Data Preparation & Ethical Data Handling

AI Masters Capstone Project - Presentation 2

Jonathan Agustin

November 2024

What We'll Cover Today

- Why ethical data handling is essential, not optional
- Automated preprocessing: cleaning, transforming, validating data
- Ensuring data quality and reliability with systematic checks
- Respecting privacy, complying with regulations (e.g., GDPR)
- Detecting and mitigating bias to build fair, trustworthy models

By the end, you'll have concrete practices to ensure data is a strength, not a liability.

Why Ethical Data Handling Matters

- Data represents real people, not just abstract entries
- Ethical handling protects privacy, dignity, and fairness
- Avoiding biases and breaches maintains trust and credibility

Ethical data stewardship is the foundation of user trust and sustainable ML solutions.

Automated Data Preprocessing

- Handle missing values, outliers, and inconsistencies automatically
- Uniformly transform data, standardizing features and formats
- Reduce human error and foster reproducible, transparent data workflows

Automation shifts focus from manual cleanup to building reliable, ethical pipelines.

Practical Preprocessing Techniques

- Missing data: Impute (mean, median), flag or remove strategically
- Outliers: Apply robust statistical checks or domain-specific thresholds
- Feature standardization: Convert formats, ensure consistent units, encode categories

Well-defined preprocessing rules build confidence and clarity into data pipelines.

Data Quality Assurance

- Implement validation checks: schema conformity, range checks, uniqueness tests
- Detect anomalies early: unusual patterns, distribution shifts, missing critical features
- Maintain logs and reports to track data quality trends over time

Robust validation turns raw data into a trusted resource for responsible AI.

Privacy Protection and Compliance

- Implement anonymization or pseudonymization for sensitive identifiers
- Follow GDPR and other regulations: user consent, data minimization, right to erasure
- Restrict access, log data handling actions, and maintain thorough documentation

Privacy isn't just legal compliance—it's respecting human rights and autonomy.

Detecting and Mitigating Bias

- Use fairness metrics: parity checks across demographic segments
- Monitor representation: ensure no group is disproportionately excluded
- Adjust preprocessing steps or sampling strategies to rebalance the data

Catching bias in the data stage prevents deploying models that perpetuate discrimination.

A Holistic Data Strategy

- Combine automation, validation, privacy measures, and bias checks into one pipeline
- Document every step for transparency and accountability
- Set the stage for fair, impactful ML models built on trustworthy data

A robust, ethical data pipeline empowers us to build AI that genuinely benefits everyone.

Next Steps

- Next Presentation: Automating Model Training & Ensuring Fairness
- Building on our ethical data pipeline to create responsible models

From well-prepared data to ethically trained models—the journey continues.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, becau LaTeX now knows how many pages to expect for this document.