# Report

## Approach:

The task was to enhance the existing lead generation tool (saasquatchleads.com) within a 5-hour coding limit. I focused on a quality-first approach by building a lead scoring tool that analyses uploaded data and prioritizes leads based on their potential. The model auto-detects the target column, preprocesses the data, and classifies leads as Low, Medium, or High. This helps businesses act on the most valuable leads without additional scraping or integrations. It's efficient, impactful, and built using accessible tools like Streamlit and scikit-learn.

## Model Selection:

I chose the Random Forest Classifier for its reliability, ability to handle both numerical and categorical data, and strong performance on classification tasks. It provides probabilistic outputs, which are ideal for ranking leads by confidence levels. Random Forest is also resistant to overfitting and works well with limited feature engineering, making it a practical choice within the given 5-hour development window.

## Data Preprocessing:

Uploaded data are first copied to preserve the original. The target column—detected automatically by low-cardinality is cast to string so all labels are treated as categories. Every categorical field (including the target) is then one-hot-encoded with "pd.get_dummies", converting mixed data types into a fully numeric matrix suitable for scikit-learn. Missing values are left to the tree-based model to handle implicitly, avoiding extra imputation steps within the tight time budget. Finally, the dataset is split 70 %/30 % into training and test sets with "train_test_split", ensuring unbiased performance evaluation.

## Performance Evaluation:

The performance of the lead scoring model was evaluated using the Random Forest Classifier, a robust ensemble learning method known for its accuracy and ability to handle diverse feature types. The dataset was split into training and testing subsets with a 70/30 ratio to ensure unbiased evaluation on unseen data. The primary metric used to assess classification accuracy was Accuracy Score, which measures the proportion of correctly predicted lead categories. Additionally, where applicable, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was calculated to evaluate the model's ability to distinguish between classes in a multiclass setting using the One-vs-Rest (OvR) strategy. The Random Forest model demonstrated strong predictive capability, balancing bias and variance, and providing reliable probabilistic outputs essential for effective lead scoring. Feature importance analysis further confirmed that the model leveraged meaningful predictors, supporting interpretability and actionable insights. Overall, this evaluation framework ensured that the model's performance is both statistically sound and practically useful for prioritizing leads.