**Module1 Critical Thinking**

**Ethical Leadership in Software Development**

Student Name: Sankara Narayana Katabathina

Colorado State University Global

Course Code: **CSC502**

Instructor: Dr [Shaher Daoud](https://csuglobal.instructure.com/courses/110368/users/83390)

Due Date: 06/15/2025

# Introduction

The process of preparing data known as data preprocessing plays a critical role in ensuring analytical accuracy and reliability. Preprocessing transforms raw data into a clean, usable format through tasks such as normalization, imputation, filtering, and encoding. However, the increasing influence of data in areas like healthcare, finance, and policing raises important ethical considerations. This paper explores key ethical questions that must be addressed before and during data preprocessing, including fairness, privacy, transparency, and consent. It also examines two real-world examples of ethical conflicts that arose during data preprocessing activities, illustrating the tangible impact of these issues.

# Key Ethical Questions in Data Preprocessing

* **Is the data collected with informed consent?**

One of the most fundamental ethical principles in data handling is respect for individuals autonomy. Consent must be freely given, specific, informed, and unambiguous. Preprocessing data without verifying the validity of consent especially in secondary use cases raises ethical red flags.

* **Are there risks of bias or discrimination?**

Biased datasets often reflect structural inequalities. Preprocessing that fails to identify and mitigate such bias can perpetuate discrimination. Techniques like oversampling or under sampling to balance classes should be applied cautiously to avoid distorting underlying truths.

* **Is personal information properly anonymized or protected?**

Data preprocessing often involves handling sensitive personally identifiable information (PII). Failing to de-identify this data appropriately not only raises legal risks (e.g., under GDPR or HIPAA) but also undermines public trust.

* **Are preprocessing choices transparent and documented?**

Many preprocessing decisions are invisible to downstream users of the data. Ethical practice demands that such transformations be fully documented and justifiable, especially when models will influence real world decisions.

* **Are stakeholders aware of the potential downstream consequences?**

Preprocessing affects outcomes in AI and ML systems. Ethical data handling involves evaluating how changes at the preprocessing stage might influence model bias, fairness, and interpretability.

# Ethical Conflict Example 1 – COMPAS Risk Assessment Tool

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool was developed to predict recidivism risks. An investigative report by ProPublica (2016) found that the preprocessing steps used to engineer features from criminal history disproportionately flagged Black defendants as higher risk compared to white defendants with similar profiles.

**Ethical Violations:**

Biased feature engineering: The tool included variables like arrest history, which reflected systemic policing biases rather than true criminal behavior.

**Lack of transparency:**

The preprocessing steps were proprietary and not subject to external audit, making it impossible to evaluate the fairness of the transformations.

**Takeaway:**

Without careful preprocessing that accounts for structural bias, even legally compliant systems can reinforce systemic injustice.

# Ethical Conflict Example 2 – Google Flu Trends

Google Flu Trends aimed to predict flu outbreaks based on search query data. However, it overestimated flu activity due to flawed preprocessing—such as improper handling of seasonal search patterns and changes in user behavior influenced by media.

**Ethical Concerns:**

Lack of data validation: Google relied on volume-based proxies without cross-validating against clinical data.

**Reidentification risk:**

Although data was aggregated, search logs are inherently personal, and insufficient anonymization posed risks.

**Takeaway:**

Preprocessing without considering data context and external validation can result in misleading outputs that harm public health decisions.

# Conclusion and Recommendations

Data preprocessing is not merely a technical step; it is a value-laden process that can significantly influence the ethical trajectory of any data-driven system. From bias in feature selection to inadequate de-identification and poor documentation, each choice made during preprocessing carries ethical weight.

**Recommendations:**

Incorporate ethics checklists into preprocessing workflows.

Conduct bias audits and apply fairness-aware data preprocessing techniques.

Provide complete preprocessing documentation and justification.

Involve ethicists and domain experts during data preparation.

Ethical preprocessing is not only about protecting organizations from regulatory penalties it is also about safeguarding the trust, dignity, and rights of the people represented in the data.

# References

Barocas, S., Hardt, M., & Narayanan, A. **Fairness and machine learning**: Limitations and opportunities. <https://fairmlbook.org/>

**Machine Bias**: There’s software used across the country to predict future criminals. And it’s biased against blacks.

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>