

## **DEDA CHECKLIST**

### **Privacy and consent:**

- ☐ Determine if the data being gathered contains personally identifiable information (PII) and whether consent is required.
- ☐ When required, get informed consent from individuals and explicitly express the purpose and scope of data collection.

Evidence:

*None of the dataset we collected - Crime, Public Nuisance, BOA, Population, Personen met uitkering, Green Index, Heat Stress ([datasets](#)) - contains personally identifiable information, therefore consent was not required.*

### **Fairness and bias:**

- ☐ Identify and mitigate any data biases to ensure fairness in algorithm development and execution.
- ☐ Assess the algorithm's influence on various demographic groups and correct any potential biases or prejudice.

Evidence:

*During the creation of the Public Nuisance prediction model, the migration-related data was omitted to reduce the risk of potential racial bias.*

### **Legal Obligation:**

- ☐ Ensure that appropriate rules, regulations, and ethical norms, such as data protection and privacy legislation, are followed.
- ☐ Consult with legal specialists to understand the legal consequences of building and employing the prediction algorithm.

### **Transparency and comprehensibility:**

- ☐ Document the data sources, methods of collection, and any data pretreatment or cleaning activities that were used.
- ☐ Explain to stakeholders the algorithm's approach, variables, and decision-making processes in detail.

Evidence:

*The sources of all the data used have been documented in the [data\\_sources.md](#) file. The data pretreatment and cleaning activities can be found in the Group 10's GitHub repository in the [EDA folder](#).*

*The algorithm's approach, variables, decision-making and other related processes have been described in detail to the Product Owner and the team members during weekly Sprint reviews.*

**Data security**

- ☐ Encrypt sensitive data while it is being sent and stored.
- ☐ To safeguard confidentiality, establish standards for data exchange both inside the project team and with external stakeholders.

Evidence:

*The team has established a data management standard ensuring all the data used is stored, shared and exchanged only to the Group 10's GitHub repository, where access is limited only to the valid team members.*

**Risk assessment:**

- ☐ Conduct a thorough risk assessment to detect any negative repercussions or unforeseen outcomes of the algorithm.
- ☐ Implement risk-mitigation strategies, such as monitoring for algorithmic bias and assessing and updating the model on a regular basis.

Evidence:

*The risk assessment has been conducted and appropriate risk-mitigation strategies have been devised in the [Group 10's Trello Risk Assessment board](#).*

**Communication:**

- ☐ Communicate the algorithm's aim, advantages, and restrictions to stakeholders, including the municipality and the general public.
- ☐ Provide transparency about the algorithm's intended application and answer any concerns or inquiries from stakeholders.

Evidence:

*The algorithm's aim, advantages, restrictions and intended application have been presented to the Breda municipality representatives during the [Sounding board session](#) and to Product Owner weekly at the Sprint reviews.*

**Evaluation:**

- ☐ Monitor the algorithm's performance on a regular basis to guarantee its correctness, fairness, and efficiency in anticipating and distributing resources.
- ☐ Evaluate the algorithm's societal impact and make any required improvements to mitigate any unforeseen repercussions.

### **Applicable and crucial elements:**

**Algorithms:** In order to maintain ethical and transparent processes when dealing with data, it is essential to consider the legal and ethical implications of the project's use of a machine learning model/algorithm. Users, stakeholders, and the general public are able to comprehend an algorithm's behavior and decision-making process when they are given explanations of it. Algorithm explanations could prove useful for deciphering the reasoning, spotting mistakes and tasks, recognizing biases and discrimination, fostering user participation and trust, and encouraging responsibility.

**Source:** The algorithm we developed throughout this project is highly dependent on data to function at its highest potential. Meaning that the quality, reliability, and validity of the information that is gathered are all directly impacted by the source of the data, making it crucial to consider while gathering data. The reliability of the source, which ensures that the data is accurate, current, and relevant, determines how trustworthy the information is. The methodology, data-gathering methods, and biases used by various sources may also differ, which can affect how the data is interpreted and potentially introduce biases. The integrity and applicability of the data collected for the analysis and decision-making processes could be ensured by assessing the source.

**Visualizations:** Visualizations are essential in machine learning models as they enable a better understanding of the data. Patterns, trends, and linkages become more obvious and intuitive when complicated information is presented in a graphic style. Visuals are also vital to our project, especially when communicating the predictions made by our machine learning model since this will make it easier for stakeholders to read and comprehend the results.

**Responsibilities:** Responsibility is of utmost importance when handling data for this project. It involves ethical considerations, legal compliance, and ensuring the privacy and security of the data. Responsible data management guarantees that information is utilized fairly and objectively while taking into account any possible effects on particular individuals or communities. In order to preserve trust, preventing unexpected effects, and sustaining professional standards throughout the data handling process, requires adherence to ethical principles. Therefore, including this item in our checklist highlights our commitment to ethical data management procedures and strengthens the moral framework of our work.

### **Communication:**

For a variety of reasons, communication is essential in teamwork. It primarily enables effective information, idea, and feedback sharing among team members. Team members may minimize misunderstandings and potential disputes by speaking freely and explicitly about their obligations, objectives, and expectations. To reinforce this, all of us met each weekday at 11:30 to discuss our respective days' work, our plans for the day, and any obstacles that would prevent us from finishing our tasks. This approach has received unanimous support and has shown to be highly beneficial since it guarantees that everyone is on the same page, prevents effort duplication, and fosters effective teamwork. Communication is vital for ensuring compliance with legal regulations and addressing ethical considerations in a data project.

**Transparency:**

As we seek to estimate the number of BOAS required to avoid public nuisance, it is undoubtedly feasible that the project may arouse public concern or fury. Several variables, including bias and discrimination, privacy concerns, openness and accountability, and community involvement, can affect how the public feels and behaves. Therefore, it is essential to have an open dialogue with stakeholders, uphold openness, and address privacy and security issues in order to reduce these potential hazards. By putting in place the proper protections, such as doing frequent impact assessments, getting feedback from the public, and abiding by ethical standards, one may address possible issues and lessen the possibility of public uproar.

**Privacy:** Since the goal of this project is to estimate the number of BOAS required to avoid public disturbance, we don't need personal data; yet, we manage a lot of sensitive data. We used non-public SQL server data regarding various types of annoyance in different areas of the city to build our model. As a result, privacy is a key notion in our project since we must treat the data responsibly as it is critical in safeguarding individual rights, trust, and the ethical use of non-public data.

**Bias:**

Although we do not utilize personal data in this research, bias can manifest itself in a variety of ways, such as selection bias, measurement bias, or contextual bias.

Addressing and mitigating these risks is critical to ensuring that the initiative supports justice, inclusion, and nondiscrimination. Controlling for these factors and reducing bias can be accomplished by employing proper statistical tools and procedures. Transparency in the analytic process, such as recording all data cleaning and preparation stages, can also assist in identifying and correcting any unintended biases. Seeking input and comments from colleagues or subject matter experts on a regular basis can give significant insights and aid in the identification of potential biases.

### **Non-crucial and non-applicable elements:**

#### **Access:**

Since the data that is utilized is already publicly available, access is not a critical ethical and legal component of our initiative because it is available to anyone without constraints.

#### **Anonymization:**

Anonymization is irrelevant to our research since the data utilized already had been anonymized and did not contain any sensitive or personal information. Anonymization is not a crucial factor since the project merely focuses on examining patterns or trends at a larger level with no aim of identifying individuals.

#### **Sharing, reusing and repurposing:**

The majority of the data we used for this project is publicly available and may be reused by anyone who desires to. However, strict confidentiality obligations or regulatory restrictions hinder the sharing or reuse of BOA's data. The project's emphasis is constrained to certain aims or particular datasets, making sharing or reuse unnecessary. Furthermore, if the data is highly specialized or private, there may be not much advantage in sharing or reusing it beyond the scope of the project. As a result, the element of sharing, reusing, and repurposing may not be relevant or necessary to the successful completion of our unique data project.