Task 2: Business Understanding

Project Title: Traffic Accidents in Estonia

Team Members:

- Karl Arthur Kade
- Aaron Anders Jaani
- Mihkel Kulu

Repository Link: https://github.com/karlkade/IDS-Traffic_Accidents

---

Identifying Business Goals

Background:

Traffic accidents remain a significant concern in Estonia, impacting public safety and resulting in economic and social consequences. Between 2011 and 2021, numerous traffic accidents involving human casualties have been recorded. Understanding the factors contributing to these accidents, especially those leading to fatalities, is crucial for developing effective prevention strategies and enhancing road safety measures.

Business Goals:

1. **Risk Assessment Model:** Develop a predictive model to estimate the likelihood of fatalities in traffic accidents by analyzing factors such as road type, driver profile, environmental conditions, and other relevant variables.

2. **High-Risk Location Identification:** Identify geographic locations with a higher predicted risk of fatal accidents to assist in targeted interventions and resource allocation.

3. **Driver Profiling:** Analyze driver demographics, including age, vehicle type, and license status, to profile drivers who are more prone to being involved in accidents, thereby aiding in the creation of targeted educational and enforcement campaigns.

Business Success Criteria:

- **Model Accuracy:** Achieve a predictive model with high accuracy and reliability in estimating fatality likelihood, validated through appropriate performance metrics.

- **Actionable Insights:** Provide clear identification of high-risk locations and driver profiles that can inform policymakers, law enforcement, and road safety organizations.

- **Stakeholder Utility:** Deliver findings that are directly applicable for stakeholders to implement safety improvements and reduce the number of fatal accidents.

Assessing the Situation

Inventory of Resources:

- **Data:** A comprehensive dataset containing detailed information on traffic accidents with human victims in Estonia from 2011 to 2021 (size: 7.66 MB).
- **Team Expertise:** Skills in data analysis, statistical modeling, machine learning, and data visualization.
- **Computational Resources:** Access to computers and software necessary for data processing and analysis (e.g., Python, data analysis libraries).
- **Guidance:** Potential consultation with our instructor, Friedrich Krull, if necessary.

Requirements, Assumptions, and Constraints:

- **Data Quality:** The analysis assumes that the provided dataset is accurate, complete, and representative of all traffic accidents involving human victims in Estonia during the specified period.
- **Data Privacy:** Compliance with data protection regulations, ensuring that any personal or sensitive information is handled appropriately.
- **Scope Limitation:** The project will be confined to the provided dataset without incorporation of external data sources, due to project scope and resource constraints.

Risks and Contingencies:

- **Incomplete Data:** Potential missing or inconsistent data fields that could affect model performance. Contingency plans include data cleaning and imputation techniques.
- **Model Overfitting:** Risk of overfitting due to the dataset size. This will be mitigated by using cross-validation and regularization methods.
- **Interpretability Challenges:** Complex models may be difficult to interpret. Preference will be given to models that balance accuracy with interpretability.

Terminology:

- **Fatality:** A death resulting from a traffic accident.
- **Risk Assessment Model:** A statistical or machine learning model used to estimate the probability of an outcome, such as a fatality in a traffic accident.
- **High-Risk Locations:** Specific geographic areas identified as having a higher incidence of accidents or fatalities.
- **Driver Profile:** Demographic and behavioral characteristics of drivers involved in accidents.

Costs and Benefits:

- **Costs:** Time and effort invested by team members in data processing, analysis, model development, and documentation.
- **Benefits:** Insights that could contribute to reducing traffic fatalities, informing policy decisions, and improving public safety in Estonia.

Defining Data-Mining Goals

Data-Mining Goals:

- **Predictive Modeling:** Develop a robust model to estimate the likelihood of fatalities in traffic accidents based on variables such as road conditions, driver demographics, and environmental factors.
- **Spatial Analysis:** Utilize geographic data to identify and visualize high-risk locations through methods like heatmaps.
- **Driver Analysis:** Examine driver-related data to identify profiles associated with higher accident involvement.

Data-Mining Success Criteria:

- **Model Performance:** Achieve satisfactory predictive performance metrics (e.g., accuracy, AUC-ROC scores) that indicate the model's reliability.
- **Insight Generation:** Successfully identify high-risk locations and driver profiles that are statistically significant and practically relevant.
- **Stakeholder Relevance:** Ensure that the findings are actionable and can be utilized by stakeholders for improving road safety measures.

Task 3: Data Understanding

Gathering Data

Outline Data Requirements:

- **Temporal Coverage:** Accident data from 2011 to 2021 involving human casualties.
- **Variables Needed:**
    - Accident specifics: Date, time, accident type, number of fatalities and injuries.
    - Location details: Address, GPS coordinates (X and Y), county, municipality.
    - Driver information: Age, license status, involvement of specific driver categories (e.g., novice drivers, elderly drivers).
    - Vehicle details: Type and number of vehicles involved.
    - Environmental conditions: Weather, lighting, road surface conditions, road type, curvature.

Verify Data Availability:

- The dataset is available from the Estonian open data portal (link).

- Initial examination confirms that the dataset includes the required variables and covers the specified time frame.
- The dataset appears to be comprehensive, with 7.66 MB of data encompassing numerous accident records.

Define Selection Criteria:

- **Inclusion:** All accident records involving human casualties between 2011 and 2021.
- **Exclusion:** Records with critical missing data (e.g., missing GPS coordinates, unknown driver age) that cannot be reasonably imputed.

Describing Data

The dataset comprises multiple fields:

- **Accident ID (`Juhtumi nr`):** Unique identifier for each accident.
- **Occurrence Time (`Toimumisaeg`):** Date and time of the accident.
- **Persons Involved (`Isikuid`):** Number of individuals involved.
- **Fatalities (`Hukkunuid`):** Number of deaths resulting from the accident.
- **Injuries (`Vigastatuid`):** Number of injured persons.
- **Vehicles Involved (`Sõidukeid`):** Number of vehicles involved.
- **Address and Location Data:** Including street names, house numbers, intersecting streets, counties, municipalities, settlements, and GPS coordinates (`GPS X`, `GPS Y`).
- **Accident Type (`Liiklusõnnetuse liik`):** Classification of accidents, both general and detailed.
- **Driver and Participant Involvement:** Indicators for involvement of pedestrians, cyclists, motorcyclists, novice drivers, elderly drivers, etc.
- **Road Characteristics:** Road type, road element (e.g., straight, curve), road surface, curvature, incline.
- **Environmental Conditions:** Weather conditions, lighting conditions, road surface conditions.
- **Speed Limits and Road Numbers:** Information about the speed limit at the accident location and road identifiers.

Exploring Data

- **Temporal Patterns:**
    - **Yearly Trends:** Analyze the number of accidents per year to identify any trends over the decade.
    - **Seasonal Variations:** Examine accidents by month to detect seasonal effects, such as increased accidents during winter months due to weather conditions.

    - **Weekly and Daily Patterns:** Identify peak days of the week or times of day when accidents are more frequent.
- **Spatial Distribution:**
    - **Accident Hotspots:** Use GPS coordinates to map accidents and identify clusters or hotspots.
    - **Urban vs. Rural:** Compare accident rates in urban areas versus rural locations.
- **Fatality Analysis:**
    - **Fatality Rates:** Calculate the proportion of accidents resulting in fatalities.
    - **Contributing Factors:** Explore correlations between fatalities and variables like road type, weather conditions, and driver age.
- **Driver Profiling:**
    - **Age Distribution:** Analyze the age distribution of drivers involved in accidents.
    - **License Status:** Examine the involvement of novice drivers (`Esmase juhiloa omaniku osalusel`) and drivers without valid licenses.
    - **Vehicle Types:** Assess whether certain vehicle types are more frequently involved in accidents.
- **Environmental Impact:**
    - **Weather Conditions:** Investigate how different weather conditions (`Ilmastik`) affect accident frequency and severity.
    - **Lighting Conditions:** Assess the impact of lighting (`Valgustus`) on accidents, distinguishing between daylight and darkness.

Verifying Data Quality

- **Missing Data:**
    - **Assessment:** Identified missing values in certain fields, such as environmental conditions and driver details.
    - **Handling:** Plan to handle missing data through imputation where appropriate or exclude records if critical data is missing.
- **Data Consistency:**
    - **Format Consistency:** Ensured that date and time fields are in a consistent format for temporal analysis.
    - **Categorical Variables:** Standardized categorical variables to ensure consistency in categories (e.g., accident types).
- **Outliers and Anomalies:**
    - **Extreme Values:** Detected records with unusually high numbers of fatalities or injuries and will investigate these cases for data entry errors.

    - **Location Accuracy:** Verified that GPS coordinates correspond to valid locations within Estonia.
- **Duplicate Records:**
    - **Check for Duplicates:** Performed checks to identify any duplicate accident records that could bias the analysis.
- **Data Completeness:**
    - **Coverage:** Confirmed that the dataset provides comprehensive coverage of accidents involving human casualties for the specified period.

Overall, the data appears to be of sufficient quality for the intended analysis, with some data cleaning required to address identified issues.

Task 4: Planning the Project

Project Tasks and Time Allocation:

1. **Data Preprocessing and Cleaning** (8 hours per team member)
    - Handle missing values and correct inconsistencies.
    - Convert data types and standardize formats.
2. **Exploratory Data Analysis (EDA)** (6 hours per team member)
    - Conduct statistical analyses to understand distributions and relationships.
    - Create visualizations to identify patterns and trends.
3. **Feature Engineering and Selection** (5 hours per team member)
    - Create new features from existing data (e.g., time of day categories).
    - Select relevant variables for modeling based on EDA findings.
4. **Model Development** (7 hours per team member)
    - Build predictive models (e.g., logistic regression, decision trees) for fatality risk assessment.
    - Tune model parameters for optimal performance.
5. **Spatial Analysis and Visualization** (4 hours per team member)
    - Generate heatmaps using GPS data to identify high-risk locations.
    - Utilize spatial analysis tools for deeper insights.
6. **Driver Profiling Analysis** (3 hours per team member)
    - Analyze driver demographics to profile accident-prone drivers.
    - Identify key characteristics associated with higher risk.
7. **Model Evaluation and Validation** (4 hours per team member)
    - Evaluate models using metrics like accuracy, precision, recall, and AUC-ROC.

    - Perform cross-validation to ensure model robustness.
8. **Reporting and Documentation** (3 hours per team member)

    - Prepare comprehensive reports summarizing methods and findings.
    - Document code and analytical processes for transparency and reproducibility.

Methods and Tools:

- **Programming Language:** Python
- **Data Manipulation:** Pandas, NumPy
- **Statistical Analysis:** SciPy, StatsModels
- **Machine Learning Libraries:** Scikit-learn for model development
- **Visualization Tools:** Matplotlib, Seaborn for plots; Folium or GeoPandas for geographical visualizations
- **GIS Tools:** QGIS for advanced spatial analysis if required
- **Version Control:** Git and GitHub for collaboration and code management

Comments:

- Regular team meetings will be scheduled to discuss progress, troubleshoot issues, and ensure alignment.
- Emphasis will be placed on model interpretability to make findings actionable for stakeholders.
- All findings will be communicated effectively through visualizations and clear reporting.
- The project plan is flexible to accommodate any unforeseen challenges, with contingency time built into each phase.