PREDICTING THE OUTCOME OF DRIVING EXAMS IN ESTONIA

Lisette Pihor & Vanessa Apuhtin
Institute of Computer Science, University of Tartu

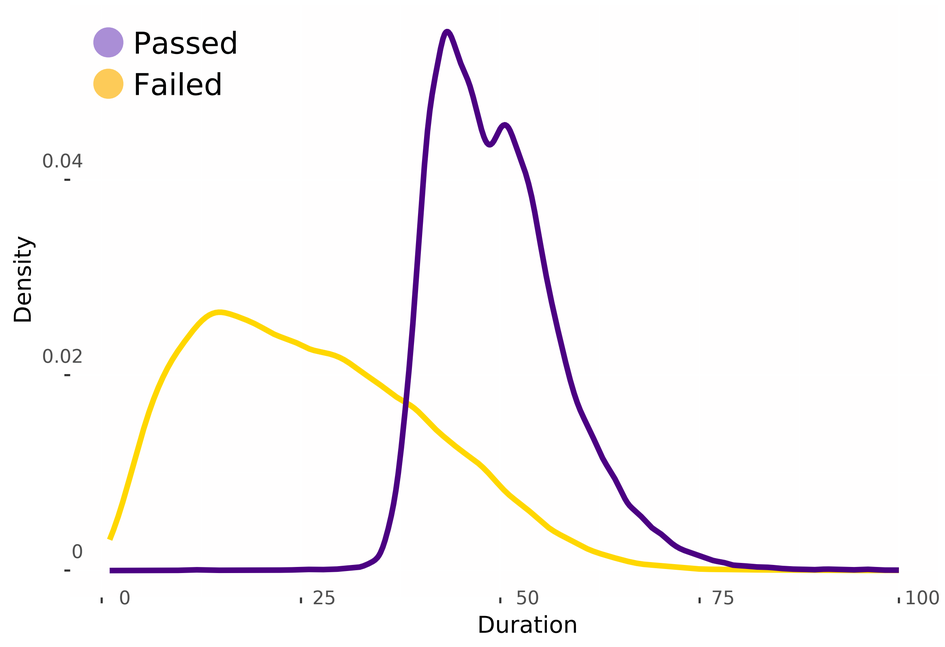
INTRODUCTION

One of The Estonian Transport Administration (ETA) functions is to conduct **practical driving tests** for people who want to obtain a driving licence. There are many myths surrounding how or where to do the driving exam. These beliefs create uncertainty and speculation surrounding the driving test process. Our aim is to find features that affect the outcome the most and develop a **predicting model**.

DESCRIPTION OF THE DATA

- The dataset is from the ETA, consisting of driving exams that took place in Estonia from 2021 to 2023. [1]
- It has 13 attributes and about **130 000 records.**
- It is quite **balanced dataset**: there are 54% passed cases (70 087) and 42% failed cases (55 298) and 3% of other outcome what we didn't consider.
- The dataset includes information on the examinees, examiners, exam locations, exam categories, exam results, exam durations, dates, and comments.
- We **got rid of 3% of the data**, caused by missing or suspicious values. Additionally, we removed the attributes, that did not contribute meaningful insights to our project.
- The dataset had mostly **categorical values.**

Density of duration based on outcome



METHODS & MODELS

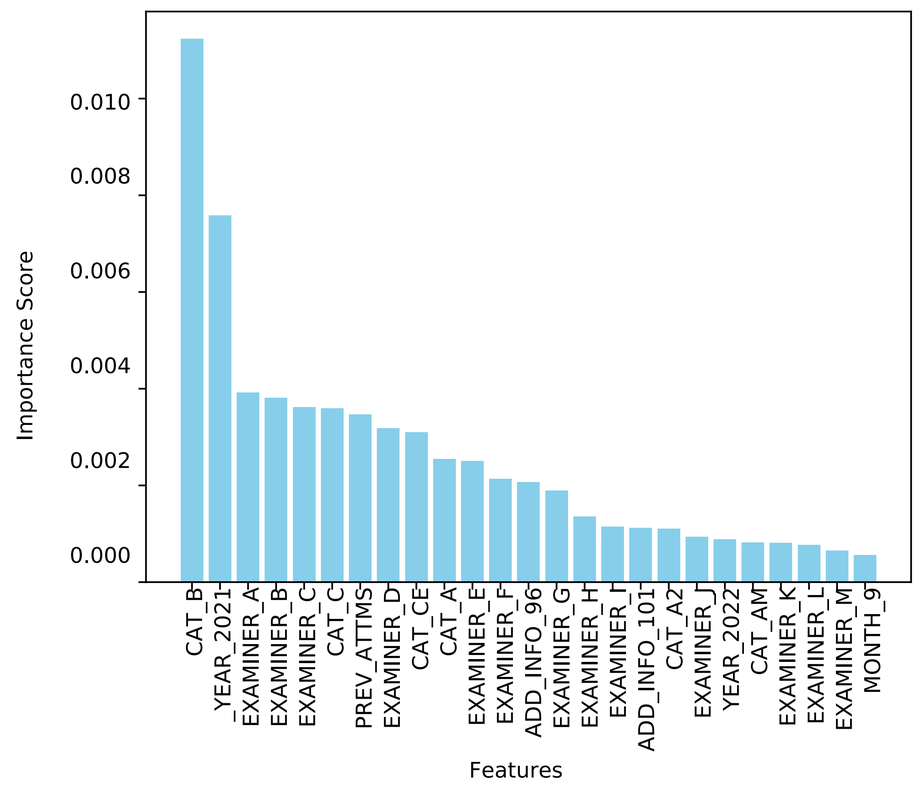
- We transformed categorical variables into binary matrix format with **one-hot encoding**.
- Due to the dataset's categorical nature we deployed **decision tree, random forest, and SVM models**.
- Methodology involved **Hold-out validation, K-fold cross-validation and Grid-search** for optimal model performance.

THE RESULT

- The outcomes of the models were rather similar. The top 6 ones are shown in the table. Among models, using **Decision Tree** with the criterion 'gini' and a maximum depth of 10 yielded had the **highest accuracy of 0.895**, recall of 0.973, and precision of 0.858.
- The second goal was to identifying the most important features in our model.
- **Duration** was overwhelmingly the most **important attribute**, the distribution of duration is shown in the table on the left. Next 25 most important features are shown in the last table

Model (criterion, max_depth)	Accuracy	Recall	Precision
Decision Tree (gini, 10)	0.895398	0.973284	0.858322
Decision Tree (entropy, 9)	0.894760	0.968141	0.860727
Decision Tree (entropy, 8)	0.893085	0.977713	0.852454
Random Forest (entropy, None)	0.890134	0.947925	0.867547
Random Forest (gini, None)	0.889496	0.946425	0.867649
Random Forest (entropy, 10)	0.885548	0.983642	0.839071

Top 25 Feature Importance from Decision Tree Model without duration



REFERENCES

[1] https://avaandmed.eesti.ee/datasets/toimunud-soidueksamid-eestis