clustering the data points:

Algorithm	**Justification**	**Drawback**
K-Means	Efficient Algorithm and handles outliers effectively	Determining the ideal number of clusters is tedious
Mean Shift	Dynamic clustering algorithm	Struggles handling outliers

Clustering



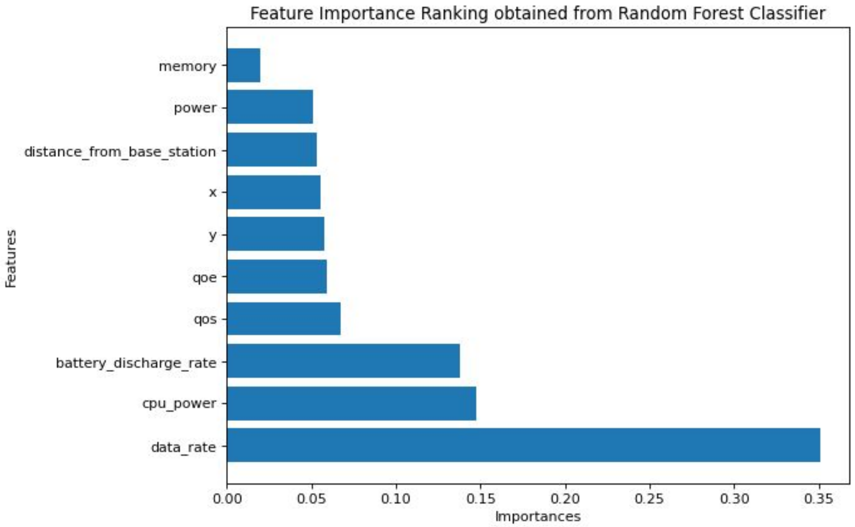
Classification

- Classification is the **second stage** of the machine learning engine and aims to categorize the UEs in the clusters as **eligible** to become a UE-VBS and **non-eligible** to become a UE-VBS
- UEs are assessed on the basis of **QoS, QoE, data rate, distance from the BS, power/battery, and processing power to determine their eligibility**
- The characteristics used to label the UEs as an eligible device to become a UE-VBS:
    - having a high data rate/spectral efficiency
    - having a good QoS & QoE value
    - good signal quality
    - close to the BS
- Eligible UEs have the **target variable** '**eligibility**' set to 1 and those which don't have the target variable 'eligibility' set to 0
- **AdaBoost**, **Gradient Boosting** and **Random Forest** were studied and compared

Parameters Considered for Classification

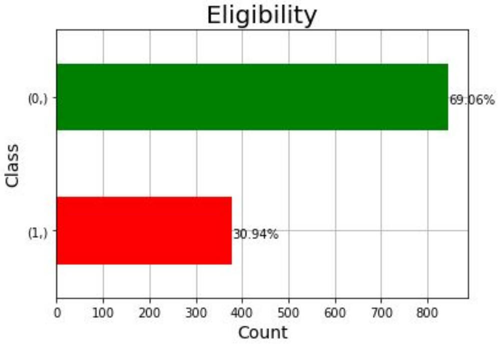
After clustering, the process of cluster head selection has to be carried out, i.e., selection of UE to act as VBS for every cluster. For selection of eligible UEs the communication parameters considered are:

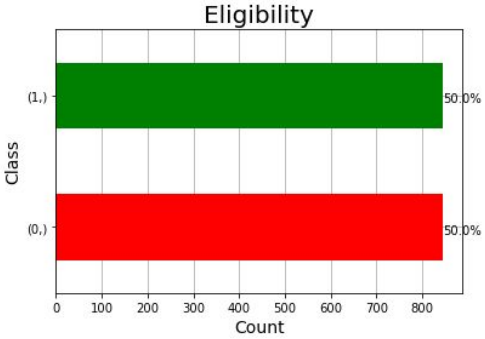
1. Battery discharge rate
2. Received power
3. Memory of UE
4. Processing Power of UE
5. Distance from Base Station
6. Date Rate
7. Quality of Service
8. Quality of Efficiency



Overcoming Class Imbalance - SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling technique which works by selecting examples that are close in the feature space, deriving a line between the examples in the feature space and drawing a new sample at a point along that line. SMOTE is employed to balance the training dataset as it synthesises data points that have smooth variation and high correlation with the existing dataset.





Results

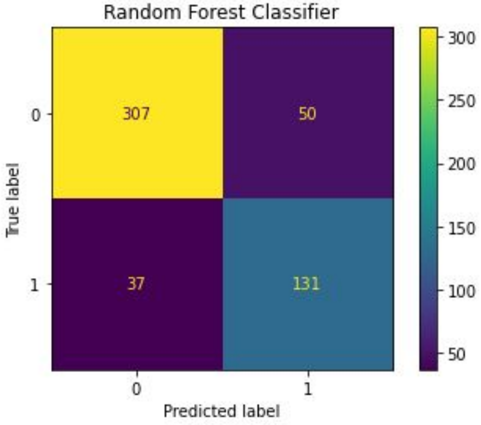
- Model Evaluation Metrics
    - Classification
    - Clustering
- Statistical Analysis

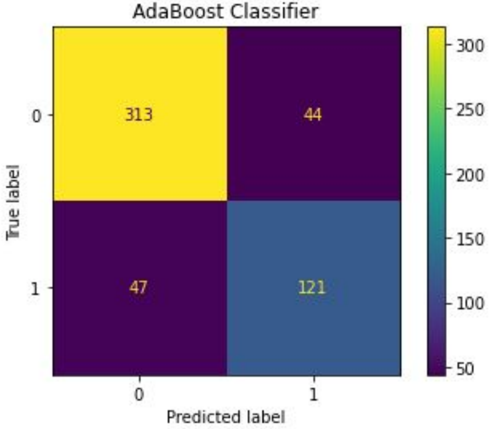
Model Evaluation Metrics for Classification

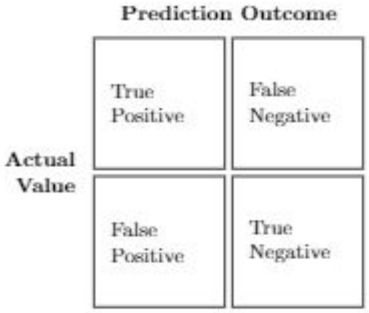
- Confusion Matrix
- Precision, Recall, F1-score and Accuracy - before & after SMOTE
- Area under Receiver Operator Characteristic

Confusion Matrix

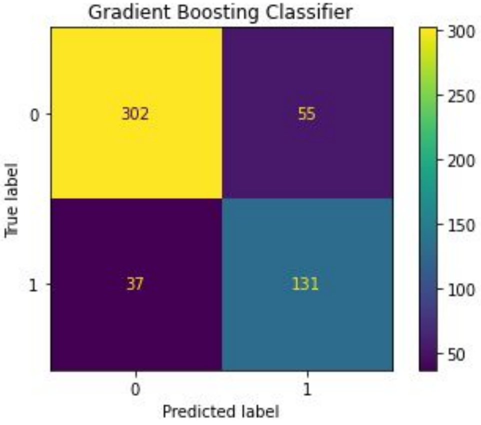
- **True Positive (TP):** Eligible devices that can become a UE-VBS correctly classified as ‘eligible’
- **False Positive (FP):** Ineligible devices that cannot become a UE-VBS incorrectly classifie as ‘eligible’
- **True Negative (TN):** Ineligible devices that cannot become a UE-VBS correctly classified ‘ineligible’
- **False Negative (FN):** Eligible devices that can become a UE-VBS incorrectly classified as ‘ineligible’


Random Forest Classifier


AdaBoost Classifier



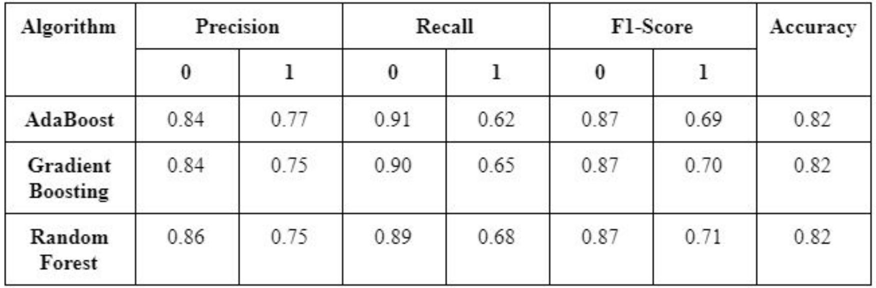
	Prediction Outcome	
Actual Value	True Positive	False Negative
	False Positive	True Negative


Gradient Boosting Classifier

Precision, Recall, F1-Score, Accuracy

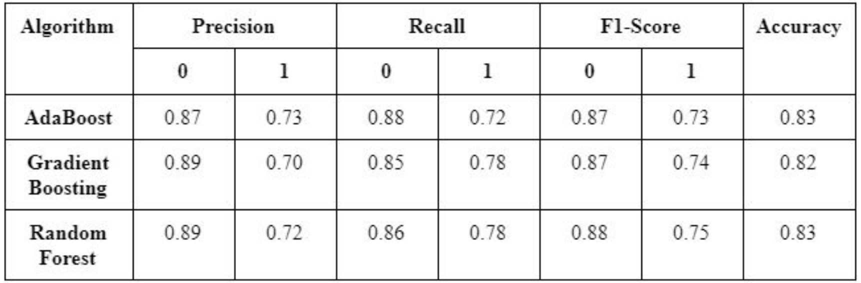
For imbalanced datasets, **accuracy** is **not a reliable metric** as it simply c**aptures the proportion of correctly classified instances**. In classification problems, the **errors made** and the **target class** are usually the area of interest. Therefore, other reliable measures such as precision and recall are used to evaluate the performance of the models. The models obtained higher recall and F1 score for the negative class (class 0) when compared to the positive class (class 1). The poor performance of the classifiers on class 1 relative to class 0 on the dataset is attributed to the **inherent skew** towards the class of devices 'ineligible' to become a UE-VBS.

Before SMOTE



Algorithm	Precision		Recall		F1-Score		Accuracy
	0	1	0	1	0	1	
AdaBoost	0.84	0.77	0.91	0.62	0.87	0.69	0.82
Gradient Boosting	0.84	0.75	0.90	0.65	0.87	0.70	0.82
Random Forest	0.86	0.75	0.89	0.68	0.87	0.71	0.82

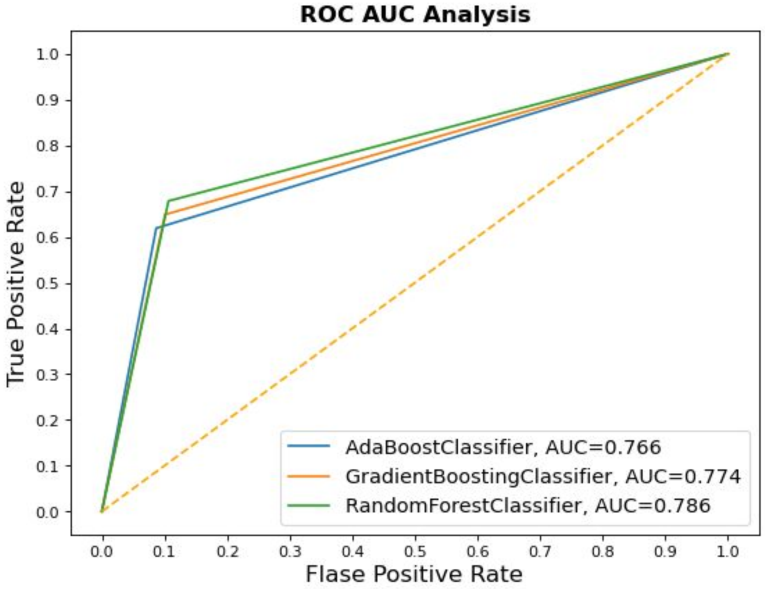
After SMOTE



Algorithm	Precision		Recall		F1-Score		Accuracy
	0	1	0	1	0	1	
AdaBoost	0.87	0.73	0.88	0.72	0.87	0.73	0.83
Gradient Boosting	0.89	0.70	0.85	0.78	0.87	0.74	0.82
Random Forest	0.89	0.72	0.86	0.78	0.88	0.75	0.83

Area under Receiver Operator Characteristic

AUC Value	Inference
AUC = 0	The classifier is predicting all negatives as positives, and all positives as negatives
AUC = 0.5	The classifier is unable to distinguish the positive and negative class points thereby predicting a random or constant class for all the data points
Between 0.5 and 1	There is a high chance that the classifier will be able to distinguish between the two classes
AUC = 1	The classifier is able to distinguish between all the positive and negative class points correctly



Model Evaluation Metrics for Clustering

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observations belonging to all the clusters:

- Mean distance between the observation and all other data points in the same cluster. This distance can also be called a **mean intra-cluster distance.** The mean distance is denoted by **a**
- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a **mean nearest-cluster distance.** The mean distance is denoted by **b**

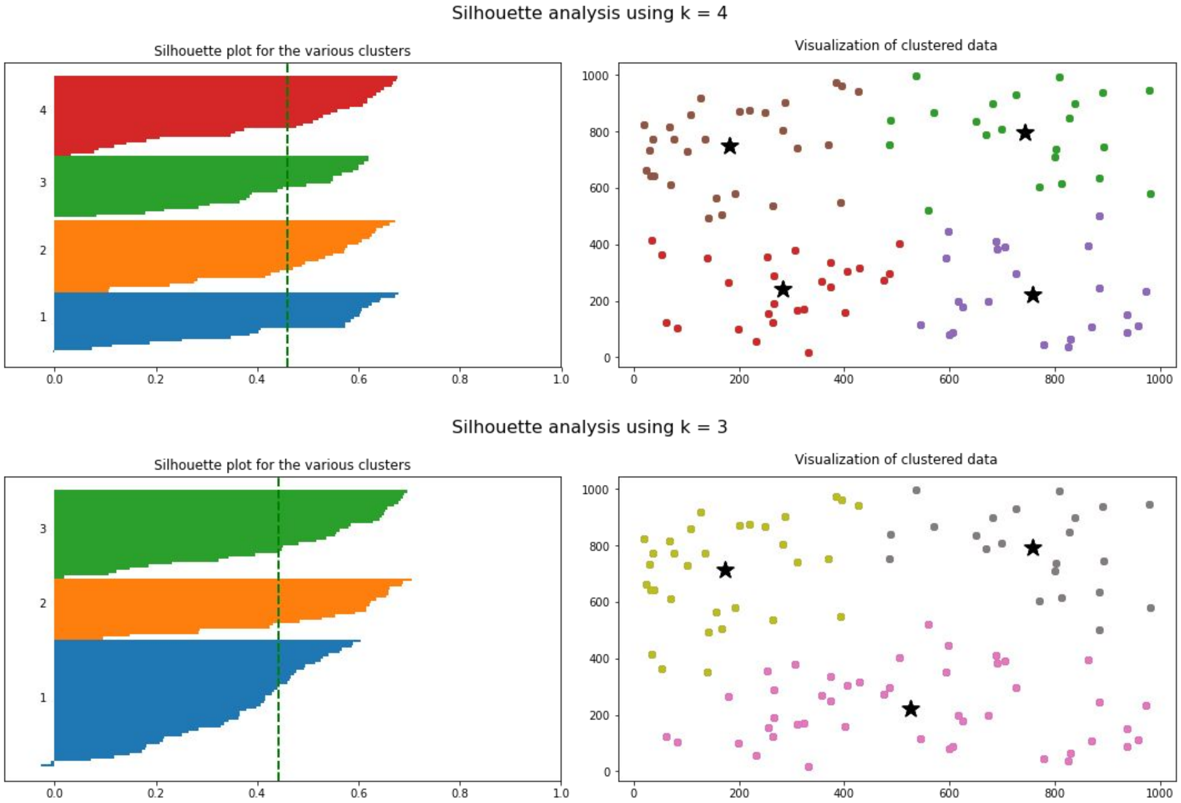
Silhouette score, **S,** for each sample is calculated using the following formula:

$$S = (b - a) / \max(a, b)$$

The value of the Silhouette score varies from -1 to 1. If the score is 1, the cluster is dense and well-separated than other clusters. A value near 0 represents overlapping clusters with samples very close to the decision boundary of the neighboring clusters. A negative score [-1, 0] indicates that the samples might have got assigned to the wrong clusters.

Model Evaluation Metrics for Clustering

Algorithm	Silhouette Score
K-Means	0.4601
Mean-Shift	0.3384



Silhouette analysis using k = 4

Silhouette plot for the various clusters

Visualization of clustered data

Silhouette analysis using k = 3

Silhouette plot for the various clusters

Visualization of clustered data

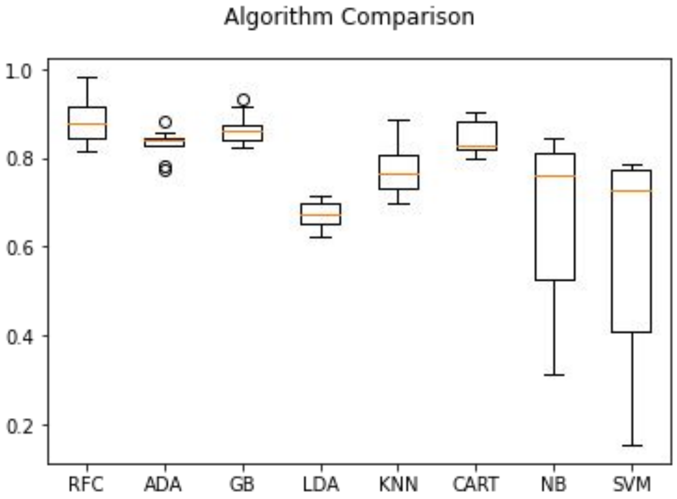
Statistical Analysis

<u>K-Fold testing</u>

Cross-validation is a statistical method used to estimate the skill of machine learning models and to compare and select a model for a given predictive modeling problem.

In *k*-fold cross-validation, the original sample is randomly partitioned into *k* equal sized subsamples. Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining *k* − 1 subsamples are used as training data. The cross-validation process is then repeated *k* times. The *k* results is then averaged to produce a single estimation.

Algorithm	Mean	Variance
RFC	0.89	0.053
ADA	0.83	0.031
GB	0.86	0.035
LDA	0.67	0.029
KNN	0.78	0.057
CART	0.84	0.037
NB	0.66	0.194
SVM	0.59	0.25



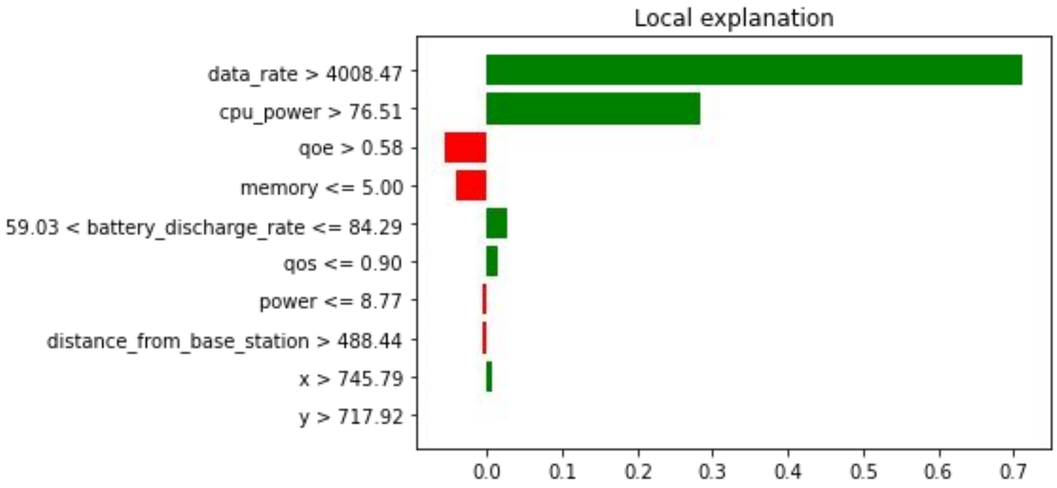
Statistical Analysis

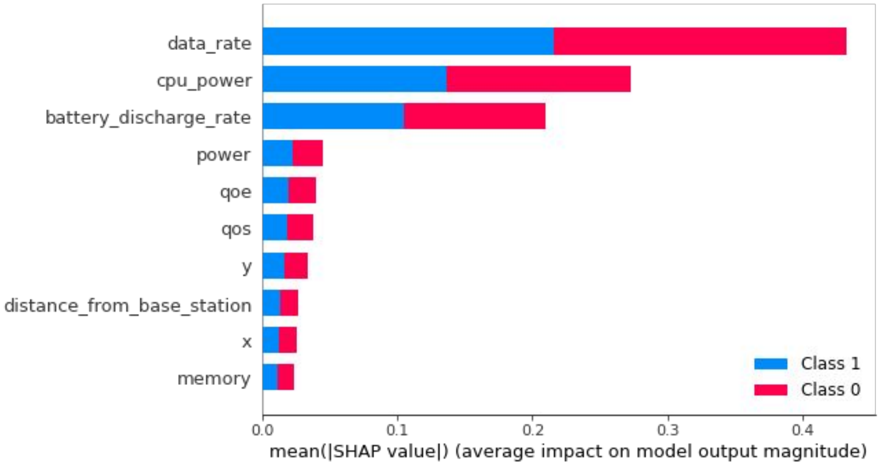
SHAP and LIME values

LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally

SHAP (SHapley Additive exPlanations) assigns each feature an importance value for a particular prediction.

Both are model agnostic. Due to its theoretical guarantees and simplicity, SHAP is widely used and more acceptable





Conclusions

The UE geographical coordinates were simulated as a **non-uniform distribution** in the time and space domain. A dataset was generated and preprocessed based on this simulation to train a two-stage machine learning model that can dynamically select a UE-VBS for clusters around the primary BS. Initially, the UEs distributed around the primary BS were clustered into groups using K-Means. **K-means** obtained a **silhouette score** of **0.46**. Subsequently, a classifier to determine the eligibility of the UEs to become a UE-VBS for each cluster was pipelined. Due to class imbalance, there was an inherent bias towards the majority class, 'Ineligible' to become a UE-VBS (class 0). The data was sampled using SMOTE before classification to overcome the bias. Out of several classification algorithms, the **Random Forest classifier** gave the best **F1 score (0.75)** and **Recall (0.78)** for the devices eligible to become a UE-VBS. In conclusion, the dynamic selection of a UE-VBS for each cluster around the primary BS was efficient and the Machine Learning model exhibited good performance.

Further Improvements

- In this work, the non-uniform random distribution of UEs geographically was modeled and simulated. Another asymmetric distribution can be simulated using different values and studied.
- The resulting distribution was subject to inspection using K-Means and Mean-Shift Clustering. Other clustering algorithms such as Spectral Clustering, Density-Based Spatial Clustering of Application With Noise (DBSCAN) and Hierarchical Clustering can be investigated.
- The results of clustering using K-Means was subject to inspection using the AdaBoost Classifier, the Gradient Boosting Classifier and the Random Forest Classifier. Other classification algorithms can be investigated.