**Iris Flower Species Prediction – Project Introduction**

**Course:** CISD 43 – Big Data (Spring 2025)

**Team:** Kendra Ling

**Overview:**

The goal of this final project is to apply big data analysis techniques to Ronald Fisher’s Iris Species dataset, which is a staple resource for machine learning dataset. This dataset was used to accurately predict flower species based on measurable attributes. Through exploratory data analysis (EDA), machine learning, big data tools, and NoSQL database management, raw data is transformed to meaningful predictions.

**Objectives:**

* Conduct thorough **Exploratory Data Analysis (EDA)** to understand the dataset's structure, handle any missing values, transform variables, and create meaningful visualizations through easy-to-read charts and graphs.
* Use **machine learning algorithms** (**K-Nearest Neighbors (KNN)** and **K-Means Clustering**) to classify and cluster iris flower species.
* Use **RapidMiner** as a big data tool to visually build and execute ML workflows.
* Store and query the iris species data using **MongoDB**, demonstrating proficiency in NoSQL database operations.
* Present a clear and comprehensive summary of findings and insights derived from the analysis.

**Tools Used:**

* **Python (Jupyter Notebook)** – coding, data analysis, and model implementation
* **RapidMiner** – visual machine learning workflows
* **MongoDB** – document-based storage and aggregation queries
* **Pandas, Seaborn, Matplotlib** – EDA and data visualization