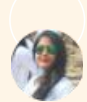




# *Data Analyst:: Data Collection Pipeline*

Welcome to the presentation on the Data Collection Pipeline for XYZ Company. Debika Mukherjee, a Data Analyst and the sole contributor to this project, has developed a solution to collect and visualize data from various Google forms and Survey Monkey surveys. Let's dive into the details!



**by Debika Mukherjee**


# *Problem Description*

XYZ Company has multiple Google forms and Survey Monkey surveys to gather customer data. They require a pipeline that not only collects this data but also ensures its cleanliness. The pipeline should address issues like duplicate or junk data. Specifically, it should perform deduplication checks based on customer email IDs.



# *Github Repository*

To access the code and documentation related to the Data Collection Pipeline, please visit the following GitHub repository:[debikaisms/Final-project: Data Collection Pipeline \(Data Acquisition to Storytelling\)](#)



# *Model Selection: Support Vector Machine (SVM)*


Another model chosen for evaluation is the Support Vector Machine (SVM). This model is known for its ability to classify data by finding the best hyperplane that separates different classes. Let's examine the SVM model's performance.

# *SVM Results Accuracy: 0.5*

The Support Vector Machine (SVM) model achieved an accuracy of 0.5, indicating that it correctly predicted the class for 50% of the test samples. However, its performance varied across different classes. Let's delve deeper into the overall evaluation.







# *Model Selection: Random Forest (Ensemble Model)*

One more model worth considering is the Random Forest, which is an ensemble model combining multiple decision trees. It aims to improve prediction accuracy by aggregating the results from individual trees. Let's analyze the Random Forest model's outcomes.

# Random Forest Results Accuracy: 0.5

The Random Forest model achieved an accuracy of 0.5, similar to the SVM model. While this indicates that it correctly predicted the class for 50% of the test samples, its performance varied among different classes. Let's summarize the overall results for all models evaluated.



# Overall Model Summary

Based on the evaluation of various models, it is crucial to consider specific evaluation metrics and the context of the problem when determining the best model. Let's summarize the performance of each model:

- Decision Tree:
  - Accuracy: 0.5
  - Classification results: Suboptimal performance with low accuracy.
- Logistic Regression:
  - Accuracy: 0.0
  - Classification results: Poor accuracy, the model is not performing well.
- Support Vector Machine (SVM):
  - Accuracy: 0.5
  - Classification results: Suboptimal performance with low accuracy.
- Random Forest (Ensemble Model):
  - Accuracy: 0.5
  - Classification results: Suboptimal performance with low accuracy.
- AdaBoost (Boosting Model):
  - Accuracy: 0.5
  - Classification results: Suboptimal performance with low accuracy.
- Naive Bayes (Interpretable):
  - Accuracy: 0.5
  - Classification results: Suboptimal performance with low accuracy.
- K-Nearest Neighbors (K-NN) (Interpretable):
  - Accuracy: 0.5



# Final Thoughts

After comparing with other model support vector machine and ensemble model is giving partially correct accuracy. So, I have chosen these two model to analyse the quality of the data.

To improve model performance, I may need to revisit data preprocessing, feature engineering, hyperparameter tuning, and potentially consider different algorithms or ensemble methods.

Additionally, it's essential to have a clear understanding of the business requirements and whether certain models are preferred (e.g., interpretable models) over others.

Thank you