Diabetes Prediction with Azure Machine Learning Studio

Overview

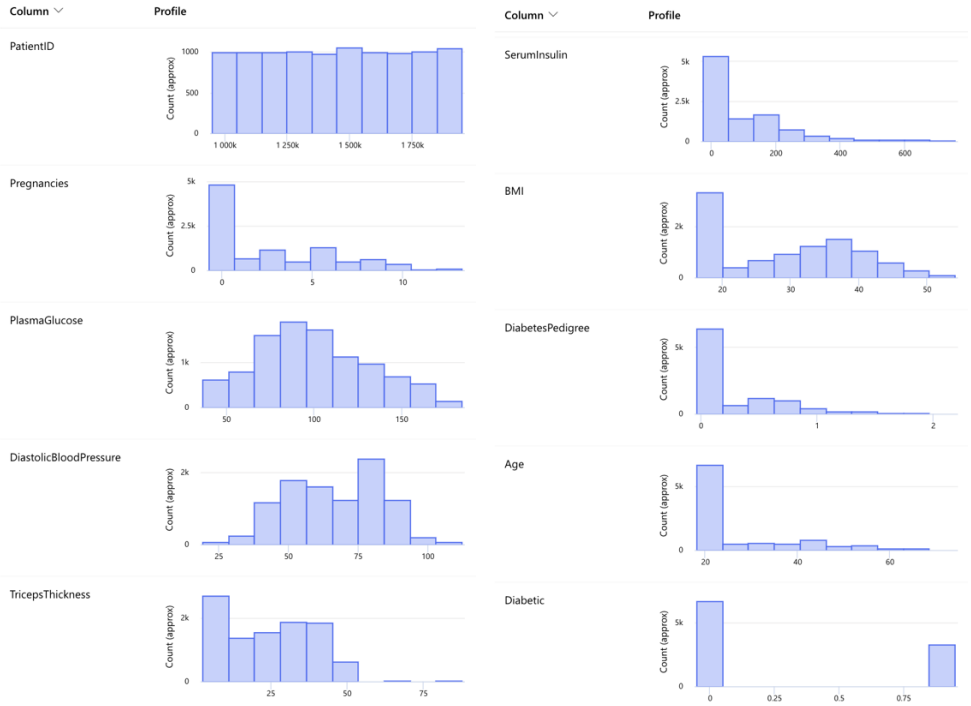
In this project, we utilized Microsoft Azure Machine Learning Designer to develop and deploy a classification model for predicting diabetes outcomes. The workflow involved setting up the Azure workspace and compute resources, constructing and evaluating a classification model pipeline, and creating an inference pipeline for real-time predictions. The predictive service was subsequently deployed to an Azure Container Instance and tested to ensure accurate classification of diabetes based on patient data.

Dataset

The dataset used in this project is the "diabetes-data" dataset downloaded from https://aka.ms/diabetes-data. This dataset includes various medical features related to diabetes, which are crucial for training the classification model to predict the likelihood of diabetes based on these attributes.

Number of columns: 10 Number of rows: First 50

PatientID	Pregnancies	PlasmaGlucose	DiastolicBloodPressure	TricepsThickness	SerumInsulin	BMI	DiabetesPedigree	Age	Diabetic
1354778	0	171	80	34	23	43.51	1.213	21	0
1147438	8	92	93	47	36	21.241	0.158	23	0
1640031	7	115	47	52	35	41.512	0.079	23	0
1883350	9	103	78	25	304	29.582	1.283	43	1
1424119	1	85	59	27	35	42.605	0.55	22	0
1619297	0	82	92	9	253	19.724	0.103	26	0
1660149	0	133	47	19	227	21.941	0.174	21	0
1458769	0	67	87	43	36	18.278	0.236	26	0
1201647	8	80	95	33	24	26.625	0.444	53	1
1403912	1	72	31	40	42	36.89	0.104	26	0

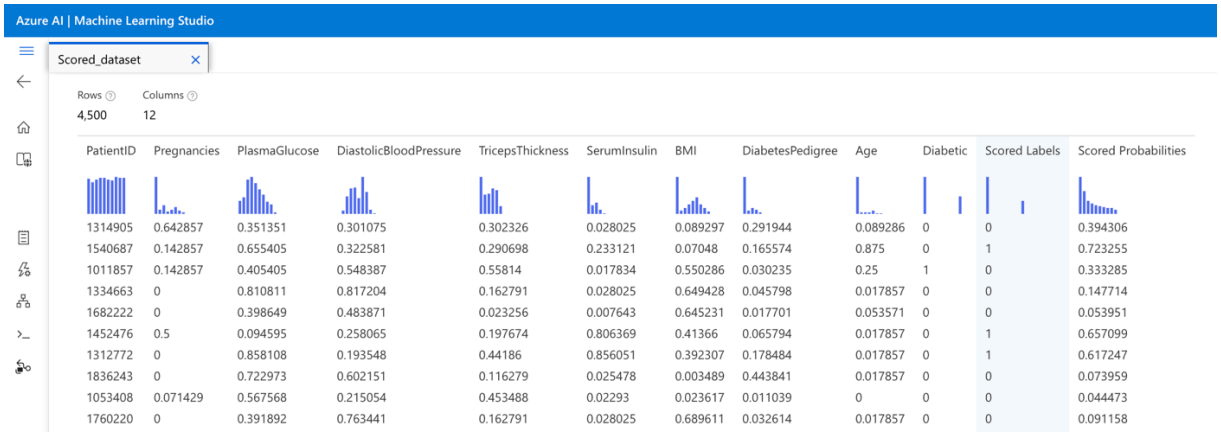


1- Setting Up Azure Machine Learning Workspace and Compute Resources

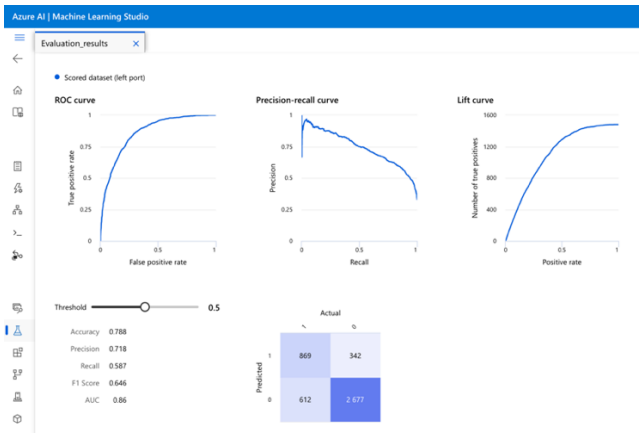
Established a Microsoft Azure Machine Learning workspace to manage data, compute resources, and models efficiently. Following this, compute targets were created to provide the necessary processing power for model training and deployment. A compute instance was set up as a development workstation, and a compute cluster was configured to handle the scalable processing of machine learning tasks, ensuring a robust environment for data science activities.

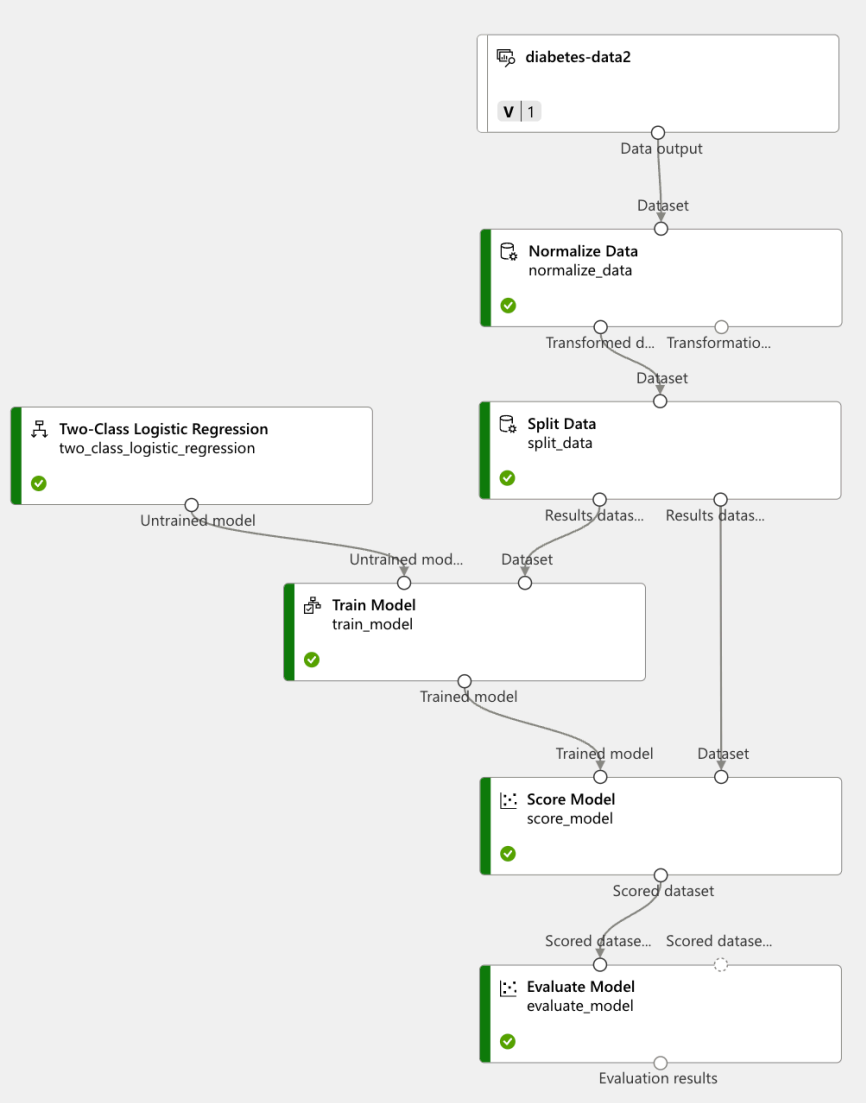
2- Classification Model Pipeline Creation and Evaluation

- **Pipeline Setup:** Created a new pipeline in Azure Machine Learning named "Diabetes Training" and selected the appropriate compute target.
- **Data Preparation:** Created, dragged and explored the "diabetes-data" dataset, addressing missing values and normalizing numeric columns using data transformations.
- **Model Training:** Configured the pipeline to split the data into training and validation sets, trained a “Two-Class Logistic Regression” model, and scored the model's predictions.
- **Model Evaluation:** Added an evaluation module to assess the model's performance using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC.



PatientID	Pregnancies	PlasmaGlucose	DiastolicBloodPressure	TricepsThickness	SerumInsulin	BMI	DiabetesPedigree	Age	Diabetic	Scored Labels	Scored Probabilities
1314905	0.642857	0.351351	0.301075	0.302326	0.028025	0.089297	0.291944	0.089286	0	0	0.394306
1540687	0.142857	0.655405	0.322581	0.290698	0.233121	0.07048	0.165574	0.875	0	1	0.723255
1011857	0.142857	0.405405	0.548387	0.55814	0.017834	0.550286	0.030235	0.25	1	0	0.333285
1334663	0	0.810811	0.817204	0.162791	0.028025	0.649428	0.045798	0.017857	0	0	0.147714
1682222	0	0.398649	0.483871	0.023256	0.007643	0.645231	0.017701	0.053571	0	0	0.053951
1452476	0.5	0.094595	0.258065	0.197674	0.806369	0.41366	0.065794	0.017857	0	1	0.657099
1312772	0	0.858108	0.193548	0.44186	0.856051	0.392307	0.178484	0.017857	0	1	0.617247
1836243	0	0.722973	0.602151	0.116279	0.025478	0.003489	0.443841	0.017857	0	0	0.073959
1053408	0.071429	0.567568	0.215054	0.453488	0.02293	0.023617	0.011039	0	0	0	0.044473
1760220	0	0.391892	0.763441	0.162791	0.028025	0.689611	0.032614	0.017857	0	0	0.091158

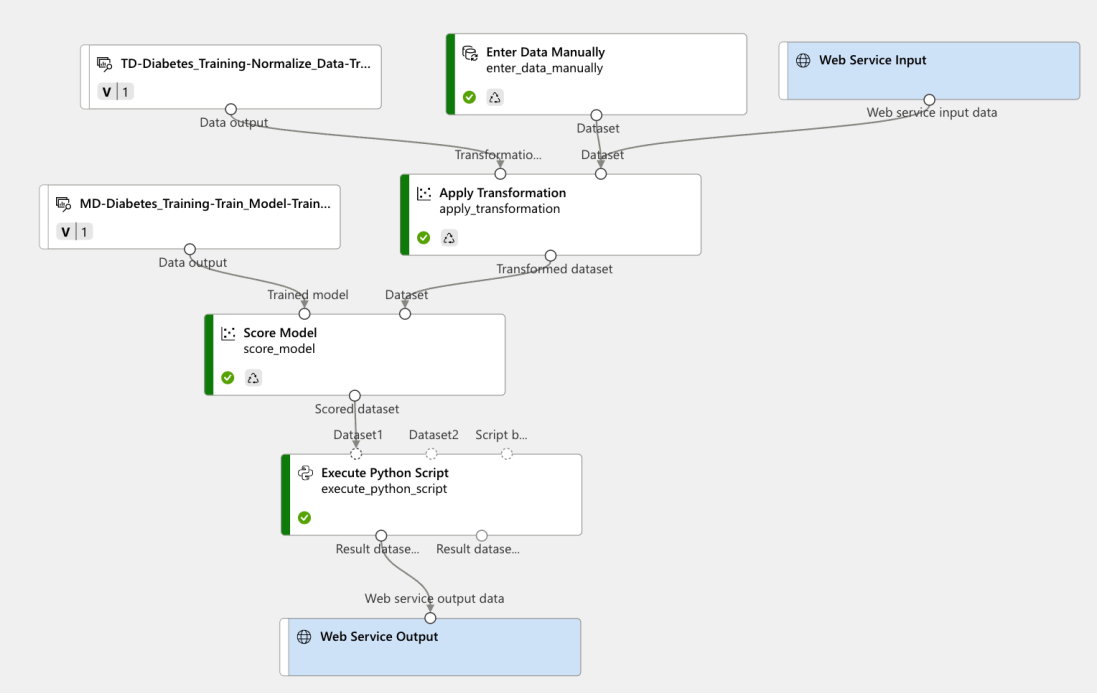




3- Inference Pipeline Creation

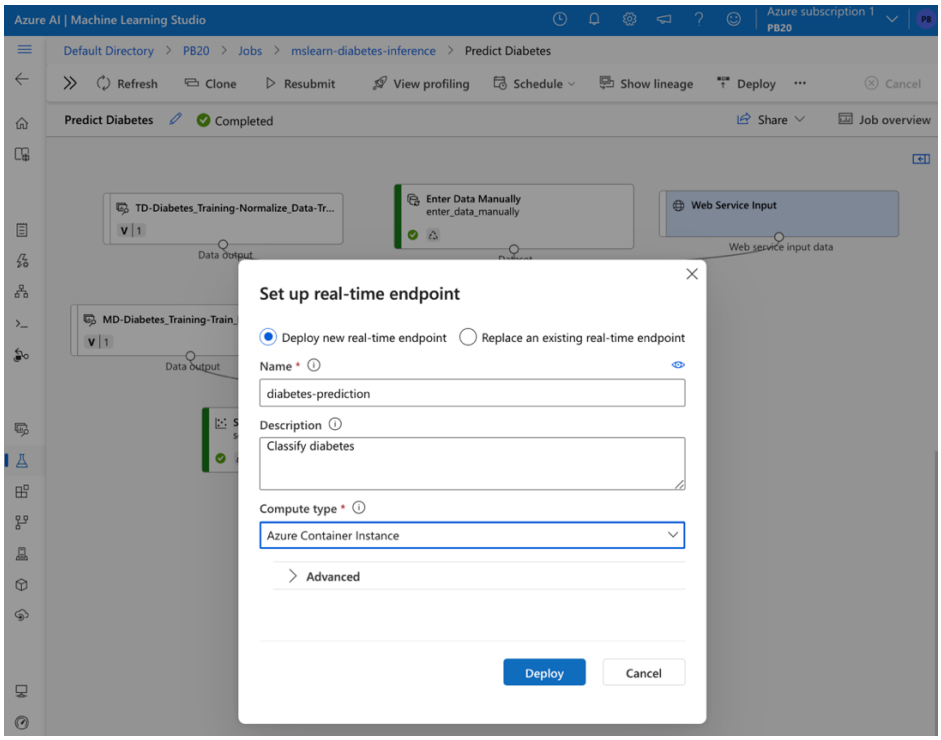
- **Pipeline Setup:** Opened the previously created "Diabetes Training" pipeline and selected the "Real-time inference pipeline" option, creating a new pipeline named "Predict Diabetes".
- **Data Transformation and Pipeline Modifications:** Replaced the dataset with an "Enter Data Manually" module for new data features, updated column selections, and removed unnecessary modules. Added an "Execute Python Script" module to extract and rename predicted labels, connecting it to the web service output.
- **Execution and Validation:** Submitted the pipeline as a new experiment on the compute cluster. Verified predictions by visualizing the output of the "Execute Python Script" module.

The inference pipeline prepares new data and applies the trained model to predict diabetes.



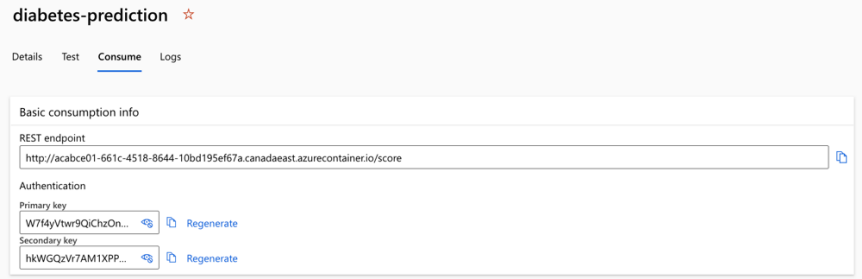
4- Deploying a Predictive Service

Deployed the "Predict Diabetes" inference pipeline by selecting "Deploy" and creating a new real-time endpoint named "diabetes-prediction" on Azure Container Instance (ACI). This deployment allows for development and testing purposes.

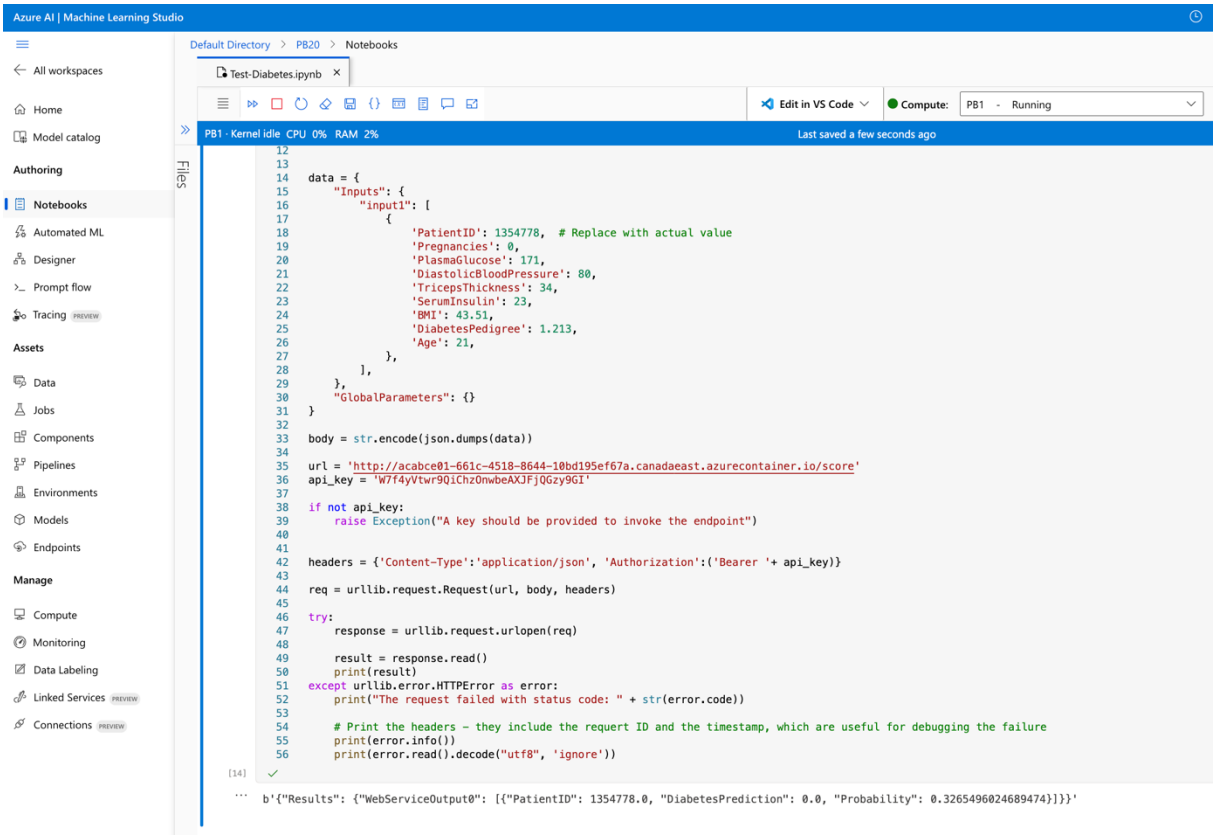


5- Real-Time Endpoint Testing

Accessed the "diabetes-prediction" endpoint on the Endpoints page to retrieve the REST endpoint and Primary Key.



Tested the deployed service by retrieving the REST endpoint and Primary Key, then using these details in a new notebook within Azure Machine Learning Studio to run a test and confirm that the service accurately predicts diabetes.



The deployment and testing process ensures that the predictive service is operational and accessible for client applications, delivering real-time diabetes risk classifications based on the trained model.