

# James Kocak

(980) 800-6120 | [LinkedIn](#) | [jamkocak88@gmail.com](mailto:jamkocak88@gmail.com) | [GitHub](#) | [Portfolio](#)

## SUMMARY

Passionate Computer Science professional specializing in Data Science and Data Engineering, experienced in building data pipelines, statistical analysis, and machine learning models. Skilled in Python, SQL, Databricks, Apache Spark, and modern data frameworks.

## EDUCATION

**North Carolina State University** | Raleigh, NC

May 2025

Bachelor of Science in Computer Science

GPA: 3.6/4.0

**Relevant Courses:** Data Structures & Algorithms, Database Management Systems, Software Engineering, Operating Systems, C/Software Tools, Intro to Artificial Intelligence, Introduction to Responsible Machine Learning, Trustworthy and Efficient Deep Learning

## SKILLS

**Languages:** Java, Python, C, C++, C#, SQL, JavaScript, HTML, CSS

**Frameworks & Libraries:** React, Spring, Django, REST, Hibernate, JUnit, NumPy, Pandas, matplotlib, PyTorch, TensorFlow, STL (C++)

**Databases & Data Technologies:** MySQL, MongoDB, PostgreSQL, SimpleDB, Apache Spark, Apache Maven

**Tools & Platforms:** Databricks, Git/GitHub, Jenkins, GitHub Actions, Docker, Makefile, CMake, Eclipse, Visual Studio, VSCode, Linux

## CERTIFICATIONS & EXTRACURRICULARS

**Databricks Certified Data Engineer Associate** | Databricks

March 2025

**Microsoft Azure AI Essentials Professional Certificate** | Microsoft

In Progress

**NCSU Hackathon Participant 2025** | NC State University

February 2025

## PROJECTS

**Machine Learning Pipeline Exploration** | *Introduction to Responsible Machine Learning*

January 2025 - April 2025

- Created detailed dataset descriptions, identifying 15+ key features, data types, and potential biases; clearly defined a hypothetical ML use-case by specifying the problem statement, stakeholders involved, and potential ethical implications related to fairness
- Built a trustworthy ML pipeline encompassing data preprocessing (5+ preprocessing techniques), targeted feature selection methods, and model comparisons (including logistic regression and decision trees), while systematically evaluating performance
- Integrated 2+ trustworthiness interventions (including fairness metrics and interpretability tools) within the ML pipeline, systematically addressing potential biases, transparency concerns, and robustness through comprehensive evaluation techniques

**Data Product Catalog** | *Statistical Analysis System (SAS) Senior Design Project*

January 2025 - April 2025

- Designed and implemented a scalable data ingestion pipeline capable of processing and managing metadata files with thousands of records, ensuring efficient storage and rapid access in a structured PostgreSQL database
- Developed an identification algorithm capable of analyzing millions of data records to suggest optimal matches for data product blueprints, achieving execution times consistently under 1 minute
- Implemented full CRUD functionality and metadata ingestion for data product blueprints and instances, supporting seamless management of 10,000+ data assets

**Synthea Data Generation** | *Personal Project*

March 2025

- Engineered synthetic healthcare datasets using Synthea API and Java to generate patient records in CSV/JSON
- Simulated realistic healthcare scenarios to improve ML evaluation and reduce false positives
- Optimized data workflows by orchestrating ingestion, validating pipelines, and designing privacy-friendly simulations

## WORK EXPERIENCE

**Target** | Front of Store Attendant

December 2020 - December 2022

- Leveraged analytical and problem-solving skills in a fast-paced environment, significantly streamlining workflow efficiency