Performance Report

Introduction

This report addresses the multilevel prediction of corruption in countries (SDG 16: Peace, Justice, and Strong Institutions), using classification models and socioeconomic, institutional, and development level indicators.

As mentioned in the problem definition, corruption distorts institutional functioning and impedes equitable and transparent development, predicting its level based on "hard" variables such as press freedom, resource dependence, education, and the internet is valuable for governments, NGOs, and investors.

Consequently, a tiered classification approach ("low," "medium," "high," "very high") was chosen, as justified by the theoretical framework, to more realistically reflect the underlying information and reduce the risk of arbitrary interpretation.

This report details the performance of the applied models, the effect of feature engineering, and hyperparameter tuning, using relevant visualizations that help to interpret and highlight the most important and advanced findings achieved during the process.

Data Used

The database integrates indicators from 2 recognized sources (World Bank and OurWorldInData) including:

- GDP per capita,
- Average years of schooling,
- Inflation,
- Internet use,
- Press freedom,
- Dependence on natural resources,
- Unemployment rates.

This selection ensures relevance and multidimensional coverage of the phenomenon, allowing the corruption problem to be modeled from several complementary perspectives.

The data combines sufficient classes and periods (2012-2021) to avoid distribution bias and ensure the generalization of the results.

Methodology and Models

Various classification algorithms were tested, including Random Forest, Logistic Regression, SVC, Gradient Boosting, KNeighbors, and XGBoost.

Initially, the first model was trained with only one engineered feature (Figure1), and subsequently, transformations and new attributes were implemented (Figure 2, 3), followed by hyperparameter tuning and cross-validation (Figure 4, 5). As a result of these changes, a direct impact on the models' performance was observed. Temporal partitioning of the data was split based on years to avoid contamination between training and testing, ensuring a robust evaluation.

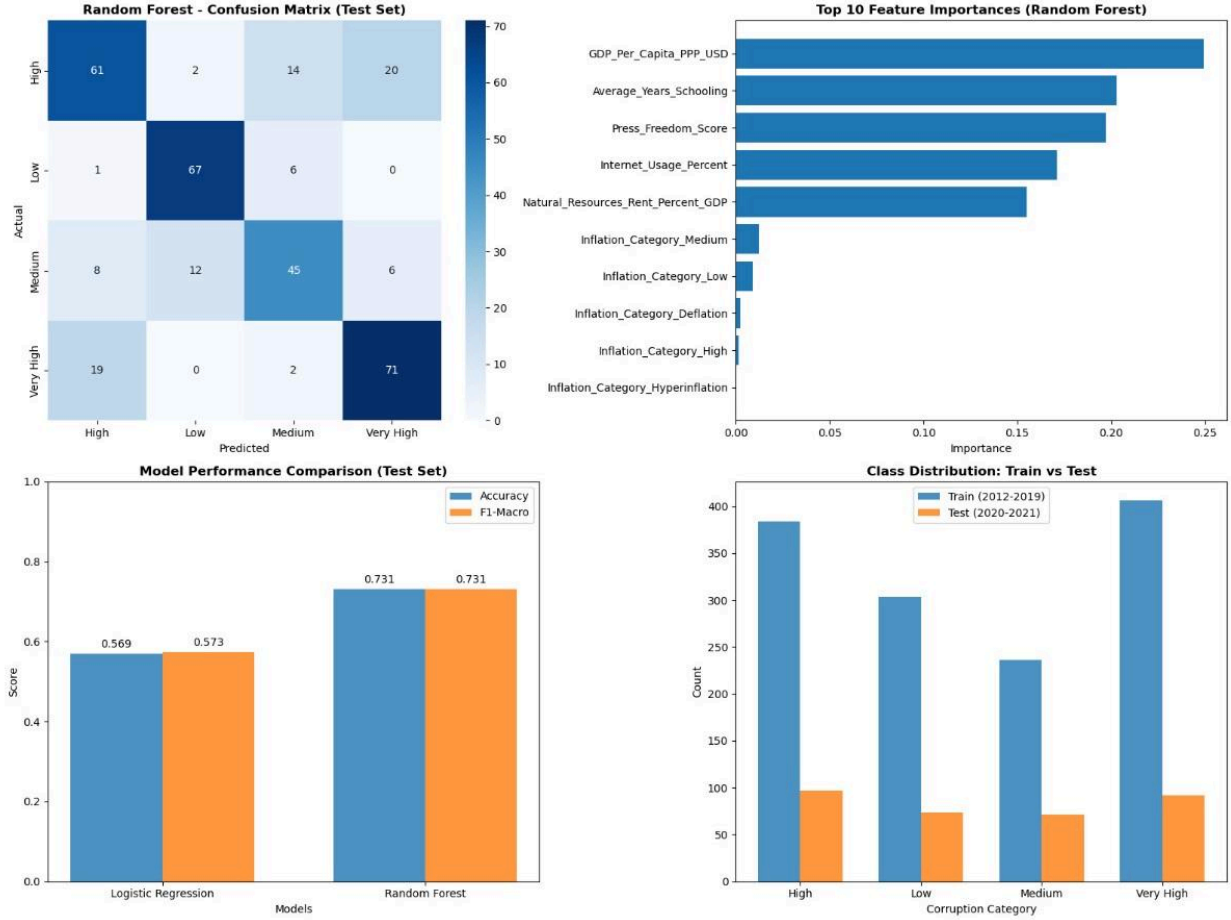

Figure 1

Initial Performance

The feature importance bar chart (Figure 1) shows that **the indicators of the country's overall development are the main predictor of the level of corruption** which aligns with the initial hypothesis, while inflation as a factor in economic stability is less influential. GDP per capita is the

most important predictor: **the development of an economy has a bidirectional relationship with corruption - more wealth can reduce corruption (if accompanied by the development of the institutions) and more corruption slows down growth because it erodes the institutions.**

The confusion matrix and the precision, recall, and F1 values (*Figure 1*) show that the base model only achieves an accuracy of around 57%.
An analysis of the performance table shows clear ranges between the basic models; for example, Random Forest exceeds 73% accuracy, while Logistic Regression lags behind. This indicates nonlinear connection of features with the target - corruption.

Figure 1: The confusion matrix reveals that, although the Random Forest model correctly predicts most “High”, “Low” and “Very High” cases, there are systematic errors in intermediate (“Medium”) classes.

Impact of Feature Engineering

In the second models (*Figure 2*, *Figure 3*), new variables and transformations are incorporated (logarithm of GDP per capita, composite index of economic stability: inflation + unemployment rate, composite index of resource curse: press freedom x resource rent as % of GDP).

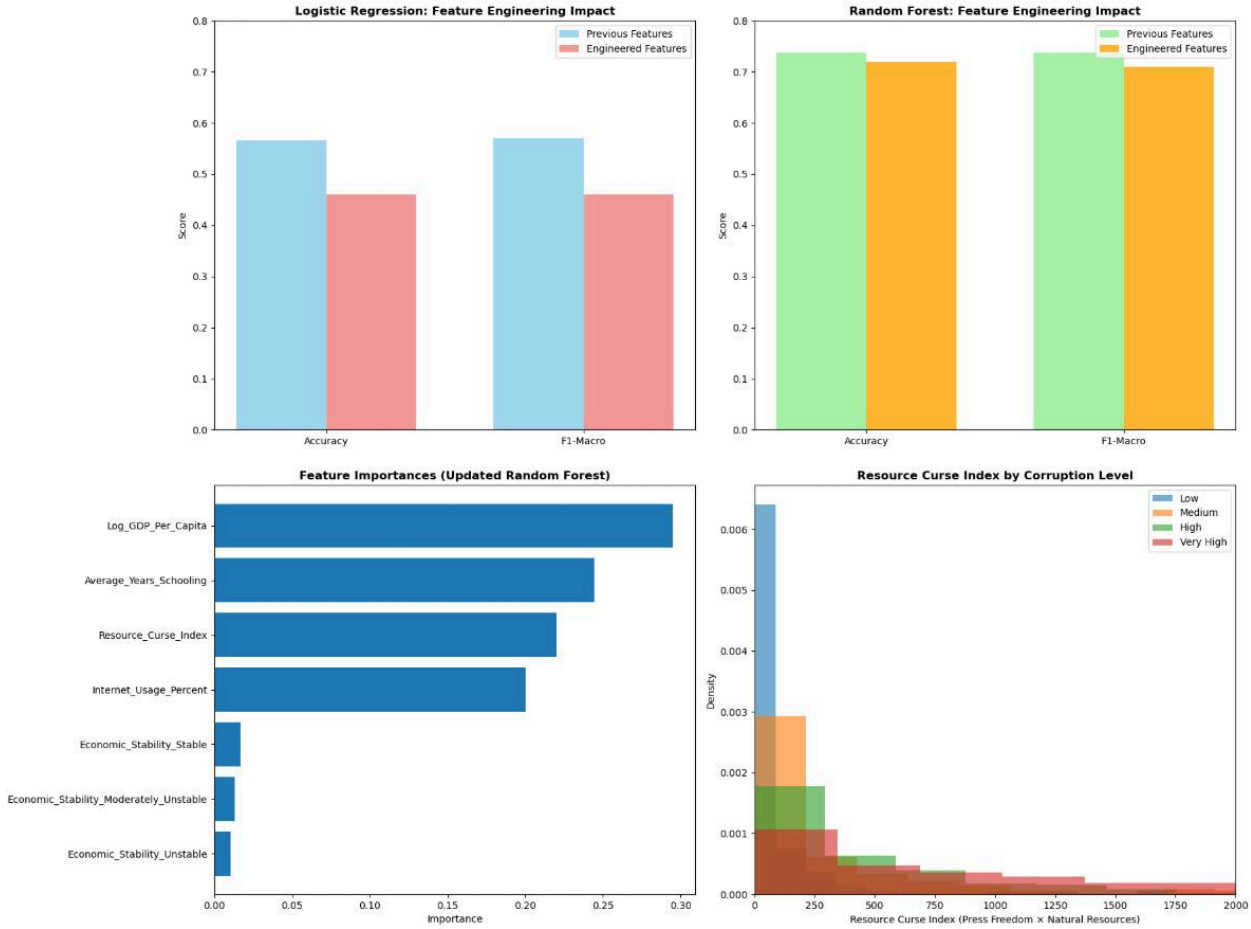

Figure 2

Variables such as Log GDP Per Capita and Average Years Schooling now appear as the most relevant (Figure 2). The variable importance plots show that the refined model (updated Random Forest) prioritizes economic development and education attributes to distinguish different levels of corruption (Figure 3).

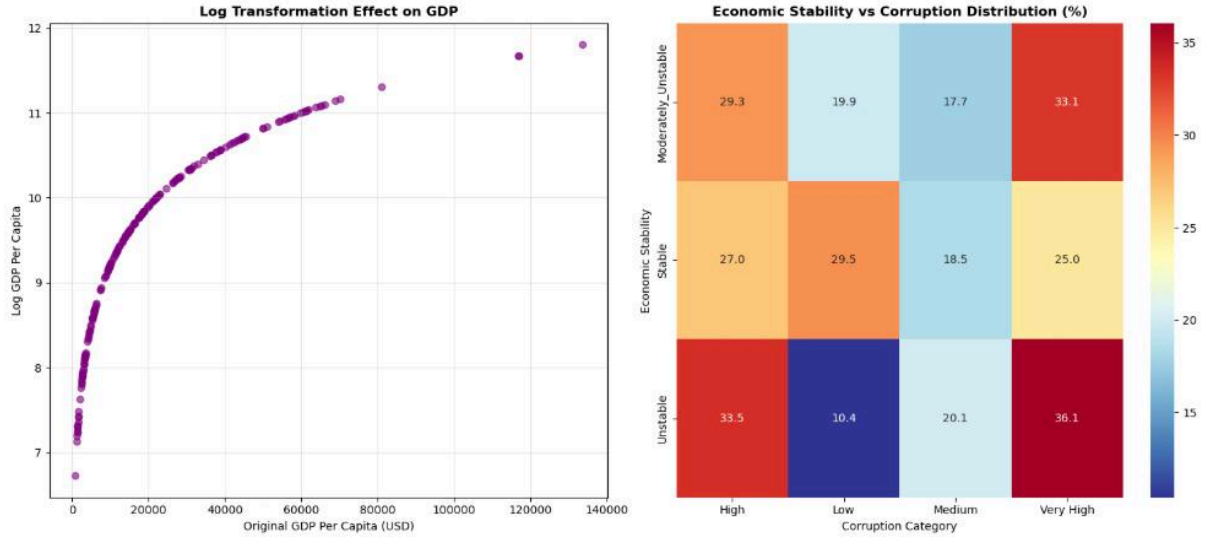

Figure 3

However, the side-by-side comparison of accuracy and F1 score (clustered bars, Figure 2) indicates that feature engineering does not produce improvements in all models; in the case of Logistic Regression. Apparently, **corruption has deeper roots and is less susceptible to temporary fluctuations in inflation and unemployment that were used as an indicator of economic stability.**

Hyperparameter Tuning and Cross-Validation

When performing hyperparameter tuning, parameters and applying raw feature sets (Figure 4, Figure 5), the best results are obtained using Random Forest and XGBoost - both show highest accuracy and achieve greater equity in precision and recall between classes ("Medium" and "Very High"). Hyperparameter tuning significantly improved the performances of GradientBoosting (+10%) and SVC (+5%) and gave a noticeable improvement to KNeighbors (+3.5%).

The accuracy and F1-macro comparison graph on the test set (Figure 5) demonstrates that the nonlinear models (Random Forest and XGBoost) provide better generalization compared to Logistic Regression.

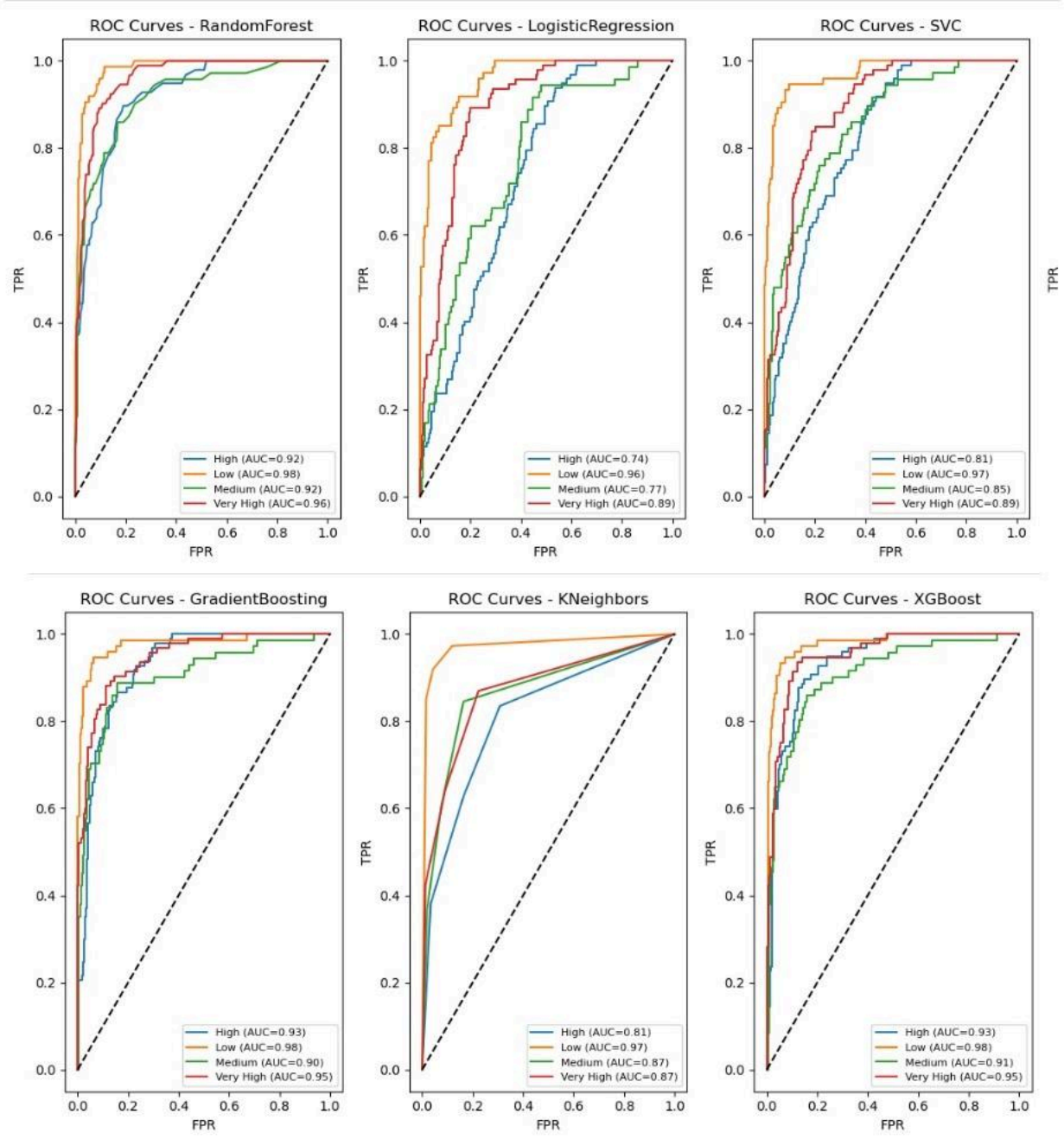
ROC Curves - RandomForest
High (AUC=0.92)
Low (AUC=0.98)
Medium (AUC=0.92)
Very High (AUC=0.96)
ROC Curves - LogisticRegression
High (AUC=0.74)
Low (AUC=0.96)
Medium (AUC=0.77)
Very High (AUC=0.89)
ROC Curves - SVC
High (AUC=0.81)
Low (AUC=0.97)
Medium (AUC=0.85)
Very High (AUC=0.89)
ROC Curves - GradientBoosting
High (AUC=0.93)
Low (AUC=0.98)
Medium (AUC=0.90)
Very High (AUC=0.95)
ROC Curves - KNeighbors
High (AUC=0.81)
Low (AUC=0.97)
Medium (AUC=0.87)
Very High (AUC=0.87)
ROC Curves - XGBoost
High (AUC=0.93)
Low (AUC=0.98)
Medium (AUC=0.91)
Very High (AUC=0.95)
TPR
FPR

Figure 4

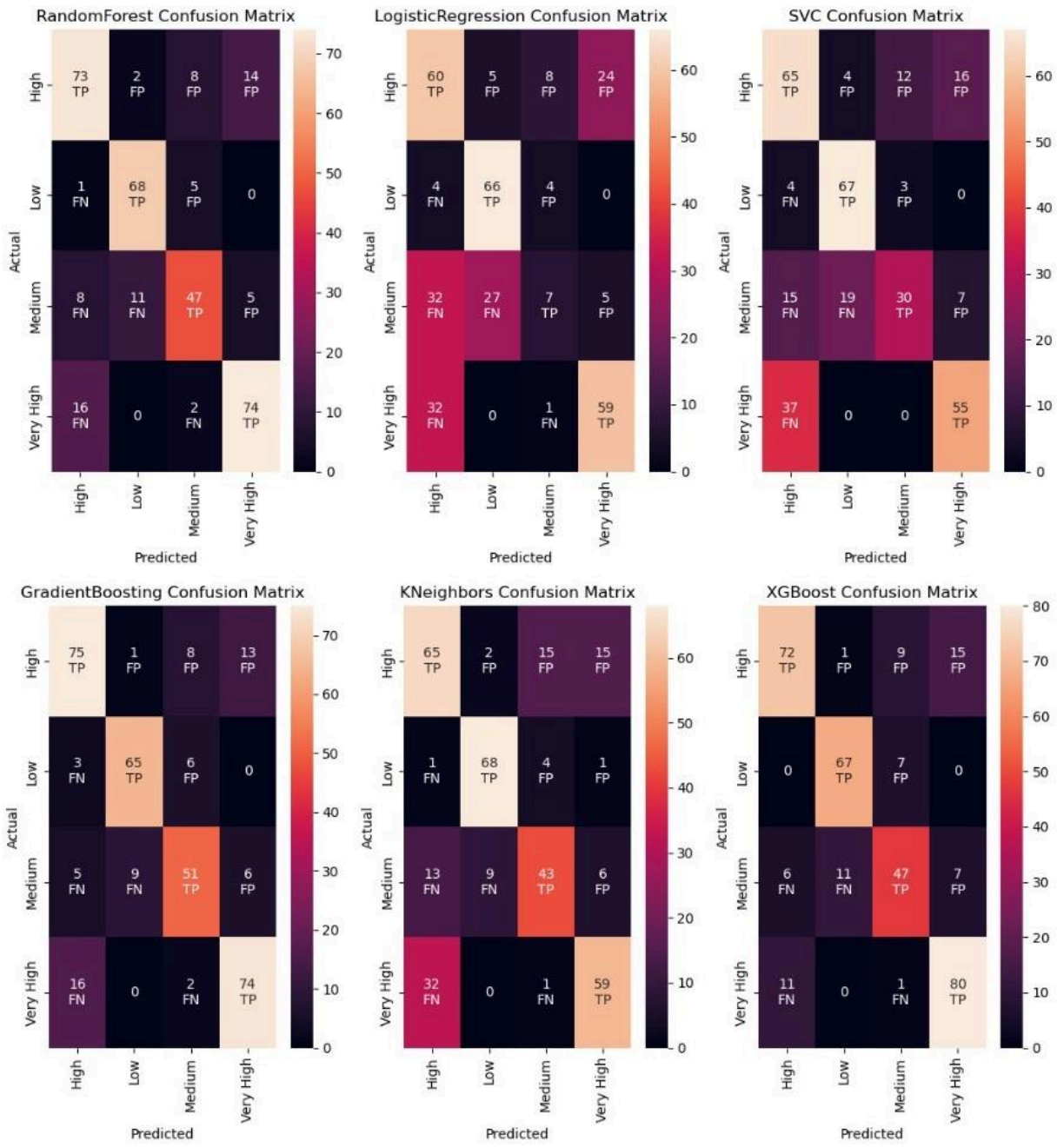

Figure 5

Discussion and Advanced Findings Visualizations

Appropriate tuning and feature selection increase the model's ability to correctly classify “Very High” and “Very Low” without sacrificing performance in the other classes (Figure 4).
All models struggle to classify the “medium” level of corruption, especially Logistic Regression that is only correct in 7% of cases for the “medium” class. This probably happens due to:

- Class imbalance - “Medium” is less common than other classes. The model simply learns to guess the more frequent labels.
- Blurred boundaries - In reality, “medium corruption” is difficult to distinguish from adjacent categories based on features. If economic, political, and press freedom indicators don't show clear jumps, the models confuse “medium” with “low” or “high.”
- Characteristics of linear models - Nonlinear models show better performance when it comes to predicting the “Medium” class. Logistic Regression constructs a simple boundary in the feature space. If classes overlap, a linear model will perform poorly with the middle category.
- Metrics - The model almost always predicts something else, even if the case is truly “medium”.

In general “medium” is the area of confusion here: it's too similar to its neighbors, and the data doesn't provide a strong signal.

Classical economic variables remain the most robust and universal predictors, amplifying the model's interpretive value for analysts, governments and investors.

Finally, time segmentation (train/test split by year) is crucial to avoid data leakage and simulate realistic deployment, demonstrating that the results are replicable and generalizable.

Analysis of models’ mistakes

The worst case is when real corruption is high or very high, but the model assigns “low”.
This risk is that the model underestimates the problem. A less harmful scenario is confusing overestimating corruption level and predicting classes that are close to each other, for example, “medium” corruption predicted as “high”. This error would not be as costly as the previous example.

Logistic Regression is the most dangerous model: many “Very High” (Figure 5) values have migrated to “High” and even “Medium.” SVC also confuses, but more often within close categories (“Medium” migrates to “High”). RandomForest, GradientBoosting, and XGBoost generally maintain boundaries better: they almost never send “Very High” to “Low” or vice versa. KNeighbors is similar to XGBoost, but sometimes confuses more at the “High” and “Very High” boundary.

From the perspective of minimizing social harm from errors, boosting models (GradientBoosting, XGBoost) and RandomForest are better: they don't make fatal errors after two levels. Logistic Regression, on the other hand, often underestimates “Very High,” which is better in practical terms.

Limitations:

- Models are trained exclusively on 21st-century data. Their predictions are not transferable to historical contexts where political, economic, and institutional structures operated under different conditions.
- Corruption is not a directly measurable quantity. Categories like “low,” “medium,” “high,” are defined based on perception indexes and expert assessments which means subjectivity. This is particularly noticeable in the “medium” class, where the boundaries with close categories are blurred.
- Datasets used for training can contain reporting errors, inconsistencies, or biases in the methods of collecting and aggregating corruption indicators. These issues can spill over into the models, limiting the reliability of their results.
- More complex models, such as gradient boosting or XGBoost, achieve higher accuracy but are harder to interpret. This limits transparency in explaining the reasons behind a particular prediction, which is relevant for sensitive topics such as governance and corruption.
- Widely used corruption indices are often based on surveys of experts or entrepreneurs. This can lead to geographic, cultural, or methodological biases that the models will then inherit.
- Even in the 21st century, the dynamics of corruption are changing. Indicators that were predictive in one decade may become less relevant in another, reducing the stability of long-term forecasts.

Final Conclusions

The report demonstrates that nonlinear models such as Gradient Boosting, XGBoost, and Random Forest are more suitable for corruption classification tasks which means corruption and the used features have nonlinear relationships. Those models consistently outperform linear methods such as Logistic Regression.

At the same time, the analysis reveals a systematic difficulty for all models in correctly identifying the “medium” level of corruption. This reflects both the imbalance of classes in the dataset and the blurred boundaries between categories. Future research could focus on improving class balance, integrating additional institutional or historical variables, and exploring interpretable ML approaches to increase the confidence in forecasts.