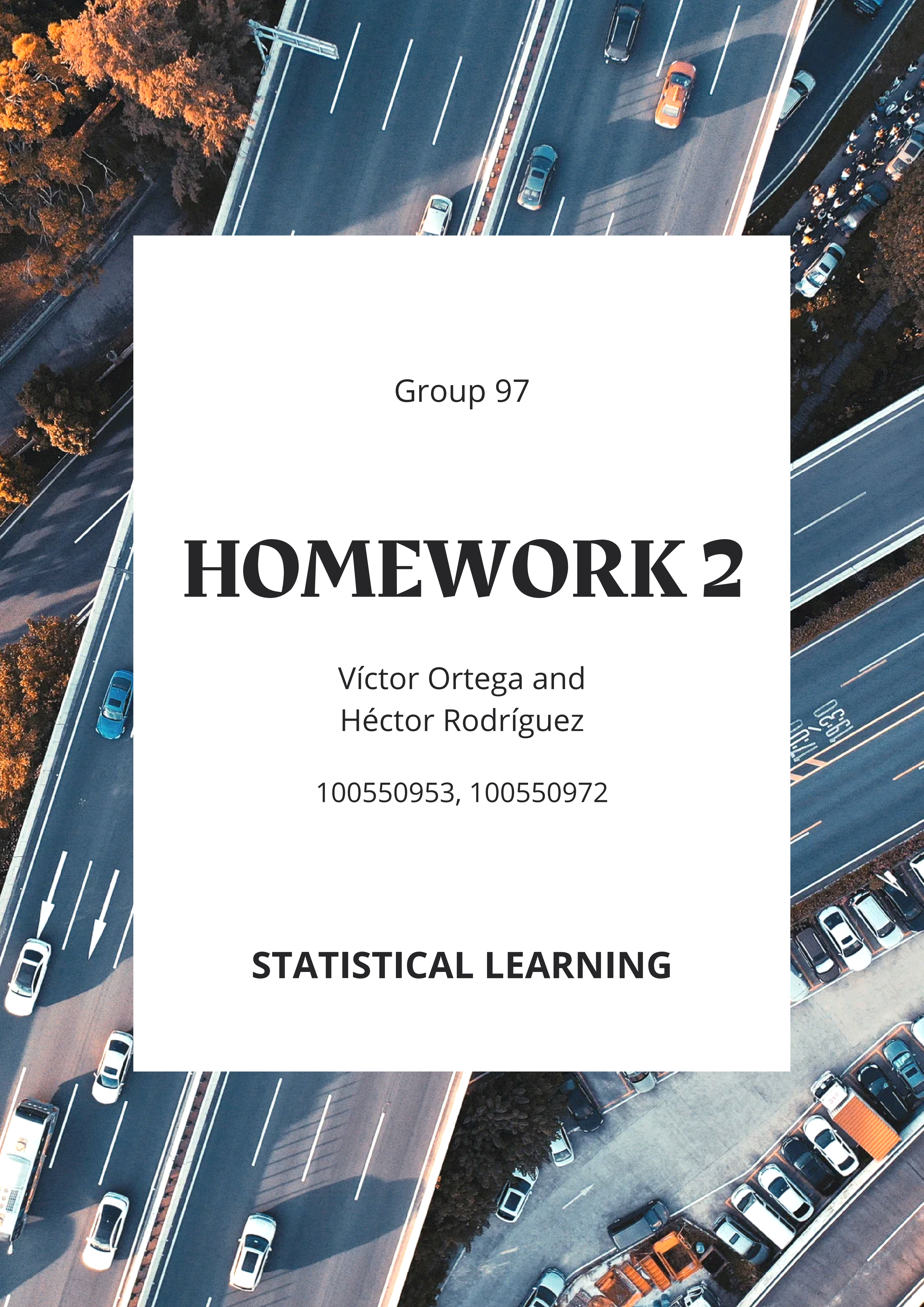


An aerial photograph of a multi-lane highway. Several cars are visible on the road, including a white car in the bottom left, a blue car in the middle left, and a red car in the top right. The road is flanked by trees with autumn-colored leaves on the left and a parking lot with several cars on the right. The overall scene is captured from a high angle, showing the flow of traffic and the surrounding environment.

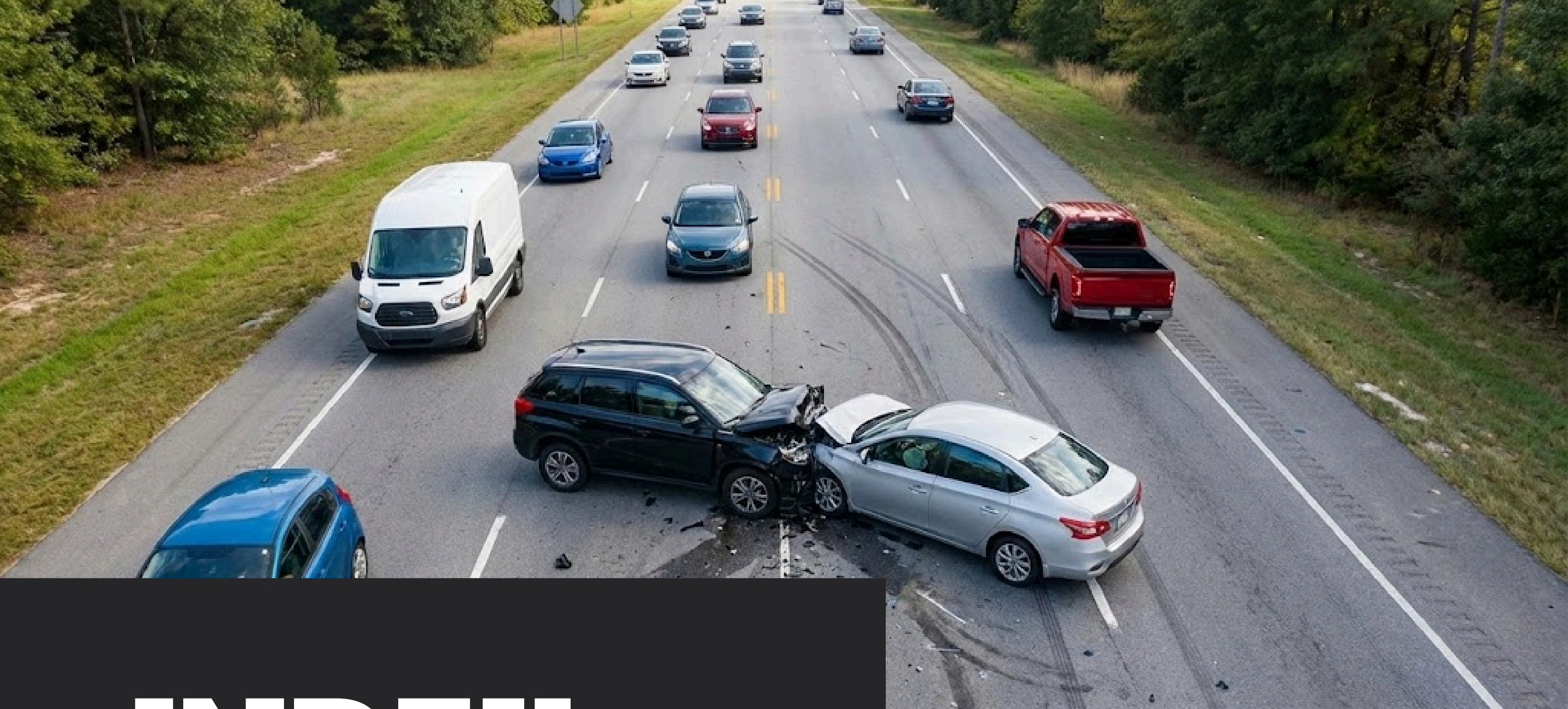
Group 97

HOMEWORK 2

Víctor Ortega and
Héctor Rodríguez

100550953, 100550972

STATISTICAL LEARNING



INDEX



- 1 Data preprocessing and visualization tools
- 2 Classification (emphasis on prediction)
- 3 Classification (emphasis on interpretation)
- 4 Feature Selection and comparison of predictor sets

INTRODUCTION



In this project, we study traffic accidents in Madrid during 2025. Our main goal is to understand what factors make an accident severe. We use different models to predict severity and SHAP values to explain how these models work. Finally, we compare a complex model with a simpler one to see if we can still get good results with fewer variables.

01

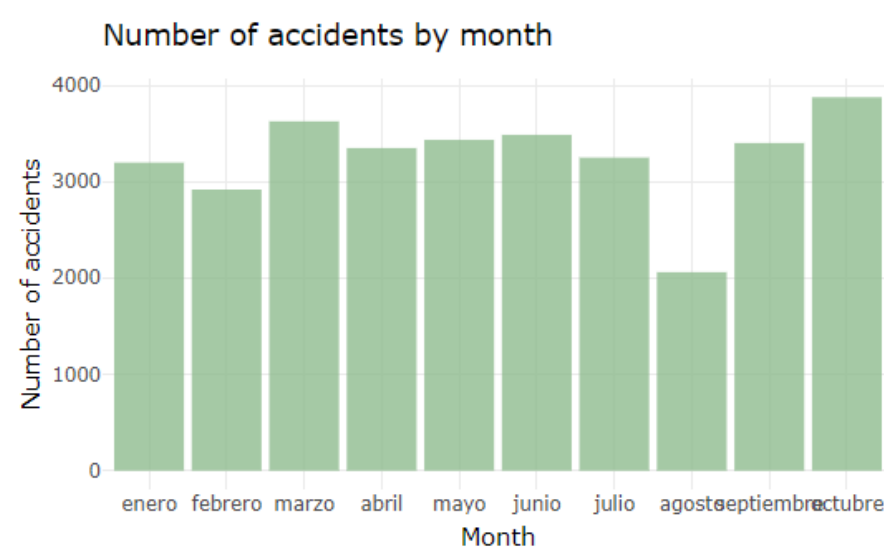
DATA PREPROCESSING AND VISUALIZATION TOOLS

First, we cleaned the dataset by removing duplicates and useless columns like coordinates. We fixed missing values and blank spaces, especially in weather and vehicle types. We also created new groups for injury severity and vehicle categories to make the data easier to analyze. To handle the fact that severe accidents are rare, we used stratified sampling and downsampling techniques.

DATA PREPROCESSING AND VISUALIZATION TOOLS

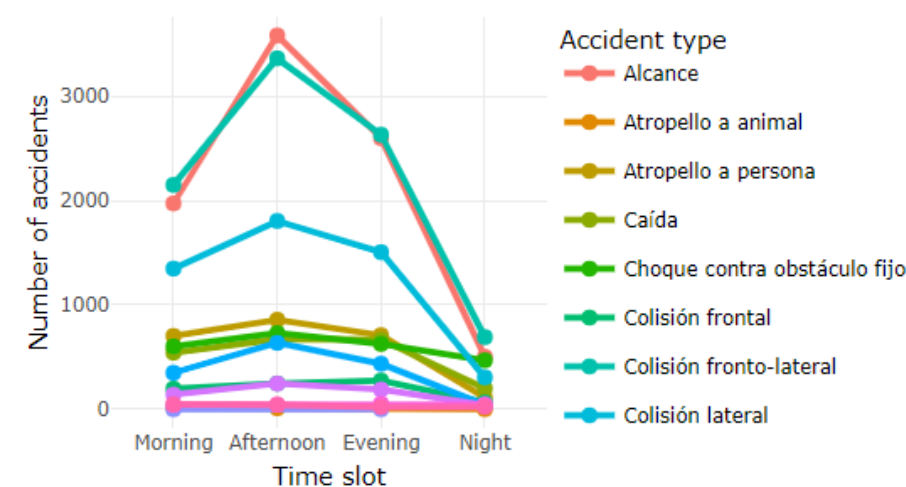
First, we cleaned the dataset by removing duplicates and useless columns like coordinates. We fixed missing values and blank spaces, especially in weather and vehicle types. We also created new groups for injury severity and vehicle categories to make the data easier to analyze. To handle the fact that severe accidents are rare, we used stratified sampling and downsampling techniques.

For the visualization part, we use eight different plots to explore the dataset and find the most important patterns and causes behind traffic accidents in Madrid.



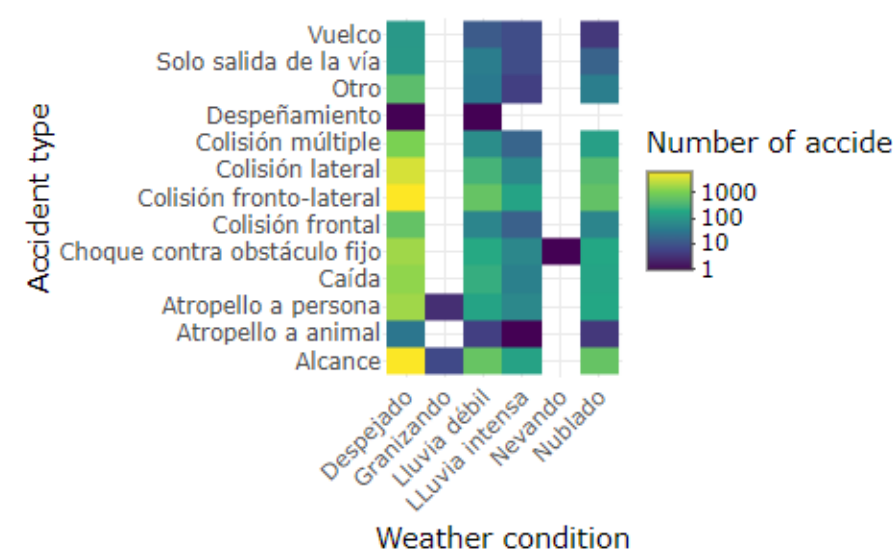
Accidents by month

This bar chart shows how many accidents happened each month. March and October have the most accidents. There is a big drop in August because many people leave Madrid for summer holidays.



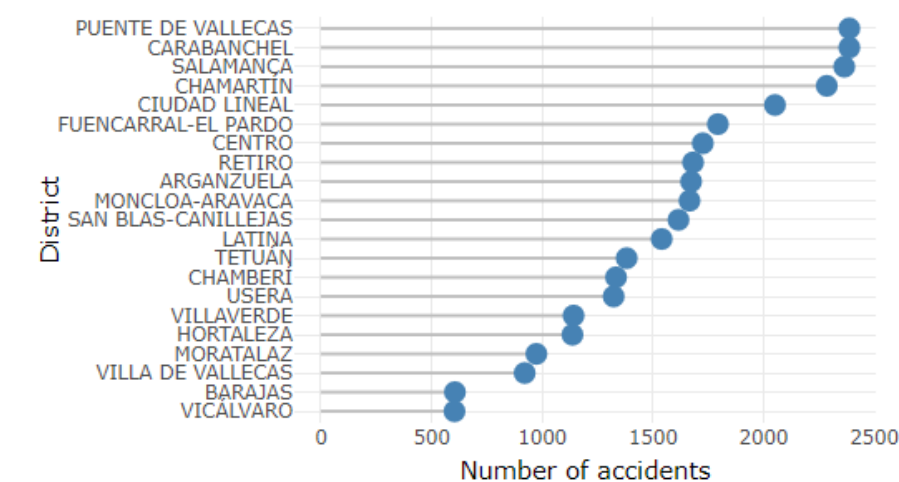
Accidents across time slots

A line chart showing accidents during Morning, Afternoon, Evening, and Night. Accidents peak in the afternoon when traffic is highest. The number of crashes drops significantly during the night.



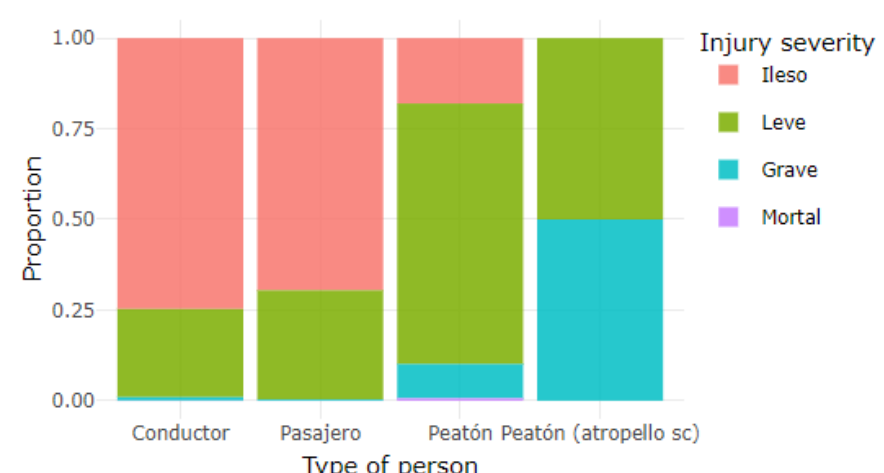
Weather and accident types

A heatmap comparing weather conditions with the type of accident. Most accidents happen when the weather is clear. This suggests that drivers relax and pay less attention when conditions seem safe.



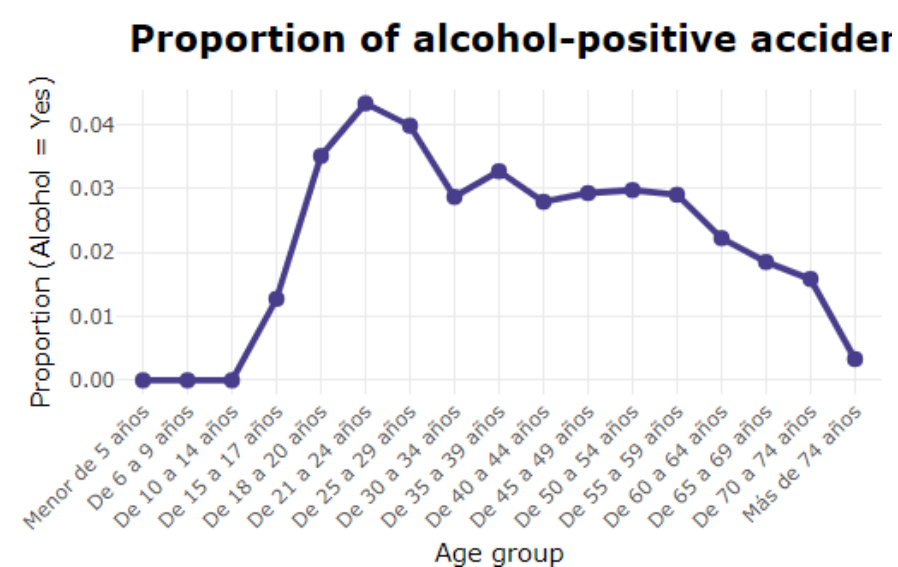
Distribution by district

A lollipop chart ranking Madrid districts by the number of accidents. Traffic accidents are not distributed equally. Some districts have much more activity and risk than others due to urban density.



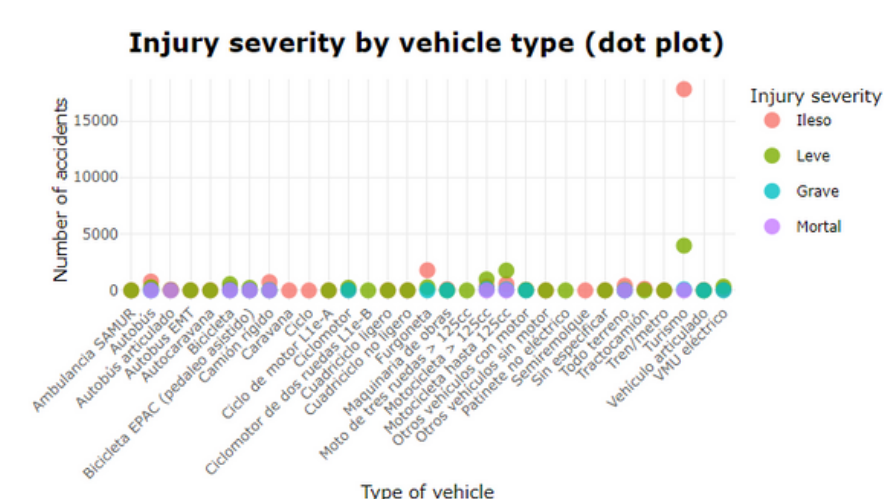
Severity by type of person

This plot shows the proportion of injury levels for drivers, passengers, and pedestrians. Pedestrians are the most vulnerable group. They have a much higher percentage of severe and fatal injuries compared to people inside vehicles.



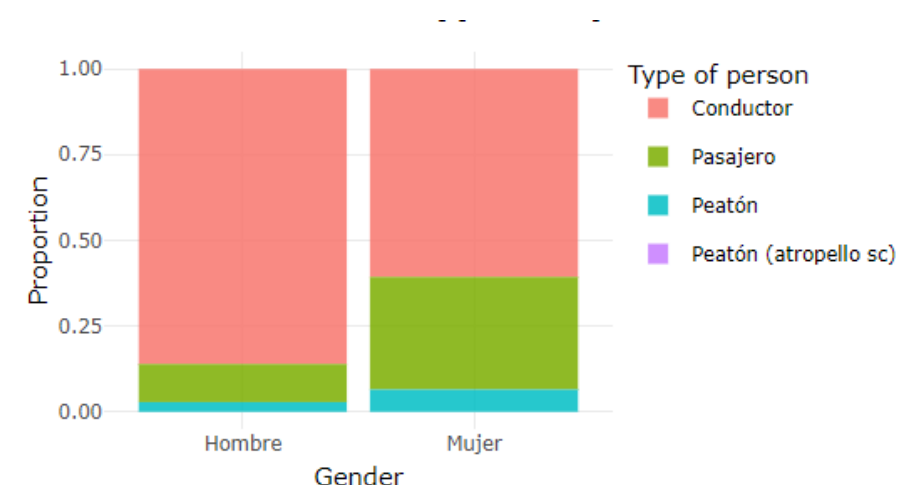
Alcohol by age group

This line plot shows the proportion of positive alcohol tests in different ages. Alcohol involvement is much higher in young adults and decreases as drivers get older. Children and seniors have almost zero cases.



Accidents by vehicle type

A dot plot showing the number of accidents for different vehicles. Passenger cars are involved in the majority of accidents, but we see many different vehicle types that need specific safety focus.



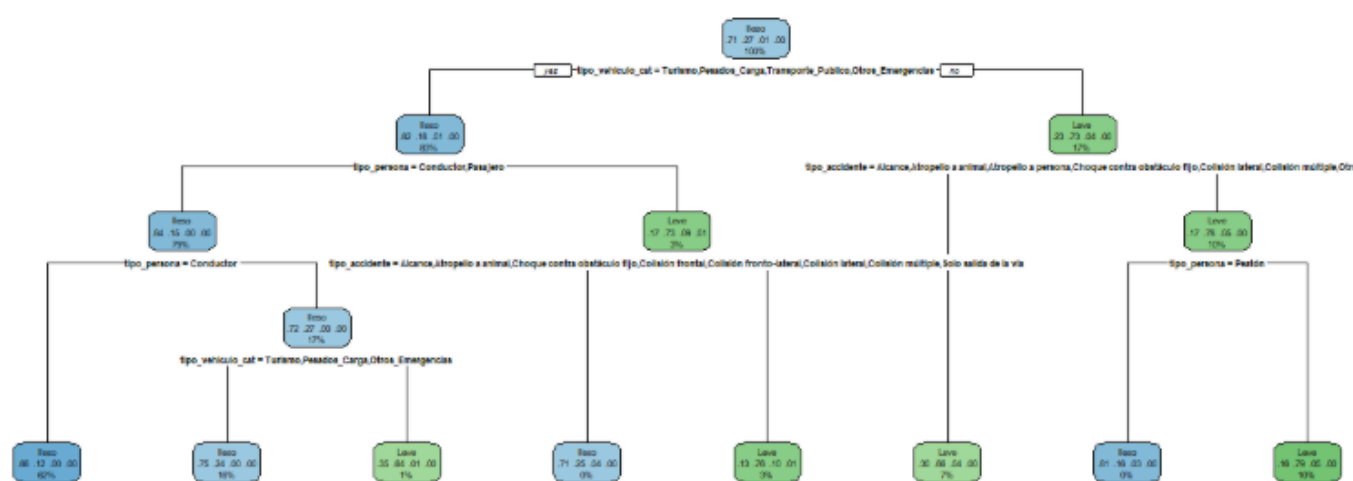
Gender and person roles

A bar chart comparing the roles (driver, passenger, pedestrian) between men and women. In both genders, most are drivers. However, women are passengers or pedestrians more often than men, showing different mobility patterns.

CLASSIFICATION (EMPHASIS ON PREDICTION)

The objective of this section is to evaluate different supervised classification models to predict traffic accident severity in Madrid. The goal is to identify which modeling approach provides the best predictive performance for detecting severe accidents, while accounting for the imbalanced nature of the target variable.

As probabilistic learning approaches, we applied Linear Discriminant Analysis (LDA), Naive Bayes, and a simple benchmark model. These methods aim to maximize overall classification accuracy, treating all errors equally. In this setting, the cost of misclassifying a severe accident is not considered, as the objective is simply to achieve the highest possible accuracy.



We next focus on machine learning models, including Decision Trees, Random Forest, and Gradient Boosting. These methods can capture complex and non-linear relationships between the predictors and accident severity.

Unlike probabilistic approaches, machine learning models allow us to take into account the cost of misclassification. In particular, failing to detect a severe or fatal accident is considered a critical error. For this reason, we also evaluated a Random Forest model combined with SMOTE, which improves the detection of severe cases under class imbalance. As an example of these methods, we present the Decision Tree model used in this analysis.

Finally, we built a table to compare the accuracy and the cost of all the tested algorithms. This comparison allows us to evaluate not only how often each model predicts correctly, but also the impact of misclassifying severe and fatal accidents. Based on the results shown in the table, the Random Forest model with SMOTE stands out as the best option for this project. It achieves a high accuracy while presenting the lowest cost, ensuring good overall performance and a better treatment of the imbalanced data. In most other models, severe and fatal cases are rarely detected, whereas the SMOTE-based Random Forest is able to account for these high-risk situations more effectively.

Model	Accuracy	Cost
Benchmark	0.715	16.31
Decision Tree	0.824	7.71
Random Forest	0.822	8.20
Gradient Boosting	0.819	7.99
Random Forest + SMOTE	0.821	7.71
Naive Bayes	0.808	NA
LDA	0.802	NA

03

CLASSIFICATION (EMPHASIS ON INTERPRETATION)

In the previous section, we focused on making a model that predicts well. However, models like Random Forest are 'black boxes', meaning it is hard to see why they make a decision. In this part, we use SHAP (SHapley Additive exPlanations) values to break down the model and see which variables are actually driving the predictions for severe accidents.

To make the interpretation clearer, we simplified the problem into a binary classification: 'Severe' (Grave/Mortal) vs. 'Non-Severe'. Before interpreting, we checked the Confusion Matrix. It helps us see if the model misses severe accidents (False Negatives), which is the most dangerous error in road safety. We ensured our model focuses on minimizing these misses to identify as many high-risk situations as possible.

```
> confusionMatrix(preds_rf, test_set$target)
Confusion Matrix and Statistics

          Reference
Prediction NonSevere Severe
NonSevere    2965     13
Severe       586     32

      Accuracy : 0.8334
      95% CI   : (0.8208, 0.8455)
No Information Rate : 0.9875
P-Value [Acc > NIR] : 1

      Kappa : 0.075

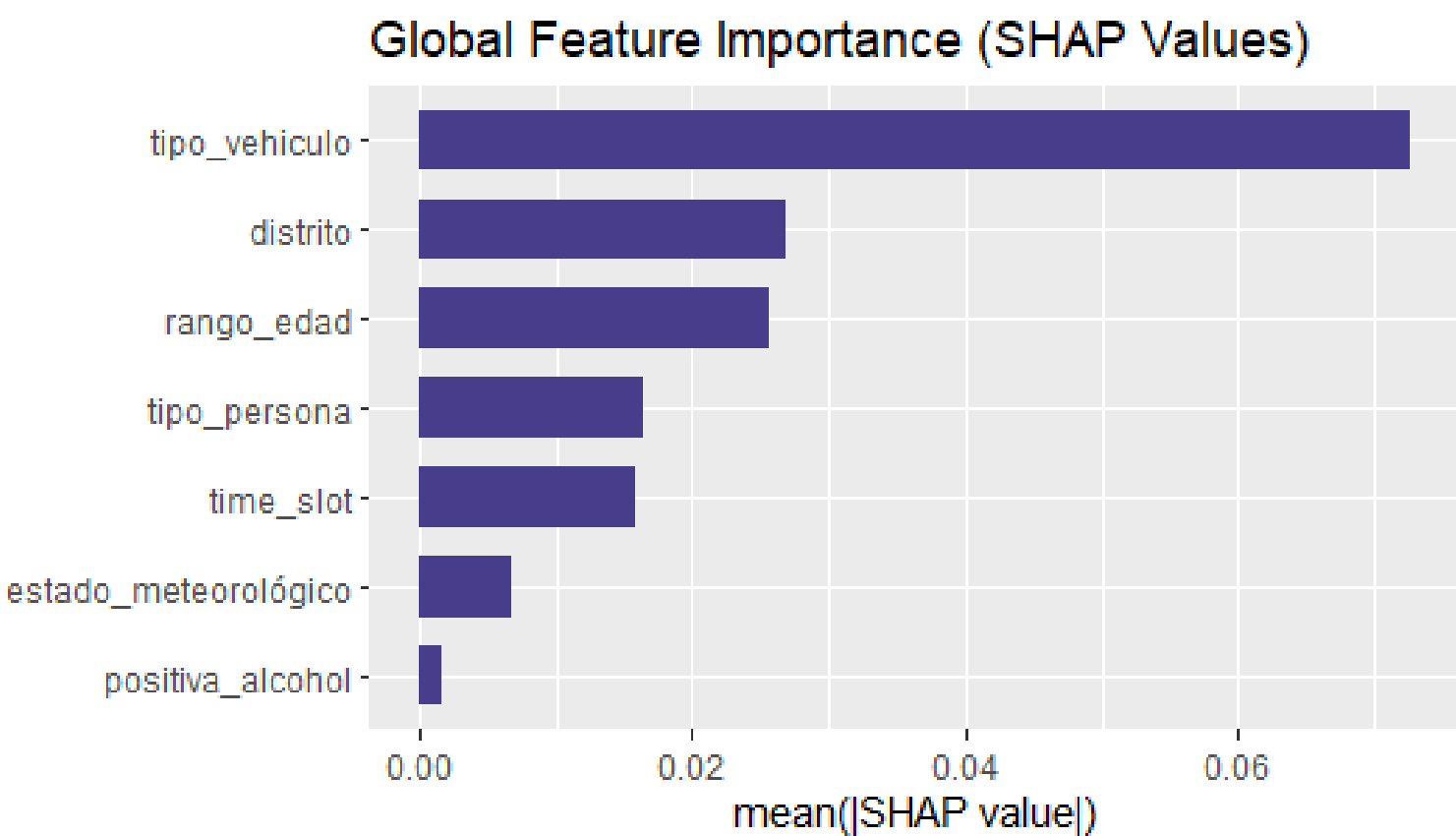
McNemar's Test P-Value : <2e-16

      Sensitivity : 0.83498
      Specificity : 0.71111
Pos Pred Value : 0.99563
Neg Pred Value : 0.05178
Prevalence : 0.98749
Detection Rate : 0.82453
Detection Prevalence : 0.82814
Balanced Accuracy : 0.77304

      'Positive' Class : NonSevere
```

Looking at our SHAP summary plot, we can see a clear ranking of variables:

- Vehicle Type (tipo_vehiculo): This is the most important factor. The model relies heavily on the type of vehicle involved to determine the severity, likely because motorcycles and bicycles have much higher risks than cars.
- Age Range (rango_edad): Age is the second most important predictor. This shows that certain age groups (like very young or very old people) are more vulnerable in traffic accidents.
- District (distrito): The location of the accident also plays a big role, showing that some areas in Madrid are more dangerous than others.



A very interesting result is that 'Alcohol' and 'Drugs' appear at the bottom of the importance list. This doesn't mean they aren't dangerous. In our Logistic Regression, they have high risk coefficients. However, because alcohol-positive cases are very rare in the total dataset, the Random Forest doesn't use them as 'global' predictors. This shows the difference between individual risk and global impact.

04

FEATURE SELECTION AND COMPARISON OF PREDICTOR SETS

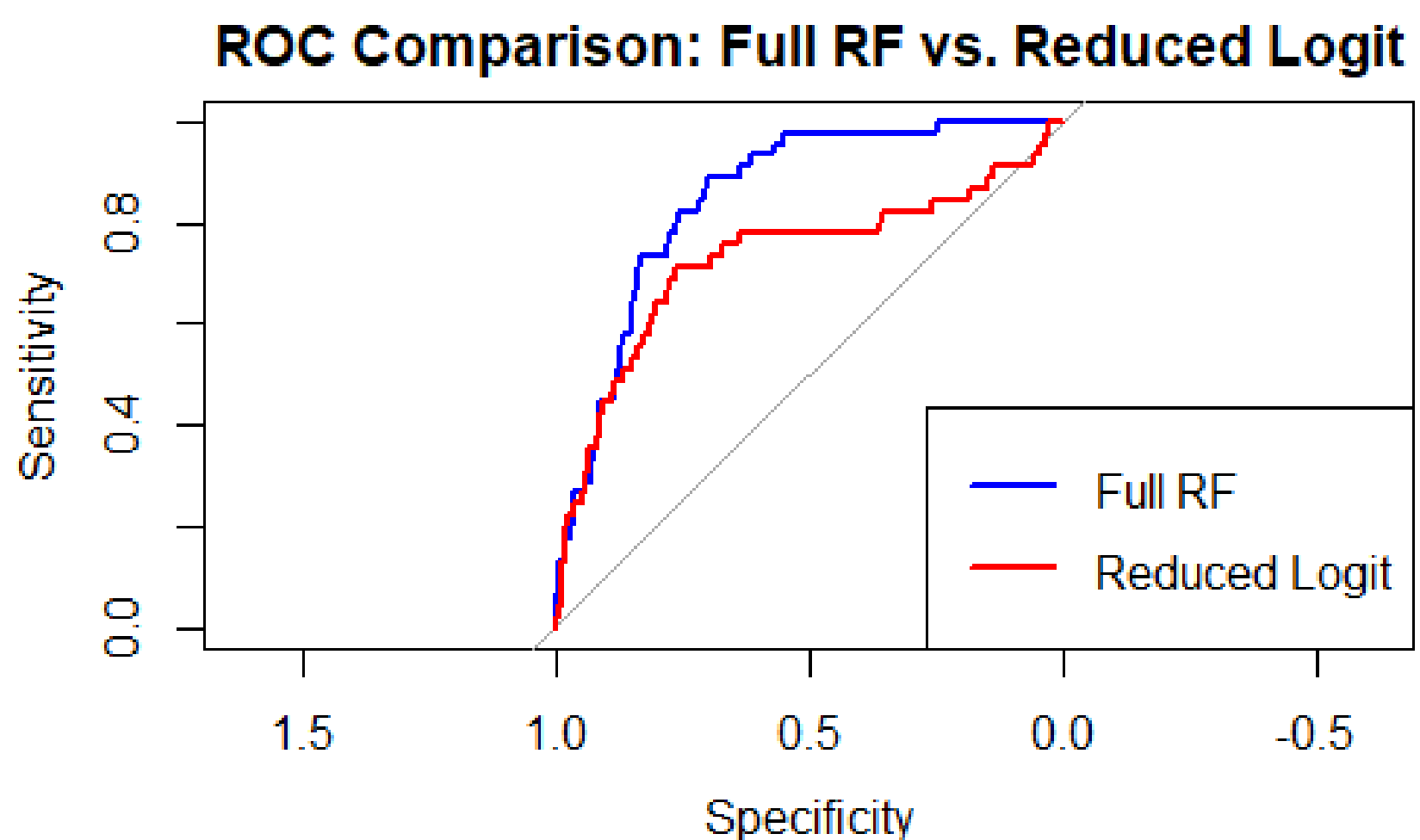
In data science, a simpler model is often better if it performs well. It is easier to explain to the authorities and faster to implement. Our goal here was to see if we could create a 'Reduced Model' using only the top 3 variables identified by SHAP (Vehicle Type, Age, and District) and still get good results.

We compared two different approaches to see if high complexity is truly needed:

- Full Model: A tuned Random Forest using all available variables to capture complex and non-linear patterns.
- Reduced Model: A Logistic Regression using only the 3 key factors identified by SHAP (Vehicle Type, Age, and District).

By doing this, we test if the complexity of the Random Forest is actually necessary or if these main risk factors are enough to explain the majority of severe accidents.

The best way to compare these models is the ROC Curve. The Area Under the Curve (AUC) provides a value between 0 and 1 that tells us the probability that the model correctly distinguishes between a severe and a non-severe accident:



- The Full Random Forest (Blue line) achieved an AUC of around 0.76.
- The Reduced Logistic Model (Red line) achieved an AUC of around 0.74. The difference is very small (only 2%). This is a key finding because it shows that most of the 'predictive power' comes from just three variables.

This comparison proves we don't need hundreds of variables to identify high-risk situations. By focusing only on vehicle safety, specific age groups, and dangerous districts, authorities can address the majority of severe accidents effectively. For organizations like the DGT, a simpler model is easier to justify and implement. While the Random Forest is slightly more accurate, the Logistic model is transparent and easier to explain to the public. Ultimately, our analysis shows that in road safety, a model that is easy to act upon is often more valuable than the most complex algorithm.