

Kush Patel

Chicago, IL, USA | +1-872-288-5249 | patel.h.kush@gmail.com | in/kush-patel-2416 | [Github](#) | [Portfolio](#)

SUMMARY

Data Science graduate student at Illinois Institute of Technology specializing in predictive modeling and NLP. Experienced in building end-to-end machine learning pipelines in Python using scikit-learn and TensorFlow, with a project portfolio spanning LLM-based RAG clinical chatbots to ensemble models for biological age prediction. Focused on engineering interpretable, robust algorithms and developing automated workflows to deliver reliable, scalable technical solutions

SKILLS

- **Programming Languages:** Python, R, SQL, C, JavaScript
- **Data Analysis Tools/ Libraries:** Pandas, Numpy, Scikit Learn, Matplotlib, Excel, Tableau, Power Bi, Seaborn, MYS, PostgreSQL
- **Big Data Technologies:** Hadoop, Spark, AWS
- **Machine Learning/AI Libraries/Frameworks:** NLP, PyTorch, TensorFlow, Keras, FastAPI, LLM, RAG Agents, Google AI Studio, Pinecone

EXPERIENCE

Internshipstudio | *Machine Learning Intern*

May 2023 - Jun 2023

- Engineered predictive regression models (Random Forest, Keras/TensorFlow) to forecast digital asset monetization based on audience interaction signals and content metadata, outperforming the baseline decision-tree model by 15% in Mean Absolute Error (MAE) to support targeted promotional strategies.
- Architected a data preprocessing pipeline in Python using Pandas, NumPy, and Spark to handle corrupted data flags, encode complex categorical taxonomies, and filter statistical anomalies in engagement metrics, which streamlined data preparation and enabled downstream models to train on clean data without manual intervention
- Developed interpretability frameworks using Random Forest feature importance to demystify complex predictive outputs, identifying the primary drivers of user engagement and translating model behavior into actionable content strategies for stakeholders.

Traninity | *Data Analyst Intern*

Nov 2022 - Dec 2022

- Engineered data sanitization workflows using SQL and Excel on a 300,000+ row loan portfolio, eliminating 41 high-null features and imputing missing financial data to establish a reliable baseline for risk modeling.
- Conducted segmented bivariate analysis and designed KPI-driven Power BI dashboards for a highly imbalanced credit dataset (92% non-default vs. 8% default), visualizing demographic distributions to uncover hidden risk correlations in income and employment.
- Presented a data-driven risk mitigation strategy to business stakeholders, specifically highlighting the Transport sector's 16% default rate, to propose dynamic interest rate adjustments and optimize the underwriting process.

PROJECT

CareBot: Clinical-Support Medical Chatbot | <https://github.com/kush-pixel/CareBot>

May 2025 - Nov 2025

- Architected a context-aware conversational AI for preliminary symptom analysis, integrating LangChain and Gemini LLMs with Pinecone vector databases to perform high-speed semantic search across chunks of clinical text from the Gale Encyclopedia of Medicine.
- Engineered a Retrieval-Augmented Generation (RAG) pipeline to parse and embed complex medical guidelines utilizing Qwen3-Embedding-0.6B (1024-dimensional vectors), anchoring the LLM's responses in verified literature to actively prevent hallucination.
- Designed an automated triage-support workflow that synthesizes patient queries with retrieved protocols, applying recursive character text splitting (800-character chunks, 80-character overlap) to maximize context retention and reduce manual document review overhead.

Diabetes Prediction | <https://github.com/kush-pixel/CSP-571-Summer>

Jun 2025 - Aug 2025

- Engineered a robust data preprocessing and feature selection pipeline for a 100,000-record clinical dataset, applying Lasso regularization and one-hot encoding to isolate critical physiological predictors (e.g., HbA1c, BMI) while systematically mitigating class imbalance.
- Developed and hyperparameter-tuned an ensemble of classification models (XGBoost, Random Forest, Logistic Regression) utilizing 5-fold cross-validation, with the optimized XGBoost model achieving an industry-leading 0.97 AUC for predicting diabetes onset.
- Conducted comprehensive feature importance analysis to bridge the gap between advanced predictive power and clinical interpretability, translating XGBoost outputs and Logistic Regression coefficients into actionable risk assessments for healthcare stakeholders

Biological Age Prediction | <https://github.com/kush-pixel/Biological-age-prediction>

Mar 2025 - May 2025

- Developed an end-to-end, reproducible ML pipeline to estimate chronological age from DNA methylation (DNAm) arrays, often used as a proxy for biological aging.
- Implemented interpretable baseline models (Ridge, PCR, PLS, Random Forest) and a light stacked ensemble for improved accuracy and calibration.
- Built robust data alignment and validation workflows to ensure cross-study generalization on an external cohort (GSE157131).

Auto-Grading Subjective Test Platform | <https://github.com/kush-pixel/Automatic-answer-grading-using-ML>

Feb 2021 - Apr 2021

- Engineered a full-stack automatic subjective grading platform (DashAsk) using Django, integrating role-based authentication (Admin, Teacher, Student) and interactive assessment dashboards to streamline online examination workflows.
- Architected an advanced Natural Language Processing pipeline leveraging the BERT model and Sentence Transformers to generate dense semantic embeddings, utilizing cosine similarity to autonomously evaluate unstructured student responses against reference answers.
- Optimized the semantic comparison engine to achieve a 0.87 F1 score, outperforming traditional ML baselines by 10%, while designing a 'human-in-the-loop' algorithmic override feature that empowered educators to maintain final assessment authority.

EDUCATION

Illinois Institute of Technology

Master of Applied Science, Data Science (GPA: 3.77)

- **Achievements:** Event Coordinator in Indian Student Association, Academic Resource Center Math Tutor

May 2026

Chicago, IL

University of Mumbai

B.Tech, Computer Science and Engineering (GPA: 8.6)

May 2023

India

Publications

- Automatic Subjective Answer Grading Software Using Machine Learning (<https://ieeexplore.ieee.org/document/10125786>) .2023 *Built an ML + NLP-based subjective answer grading system achieving 85% grading accuracy by computing semantic similarity between student and ideal responses. Developed a Django web application supporting 100+ concurrent users with <2s grading latency, enabling near real-time evaluation during online exams. Reduced manual grading effort by 60–70% using a human-in-the-loop workflow, allowing admins to review and override automated scores.*