John Baik

Frederick, MD jbaikbmail@gmail.com ☎ 240 893 5379 in john-baik-umd jbaik414

Education

University of Maryland *Expected: May 2026*

BS in Computer Science and Statistics

- GPA: 3.51/4.0
- **Coursework:** Intro to Machine Learning, Intro to Data Science Algorithms, Intro to Probability Theory, Linear Algebra, Applied Probability and Statistics, Advanced Data Structures

Experience

Research Intern *Bethesda, MD*

National Cancer Institute *June 2024 – Aug 2024*

- **Designed and developed a neural network using PyTorch to classify pancreatic cell types (alpha, beta, ductal, acinar), using DNA transcription sequence data from genomes.**
- Engineered and preprocessed a large-scale biological dataset using Pandas and NumPy, implementing data normalization, feature extraction, and dimensionality reduction to improve model performance.
- Trained, validated, and optimized a custom neural network, incorporating tensor operations, loss function tuning, and scikit-learn metrics for a diverse performance evaluation to account for multiple conditions.
- Collaborated with cancer genomics researchers to translate biological questions into machine learning problems

Projects

Phishing Detector *Phishing-Classifier*

- Developed a **classification model** to accurately detect phishing emails using real-world datasets and supervised learning techniques
- **Achieved 98% accuracy, 99% precision, and 99% recall**, displaying high accuracy in distinguishing phishing scams from legitimate emails
- Utilized tools including Jupyter Notebook, Pandas, NumPy, scikit-learn, and Matplotlib for data preprocessing, model training, evaluation, and visualization

Diabetes Classifier and EDA Project *Diabetes-classifier*

- **Cleaned and standardized a patient biomarker dataset** by handling missing values, encoding categorical variables, and scaling features.
- **Ingested and managed the dataset using Snowflake**, uploading cleaned CSV files into cloud tables for structured querying and reproducible analysis.
- Performed comprehensive **EDA using Pandas, Seaborn, and Matplotlib** to uncover patterns, skewness, and correlations—particularly between glucose and hemoglobin levels using **t-SNE**.
- **Trained and compared multiple classification models** (Logistic Regression, Random Forest, SVM), with the best model(Logistic Regression) achieving **93.42% accuracy and 0.80 recall for diabetic classification**.

Housing Price Predictor *Housing Regression Model*

- Implemented a **Random Forest Regression model** to analyze and predict California housing prices, identifying the top 5 most influential features impacting housing price
- **Achieved an R^2 score of 0.8153 and an average cross-validation score of 0.8156**, showing strong model generalization and performance
- Utilized Jupyter Notebook, Pandas, NumPy, scikit-learn, and Matplotlib for data cleaning, model development, feature importance extraction, and results visualization

Technologies

Languages: C, Java, Python, SQL, JavaScript, OCaml, R

Libraries/Tools: Pandas, NumPy, Matplotlib, scikit-learn, Tensor , PyTorch, Microsoft Excel

Frameworks and Environments: React, PyTorch, TensorFlow, VS Code