#### 

#### **NORTHEASTERN UNIVERSITY**

**ALY 6080- INTEGRATED EXPERIENTIAL LEARNING**

**MODULE 12**

#### **XN GROUP PROJECT: REPORT- FINAL**



**INSTRUCTOR- PROF. AJIT APPARI**

**GROUP 7**

**YU KUEI WANG**

**VARTIKA JOSHI**

**ABSTRACT**

Leveraging a data-centric methodology, the objective of this research is to delve into a more profound understanding of the multilayer plastic waste pollution problem in coastal regions of the United States, with a view to devising effective strategies for its reduction. The research uses data provided by the 5 Gyres Institute; an organization dedicated to combating pollution in the world's oceans. To gain an understanding of the scale and nature of the problem, an exploratory data analysis is conducted, highlighting the most significant contributors to plastic pollution and the roles they play.

In the Predictive Analytics section, we implemented a Random Forest model to predict layer types of waste materials. The model achieved an overall accuracy of 0.81. However, it struggled with 'multi-layer' items, having a recall of 0.51. After employing the Synthetic Minority Over-sampling Technique (SMOTE), the recall for 'multi-layer' items improved to 0.65, despite a slight decrease in overall accuracy to 0.80. Feature importance analysis indicated 'year' and 'product type' as the most influential features for predicting layer types.

Our findings underline the significance of the multilayer plastic waste issue and stress the urgent need for mitigation strategies. We advocate for concerted efforts in policy changes, technological advancements, business practice reforms, and consumer behavior modifications to tackle this environmental challenge effectively and sustainably.

**INTRODUCTION**

The substantial impact of multilayer plastic waste on marine pollution is largely due to its intricate composition that poses recycling difficulties, coupled with its resilience that allows it to persist in the environment for extended periods. Over time, these materials break down into damaging microplastics that can be consumed by marine creatures, potentially leading to a ripple effect of pollution up through the food chain. The widespread reliance on multilayer plastics for disposable items only compounds the amount of plastic waste entering our oceans. In light of these damaging effects, it's crucial that we explore inventive approaches in areas like product design and waste management, develop more efficient recycling technologies, and promote a shift towards more environmentally conscious consumer behaviors. In light of this, our focus is centered on the data related to coastal area plastic waste, which has been supplied by our sponsor, 5 Gyres.

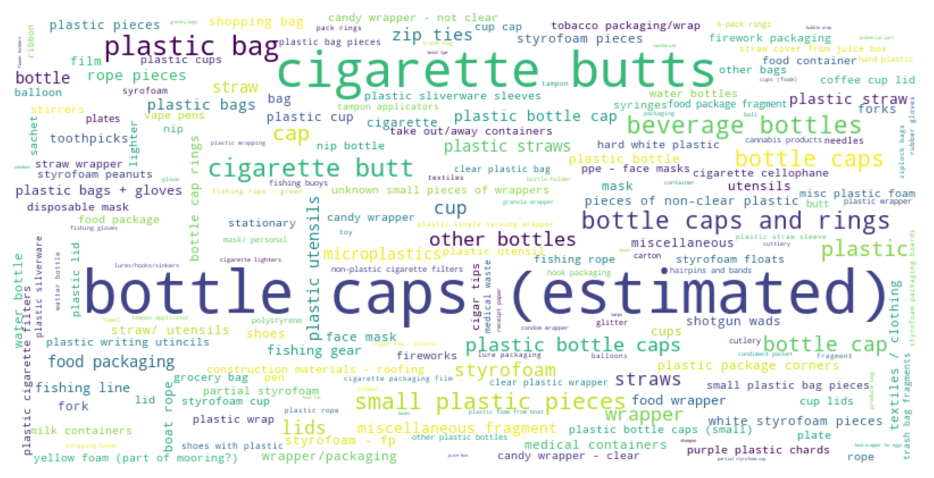
Our report is divided into three main sections: Exploratory Data Analysis, Predictive Analytics and Prescriptive Analytics. Within Predictive Analytics, we delve into the application of the Random Forest model, discuss its performance, and uncover insights from a detailed analysis of feature importance. This involves investigating intricate relationships between variables, examining trends over time, and exploring the significant impact parent companies have on the type of layer material.

Moving to the Prescriptive Analytics part, we employ the knowledge gained from our predictive analysis to propose actionable strategies and recommendations. This approach touches on a variety of areas from refining supply chain responsibility to guiding product design, and from shaping government regulations and policies to designing geographically targeted clean-up efforts. Additionally, it includes creating time-sensitive action plans and promoting education and collaboration among stakeholders.

**EXPLORATORY DATA ANALYSIS**

Our analysis of the data set revealed some interesting findings. By tallying up the unique event ids, we learned that 48 distinct organizations led a total of 93 coastal community cleanup events across the United States. But to really delve into the data and glean more substantial insights, we decided to segment the data according to certain variables, allowing for a more comprehensive exploration of the dataset.

The analysis reveals that a predominant share of 21,821 instances falls under the 'unbranded' category. This suggests a considerable segment of plastic pollution originates from products without identifiable brand or parent company affiliations. Upon sifting the original dataset to include only entries where the 'parent\_company\_name' is 'unbranded', we engaged in an illustrative word cloud analysis. This analysis considered both the 'item\_description' and 'total\_counts' columns to bring clarity to the composition of this 'unbranded' category.

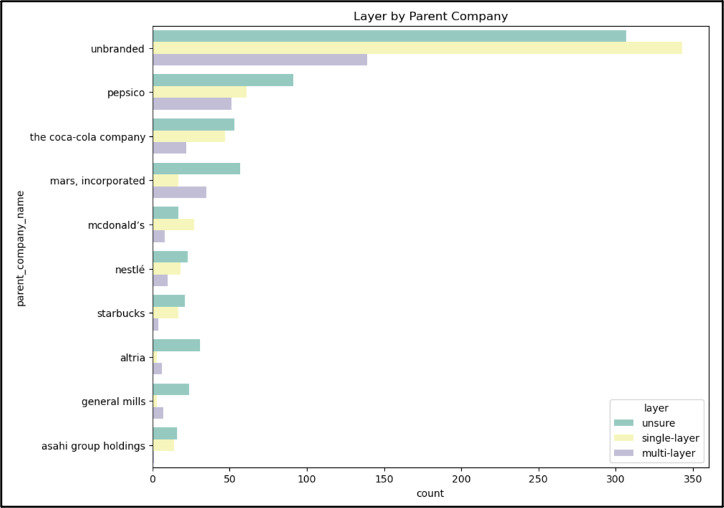


*Figure* ***1****: Word cloud of item descriptions for unbranded parent companies*

The outcome highlighted 'cigarette butts' and 'bottle caps' as the terms that recurred with the highest frequency in the item description where the parent company's name remained unidentified.

After 'unbranded', we see a significant drop in the count to the next highest contributors, which are 'Pepsico' and 'Nestlé', with 926 and 801 instances, respectively. However, it's important to note that the counts do not necessarily reflect the total amount of plastic pollution these companies are responsible for. The counts represent instances of pollution, not volume, and different products may contribute different amounts of plastic waste. For example, a single instance of a small plastic wrapper would be counted the same as a larger plastic bottle, even though the latter contributes more to total plastic volume.

**Contribution of top parent companies:**



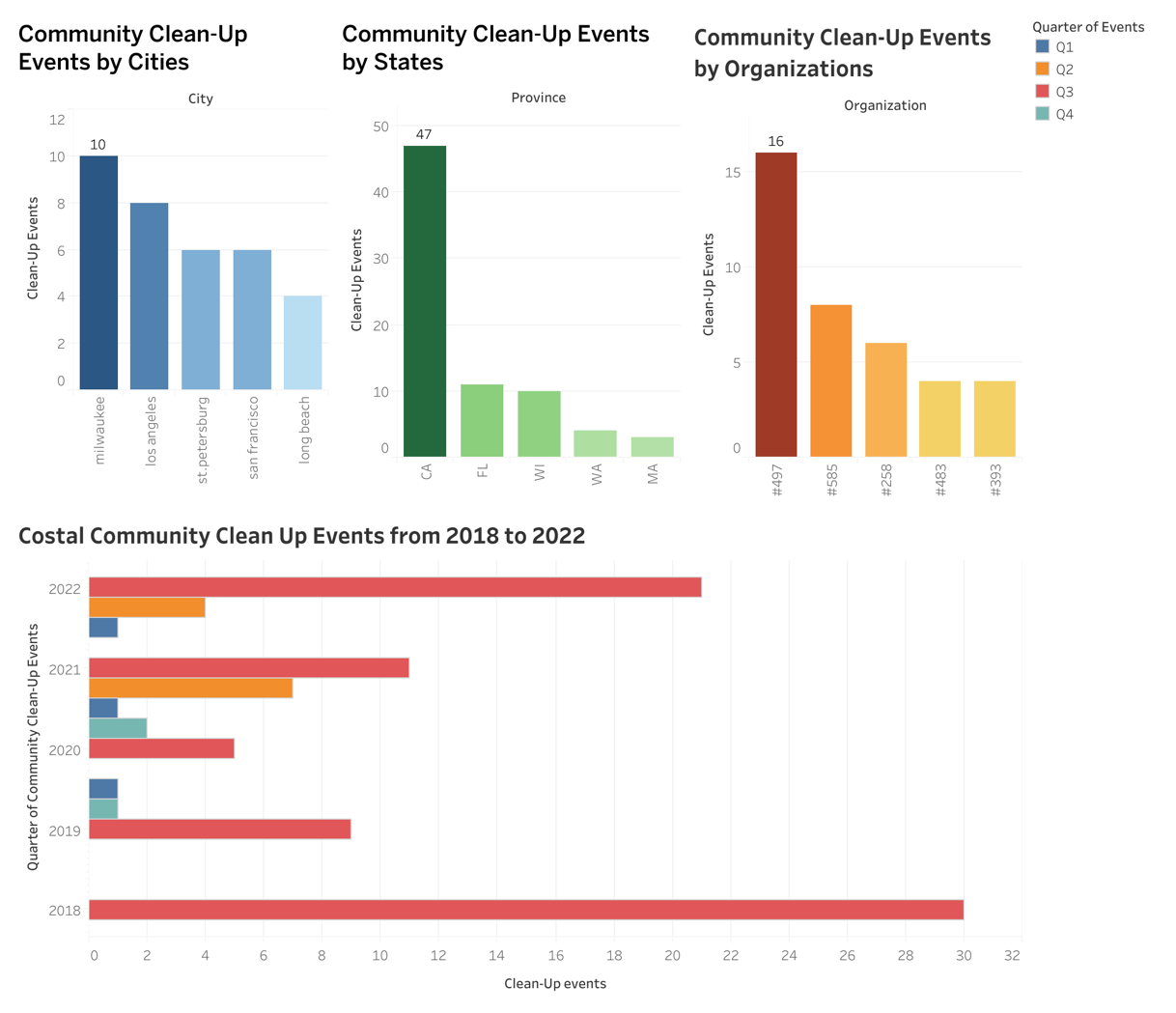
*Figure* ***2****: Bar chart showing layer distribution by parent company*

Certain companies, Pepsico for example, emerge as major contributors across all types of layer categories: multi-layer, single-layer, and unsure. Other notable contributors, such as Mars, Incorporated and The Coca-Cola Company, also show significant influence across all layer types, albeit with a higher propensity towards certain categories over others.

Meanwhile, some companies, including Asahi Group Holdings and Starbucks, exhibit a more concentrated influence on specific layer types. For instance, Asahi shows a strong correlation with single-layer types, while Starbucks leans more towards unsure types.

These findings illustrate the diversity in waste production practices across various companies. They highlight the opportunity for strategic interventions in sustainable waste management. For example, companies associated with a high proportion of 'unsure' types could be encouraged to improve their packaging and labeling information to facilitate more effective waste sorting and recycling.

**Community engagement in clean-up events:**



*Figure* ***3****: Community Clean-up Events Dashboard*

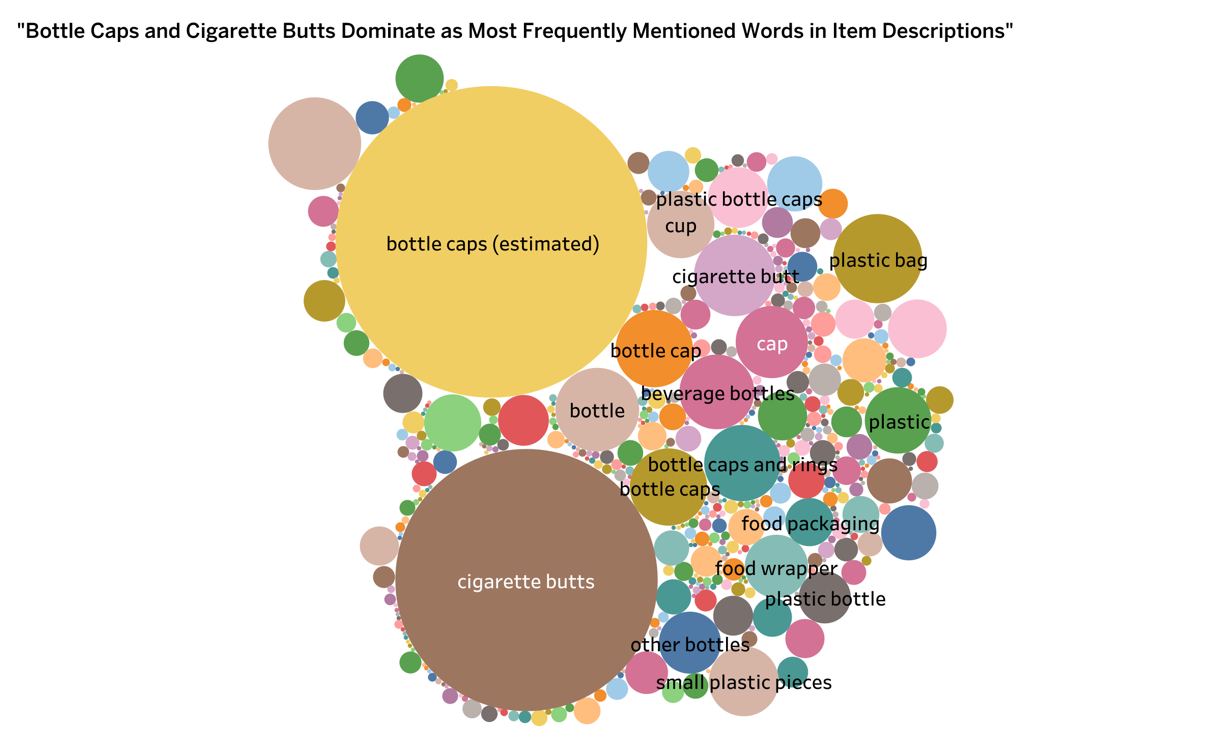
Milwaukee has the most community clean-up events, with 10. This could be due to a strong environmental awareness in the community, or a significant local pollution problem that has mobilized citizens. Los Angeles, a much larger city, comes in second with 8 events. This suggests a lower event-per-capita ratio than Milwaukee.

California (CA) leads by a significant margin with 47 cleanup events. This is not surprising, given that three of the top participating cities (Los Angeles, San Francisco, Long Beach) are in California. California's known focus on environmental issues, as well as its large population, may also be contributing factors.

The majority of these clean-up events occur in the third quarter, coinciding with the summer months when coastal visitation peaks.

It is important to note that raw counts of cleanup events may not provide a complete picture of community engagement. Factors like population size, geographic size, and the scale and impact of each cleanup event can all affect interpretation of this data.

**Frequency of Item Descriptions in the Dataset:**



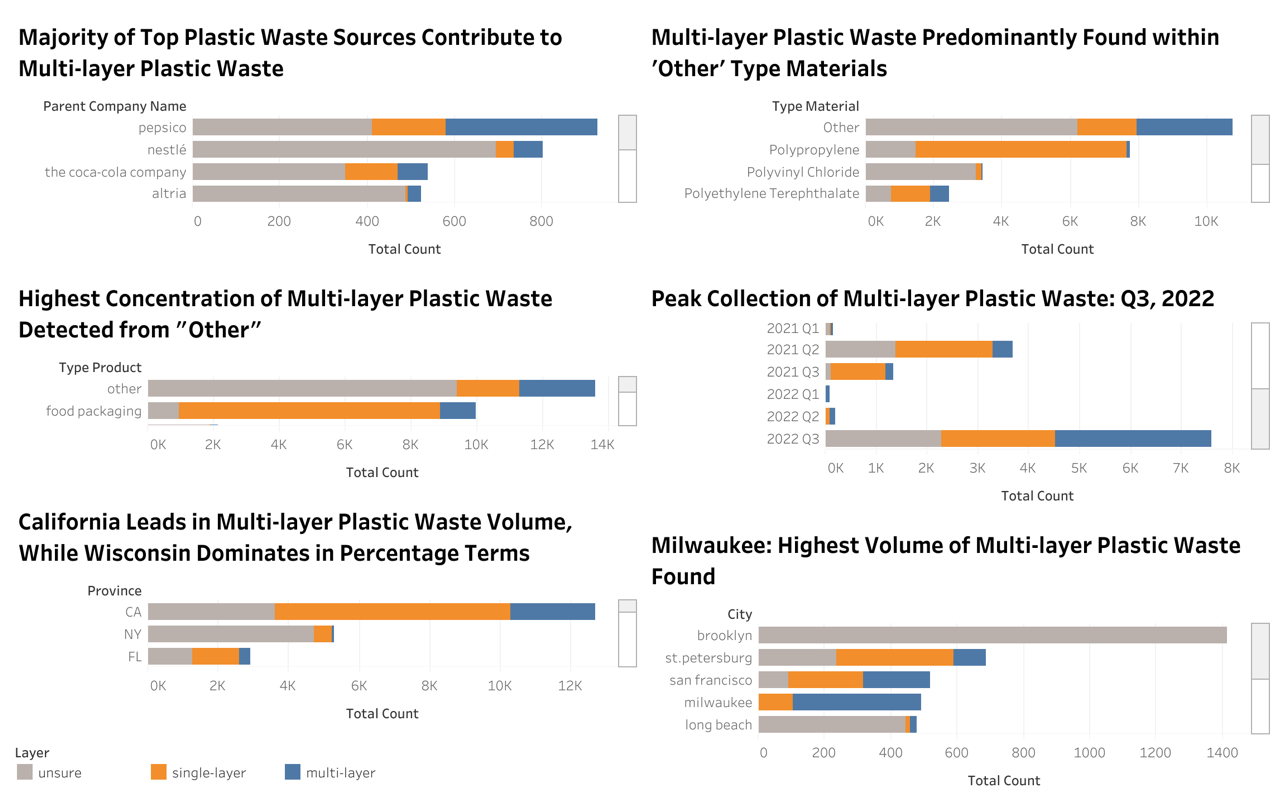
*Figure* ***5****: Bubble Chart Showing Most Frequent Word in Item Descriptions*

The analysis of Figure 5 brings to light an alarming issue of pollution in our oceans. The terms "bottle caps" and "cigarette butts" emerged as the most frequently mentioned items in the descriptions of trash found during coastal community clean-up initiatives. This suggests a prevailing problem with single-use plastic and non-biodegradable items polluting our water bodies.

Bottle caps, generally composed of plastic materials, are non-biodegradable and often discarded carelessly after use. Similarly, cigarette butts, which are often misconstrued as harmless and biodegradable, actually contain plastic filters and harmful toxins. When these items make their way to our oceans, they contribute significantly to the plastic pollution problem.

The frequent occurrence of these items in clean-up records indicates that despite recycling and waste management efforts, a considerable amount of these items still ends up in our oceans. This highlights the need for more stringent controls and better waste management practices to prevent these materials from entering our water bodies.

**Distribution of multi-layer plastic waste:**



*Figure* ***6****: Dashboard of Plastic Waste Distribution*

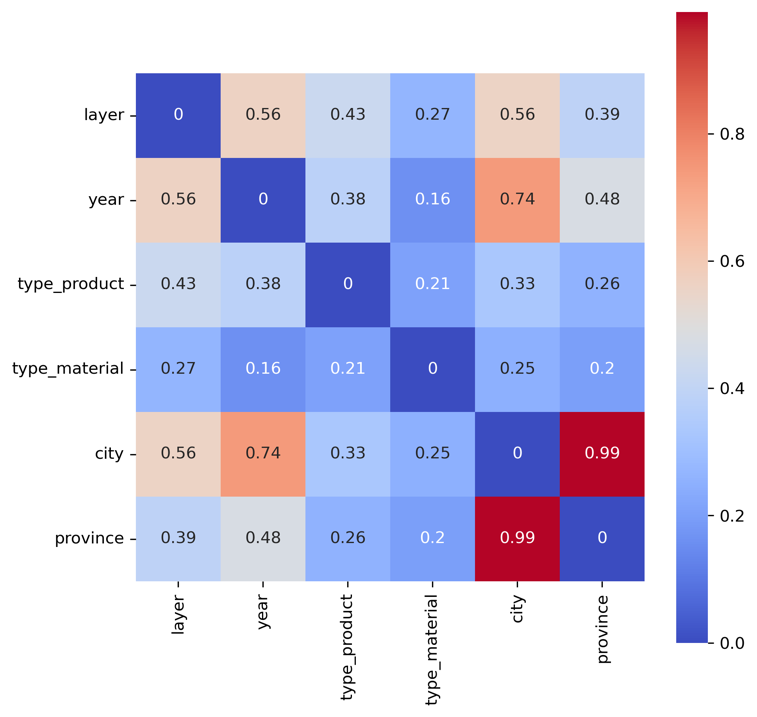
Figure 6 shifts focus to plastic waste sources, revealing that a significant portion is attributed to multi-layer plastics - often used in food and beverage packaging and notoriously difficult to recycle. The data show that this type of waste mainly falls within the "other" category of materials, indicating potential disposal issues.

Interestingly, the highest incidence of multi-layer plastic waste collection occurred in the third quarter of 2022, possibly reflecting the higher number of clean-up events during this period. At the city level, Milwaukee registered the highest volume of multi-layer plastic waste.

In the comparison of California and Wisconsin, California recorded a higher volume of multi-layer plastic waste, while Wisconsin had a greater percentage of this type of waste in its overall waste stream. This might imply California's more efficient collection of multi-layer plastic waste or a larger presence of this waste type in Wisconsin's total waste.

These dashboards offer essential insights into plastic pollution sources and their distribution across North America, underscoring the need for better waste management strategies and advancements in multi-layer plastic recycling technologies.

**PREDICTIVE ANALYSIS**



*Figure* ***7****: Correlation plot of layer with selected variables*

Figure 7 shows the correlation of various variables with the 'layer' variable in the dataset, as calculated by Cramer's V, which is a measure of association between two nominal variables.

Year - Layer Correlation (0.564264): This is the strongest correlation in the dataset, suggesting a significant association between the year and the layer. This could potentially mean that the layer of material used in products changes over the years, possibly due to changes in manufacturing technology, material availability, regulations, or consumer preferences.

Type\_product - Layer Correlation (0.429576): This indicates a moderate correlation, suggesting that the type of product is somewhat associated with the layer. This might mean that certain types of products tend to use certain types of layers more than others.

Type\_material - Layer Correlation (0.265813): This indicates a weak correlation, suggesting that the type of material is less strongly associated with the layer. This could mean that a variety of materials are used across different layers, and the choice of material does not necessarily determine the layer.

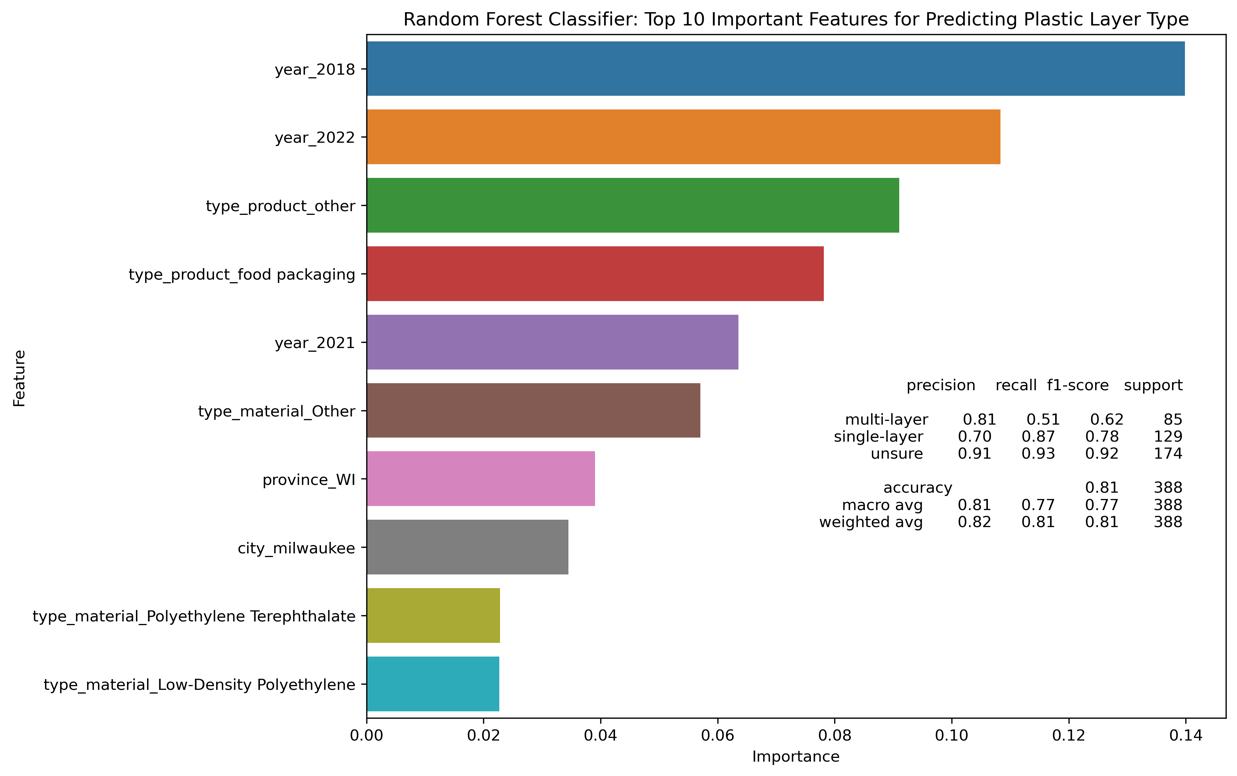
City - Layer Correlation (0.558221): This shows a strong correlation, close to the correlation between year and layer. It suggests that the layer might vary by city, possibly due to local manufacturing practices, regulations, or consumer preferences.

Province - Layer Correlation (0.390463): This shows a moderate correlation, indicating that there is some association between the province and the layer. Like with the city-layer correlation, this could be due to regional differences in manufacturing, regulations, or consumer preferences.

We utilized the Random Forest Classifier with the above selected variables, given its notable capabilities in managing high-dimensional data. Random Forest uses a collection of decision trees, each trained on different subsets of the data, and integrates their predictions. This method mitigates the risk of overfitting, a common issue with individual decision trees, thereby enhancing the model's broad applicability.

The dataset was partitioned into a training set, which included 80% of the data, and a test set, comprising the remaining 20%. This split is standard practice as it strikes a balance, ensuring that the model has ample data to learn from while retaining a significant subset for performance evaluation.

Upon training the Random Forest Classifier with the training data, we then utilized the model to predict outcomes on the test data. To assess the model's performance, we examined critical metrics such as precision, recall, and the f1-score for each category, providing a comprehensive overview of the model's performance. With an overall accuracy of 0.81, the model demonstrates a robust ability to predict the layer type based on the given features.



*Figure* ***8****: Bar chart of the top 10 influential features and corresponding model performance metrics in a random forest classifier for predicting layer type*

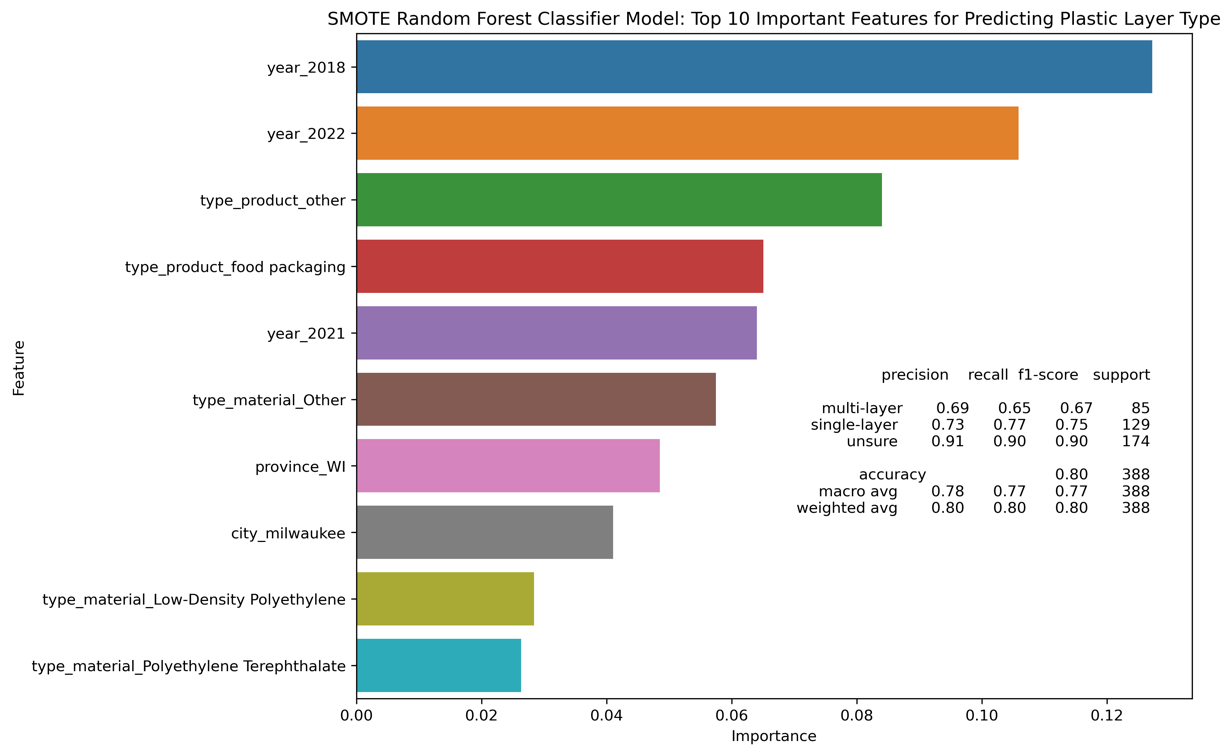
From Figure 8, we can see that the model does a relatively good job of classifying 'unsure' and 'single-layer' items (particularly in terms of recall for 'single-layer' and both precision and recall for 'unsure'), but it struggles more with 'multi-layer' items (with a recall of 0.51). This indicates that the model has more difficulties distinguishing 'multi-layer' items from the other classes.

The performance of the model for predicting the 'multi-layer' class is relatively low compared to the other classes. The model has a precision of 0.81 for 'multi-layer', which means that when it predicts an item to be 'multi-layer', it is correct 81% of the time. However, the recall is only 0.51, meaning that it only correctly identifies 51% of all true 'multi-layer' instances.

**Improve the Random Forest Classifier Model**

To enhance the prediction accuracy of multi-layer plastic waste by our model, we introduced the SMOTE technique after segmenting our data. This led to the creation of a balanced training data set, providing an equal quantity of instances for all classes. Subsequently, we trained our Random Forest Classifier on this equilibrated dataset.

Although this method could potentially enhance the model's ability to predict the 'multi-layer' class, which was initially underrepresented, it could come at the expense of reduced performance in predicting the other classes.



*Figure* ***9****: Bar chart of the top 10 influential features and corresponding model performance metrics in the SMOTE random forest classifier for predicting layer type*

When comparing this with the prior model, we observe an improved ability in predicting the 'multi-layer' class. The recall for 'multi-layer' saw an increase from 0.51 to 0.65, signaling that a larger portion of 'multi-layer' instances within the data is now correctly identified by the model. This significant enhancement implies that the model is now less likely to overlook 'multi-layer' instances.

Contrarily, the precision for 'multi-layer' witnessed a slight decrease, moving from 0.81 to 0.69. Precision represents the ratio of true positives (correct 'multi-layer' predictions) to all positive predictions. This drop in precision suggests that the model is committing more false positive errors, predicting non 'multi-layer' instances as 'multi-layer'.

The F1-Score for 'multi-layer', which provides a balance between precision and recall, also saw improvement, moving from 0.62 to 0.67. This indicates a better overall performance in predicting this class.

It's crucial to understand that while the model's overall accuracy experienced a slight decrease, moving from 0.81 to 0.80, this is not necessarily an unfavorable outcome. This could imply that the model's predictions have become more balanced across different classes, showing particular improvement in predicting the 'multi-layer' class.

**Feature Importance of the model:**

Upon employing OneHotEncoder on categorical features, each category within these features is converted into an individual binary feature. Each binary feature is regarded independently by the model, implying that the importance of each binary feature symbolizes the unique contribution of that particular category to the model's predictions. Herein is a bar chart delineating the top 10 feature importance metrics from the refined model.

Figures 8 and 9 reveal the relative significance of the features included in the model. It primarily identifies 'year' and 'product type' as the most significant predictors for determining the layer type, with specific years and product types demonstrating a remarkably high level of influence.

It's crucial to note a shift in the significance between number 9 and number 10 in the comparison of the original model to the enhanced model. Here, Low-Density Polyethylene holds a higher level of importance than Polyethylene Terephthalate within the type of material category.

**RECOMMENDATIONS:**

From the findings obtained from the Exploratory and Predictive Analytics phases, we can make the following recommendations:

**Stricter Regulations for Unbranded Plastics**: Given the high count of unbranded plastic waste, governments need to enforce stricter regulations that mandate manufacturers to clearly mark their products for easier identification and accountability. This can help in tracking down companies responsible for high amounts of plastic waste.

**Enhance Waste Management Practices**: Addressing the problem of 'bottle caps' and 'cigarette butts' in the environment could include implementing stronger recycling incentives or penalties, and creating public awareness campaigns about the proper disposal methods of these items.

**Seasonal Cleanup Events**: As the data showed, most cleanup events take place in the third quarter, which coincides with peak beach usage during the summer. Organizing more cleanup events in this period, as well as immediately following, can help to mitigate pollution more effectively.

**Target High-Pollution Cities**: California, particularly the cities of Los Angeles, San Francisco, and Long Beach, had the most cleanup events. Additional resources and efforts should be directed toward these locations for more efficient cleanup operations and to develop preventive measures.

**Improve Recycling Technologies for Multi-Layer Plastics:** Since multi-layer plastics contribute significantly to plastic waste and are notoriously difficult to recycle, efforts should be made in the research and development of efficient recycling technologies for these materials. Companies could also be incentivized to design packaging that uses fewer layers or more easily recyclable materials.

**Guided Product Design**: Companies such as Pepsico, Nestlé, and The Coca-Cola Company, which emerge as major contributors across all types of layer categories, should be advised to redesign their products to minimize waste production. For instance, they can adopt single-layer packaging over multi-layer packaging or use biodegradable materials in their product design.

**Educate and Collaborate with Stakeholders:** A collaborative effort from all stakeholders, including consumers, manufacturers, and government bodies, is necessary for combating marine plastic pollution. Campaigns to promote awareness about the impacts of plastic waste on marine ecosystems and human health can help drive more environmentally conscious consumer behaviors.

**Optimization Models for Cleanup Events:** Apply advanced analytics to design optimization models for cleanup events, which consider factors like timing (based on seasonal trends), location (based on pollution hotspots), and resource allocation (such as volunteers and equipment). By optimizing these variables, the effectiveness of cleanups could be maximized.

**Consider Geographic Differences in Policy and Practice**: The strong correlation between 'city' and 'layer' could suggest regional differences in waste practices and product availability. Understanding these differences can guide more effective policy measures and educational campaigns tailored to specific local contexts.

**Smarter Supply Chain Responsibility:** As certain companies emerge as major contributors to plastic waste, it's imperative that they take on more responsibility for the lifecycle of their products. This could involve adopting more sustainable practices in their supply chains, such as using recyclable materials and encouraging product take-back programs.

By employing these measures, we can make significant strides towards reducing marine plastic pollution, creating a safer environment for marine life, and ensuring a healthier planet for future generations.

**CONCLUSON**

Our analysis has uncovered critical insights into the problem of multilayer plastic waste. We've identified the magnitude of the issue, determined the contributing factors, and used predictive analytics to project future trends. These findings indicate a significant and escalating environmental challenge that requires urgent attention and action. By applying prescriptive analytics, we can begin to formulate strategic responses to this issue. These will involve a combination of policy changes, technological advancements, business practices reforms, and consumer behavior modification. By addressing this problem comprehensively and proactively, we can mitigate the environmental impact of multilayer plastics and move towards a more sustainable future.

**REFERENCES**

[1] Frankenfield, J. (2023, May 16). *Predictive modeling definition*. Investopedia. <https://www.investopedia.com/terms/p/predictive-modeling.asp>

[2] *What is a dashboard? A complete overview*. Tableau. (n.d.-b). <https://www.tableau.com/learn/articles/dashboards/what-is#:~:text=A%20dashboard%20is%20a%20way,easy%2Dto%2Ddigest%20form>.

[3] 5Gyres.org. (n.d.-a). <https://www.5gyres.org/>