I. INTRODUCTION

The section includes SEVERAL paragraphs summarizing your project. It is like the extended version of Abstract - you may use one paragraph for each of these parts - problem statement, dataset description, machine learning algorithms you will use to solve the problem, experimental results, and how your solutions are better as compared to existing solutions.

Please try to limit Introduction to one page.

Introduction: Waste management is a critical aspect of urban sustainability, and understanding the challenges within a city can pave the way for targeted interventions to enhance environmental practices. In the bustling metropolis of New York City (NYC), effective waste disposal and recycling are vital components of a greener, more sustainable future. As we navigate the complex landscape of waste management, this project endeavors to leverage the power of machine learning to uncover crucial insights. Our focus lies in identifying areas within NYC that face significant challenges in recycling and discerning the types of materials that are least likely to be properly disposed of.

Problem Statement- The overarching goal of this project is to employ machine learning techniques to pinpoint areas in NYC where recycling efforts encounter the most challenges. By delving into the intricacies of recycling rates and material-specific disposal patterns, we aim to provide city officials and policymakers with actionable insights. These insights can serve as a compass, guiding the allocation of resources and educational efforts towards areas that need them the most. Our mission is not merely to analyze data but to contribute to the enhancement of waste management practices, fostering a more sustainable and eco-conscious urban environment. Through this endeavor, we seek to bridge the gap between data-driven analysis and informed decision-making, creating a roadmap for a cleaner, greener NYC.

Dataset Description- The data analyzed comes from the New York City Department of Sanitation and gives monthly recycling and waste information for NYC's 59 sanitation districts that fall within 7 zones (Manhattan, Bronx, Brooklyn North, Brooklyn South, Queens West, Queens East, Staten Island) from 2016 to 2019. The data is provided in a tabular CSV format with 2832 values for each feature. Link to Source: https://www.kaggle.com/datasets/new-york-city/nyc-recycling-diversion-and-capture-rates/data

Machine Learning Algorithms- The three models we will use to solve the problem include a decision tree regression, a logistic regression model, and a random forest algorithm.

II. RELATED WORK

This section summarizes existing solutions to the problem

or similar problems. Please try to categorize these existing

techniques and provide some discussion on the pros and cons

of them. Don't forget to include references to any existing

work you mention.

The most popular approach, ANN, performs poorly due to poor generalization, local minima, and overfitting of the data. GBRT is made up of easily interpretable multiple regression trees with strong generalization, and it frequently produces superior results than ANN.

The largest challenge is the lack of historical data and other required data, particularly at the level of subdivisions like homes, buildings, or communities (Cubillos, 2020)[1], which can be linked to inefficient waste management even though ML models can accurately predict the generation of MSW.

Several researchers have used machine learning (ML) methods to achieve waste bin detection, which is essentially a classification problem, in an effort to increase the efficiency of waste collection.

For example, the decision tree model developed by the authors in [2] efficiently determines the quantity of garbage bins that are waiting for waste truck collection. The outcomes showed that the model's suggested collecting routes' length had decreased. The Recurrent neural network (RNN) classifier was created by the authors in [3] to predict the garbage bins' fill levels. However, we believe that the suggested approach is overly complex and costly to implement in actual environments (i.e., municipalities). The system does not optimize the resources of waste authorities; instead, it concentrates primarily on trash identification and sorting. Predictive models designed specifically to address time-series waste creation issues are also lacking.

The authors in [4] also used a Decision tree methodology to understand consumer electricity use based on environmental features such as temperature, pressure, and wind speed as well as date and time features to determine when consumers use the most electricity. The goal of this analysis was to provide insight to energy distribution companies on best practices for energy planning and when to perform maintenance for minimal consumer disturbance. The pros of this technique was the hyperparameter value of maximum number of splits for three trees was clearly stated and compared to the overall accuracy of the model. It was found that for 300, 200, and 100 splits resulted in accuracies of 76.8, 76.4,

and 74 percent respectively. This illustrates the balance between parameter optimization and overfitting tendencies of the model; in this example, maximum number of splits is likely optimized around 300 splits since there isn't much change between the accuracy from 200 to 300 splits. A limitation of these author's specific application is that they did not consider the influence of location on electricity consumption. They also do not discuss which features have a greater influence on electricity consumption (feature importance0 which could be very valuable to the electricity distribution companies.

References:

[1] Xia W, Jiang Y, Chen X, Zhao R. Application of machine learning algorithms in municipal solid waste management: A mini review. Waste Management & Research. 2022;40(6):609-624. doi:10.1177/0734242X211033716

[2] Londres G., Filipe N., Gama J. Optimizing Waste Collection: A Data Mining Approach. In: Cellier P., Driessens K., editors. ECML PKDD 2019: Machine Learning and Knowledge Discovery in Databases, Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Würzburg, Germany, 16–20 September 2019. Volume 1167. Springer Science and Business Media LLC; Berlin, Germany: 2019. pp. 570–578.

[3] Camero A., Toutouh J., Ferrer J., Alba E. Ibero-American Congress on Information Management and Big Data. Volume 978. Springer Science and Business Media LLC; Berlin, Germany: 2018. Waste generation prediction in smart cities through deep neuroevolution; pp. 192–204.

[4] O. Yaman, H. Yetis and M. Karakose, "Decision Tree Based Customer Analysis Method for Energy Planning in Smart Cities," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), Sakheer, Bahrain, 2020, pp. 1-4, doi: 10.1109/ICDABI51230.2020.9325644.

III. OUR SOLUTION

A. Description of Dataset

This subsection describes the dataset that you will use.

In addition to the source of the dataset, it is also expected

to include discussions on data proprocessing, e.g., statistical

properties or visualization of the raw data, observation of some

issues with the data (e.g., missing features, irrelevant features,

outliers, etc.), some feature engineering to fix the problems

(e.g., filling missing data, encoding of strings or characters

into numerical values, normalization, and other operations as

appropriate).

The data analyzed comes from the New York City Department of Sanitation and gives monthly recycling and waste information for NYC's 59 sanitation districts that fall within 7 zones (Manhattan, Bronx, Brooklyn North, Brooklyn South, Queens West, Queens East, Staten Island) from 2016 to 2019. The time (Month and Year) and location (Zone and District) CSV tabular data acts as independent feature inputs that will influence the resultant output. Output options include Diversion Rate-Total (Total Recycling / Total Waste), Capture Rate-Paper (Total Paper / Max Paper), Capture Rate- Metal, Glass, Plastic (Total MGP / Max MGP), Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100. "Diversion Rate- Total is tonnage diverted divided by the sum of tonnage diverted and disposed. Disposed materials are sent via transfer stations to landfills or waste-to-energy facilities outside of NYC. Diverted materials are sent to reuse or recycling facilities inside or outside of NYC. Capture rate is the amount of materials set out for residential recycling collection as a percentage of designated recyclable materials in both recycling and refuse streams. This ratio measures how much of the targeted materials are actually being recycled, which is a measure of how successfully such materials are recycled" (New York City Department of Sanitation).

Machine Learning analysis of this data would allow for city officials to understand if there are certain zones and districts within NYC that struggle more with proper recycling than others, or if there are certain seasons where recycling patterns are better than others (ex: does colder weather correlate to less proper recycling because residents do not want to make extra trips outside to recycle?). The yearly information can also show if there is a general improvement in recycling habits from the start of the dataset in 2016 to the end of the dataset in 2019.

Pre-processing of the data included looking for missing values, identifying and altering data types, determining if normalization was needed, and deciding on the most relevant features to analyze. There are no missing values for any of the 9 data features. The input features of Zone, District, and Month Name are Objects while Fiscal Month Number and Fiscal Year are Integers. The output features 1) Diversion Rate-Total , 2) Capture Rate-Paper, 3)Capture Rate- Metal-Glass-Plastic, and 4)Capture Rate-Total are all float64 values.

For the decision tree algorithm, limited pre-processing was needed as no normalization of the data is required for this method. The features selected for analysis were Zone, District, Fiscal Month, and Fiscal Year for the inputs and Diversion Rate-Total for the output. Month Name was removed because it was redundant to Fiscal Month which was already in the right data type format for the SciKit learn analysis limitations (no "objects" allowed). Zone and District were both converted from objects to integers. Diversion Rate-Total was chosen as the output because of its simplicity in calculation and consideration of all recycling compared to the Capture Rate data.

For the Random Forest algorithm, we first clean the data and then we have used the RandomForestRegressor(), as the data which has to be predicted is numerical and where first we needed to convert the Zones, District, Fiscal Month from objects to integers in order to use the fit() function which helps us to fit the data into the model. We have chosen Diversion-Rate-Total as the data that has to be predicted in the model. Random Forest is one of the most accurate models that is used to predict the outcome.

B. Machine Learning Algorithms

This subsection describes machine learning algorithms that you plan to use. For each ML algorithm, briefly 1) explain why it might be appropriate for the problem and 2) describe your main design. For example, if it is neural network, provide the network structure and your initial choice of some key parameters (e.g., activation function to use, number of layers, number of hidden nodes of each layer). You may change the parameters during the training process.

A regression-based decision tree is an appropriate algorithm for this problem because there is both categorical and numerical data types involved (all inputs are categorical with integer/float identifiers while the output is a numerical ratio of total recycling to total waste). The decision tree also makes it easy to communicate which input features are most relevant to the prediction of the output diversion rate with the more important features closer to the "root" of the tree. This is important because the target audience of city officials will want an easy to understand deliverable to communicate the needs of different sanitation districts over time. SciKit Learn's Decision Tree Regressor is used as the design framework for this algorithm.

A logistic regression model is selected to pinpoint the major variables that are correlated with recycling behavior. When predicting diversion rate as a binary variable, logistic regression is a suitable method.

The most significant variable, according to the logistic regression model, was the capture rate for metal, glass and plastic; lower MGP capture rates are correlated with lower diversion rates. On the test the model accuracy is 85%.

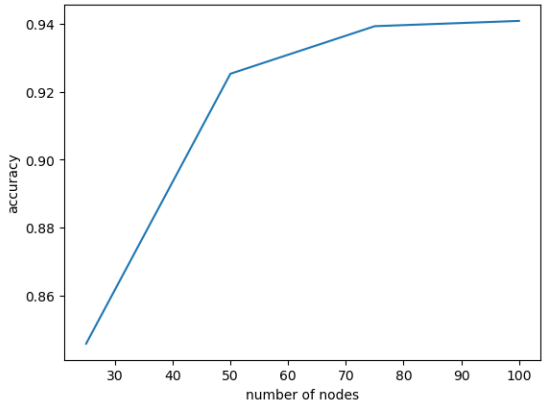
Random Forest algorithm, because of its high accuracy is used for performing the Regression analysis on the model. With the use of Random Forest Regressor, we can identify the actual factors which impact the recycling rates and based on that, the model is used to predict the output. We can also avoid overfitting data using this model.

C. Implementation Details

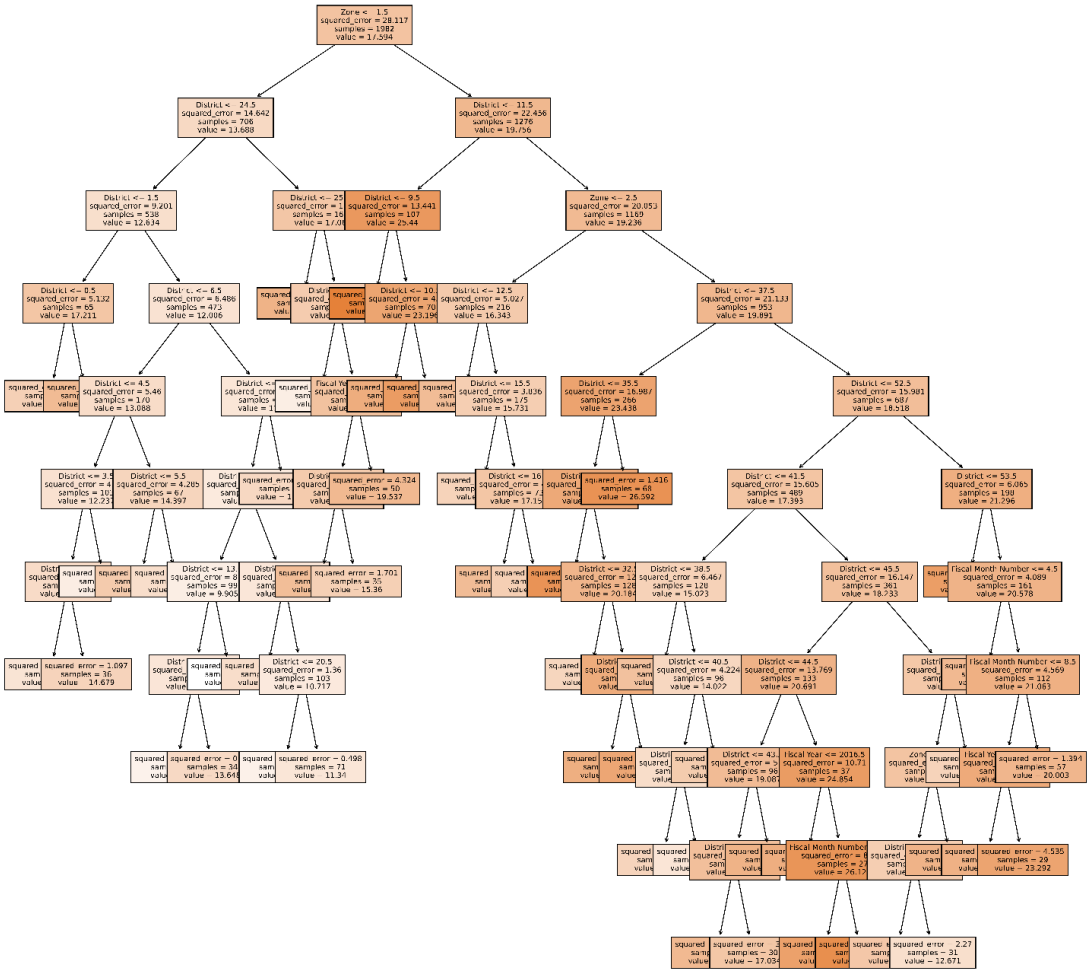
This subsection describes details of your implementation. Please focus on how you test and validate the performance, tune the hyperparameters, and select the best-performing models. Elaborate on techniques that you apply to improve the performance and explain why you use these techniques. You include few most important results/figures to illustrate your

idea but do not let figures/tables dominate the content of the report. You can include few lines of critical code if needed. But please avoid paste lengthy code in your report. Please make sure the figures/tables/code snapshots are of appropriate size including the font size.

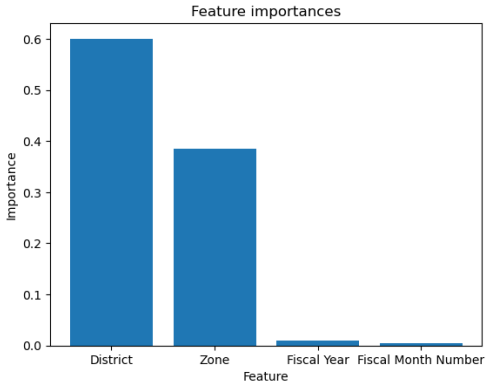
Decision Tree Regressor- The data was first split into 70% training and 30% testing with a random_state=42 using SciKit Learn's model selection train_test_split function. When fitting the tree, the default parameter of squared_error was used to split the data. A hyperparameter of max_depth=10 needed to be introduced, so the tree could load in a timely fashion. Using SciKit learn's cross_val_score function with the testing data, the accuracy of the tree was determined to be 94%. However, this tree was not readable to the viewer because it had 581 leaf nodes. Easy comprehension is supposed to be one of the key advantages of decision tree models, so the parameter of maximum number of leaf nodes was re-evaluated using GridSearchCV. The cross_val_score using the testing data was calculated for a maximum number of nodes of 25,50,75, and 100 to see how much this changed the accuracy of the model while greatly enhancing readability by cutting down the number of nodes by a factor of more than five. It was found, as shown by the figure below, that the accuracy of the model only decreased by only 2% (from 94% to 92%) with a max number of 50 nodes, so that was selected as a new hyperparameter. This can be explained by the tendency of decision trees to overfit to the training data, so less complex models can lead to similarly accurate results during testing.



The tree produced was now much more readable as shown below (some modifications still need to be determined to ensure the leaves aren't overlapping). Insights that can be gained from this tree include that the root node that provides the most information gain to the model is the Zone being Brooklyn North or not (Brooklyn North=zone 1). This means that Brooklyn North has the most unique recycling habits of all the sanitation district zones, and if you look at the first split, where the lefthand side of the tree is "TRUE" and the righthand side is "FALSE," the value is higher for zone=Brookyn North is FALSE (19 vs 13) meaning that Brooklyn North has the worst recycling to waste ratio of all the zones from 2016-2019.



Other analysis conducted thus far includes a comparison of the input influence on the decision tree using SciKit learn's feature_importances function where it was found, as illustrated in the figure below, that overall, District provides the most information gain to the tree followed by Zone. Whereas, Fiscal Year and Month have a very small (less than 1%) influence on the tree output. This means that in NYC, there was not a significant change in recycling to waste diversion on a monthly or annual basis from 2016 to 2019, and recycling behavior varied more based on location.



Logistic Regression- To forecast the binary goal variable of high vs low diversion rate, a logistic regression model is chosen. Cross validation is used to adjust important hyperparameters such as regularization and strength.

Using the pre-processed data set, the logistic regression model was trained. A threshold is used to transform the diversion rate into a binary variable. Accuracy, precision, recall, and F1 –score is used to assess the model on the test set. On the test set, it obtains an F1 score of 0.82 and an accuracy of 85%.

The MGP captures rate was once again the most significant factor. Based on its capture rates, the logistic regression model offers probability estimates for a district's high diversion rate. This understanding can direct actions to enhance New York's recycling behavior.

Random Forest- The main objective was to identify the features that had to be considered, to predict accurate results based on several factors. We use sklearn library from where we import the Random Forest model. We then preprocess the data based on the categories and split them into features to analyze the output and the target value. After this, we split the data into testing and training data. We split 30% data for training and 70% data for testing the model. After splitting the data, we convert all the objects to numerical value using the transform and OneHotEncoder function as we cannot fit the objects into the model. After splitting the data, we then import the Regression Model using the sklearn.ensamble library. The fit() function is used to fit all the data into the model. By which we train the model. We test the performance of the model by testing the accuracy of the model.