Estonian Traffic Accidents

Mia Johanna Paas

Armin Liiv

Tartu 2025

Table of Contents

1. Identifying business goals

1.1. Background

Our project, *Estonian Traffic Accidents*, examines which factors influence traffic accidents involving human victims in Estonia and forecasts future trends. Although we have no specific client, the work is relevant for transport authorities and road safety planners. Existing public statistics provide only general summaries and do not show how conditions such as weather, road type or time of day relate to accident risk.
Using accident-level data from 2011–2024 and national trend statistics, the project aims to identify circumstances linked to severe accidents and provide simple predictive insights that could support future planning.

1.2. Business goals

The main goal is to determine which factors most affect the severity and frequency of traffic accidents in Estonia. This includes analysing how conditions like weather, time, season and road characteristics relate to outcomes.
A second goal is to build a predictive model that estimates future fatal accident counts based on historical data.

Although there is no client, these goals are defined from the perspective of potential public beneficiaries such as transport authorities and road safety planners.

1.3. Business success criteria

The project is successful if it identifies the key factors that clearly correlate with severe or fatal traffic accidents and presents these findings in an understandable way. Additionally, the predictive model should reflect historical accident trends well enough to give reasonable future estimates. Since there is no client, success means producing results that could realistically support public safety planning if used by relevant organizations.

2. Assessing situation

2.1. Inventory of resources

The project uses two publicly available Kaggle datasets:

1. Data on Traffic Accidents with Human Victims (2011–2024) – detailed accident-level information. [1]
2. Traffic Accidents and Casualty Trends in Estonia – long-term national statistics. [2]

Technical resources include Python, Jupyter Notebook, and libraries such as Pandas, NumPy, Matplotlib, and others for analysis and modelling.

Human resources consist of the two team members, who have basic skills in data analysis, visualisation, and introductory machine learning. No external experts or clients are involved.

2.2. Requirements, assumptions and constraints

Requirements: Access to datasets, ability to load and process CSV files, and tools for exploratory analysis and modelling.

Assumptions: The datasets are sufficiently accurate and representative, and missing values can be handled without harming results.

Constraints: Some variables may be incomplete or inconsistent across years. Estonia's small number of fatal accidents limits prediction accuracy. Time and skill level also constrain the scope.

2.3. Risks and contingencies

Risk: Incomplete or inconsistent data. Some variables may be missing for certain years or recorded differently across datasets.

Contingency: Exclude unreliable fields or standardize formats where possible and clearly document limitations.

Risk: Unexpected data quality issues. Outliers, duplicates, or formatting errors may appear during analysis.

Contingency: Apply cleaning steps early and verify data structure before modelling.

2.4. Terminology

- Accident severity – the level of harm caused, based on injuries and fatalities.
- Fatal accident – an accident with at least one death.
- Injury accident – an accident resulting in non-fatal injuries.
- Risk factor – any condition linked to higher accident severity (e.g., weather, road state).

- Environmental conditions – weather, lighting, and road surface variables.
- Participant type – categories of road users involved (pedestrian, cyclist, driver).
- Temporal features – date, time, month, and weekday used for pattern analysis.
- Predictive model – a model forecasting future accident counts.
- Trend – long-term change over time.

2.5. Costs and benefits

Costs: No financial costs, only the team's time for analysis, modelling and reporting.

Benefits: The project offers clearer insights into factors influencing accident severity and provides reasonable forecasts of future fatal accident trends. It may support safety planning and helps the team develop practical data-analysis skills.

3. Defining data-mining goal

3.1. Data-mining goals

The first data-mining goal is to analyse the accident-level dataset to identify which variables have the strongest relationship with accident severity and frequency. This includes computing correlations, visualising distributions, and detecting meaningful patterns across conditions such as weather, time, season, and road type.

The second goal is to build a predictive model capable of estimating future fatal accident counts using historical data. The aim is not perfect accuracy, but a model that captures general seasonal and long-term trends and produces outputs that could support planning if applied by public safety stakeholders.

3.2. Data-mining success criteria

The data-mining process is considered successful if it produces:

- Clear, interpretable visualisations and statistical results showing which factors are most associated with accident severity and frequency.
- A predictive model that captures the main historical patterns of fatal accidents and provides reasonable future estimates without major systematic errors.
- Data that is sufficiently cleaned, well-structured, and documented to support reliable modelling and interpretation.

Success does not require perfect prediction accuracy but demands outputs that are technically sound, consistent with the data, and suitable for informing further analysis or practical use.

4. Gathering data

4.1. Outline data requirements

The project requires detailed accident-level data that includes date, time, location, accident type, number of injured and killed, and contextual variables such as weather, road type, and vehicle involvement. Additionally, long-term aggregated statistics on yearly or monthly casualty trends are needed to support forecasting.

4.2. Verify data availability

Both required data sources are publicly available on Kaggle. They are downloadable in CSV format and compatible with Python-based analysis.

4.3. Define selection criteria

We will use all accident records from 2011–2024 that involve human victims, and all yearly or monthly aggregated statistics available in the second dataset. Irrelevant fields, duplicated entries, or years with incompatible formats will be excluded. Only variables relevant to severity analysis and prediction will be retained.

5. Describing data

The project uses two datasets with different structures and purposes. Below is a detailed description of the fields they contain. We plan to translate all variable names into English later in the project to ensure consistency throughout the analysis.

Dataset 1: lo_2011_2024.csv — Accident-level data

This dataset contains one row per traffic accident involving human victims. It includes detailed information about the event, participants, environment, and circumstances.

Key fields in the dataset (Estonian → English):

General accident information:

- Juhtumi nr (Case number)
- Toimumisaeg (Date and time of accident)
- Aasta (Year)
- Kuu (Month)
- Nädalapäev (Weekday) – derived later
- Isikuid (Persons involved)
- Sõidukeid (Vehicles involved)
- Hukkunuid (Fatalities)
- Vigastatuid (Injured persons)
- Location information
- Aadress (Address)

- Maakond (County)
- Omavalitsus (Municipality)
- Asustusüksus (Settlement)
- Tänav (Street)
- Maja nr (House number)
- Ristuv tänav (Intersecting street)
- X koordinaat (X coordinate)
- Y koordinaat (Y coordinate)

These coordinates make it possible to perform spatial analysis if needed.

Accident type and participant roles:

The dataset includes many indicator variables specifying whether certain types of road users were involved:

- Jalakäija osalusel (Pedestrian involved)
- Jalgratturi osalusel (Cyclist involved)
- Mootorratturi osalusel (Motorcyclist involved)
- Mopeedijuhi osalusel (Moped rider involved)
- Kaassõitja osalusel (Passenger involved)
- Alaealise osalusel (Minor involved)
- Eesmase juhiloa omaniku osalusel (First-year driver involved)
- Joobes mootorsõidukijuhi osalusel (Intoxicated driver involved)
- Eaka (65+) mootorsõidukijuhi osalusel (Elderly driver 65+ involved)
- Sõiduautojuhi / Veoautojuhi / Bussijuhi osalusel
- (Car / Truck / Bus driver involved)
- Ühissõidukijuhi osalusel (Public transport driver involved)
- These fields allow modelling risk factors based on user type.
- Road and environmental conditions
- Tee tüüp (Road type)
- Tee liik (Road category)
- Tee element (Road element)
- Kurvilisus (Curvature)
- Tee tasasus (Levelness of road)
- Tee seisund (Road condition)
- Teekatte seisund (Surface condition)
- Teekate (Surface type)
- Lubatud sõidukiirus (Speed limit)
- Ilmastik (Weather)
- Valgustus (Lighting conditions)

These attributes are crucial for analysing severity patterns.

Dataset 2: TS093_20241202-120451.csv — Monthly accident trends

This dataset includes yearly rows with monthly accident counts and annual totals.

Key fields in the dataset:

- Indicator / Month (metadata about structure)
- January – December (Monthly accident counts)
- Ascending total (Cumulative total per year)
- Persons injured (Jan–Dec) (Monthly injury counts)
- Year (first column, e.g., 1990, 1991, …)

Although column names are not initially clean, the dataset clearly contains structured monthly time series suitable for forecasting.

Dataset suitability
The first dataset contains detailed accident data for analysis, while the second provides long-term monthly trends for forecasting. Core fields are mostly complete, and missing text values do not affect key tasks. All variables will be translated into English during data preparation.

6. Exploring data

The accident-level dataset contains timestamped records with casualties, locations, and environmental factors such as weather, road condition, and lighting. Fatal accidents form only a small portion of all cases, while injury counts are consistently recorded. Some descriptive fields include missing values, but core variables are usable for analysis. The monthly trend dataset shows clear long-term declines in accident numbers with noticeable seasonal variation. Together, the datasets allow examining temporal patterns, severity-related factors, and preparing features suitable for correlation analysis and forecasting.

7. Verifying data quality

The accident-level dataset is generally consistent: key fields such as dates, casualty counts, and main categorical variables are complete for most entries. Some location-related fields and detailed road descriptors contain missing or irregular values, but these do not affect core analysis. The dataset shows no structural corruption, though categories may require grouping due to high detail levels. The monthly trend dataset is clean, with complete yearly rows and no missing values in accident totals. Overall, the data is suitable for modelling after minor cleaning and standardisation.

8. Project plan

8.1. Project plan

Our project consists of five main tasks, each with the estimated workload per team member. The goal is for both members to contribute approximately 30 hours in total.

1. Data collection and initial cleaning

- Download datasets, check formats, unify date fields, remove duplicates, and inspect missing values.
- Hours: Mia 6h, Armin 6h.

2. Data exploration and visualisation

- Produce descriptive statistics, create temporal and severity plots, analyse patterns, and document findings.
- Hours: Mia 7h, Armin 7h.

3. Feature preparation

- Clean categorical variables, handle missing values, engineer time-based features, and select modelling variables.
- Hours: Mia 6h, Armin 6h.

4. Predictive modelling

- Train and compare models (e.g., regression, time-series), evaluate performance, refine models, and validate results.
- Hours: Mia 6h, Armin 6h.

5. Final reporting and presentation

- Write the report, prepare visual summaries, compile results, making a poster and finalise presentation text.
- Hours: Mia 5h, Armin 5h.

Total per member:

Mia: 6 + 7 + 6 + 6 + 5 = 30h

Armin: 6 + 7 + 6 + 6 + 5 = 30h

8.2. Methods and tools

We will use Python, Jupyter Notebook, Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn for data cleaning, exploration, feature preparation, and modelling. Time-series or regression-based forecasting methods will be used depending on data suitability. All work will be documented, and both members will contribute to interpretation and reporting. No paid software or external resources are required.

9. Sources

[1] https://www.kaggle.com/datasets/olaflundstrom/data-on-traffic-accidents-with-human-victims?select=lo_2011_2024.csv

[2] https://www.kaggle.com/datasets/mohammadzisan/traffic-accidents-and-casualty-trends-in-estonia