Client Report: Cost-Optimised Feature Selection for Conductivity Classification

Introduction

The goal of this analysis is to determine the minimum set of material measurements required to reliably classify samples as conductive or non-conductive, thereby reducing experimental cost without sacrificing predictive performance. Several machine learning models were trained and evaluated on Dataset 1 to assess feature relevance towards determining conductive properties.

Model Selection Rationale

We evaluated four classifiers: Logistic Regression, Decision Tree, Random Forest, and XGBoost. These choices reflect different modelling assumptions. Logistic Regression was chosen first due the simplicity of the model and because conductivity is a binary property, easily encoded as 0/1. Logistic regression provides smooth, interpretable decision boundaries and allows us to inspect feature weights directly. Decision Trees were included to test for non-linear relationships between features. To extend from this, Random Forests mitigate the overfitting tendencies of a single tree and produce robust feature-importance estimates. Last of all, XGBoost is a standard, high-performance model for tabular data. It was seen as a thorough model to detect potentially more subtle patterns in the data. Additionally, each model was trained both with and without standardisation to test whether performance depended on feature scaling.

Classification Performance

All models achieved 100% accuracy, regardless of the classifier used or whether the data was standardised. Furthermore, both Decision Tree and Random Forest performed perfectly, implying the class boundary is so sharp that ensemble averaging provides no additional benefit.

This was confirmed through: Train/test evaluation, 5-fold cross-validation and for Logistic Regression, regularisation sweeps (C values) were implemented across four orders of magnitude. This consistency indicates that the conductive vs non-conductive classes are perfectly separable, and that the predictive signal is extremely strong.

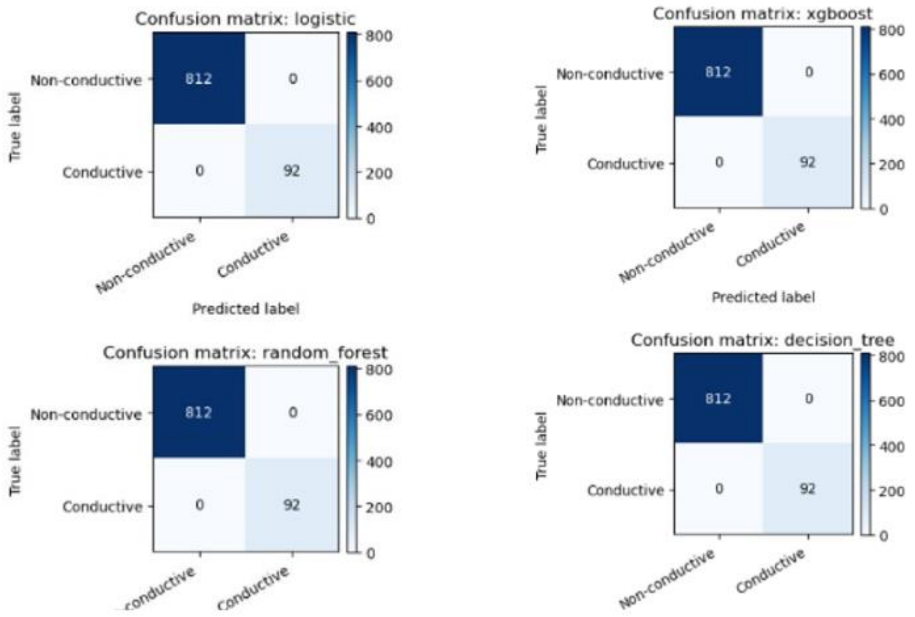

Figure 1. Confusion matrices of all four classifiers used on data (xGboost, Logistic regression, Random Forest, Decision Tree). All four models returned an accuracy of 100% regardless of cross validation.

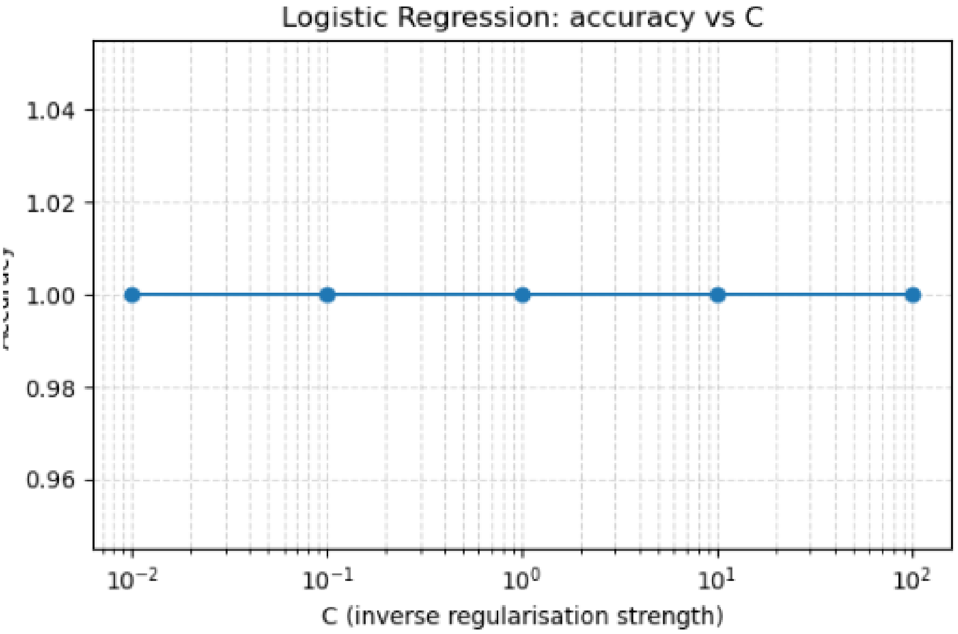

Figure 2: Logistic Regression Accuracy vs inverse regularisation strength (C) across four orders of magnitude. Model accuracy is independent of C, maintaining 100% accuracy.

Feature Importance & Physical Interpretation

Feature-importance plots across all tree-based models and coefficient magnitudes from logistic regression showed the band gap overwhelmingly dominates prediction. Regardless of model, all other features (density, hardness, crystallinity index, etc.) contribute negligible predictive value.

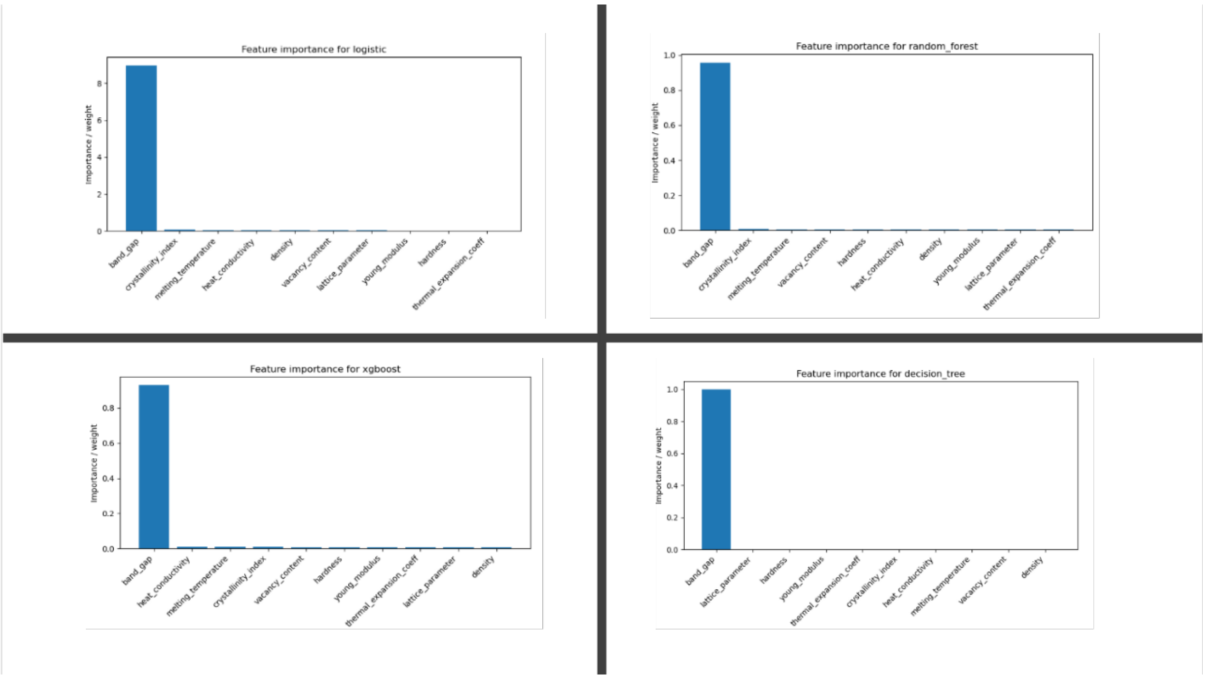

Figure 3: Feature Importance Plots for all four models (Logistical Regression, Random Forest, XGboost and Decision Tree). Amongst all models, the band gap dominates the feature importance for selection, being the practical sole determiner of classification.

Cost Reduction Recommendation

Since band gap alone provides complete predictive power, we recommend measuring only the electronic band gap. All other measurements can be removed without reducing any accuracy, enabling substantial savings in equipment use and testing time.

This recommendation provides the maximum cost reduction with zero loss in classifier performance or scientific reliability.