Sanzhar Kuandyk

Junior Data Scientist • Machine Learning • Python
Astana, Kazakhstan | sanzharkuandyk21@gmail.com | github.com/SeanKu03 | linkedin.com/in/sanzhar-kuandyk03

Profile

MSc student in Robotics Engineering at Nazarbayev University who combines machine learning with a strong mathematics and optimization background and real-time robotics software (MPC, ROS2). Built a local RAG question-answering system with measured retrieval and answer quality, and implemented core ML models from scratch (NumPy/SciPy, PyTorch). Seeking a junior Data Scientist / ML role.

Education

MSc in Robotics Engineering 2026 – present
Nazarbayev University *Astana, Kazakhstan*
BSc in Robotics Engineering August 2022 – June 2026
Nazarbayev University *Astana, Kazakhstan*

- Coursework: Machine Learning with Applications, Data Structures & Algorithms, Probability & Statistics, Control & Modeling
- Capstone: Human-Aware Robot Navigation using Model Predictive Control (MPC)
- Certificate: Backend Developer in Python, Data Science Academy (2024)

Projects

Local RAG Chat over Kazakhstan Legal Documents [GitHub] November 2026
Personal project • Python, sentence-transformers, BM25, Ollama (gemma3:4b), Streamlit

- Built a local retrieval-augmented chat over 5 military-service laws and orders (318 pages): PDF cleaning, 400-character chunks, **hybrid retrieval** (multilingual-E5 embeddings + BM25, reciprocal rank fusion), a local LLM that answers with file and page sources, Streamlit interface.
- Compared 10 retrieval configurations on 22 hand-written questions: page-level hit@5 (±1 page) rose from **0.59** (MiniLM baseline) to **1.00**, MRR from 0.45 to **0.92**.
- Graded every answer by hand against the source text: 11 of 22 fully correct, 7 partially correct, 4 wrong; 4 of 5 out-of-corpus questions correctly refused. Error analysis separated retrieval failures (4) from generation failures (7).

From-Scratch ML Models and Loan Default Prediction [GitHub] October 2025
ROBT 407: Machine Learning with Applications (team project)

- Implemented linear and logistic regression and linear SVM (primal and dual) from scratch in NumPy/SciPy and checked them against scikit-learn; compared regularized softmax regression with linear and RBF SVMs on MNIST using 5-fold cross-validation (best: RBF SVM, **92.3%** accuracy).
- Built a loan-default pipeline on 255K Kaggle records (imputation, encoding, GridSearchCV with stratified CV, class weighting) for an imbalanced target (11.6% defaults), selecting models by F1.

Titanic Survival Prediction (ROBT 407, team): PyTorch MLPs (dropout, batch norm, early stopping) vs. a logistic regression baseline, 81.6% vs. 79.9% validation accuracy [GitHub] November 2025

Research Experience

Research Assistant / Capstone Project 2025 – 2026
Nazarbayev University • Supervisor: Prof. Matteo Rubagotti

- Developed MPC-based motion planning for a mobile robot navigating among humans: cost functions, constraints, real-time constrained optimization.
- Ran simulation experiments in MATLAB/Simulink and ROS2; evaluated safety and performance metrics across scenarios and tuned controller parameters.
- Analyzed robot kinematics and dynamics data (preprocessing, visualization, quantitative evaluation) in Python and MATLAB.

Technical Skills

ML & Data: Python (pandas, NumPy, scikit-learn, PyTorch, Matplotlib), SQL (basic), Jupyter; regression, classification, SVM, regularization, cross-validation; metrics (F1, ROC-AUC, MRR)
LLM / RAG: embeddings (sentence-transformers), BM25, hybrid retrieval, Ollama, Streamlit, retrieval and answer evaluation
Mathematics: linear algebra, probability & statistics, optimization
Programming & Tools: Python, MATLAB, C++/Java (basic); Git/GitHub, Linux, ROS2, Simulink
Languages: English (C1), Russian (fluent), Kazakh (fluent)