BIG DATA ANALYSIS WITH IBM CLOUD DATABASES

PROJECT OVERVIEW:

1. Studying Big Data: Big data means a huge amount of information. It includes lots and lots of different pieces of information, like temperature records, social media posts, or other large sets of data.
2. Using IBM Cloud Databases: IBM Cloud Databases is like a special tool that helps us organize and manage all this big data.
3. Finding Important Stuff: The data need to be completely checked out and finds only the important and useful bits.
4. Making a Plan: A step by step plan should be made to work on this project.
5. Setting Up Our Tools: IBM cloud databases should be set to work efficiently.
6. Analysing and showing the results: All the data needs to be analysed carefully and it should be presented in such a way that it is easy to be understandable.

DESIGN THINKING:

1. Data Selection: Our team have chosen the Artificial Intelligence tools 2023 dataset which is at the current trend.
2. Database setup: Our team have set up the dataset in IBM Cloud Databases for storing and managing dataset.
3. Data Exploration: Our team are in the middle of developing queries and scripts to explore the datasets, extract relevant information, and identify patterns.
4. Analysis Techniques: Our team are yet to apply appropriate analysis techniques, like statistical analysis or machine learning, to uncover insights.
5. Visualization: Our team will design a visualization chart to present the analysis results in an understandable and impactful manner.
6. Business Insights: Our team will interpret the analysis findings to derive valuable business intelligence and actionable recommendations.

# ABSTRACT

Big Data analysis has emerged as a crucial tool in contemporary data-driven decisionmaking processes, providing valuable insights from vast and diverse datasets. In this context, IBM databases stand out as a potent platform for managing and analysing large-scale datasets effectively. This abstract outlines the significance and benefits of utilizing IBM databases for Big Data analysis. we highlight the practical implications of employing IBM databases, showcasing their potential to drive informed decision-making and enhance organizational efficiency in the era of data abundance.

# DATASET

A good source of dataset for big data analysis with IBM cloud databases depends on various factors, including the specific research or analysis goals, the industry or domain of interest, the tools and technologies being used, and the desired insights. We have chosen **Artificial Intelligence Tools 2023** dataset.

# PREDICTIVE ANALYSIS

In big data analysis, Predictive analysis is an essential tasks that leverage advanced machine learning algorithms to extract valuable insights and detect abnormal patterns within massive datasets. In our project, we have used 3 Machine Learning algorithms such as

* Gradient Boosting Machines (GBM)
* Random Forest
* Deep Learning Models
* Support Vector Machines (SVM)
* Ensemble Methods

## Gradient Boosting Machines (GBM)

Gradient Boosting Machine (GBM) is a powerful and versatile machine learning algorithm widely used in Big Data analysis for various applications. Its strengths lie in its ability to handle large volumes of data, complex features, and provide accurate predictions. Some of the key uses of GBM in Big data analysis are

* Predictive Modelling and Regression
* Classification and Risk Assessment
* Natural Language Processing (NLP)

## Random Forest

Random Forest is a versatile and powerful ensemble learning algorithm that is widely used in Big Data analysis for a variety of applications. Its capability to handle large and complex datasets, as well as its ability to provide accurate predictions. Some of the key uses of

Random forest in Big data analysis are

* Classification and Predictive Modelling
* Multi-modal Data Fusion

## Deep Learning Models

Deep learning models are immensely valuable in Big Data analysis due to their capability to learn intricate patterns from large and complex datasets. They excel in processing vast amounts of data and extracting high-level features, making them highly suitable for various applications in Big Data analytics. Here are several key uses of deep learning models in Big Data analysis are

* Graph Analytics
* Recommendation Systems

## Support Vector Machines (SVM)

Support Vector Machine (SVM) is a powerful machine learning algorithm known for its effectiveness in both classification and regression tasks. In the context of Big Data analysis, SVM serves several critical purposes and offers various applications due to its ability to handle high-dimensional data and nonlinear relationships. Some of the key uses of SVM in

Big Data analysis are

* Classification
* Text and Document Classification • Regression Analysis.

## Ensemble Methods

Ensemble methods are powerful techniques in machine learning that combine multiple individual models to create a stronger and more accurate predictive model. In the context of Big Data analysis, ensemble methods play a crucial role in improving predictive performance, handling diverse data, and making effective use of the vast amount of information available. Some of the key uses of ensemble methods in Big Data analysis are

* Reduced Overfitting
* Handling Imbalanced Data
* Ensemble Clustering and Data Segmentation

## LIBRARIES USED FOR PREDICTIVE ANALYSIS

In Python, there are several popular libraries and frameworks commonly used for predictive analysis and machine learning. These libraries provide a wide range of tools and algorithms to build predictive models, perform data preprocessing, and evaluate model performance.

* Pandas
* NumPy
* Scikit-learn
* Matplotlib
* Seaborn

### Pandas

Pandas, a Python library, is a powerful tool widely used in Big Data analysis for efficiently handling, processing, and analysing large volumes of data. Key uses of Pandas in Big Data analysis are

* Data Loading and Reading
* Data Cleaning and Preprocessing
* Data Exploration and Analysis

### NumPy

NumPy (Numerical Python) is a fundamental Python library for numerical computing, providing support for handling multidimensional arrays, mathematical functions, and linear algebra operations. Key uses of NumPy in Big Data Analysis are

* Numerical and Mathematical Operations
* Linear Algebra
* Random Number Generation

### Scikit-learn

Scikit-learn, often referred to as sklearn, is a popular Python library for machine learning and data analysis. it can also play a significant role in Big Data analysis, especially when combined with distributed computing frameworks. The key uses of scikit-learn in Big Data analysis are

* Machine Learning Algorithms
* Data Preprocessing and Feature Engineering
* Model Evaluation and Hyperparameter Tuning

### Matplotlib

Matplotlib is a popular Python library used for creating static, animated, and interactive visualizations, making it a valuable tool in Big Data analysis. The key uses of

Matplotlib in Big Data analysis are

* Data Visualization
* Exploratory Data Analysis (EDA)
* Statistical Analysis and Comparison

### Seaborn

Seaborn is a Python data visualization library built on top of Matplotlib, providing an interface for creating informative and visually appealing statistical graphics. It can still be valuable in Big Data analysis when combined with appropriate data sampling or aggregation strategies. The key uses of Seaborn in Big Data analysis

* Pair Plots and Pair Grids
* Statistical Data Visualization
* Cluster Maps and Heatmaps

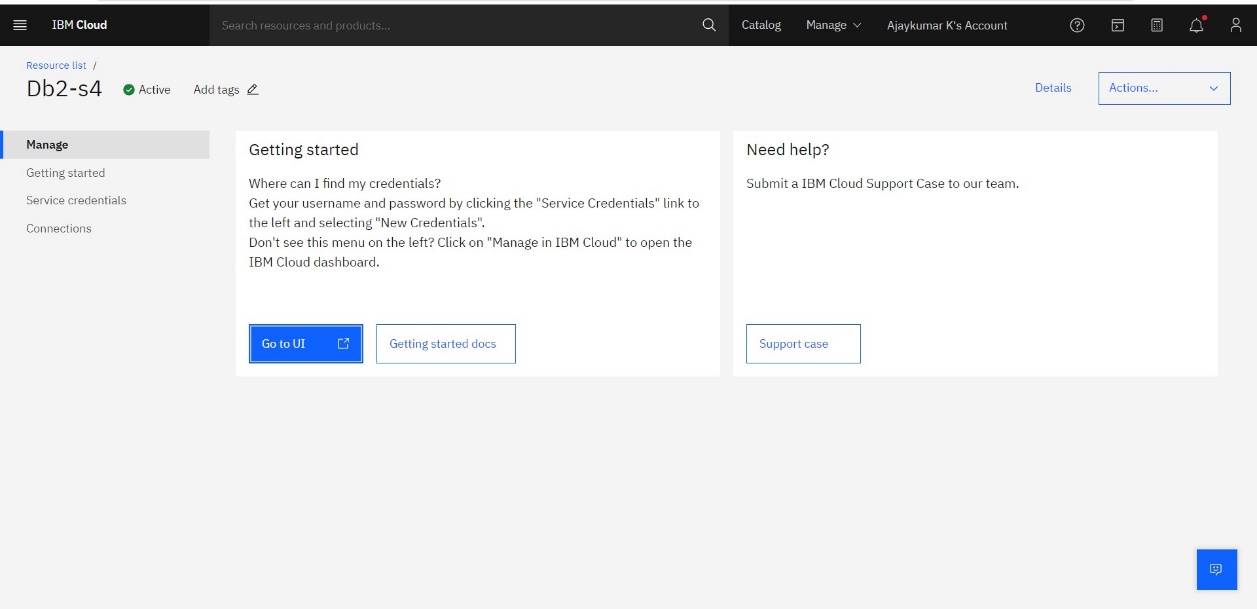
**IBM DB2**

IBM DB2 is designed for managing and storing large volumes of data efficiently and securely. Db2 is used in a variety of industries and applications for various purposes. common uses of IBM Db2:

* Relational Database Management
* Enterprise Database
* Data Warehousing
* Application Development
* Mainframe Data Storage

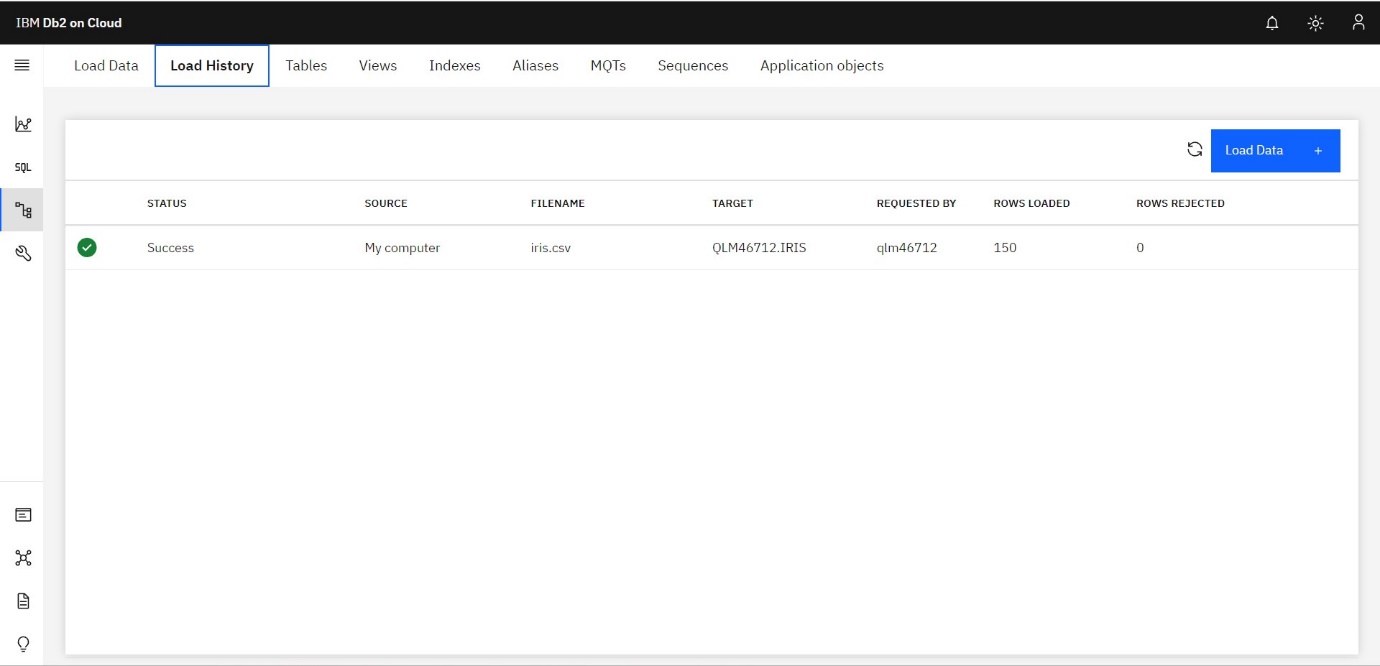
**DATABASE SETUP:**

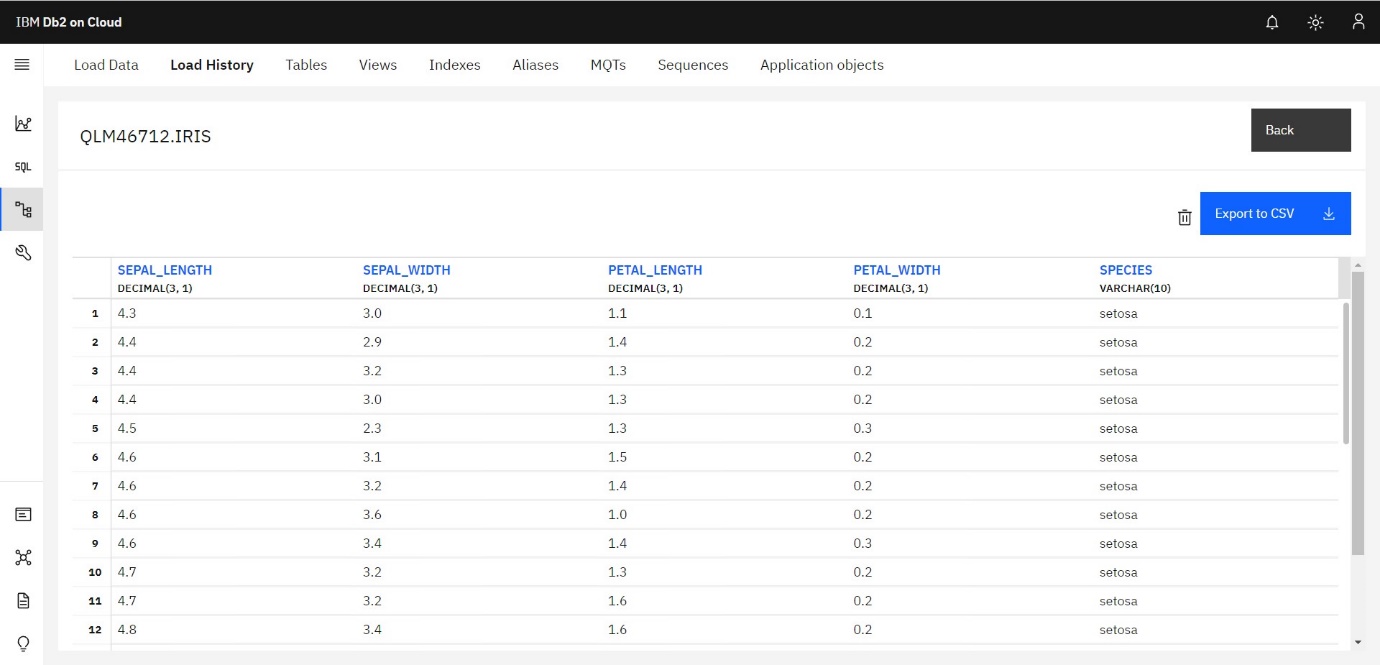
We have created an IBM Cloud account, and created the appropriate database service Db2, and had set up a database instance.



**LOADING THE DATASET:**

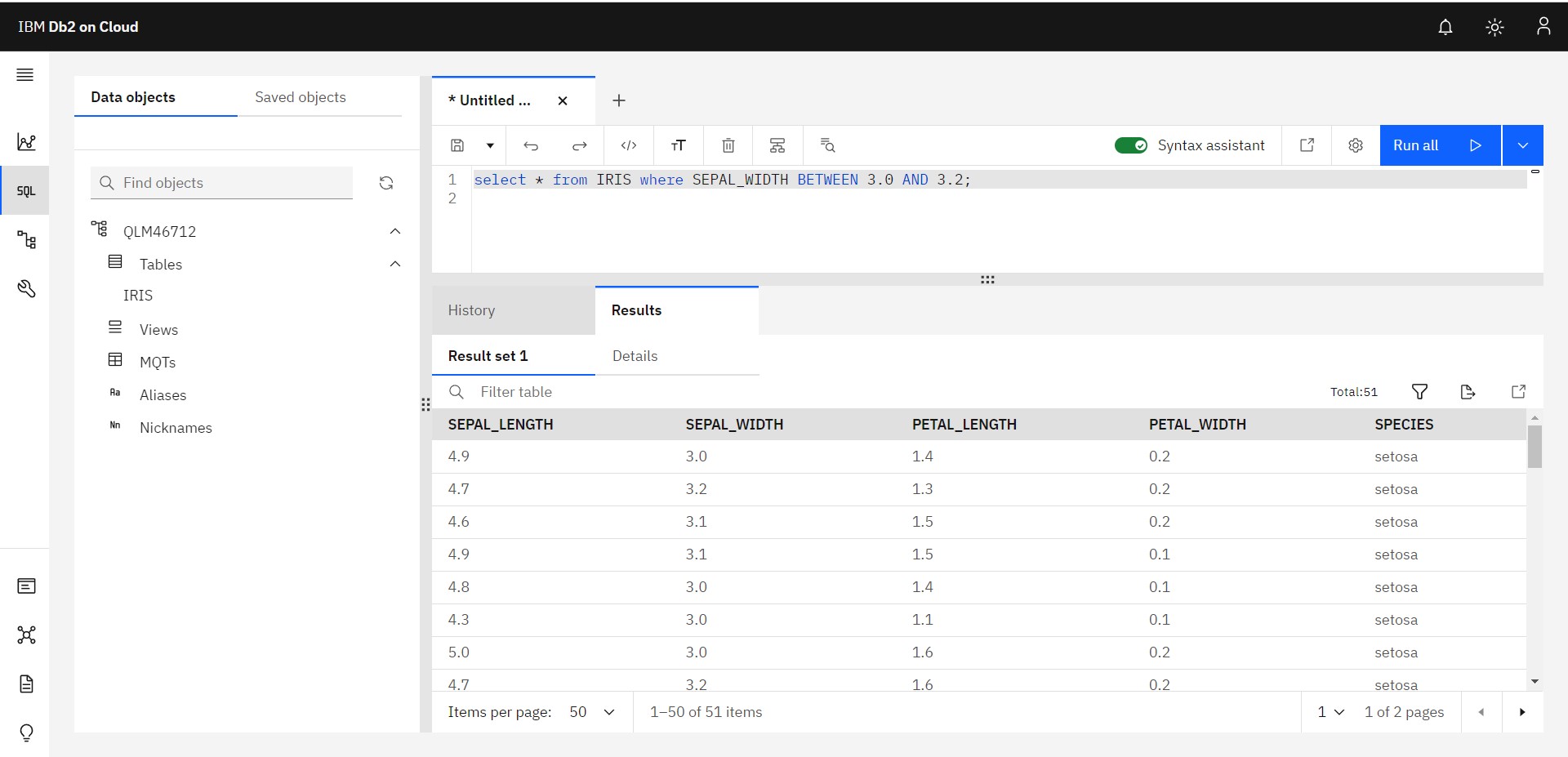
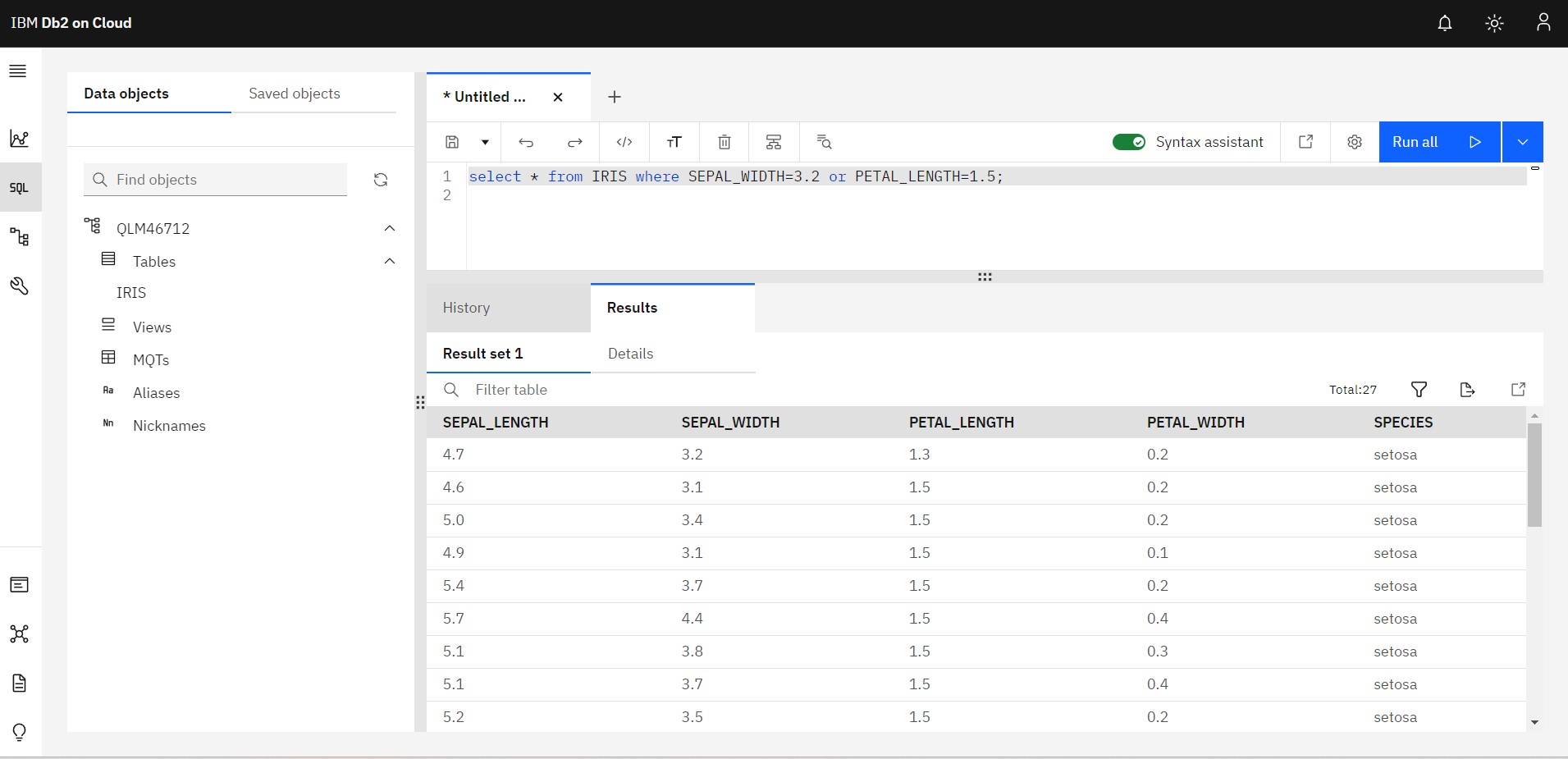
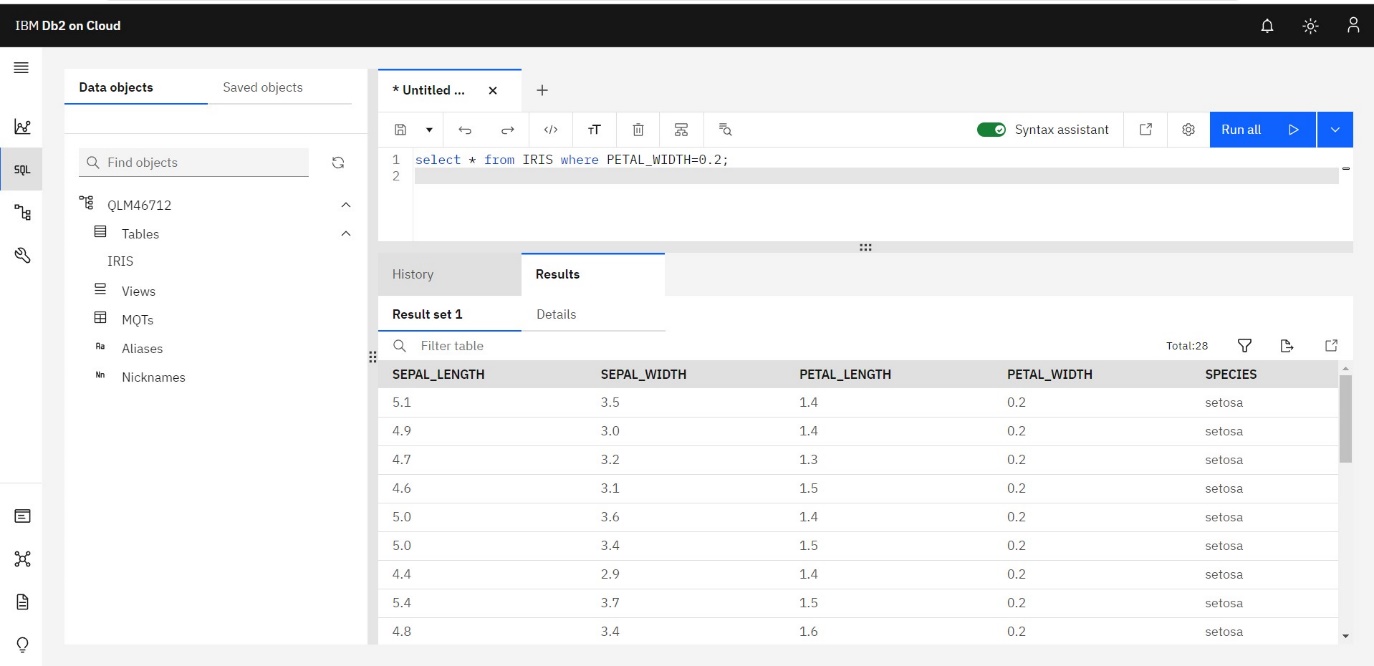
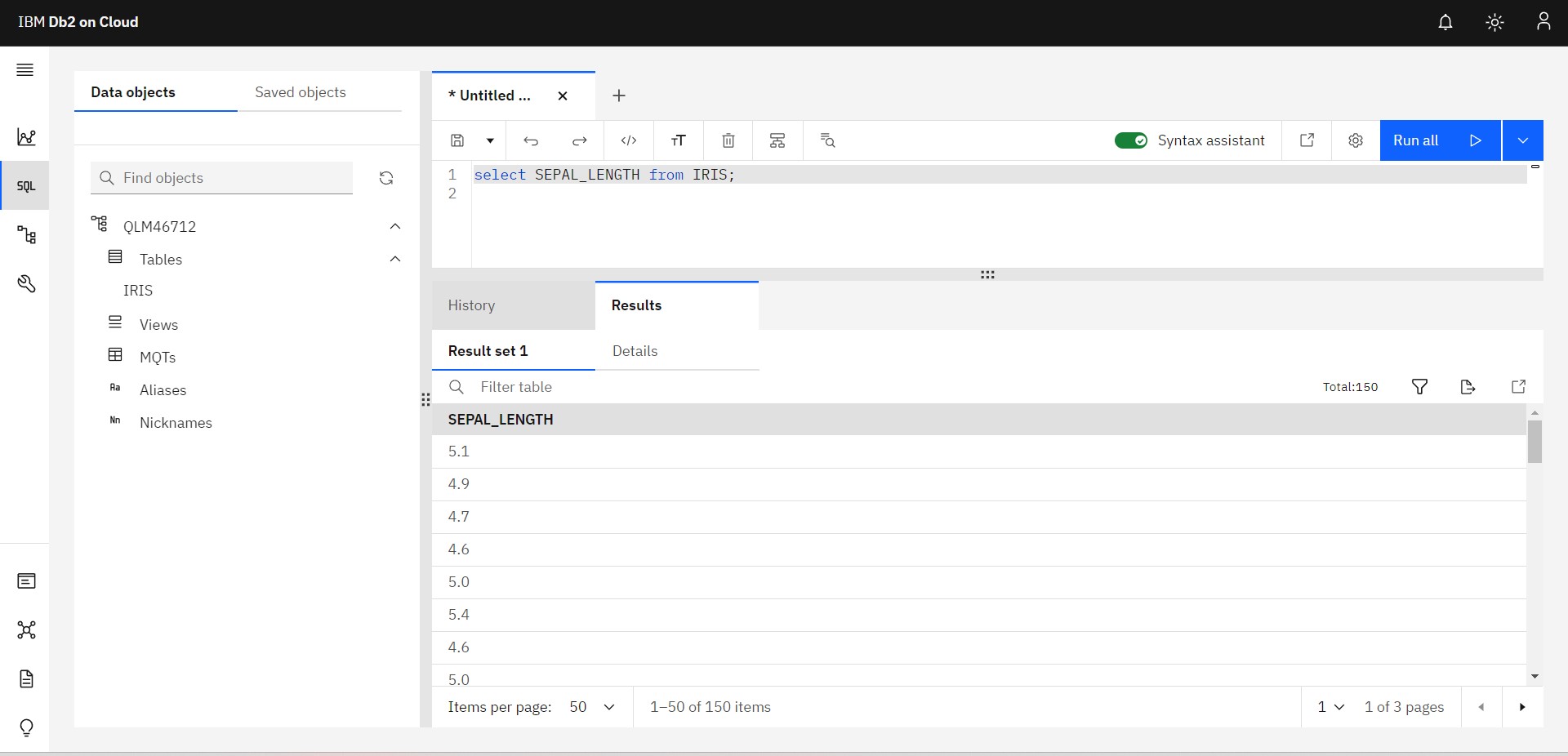
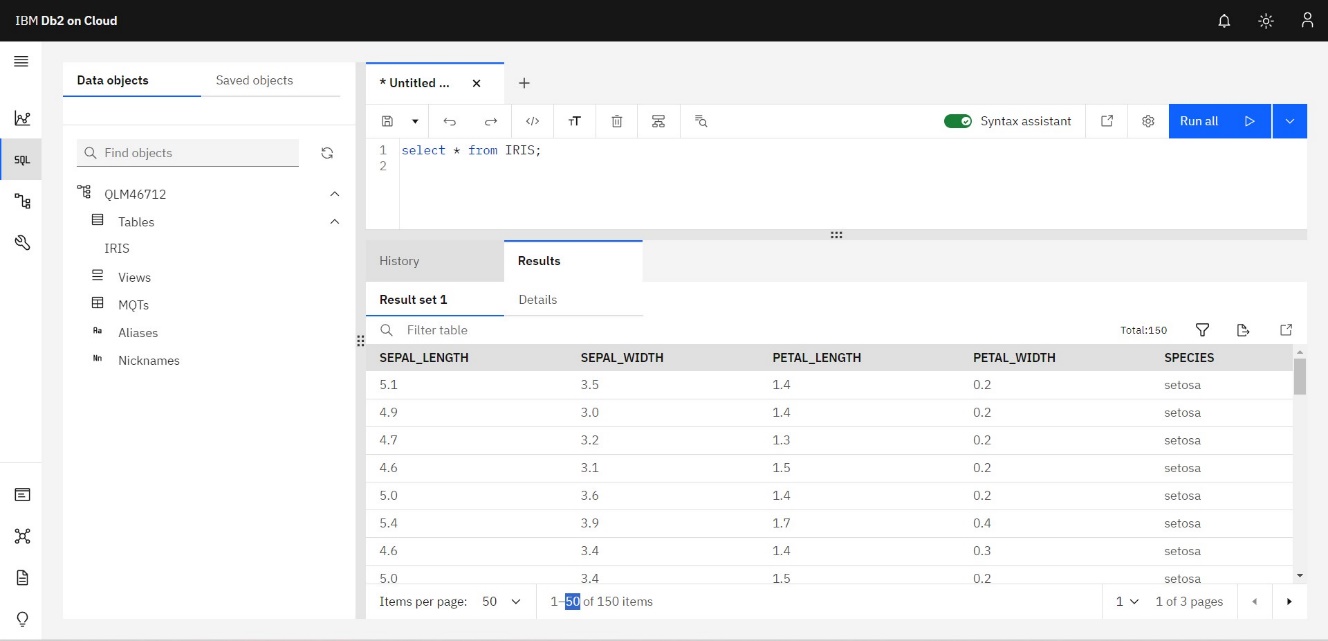
After creating the DB2 instance, we have created a table called AI in the database and loaded the dataset ‘all\_ai\_tool.csv’ in it.





**EXECUTION:**

After adding the dataset ‘languages.csv’ to the database, we have executed some basic SQL queries to check whether the dataset is connected to the database correctly.



**VISUALISATION WITH IBM WATSON STUDIO**

# IBM WATSON STUDIO

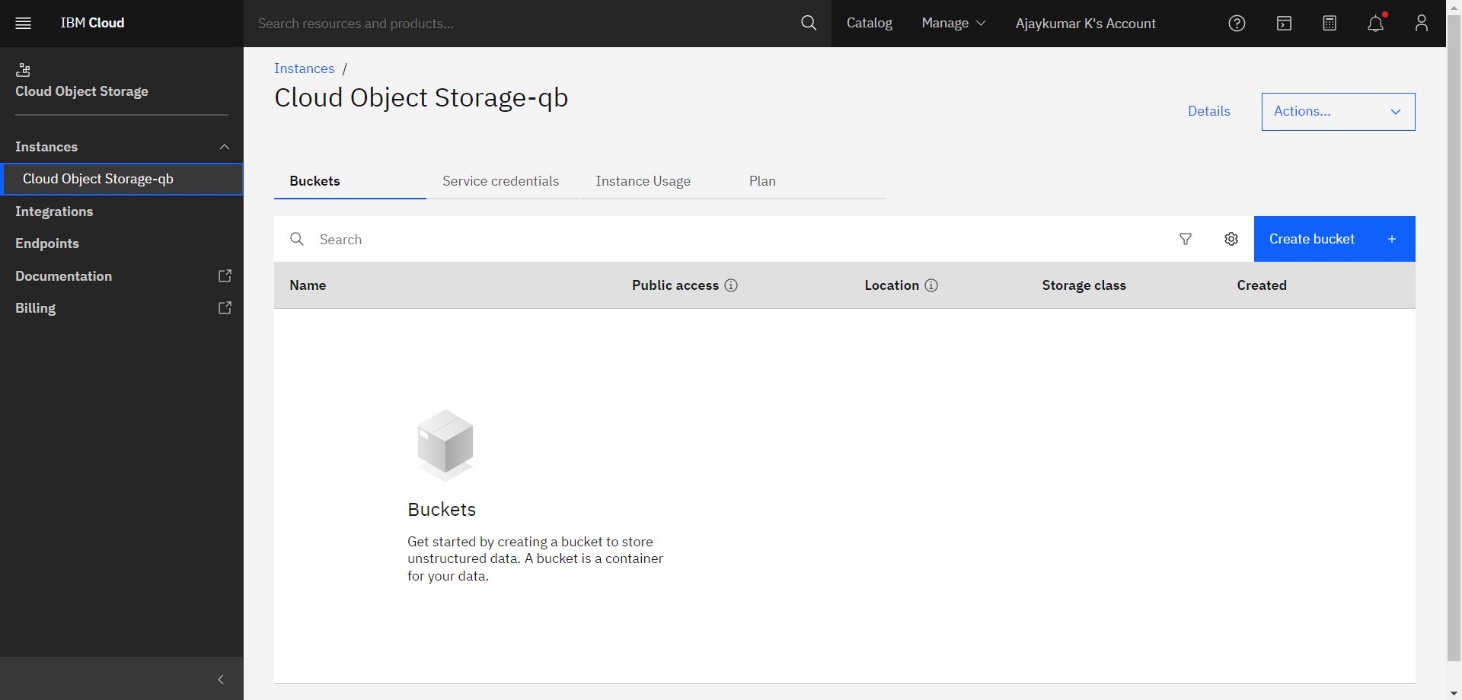
IBM Watson Studio is a comprehensive data science and machine learning platform offered by IBM. It is designed to help data scientists, developers, and other professionals collaborate and work on data analysis, machine learning, and artificial intelligence (AI) projects. Watson Studio provides a variety of tools and services to streamline the entire data science workflow, from data preparation to model deployment. Some of the key features of IBM Watson Studio are

* Data Preparation.
* Machine Learning Models.
* Model Deployment.
* Explainability.
* Data Catalog and Governance.

# WORKING WITH IBM WATSON STUDIO OBJECT STORAGE

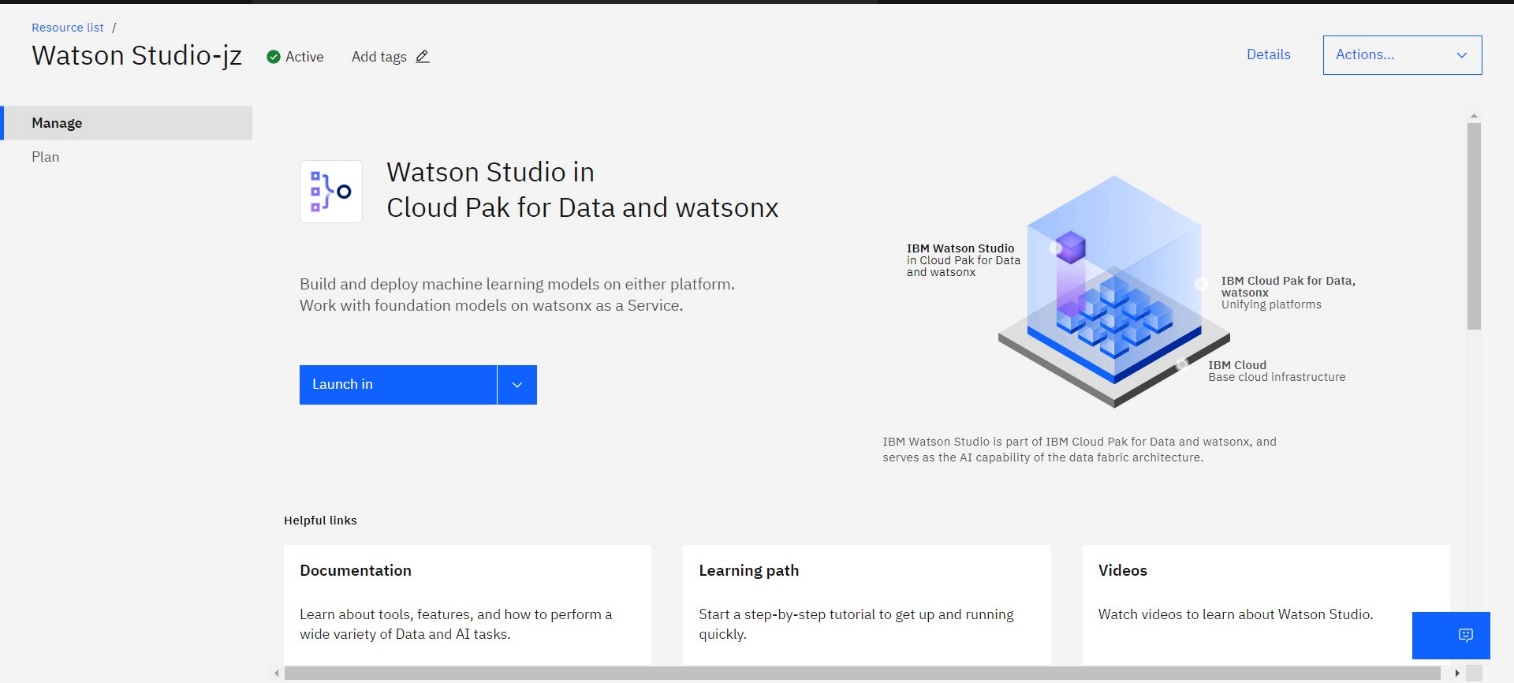
Create a cloud object storage. Create a bucket named “iris.csv”. Drop the

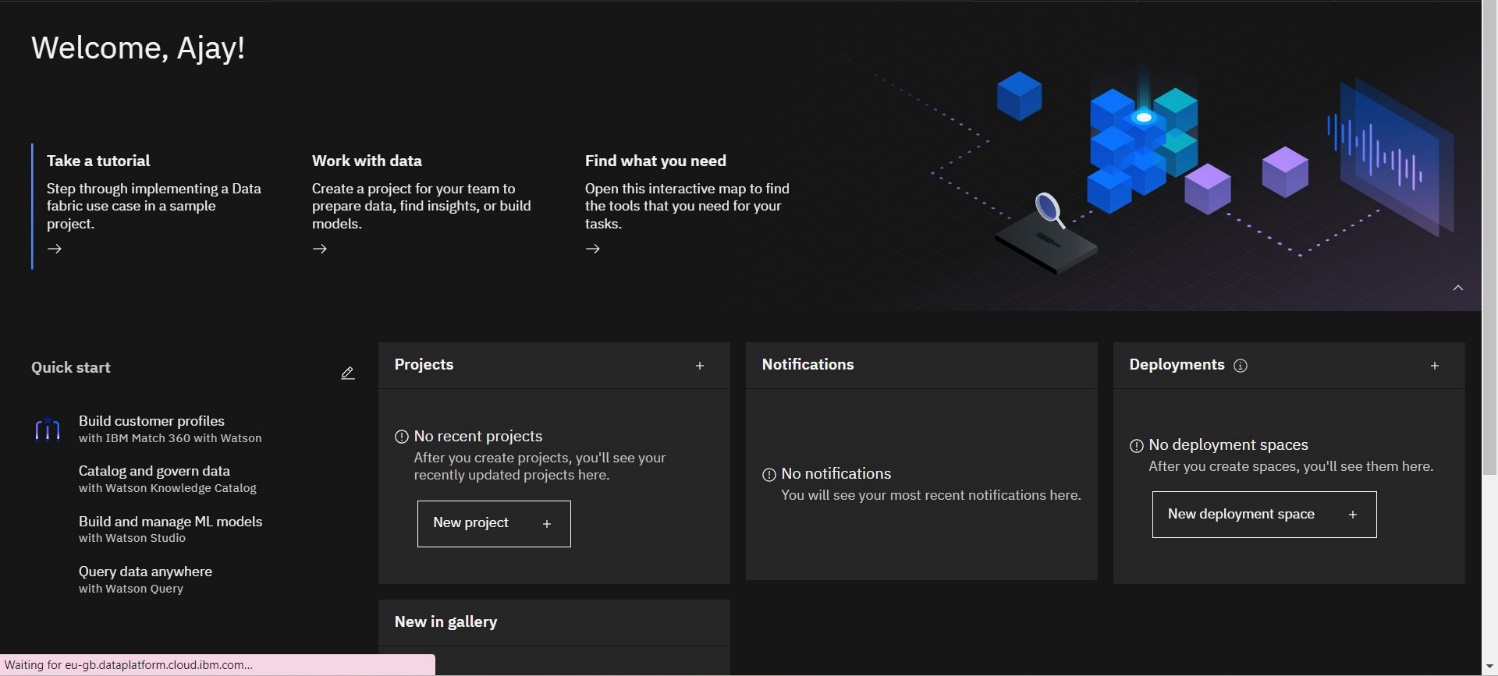
“iris.csv” dataset in the bucket.



# LAUNCHING THE WATSON STUDIO

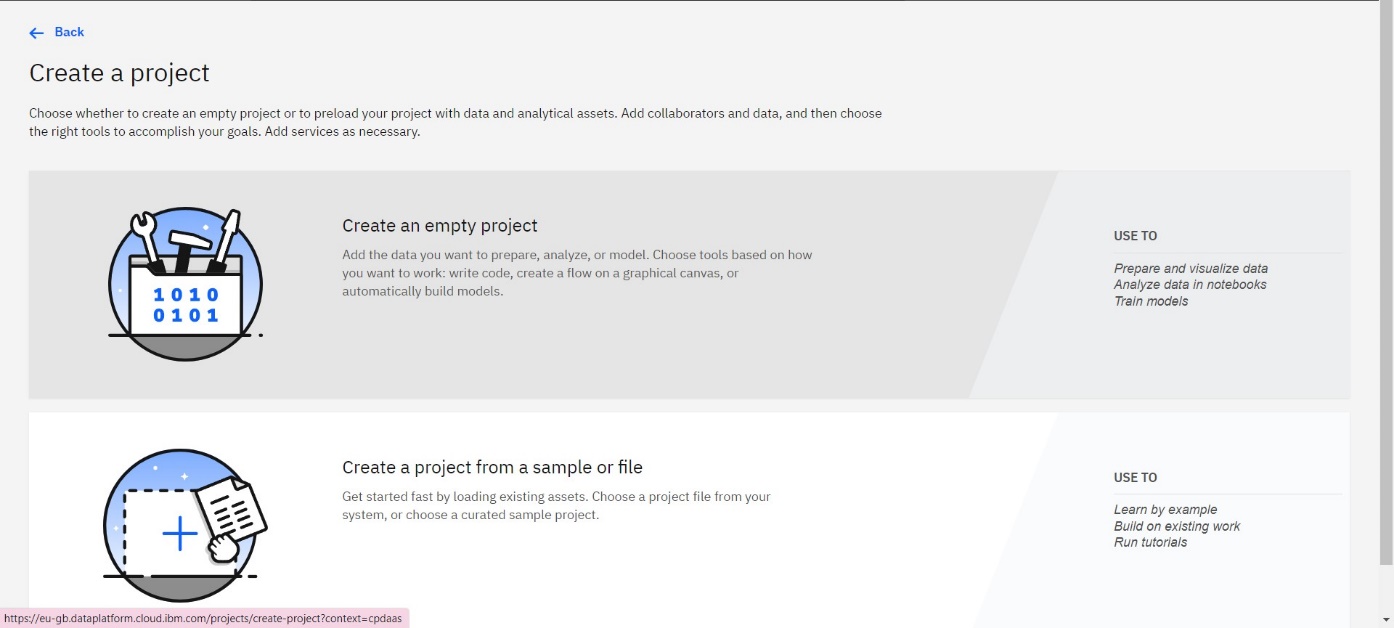
Launch the IBM Watsonx in Watson studio by providing our login credentials, and create a project called “iris”.

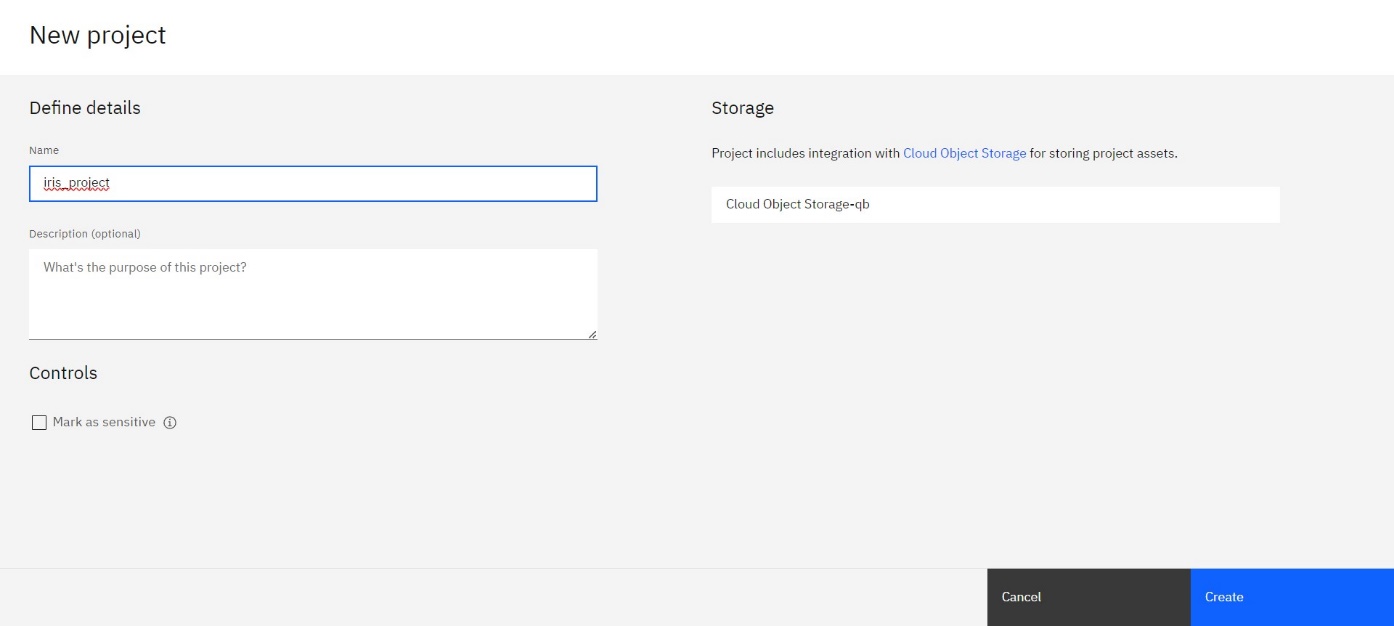


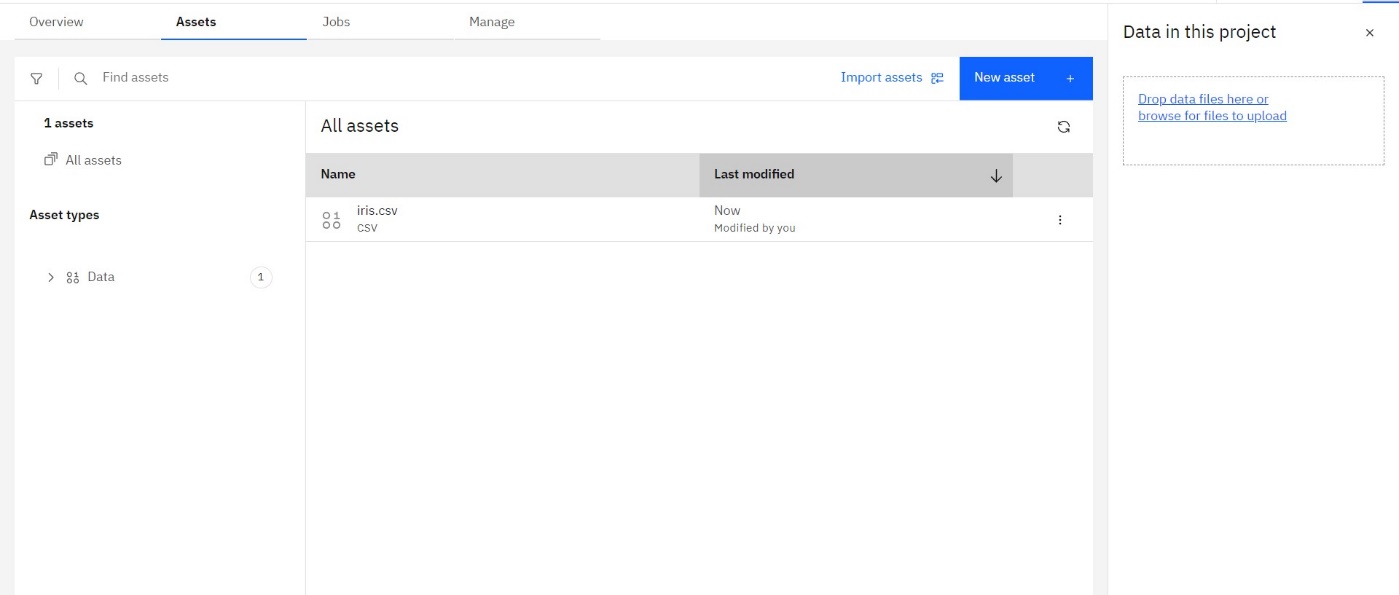


# DATABASE SETUP

Upload the dataset “iris.csv” in the asset section of the “AI-TOOL-Analysis” bucket in the Watson studio.

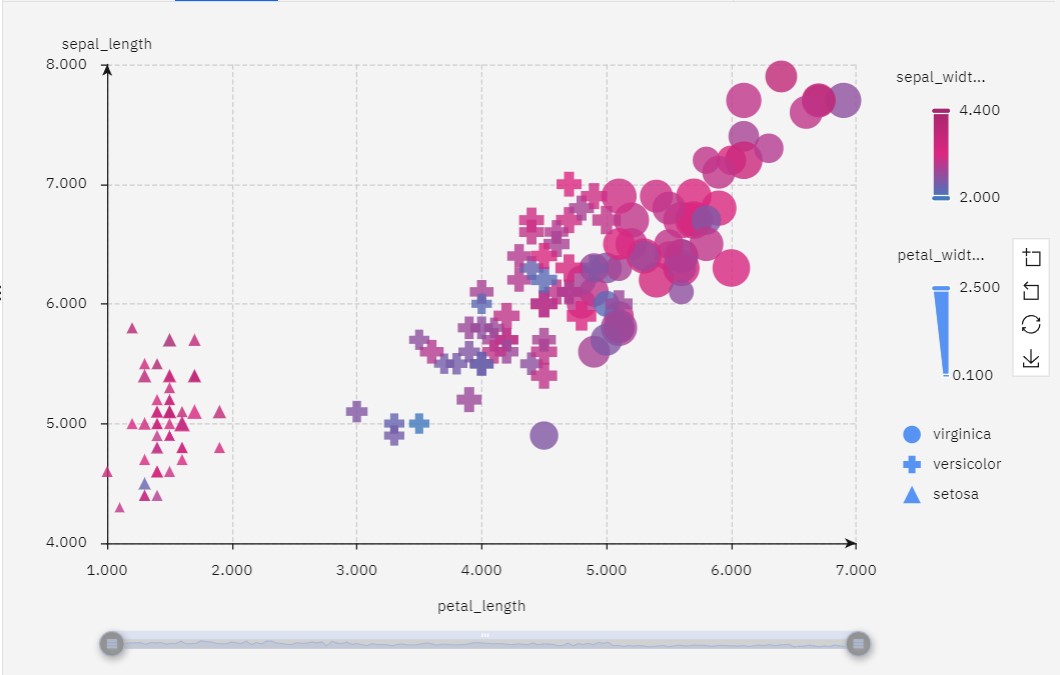




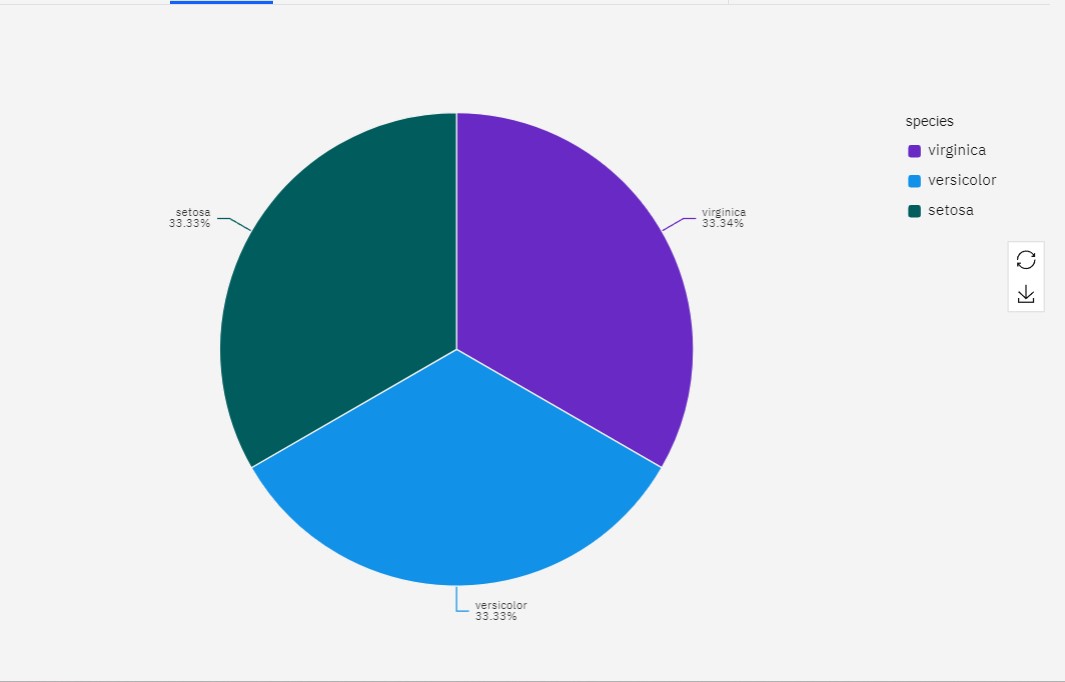


# VISUALIZING THE DATASET IN VARIOUS ASPECTS

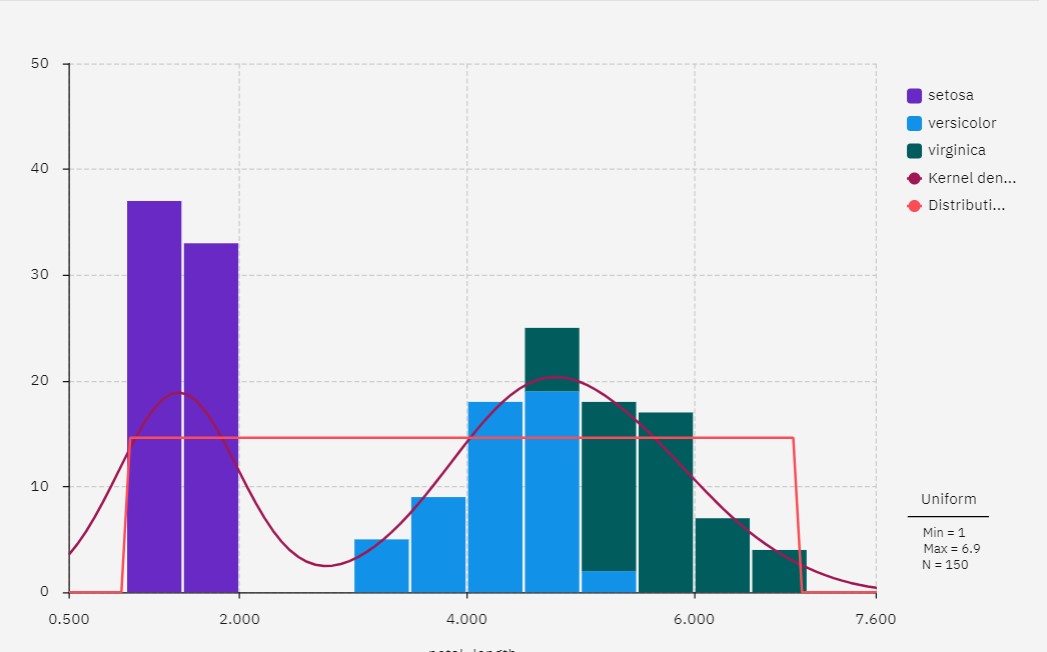
1. Scatterplot



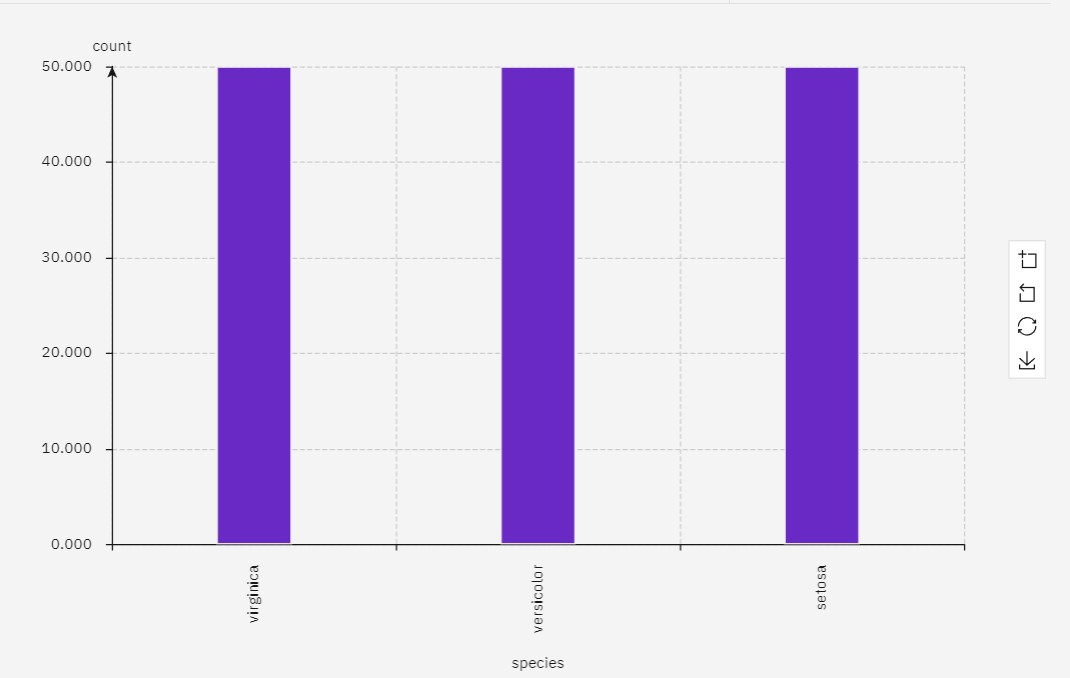
1. Piechart



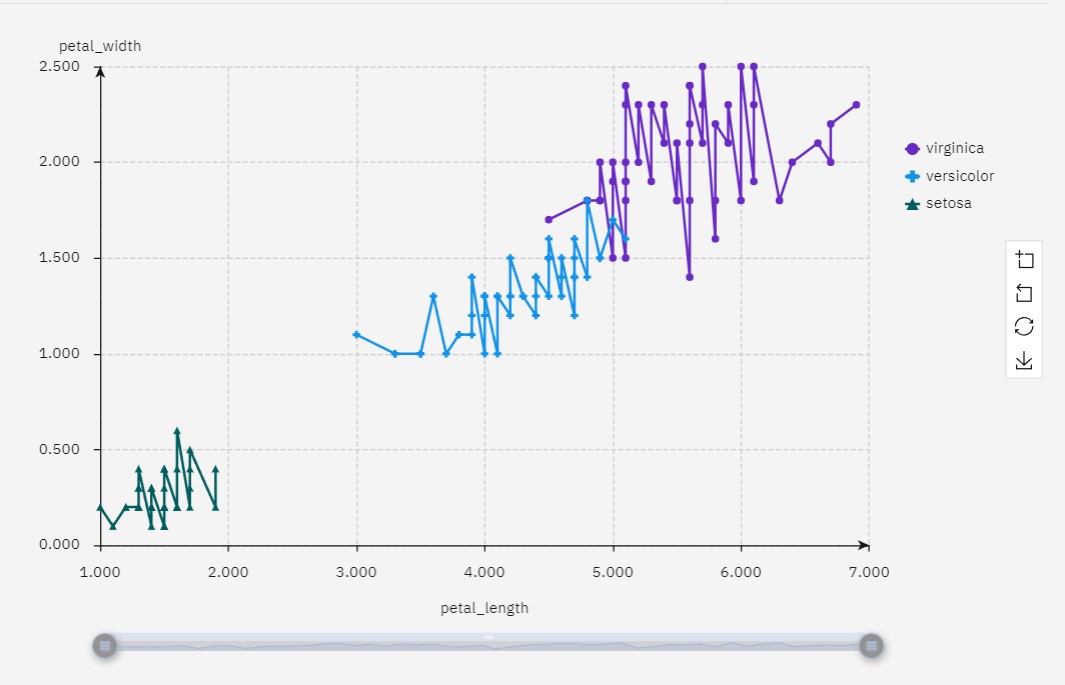
1. Histogram



1. Bar graph



1. Line chart



# MACHINE LEARNING ALGORITHMS

Machine learning algorithms are computational methods that enable computers to learn and make predictions or decisions without being explicitly programmed for a specific task. There are several types of machine learning algorithms, which can be broadly categorized into the following three groups

* Supervised Learning
* Unsupervised Learning
* Reinforcement Learning

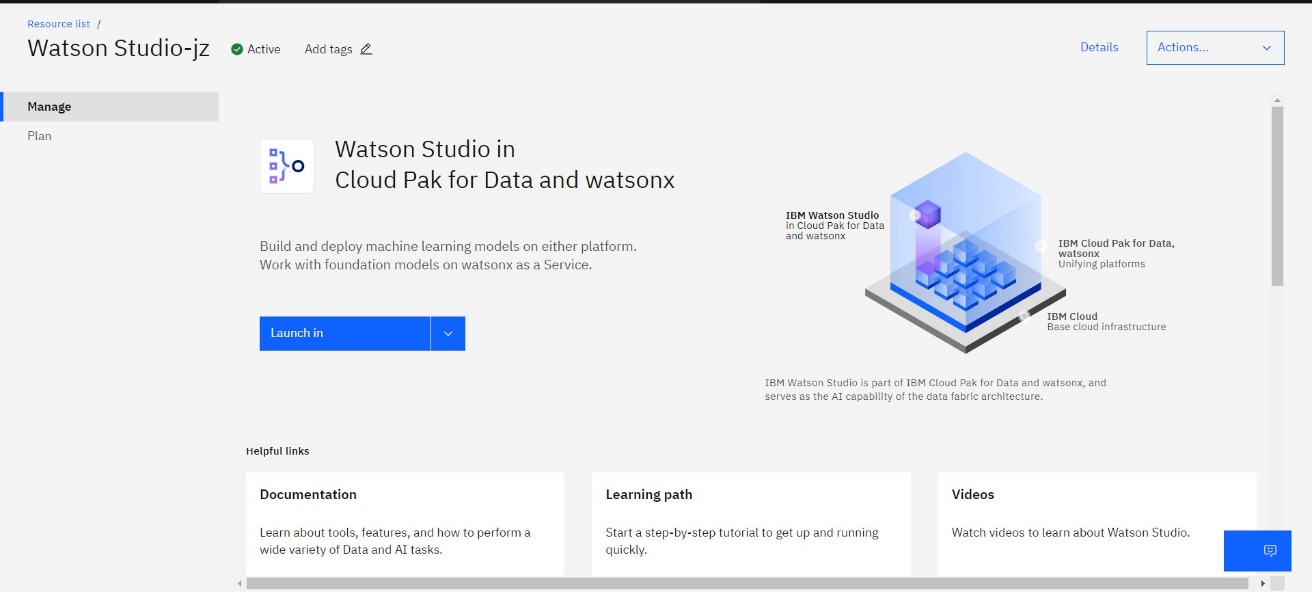
In addition to these main categories, there are other specialized machine learning techniques and algorithms, including

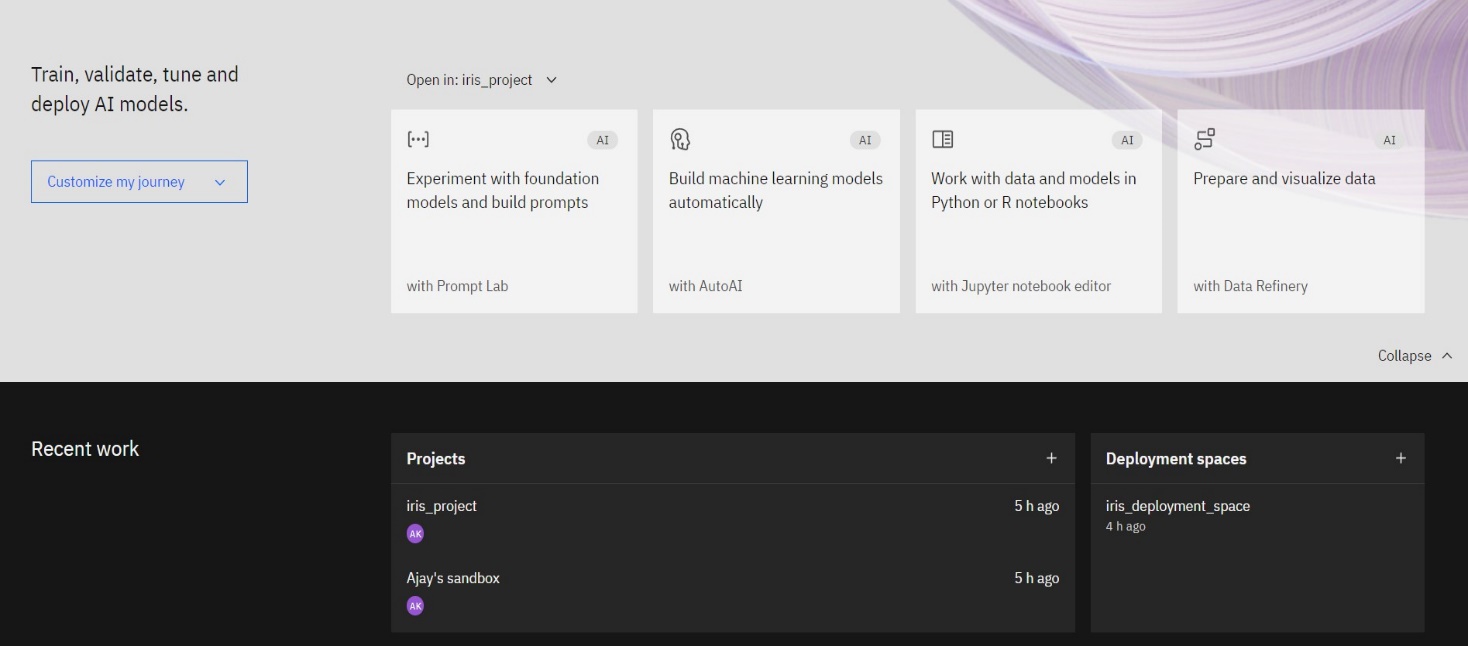
* Semi-Supervised Learning
* Transfer Learning
* Time Series Analysis
* Anomaly Detection
* Ensemble Learning

**MACHINE LEARNIG MODELS WITH IBM WATSON**

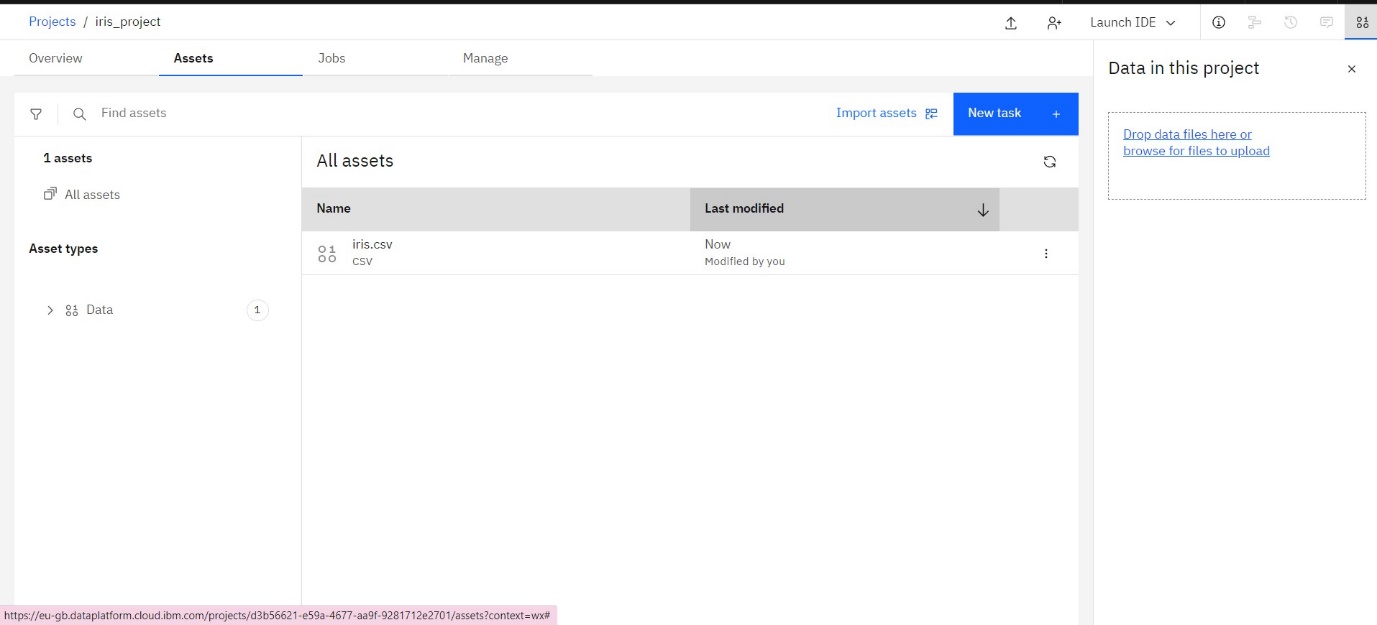
## IMPORT DATA TO A PROJECT

Create a new project in IBM Watson Machine learning platform under IBM Watsonx for data in the launch in twisty.



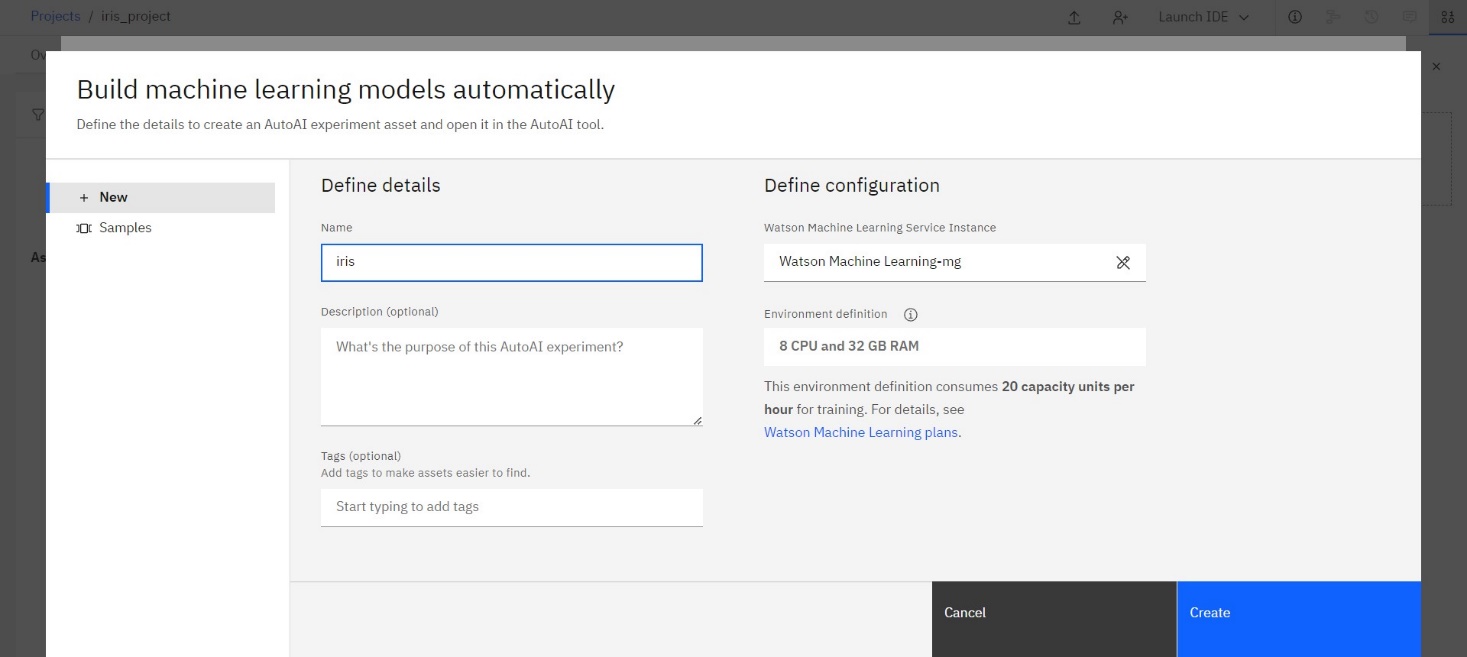


Upload the dataset “iris.csv” in the new project “iris\_project”.



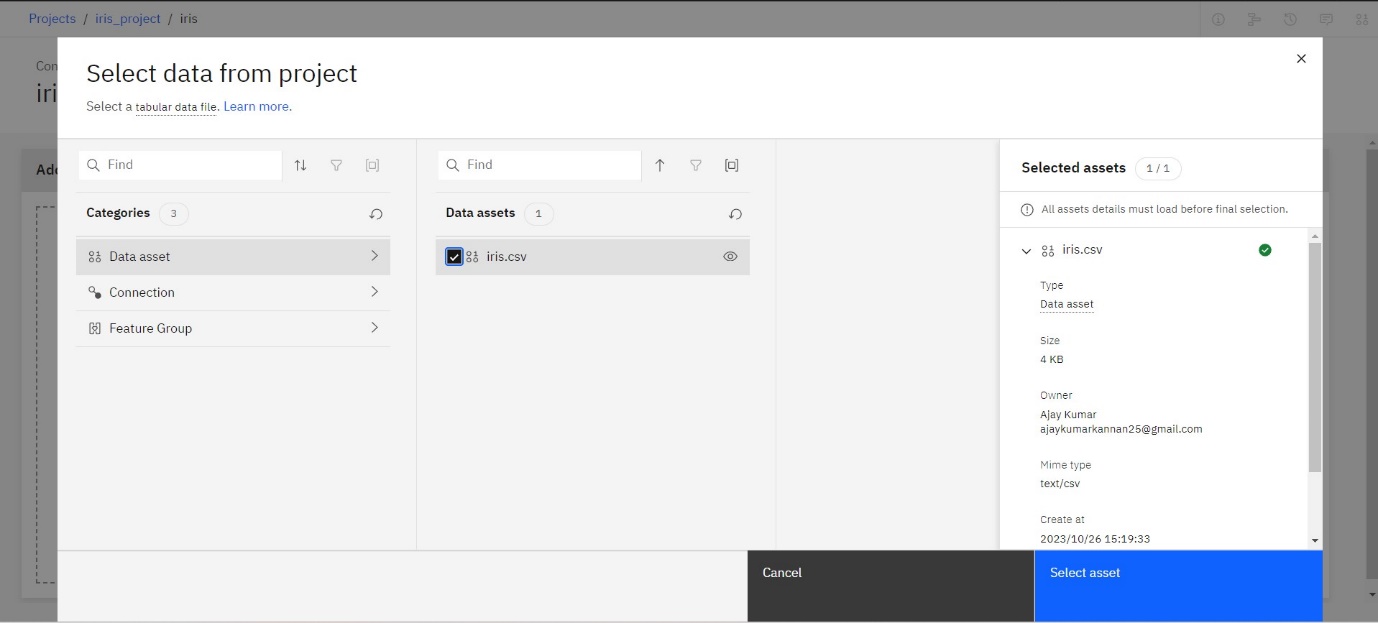
## ASSOCIATE THE MACHINE LEARNING SERVICE

Train the dataset with AutoAI tool and Associate the machine learning service within it.

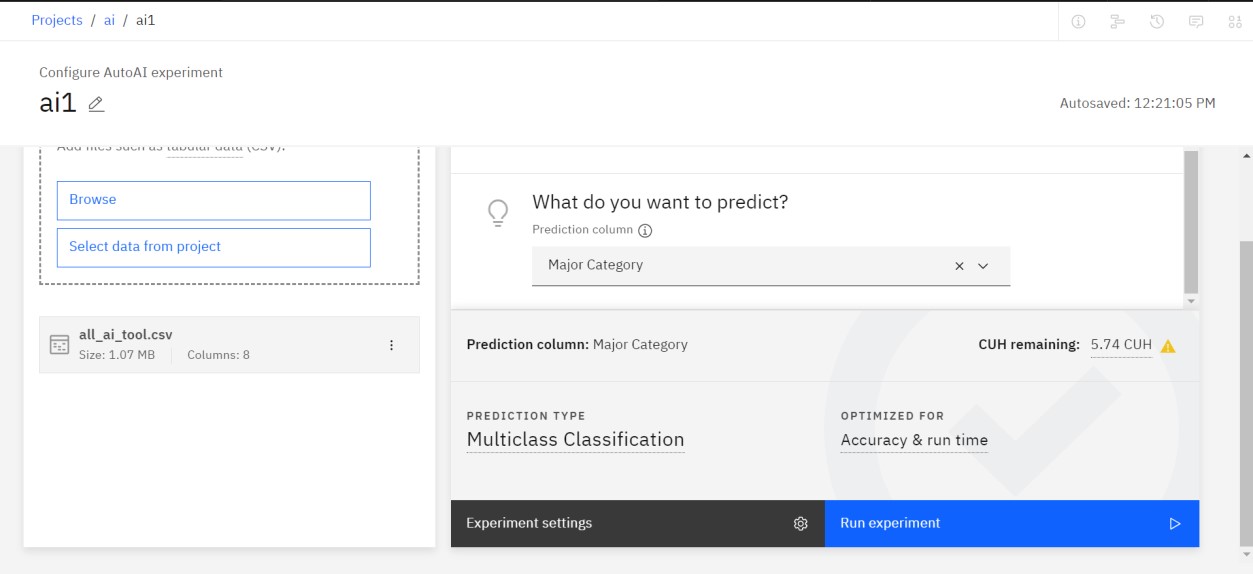


# BUILD A MACHINE LEARNING MODEL

Once the service is associated with our project, Build the model using AUTOAI tool. Drop the dataset “iris.csv” to train the data.

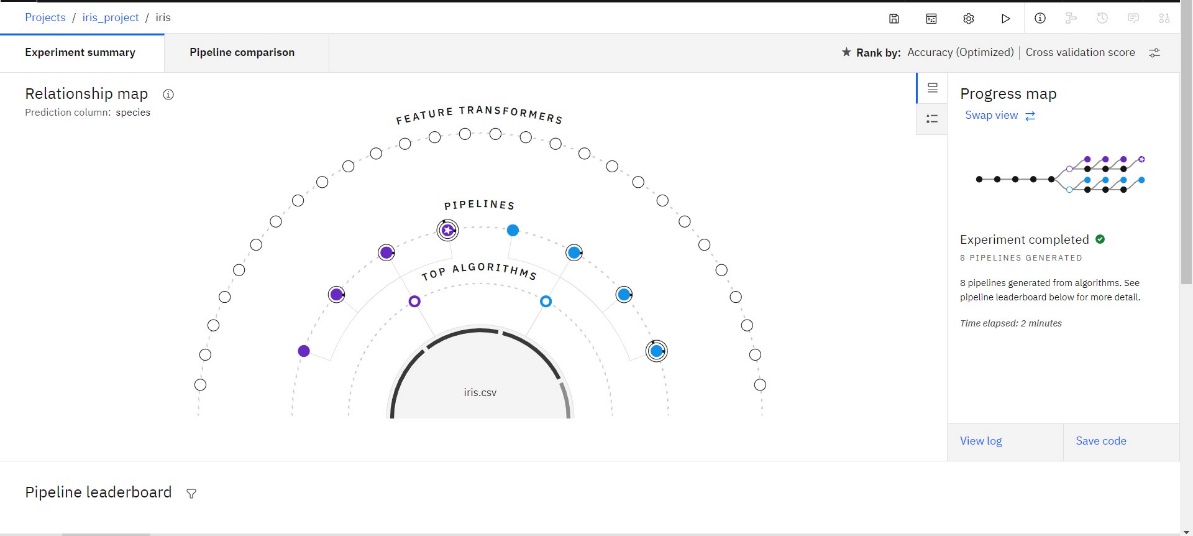


After adding the dataset, choose the column “Majority Category” to predict. Select the prediction type as “Multiple Classification” and Optimized for “Accuracy and runtime. Run the experiment.



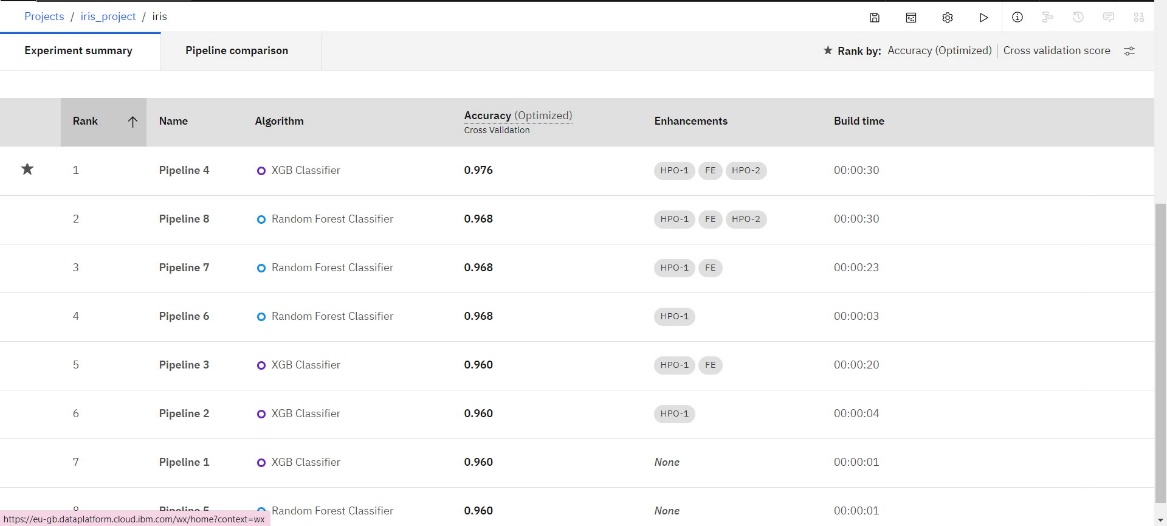
## RELATIONSHIP MAP

Processing the relationship map for the column “Major Category”

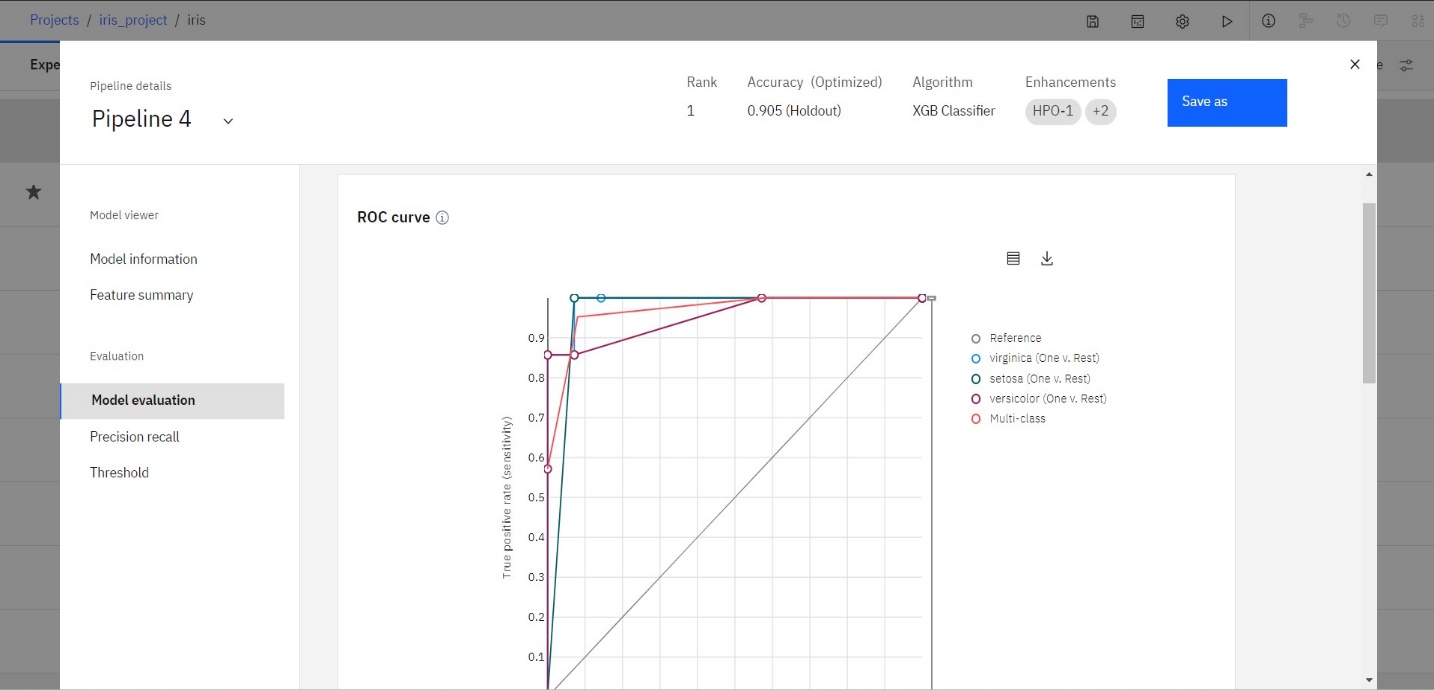


The complete relationship map

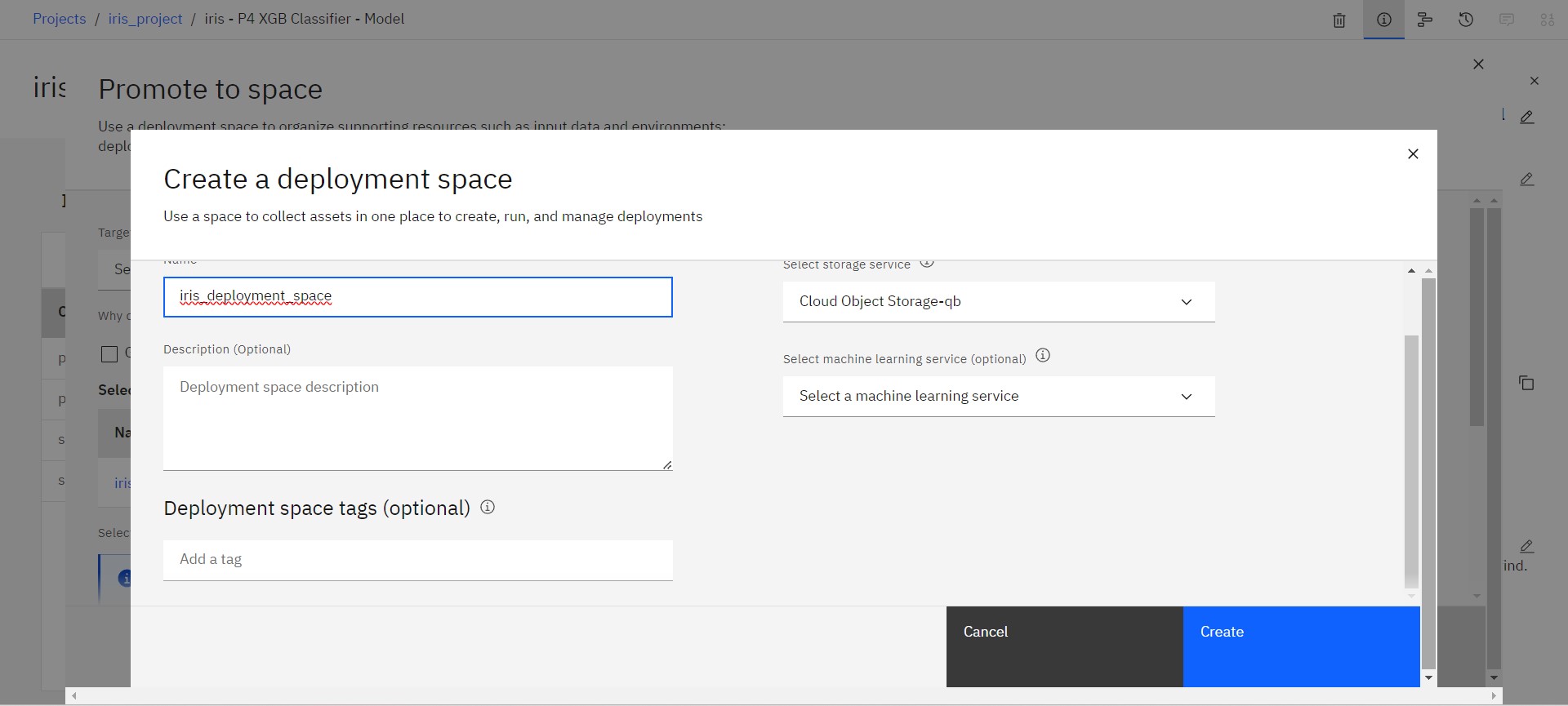
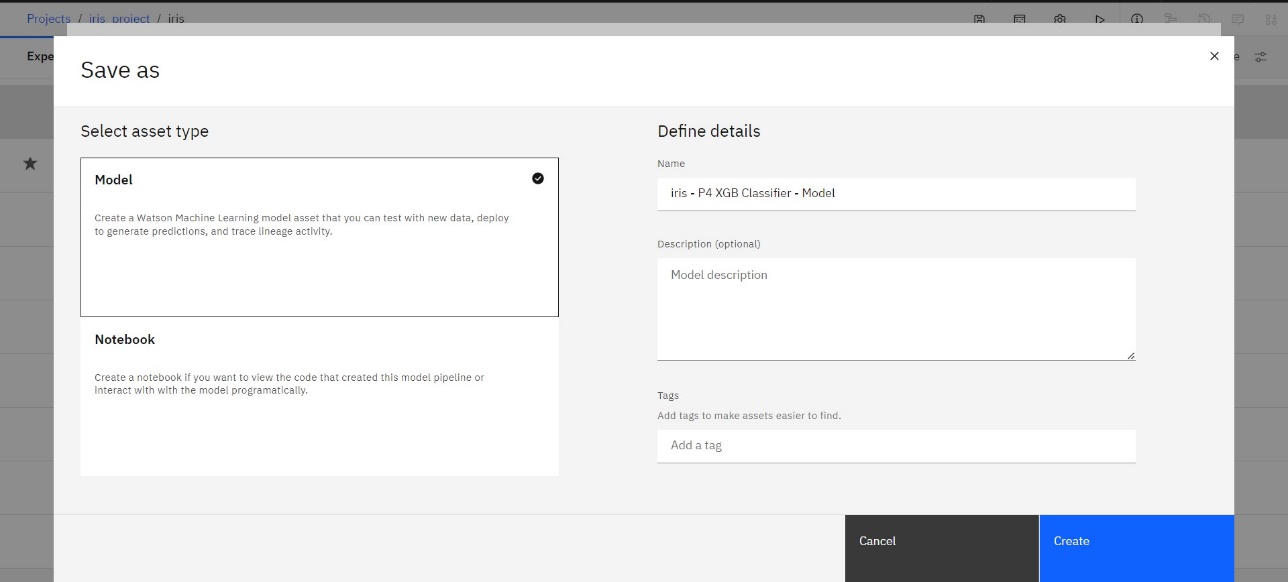
# Pipeline Leaderboard



From the leader board, Select the algorithm which is ranked as 1. Save the algorithm and deploy the dataset.

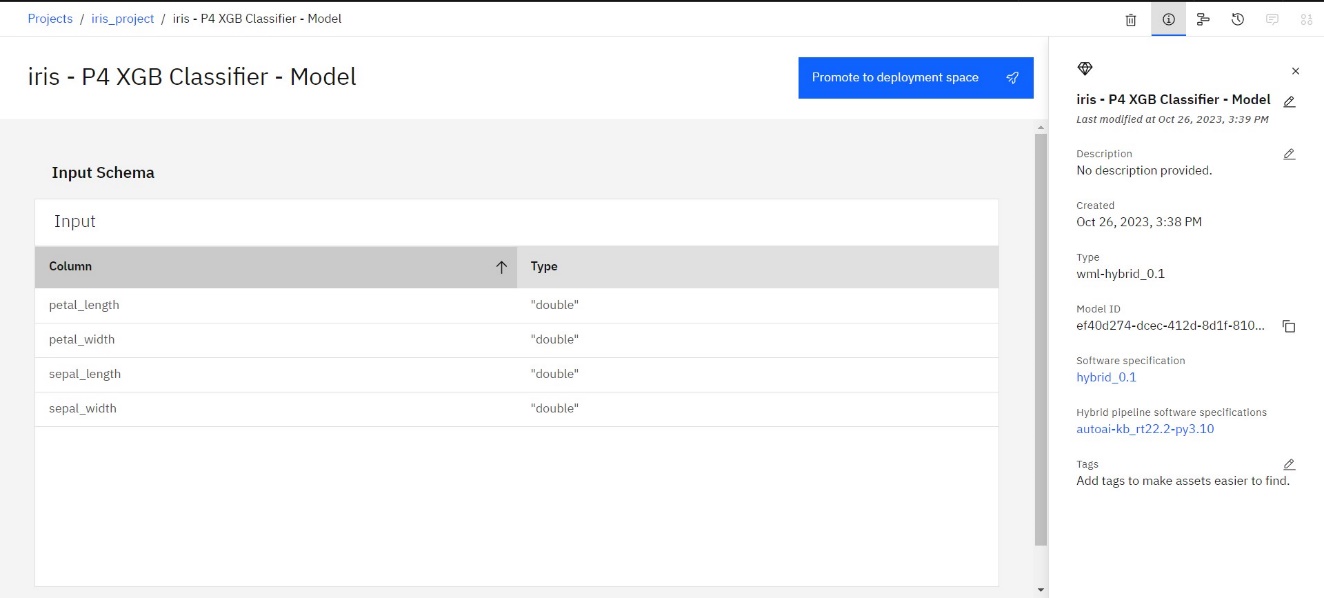


# Deployment Space:

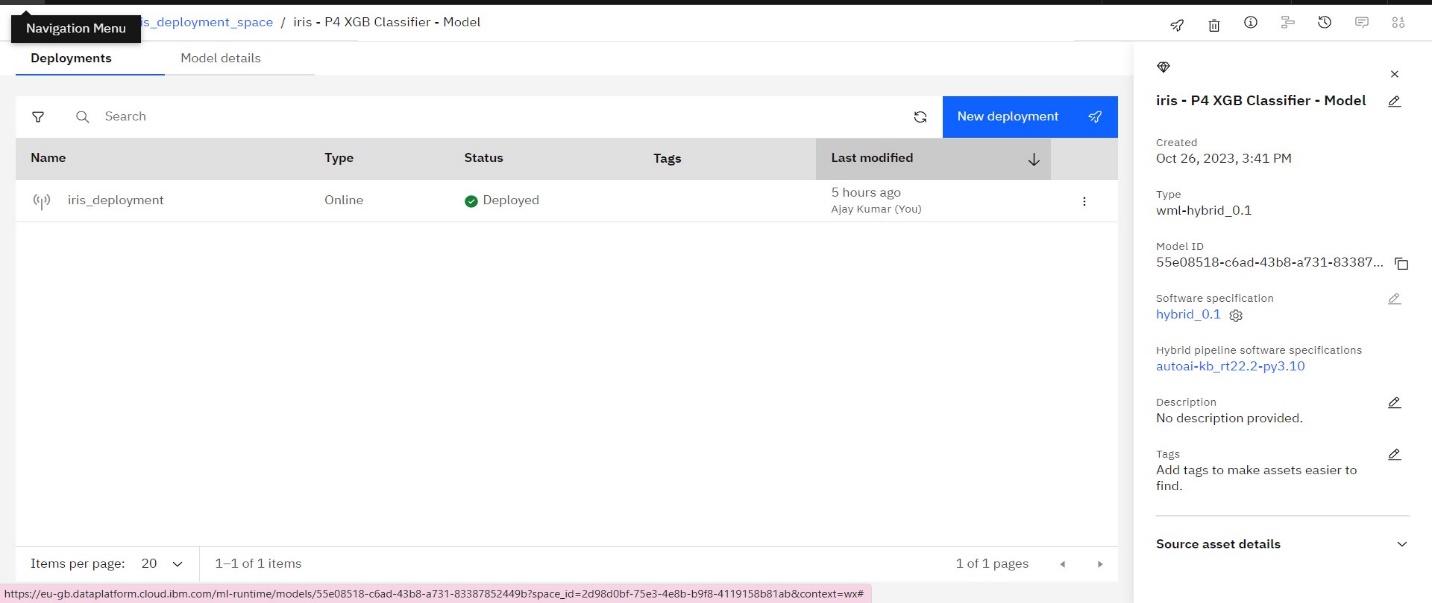


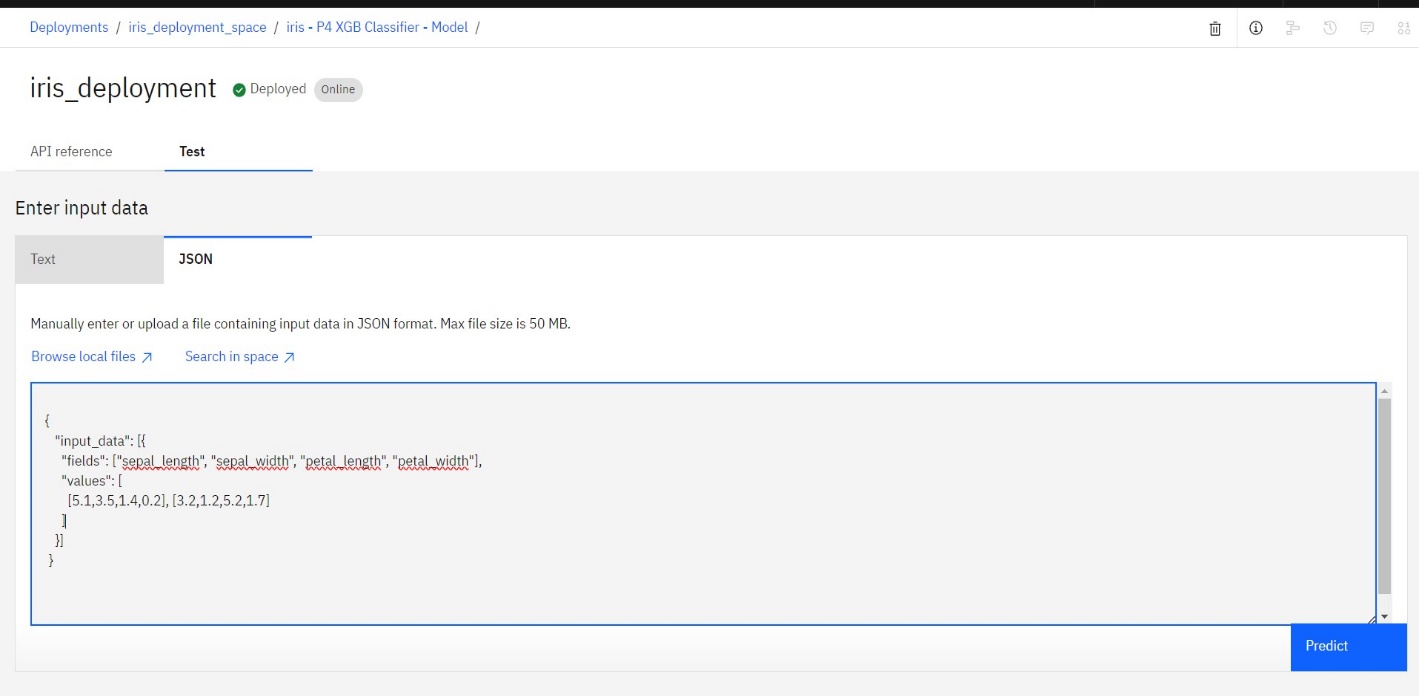
**DEPLOY AND TEST YOUR MODEL**

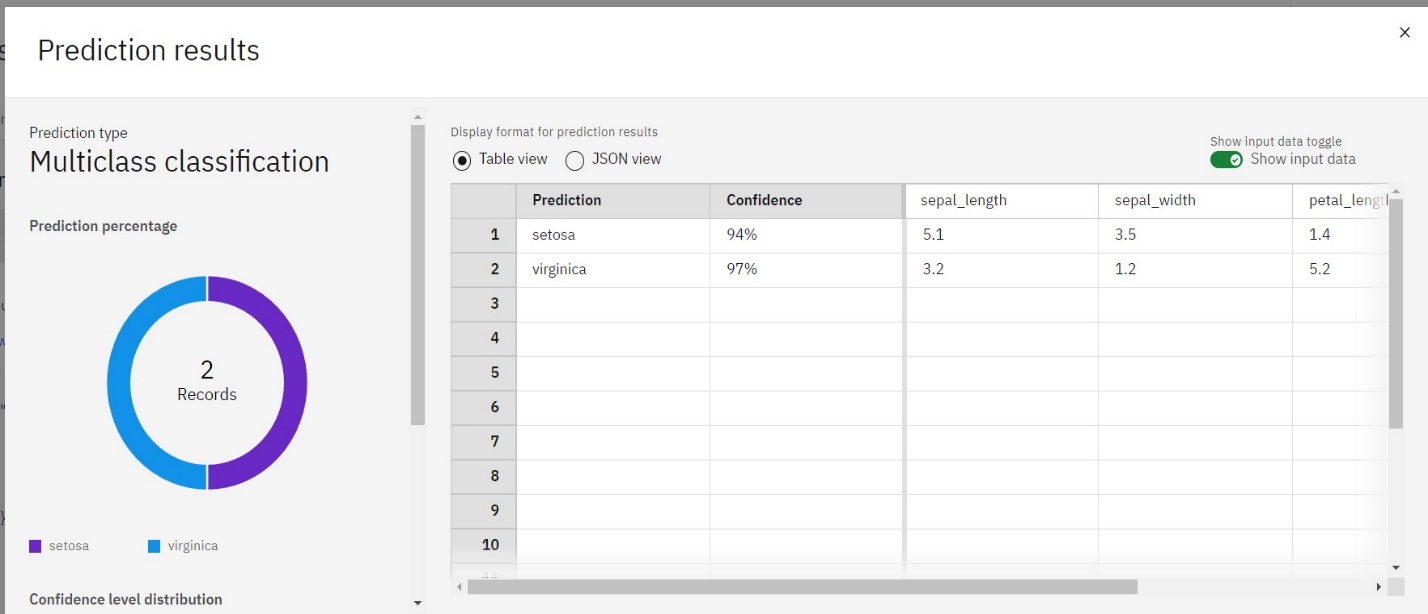
Open the deployment space created. Select the model and promote the deployment space



Promote it and find the predicted results by providing any two row of data from the dataset







## CONCLUSION

In this solution, we successfully conducted big data analysis using IBM Cloud Databases. By following the problem definition, design thinking, development, and documentation phases, we designed and implemented an analysis process that extracts insights from vast dataset. The solution utilized IBM Cloud Databases to store and manage the data, applied advanced analysis techniques, and visualized the results for valuable business intelligence. Through data-driven adventures, we explored the endless possibilities of big data and uncovered hidden insights in climate trends or social patterns.