Machine Learning with SAP Data Intelligence introduces the unique features and components offered by SAP’s new data science platform including a look at the underlying machine learning and data science concepts which will help readers understand the platform and its applicability towards building intelligent solutions.

“Machine Learning with SAP Data Intelligence”

Contents

Index

The Author

Laboni Bhowmik, Avijit Dhar, Ranajay Mukherjee

Machine Learning with SAP: Models and Applications

495 pages, 2021

ISBN 978-1-4932-1926-1

www.sap-press.com/5014
Chapter 5
Machine Learning with SAP Data Intelligence

With the SAP Data Intelligence platform, SAP took a giant leap to empowering the intelligent enterprise. In this chapter, you’ll learn about SAP’s latest data science platform, including its unique features and components. We’ll also dive deeply into the technicalities of underlying machine learning and data science concepts to help you comprehend the vast potential of the platform and its applicability to building intelligent solutions. The knowledge you’ll gain from this chapter will also help you understand the hands-on chapters that follow.

Before we start our discussion of SAP Data Intelligence or define it as a product, setting the context on the right track is vital to understanding what today’s enterprises need from a data science platform in totality and how they can effectively utilize it to apply enterprise intelligence. Today’s enterprises have realized that only a portion of the enormous transformational potential of their enterprise data is being utilized because their data science teams have lacked the right platform to analyze the data and build effective machine learning models and scenarios. Without a robust data science platform, models may ultimately lack accuracy and robustness and frequently never make it into production. Thus, the effectiveness of a data science platform is of utmost importance for enterprises hoping to take advantage of enterprise intelligence. As a holistic approach, enterprise intelligence is all about machine learning and its integration into enterprise applications to create a truly scalable, governed, and transparent way to rapidly deliver business value out of enterprise data. Enterprise intelligence incorporates artificial intelligence (AI) into enterprise contexts alongside capabilities for the management of data, the development of AI models and scenarios, and the governance of data and models across various enterprise environments. To rapidly derive value from your business data, bringing together enterprise intelligence and intelligent information management in a trusted and governed manner is a crucial requirement. The ultimate result is often an increasingly complex landscape that is well adapted so you can discover intelligence out of enterprise data coming from various sources and formats.
However, delivering all of these capabilities into a single platform can be quite challenging, especially when every stakeholder, including chief information officers (CIOs), the IT operations team, data scientists, and the DevOps team, have different points of view, as shown in Figure 5.1:

- For the CIO, the focus is on stable, scalable data science or machine learning applications at low cost.
- For the data science team, the focus is on a development environment that is compatible with enterprise requirements, while still providing room for exploration.
- For data engineers and enterprise architects, the focus is on tools to connect and manage data across the organization for multiple use cases.
- For the IT operations or DevOps team, the focus is on reusability, stable deployment, and maintainability.

With this wide diversity of outlooks, a unified data science platform that can accommodate different teams is a complex challenge, but only part of the story, as you’ll see throughout this chapter.

In this chapter, we’ll start our exploration of SAP Data Intelligence with a brief look at the typical data science project lifecycle in Section 5.1 as well as how to manage your data science projects in Section 5.2. Then, in Section 5.3, we’ll turn to SAP Data Intelligence with a comparison between this platform and SAP Data Hub and, in Section 5.3.2, a look at the architecture of SAP Data Intelligence. We’ll also provide a thorough look at the key capabilities of the platform, including its built-in operations, in Section 5.4. The final section of this chapter returns to the SAP Data Hub, showing you how to migrate to SAP Data Intelligence in Section 5.5.

**Figure 5.1 What Today’s Data Science Citizens Want**

In general, a data science project is an iterative process that combines various steps until a methodology is fine-tuned to solve a specific business case. The process often includes the following activities:

- **Understanding the business case**
  The first and most important step for starting any data science project cycle, understanding the business case determines the success of the project. Tremendous computing power and vast collections of data are not worth the cost if your formulation of business problems is inaccurate. Creating a problem statement is a typical pain point that can often be resolved by applying data science instead of a regular enterprise solution. Asking the right question in this context is really the right recipe.

- **Collection of data**
  Once you formulate the right problem statement by applying the right recipe, the next step is to gather the key ingredients. This activity is all about gathering the relevant data and breaking down the problem statement into smaller units. For the correct data collection, the data science team comes into the picture, and your data scientists will need to know which ingredients are required, how to source and collect this data, and how to prepare the data to meet the desired outcome. A typical data science project includes identifying various data sources, which could be logs from webservers, social media data, data from online repositories like US Census datasets, data streamed from online sources via application programming interfaces (APIs) or web scraping, or even data in Excel files. Data acquisition involves acquiring data from all the identified internal and external sources that can help you answer the business question.

- **Grounding of data**
  Now that you’ve gathered the relevant data from the various data silos that exist in your landscape, we can now focus on data grounding, which is all about preparing your data for performing further analysis. This methodology helps you apply data science to determine whether the data that you’ve collected can even address your business problem. If the data you have is not good enough or not relevant enough, you’ll need to go back to the collection step of the project lifecycle. To realize the potential of your data, several methods are available to explore your data through data statistics, like means and medians. Some analysts plot the data and look at its distribution through visualizations like histograms, spectrum analyses, population distributions, etc. Based on this general understanding, the data can be further prepared for analysis, which involves several other steps, including:
  - Identifying missing data
  - Removing invalid data
– Removing duplicates
– Making data ready to feed into a data model

Overall, 70%-90% of the time spent in a data science project goes into phases involving the collection, realization, and grounding of data, which are very critical for the success of any data science project.

**Modeling of data**
The hardest part of the lifecycle is already over since we’re done with data collection and grounding work. Now is the time for applying modeling, which is equivalent to tasting the curry you’re cooking after preparing the ingredients. At this point, you’ll need to determine whether the curry needs more seasoning, or is it ready to serve? The data modeling phase basically deals with training, validating, and testing the model. This entire phase is focused on the evaluation of the model in terms of accuracy and relevance. Accuracy refers to how well the model is performing, while relevance is more about how well the model answers to the context of your query. During this phase, diverse machine learning techniques are applied to the data to identify the best fit machine learning model, and competing models are trained with training datasets.

**Deployment and iteration**
We’re at the final step: Now, we’ll expose our selected model to the real world for human use. The more our deployed model runs, the more feedback is received. Capturing this feedback can be decisive for any data science project, and accurate feedback can help influence effective data model changes. As the final stage, deployment and iteration involves the retraining of your machine learning model in production whenever a change occurs in the data or the data source so that the performance of the machine learning model can remain high.

---

**5.2 Managing the Data Science Project Lifecycle**

MLOps is the logical retort to the current challenges that enterprises face when trying to put machine learning models into production.

In simple terms, as shown in Figure 5.3, MLOps is the machine learning equivalent of DevOps. While DevOps optimizes the production lifecycles of big data projects, MLOps tries to solve problems associated with implementing machine learning in production. By its very nature, machine learning operationalization is a deeply collaborative process that requires collaboration between data scientists and the operations or production team to enhance automation, minimize waste, and ultimately harvest productive and consistent insights through machine learning.

According to an analysis called “Operationalizing Your Data Science and Machine Learning Initiatives” published by Gartner Analyst and Research Director Erick Brethenoux,
while many companies develop machine learning models, only 47% of those models make it to production. Also, nearly 88% of AI initiatives in the corporate sector struggle to move past the test stage. Since machine learning involves numerous intricacies and complexities, standard DevOps processes aren’t enough to address these challenges. Thus arose the need for a more comprehensive solution for machine learning implementation—MLOps. To use MLOps, you’ll need a comprehensive data science platform that can enable MLOps and provide your data scientists all the support they need to carry out data and analytics tasks. These tasks encompass visualization, interactive exploration, deployment, performance engineering data preparation, and data access. But what is the incentive to adopt MLOps? According to the published Global AI survey by McKinsey & Company, companies applying core practices for using AI are experiencing much higher revenues as well as greater cost savings compared to other AI-adopting organizations. Automation through MLOps basically helps a company enable faster time-to-market and greatly reduces operational costs, allowing companies to be more agile and strategic in their decisions. Data science platforms that can put core MLOps methodologies into practice and that can enable the production of reproducible, traceable, and verifiable machine learning are gaining a lot of traction. A powerful tool, the SAP Data Intelligence platform ends the prolonged pursuit for an unified data science platform that can offer end-to-end solutions for developing, fine-tuning, deploying, and even tracking and versioning machine learning models. At the same time, SAP Data Intelligence enhances visibility over the resources used for easier reproducibility and easier performance monitoring of a model, and then retraining when appropriate. Now that you have a pretty good understanding of the foundational concepts, in upcoming sections, we’ll focus more on the features and capabilities available in SAP Data Intelligence.

### 5.3 SAP Data Intelligence

With the introduction of the SAP Data Intelligence platform, SAP addresses the end-to-end machine learning project lifecycle within a single data science platform where data scientists can explore, experiment, and build at maximum speed while also being able to share components of the work with colleagues and collaborate with them securely on certain tasks. AI generally requires a team effort, and coordination and cooperation will be needed between the business users who understand the needs of the organization and its customers, the data engineers that understand where the data is located and how the data is structured, the data science teams that understand how to extract value from that data, and the IT and DevOps teams that provide support. SAP Data Intelligence is a comprehensive AI platform that complements all these needs and can build intelligent enterprise solutions with its wide variety of features and services to accommodate multiple enterprise users and stakeholders, including data scientists, data engineers, developers, and administrators. Figure 5.4 shows a snapshot encapsulating the overall features and capabilities of SAP Data Intelligence.

According to SAP’s official definition, SAP Data Intelligence is a comprehensive solution for delivering data-driven innovation and intelligence across an enterprise, unifying scalable enterprise AI and intelligent information management. SAP Data Intelligence leverages Docker and Kubernetes-based container technology and allows users to query large amounts of data (in distributed storages like Amazon S3, Azure Data Lake ADL, Azure Storage Blobs, HDFS, and Google Cloud Storage) and run highly scalable data flows and data-driven applications. The primary focus of the platform is intelligent data processing and making machine learning processes operational in production landscapes so you can make the most out of different data sets through seamless integration. Simultaneously, stakeholders like data scientists can continue using their tools of choice, while IT can focus on innovative prototyping and scaling their projects.

![Figure 5.4 SAP Data Intelligence](image)

**Note**

A Kubernetes cluster is basically a collection of node machines for running containerized applications; running Kubernetes ultimately means running a cluster. At a minimum, a cluster comprises a control plane and one or more compute machines, or nodes, where the control plane solely manages the desired state of the cluster, for instance, which applications are running and which container images are used. More details can be found at https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/.

#### 5.3.1 SAP Data Intelligence Versus SAP Data Hub

In this context, as a platform, SAP Data Intelligence is the successor product to SAP Data Hub and SAP Leonardo Machine Learning Foundation, which have been combined into
a single amalgamated data science platform. In this way, SAP Data Intelligence encompasses a wide spectrum of enterprise AI building features covering data discovery, refinement, orchestration, composition, monitoring, governance, and compliance all on a single platform. Basically, SAP Data Intelligence and SAP Data Hub are the same solution delivered in different ways: SAP Data Hub is offered as a service on SAP Cloud Platform whereas SAP Data Intelligence is offered as a BYOL (Bring Your Own License) product. With the BYOL licensing model, your enterprises can use their licenses flexibly (on-premise or in the cloud) unlike traditional restrictive software licensing policies where software is tied to a specific server (e.g., on-premise) and can't be reused for deployments in the cloud without violating the licensing agreement. With true license mobility, businesses can migrate to and from services with ease—without having to worry about managing multiple licenses for the same product across multiple platforms and services.

### 5.3.2 Architecture

Our discussion so far established the basic concepts, but now is the time to take a step further and look at the components found in SAP Data Intelligence’s architecture, as shown in Figure 5.5:

- **Connecting data sources**
  Connecting remote systems or data sources is supported by a robust set of connection types, including ABAP, SAP Cloud Platform Integration, Azure Data Lake, IBM DB2, Hadoop Distributed File System (HDFS), OData, IMPAP, Google Cloud Storage, and many others. SAP Data Intelligence connection management also lets you administer connections including the ability to create, update, delete, and monitor connections. With its predefined connectors for standard systems, SAP Data Intelligence allows you to access various data sources and manage those remote systems in a central place.

- **Metadata and governance**
  A set of SAP Data Intelligence capabilities to help govern and manage metadata assets that are spread across diverse systems and disparate sources. With the Metadata Explorer, enterprise users are empowered with set of tools and services to discover reliable datasets along with rule-driven data quality monitoring and self-service data preparation capabilities. Business users and analysts can leverage the data pipelining capability in an easy-to-use spreadsheet UI to enhance and enrich datasets to extract value out of data that may reside in different silos.

- **SAP Data Intelligence modeler**
  SAP Data Intelligence modeler, designed on the SAP Data Intelligence pipeline engine, helps you create data processing pipelines or graphs and provides a runtime and design-time environment for data-driven scenarios. The tool reuses existing code libraries to orchestrate data processing in distributed landscapes. The SAP Data Intelligence modeler, with its robust set of operators, can help users define graphs, including non-terminating, non-connected, or cyclic graphs. The SAP Data Intelligence launchpad offers access to the SAP Data Intelligence modeler so you can create your own data processing pipelines. The SAP Data Intelligence modeler can help you design and execute your data-driven use cases.

- **Intelligent processing**
  Intelligent processing components are all about the platform’s machine learning capabilities, which is founded on a combination of machine learning scenario manager, the machine learning data manager, AutoML, and open source tools. Out of these tools, the machine learning scenario manager is focused on organizing data science artifacts, managing related tasks through a single cockpit, and providing easy access to Jupyter Notebook. Named for a combination of programming languages—Julia (Ju), Python (Py), and R—Jupyter Notebook can add and deploy machine learning pipelines, while the machine learning data manager enables easy access to machine learning datasets whenever you need them. AutoML progressively automates the machine learning workflow so that your data scientists can benefit from the automation of tasks like data preparation, feature selection, model choice, model parameter optimization. In this way, manual effort and the likelihood of error can be reduced. Finally, since open source tools like Jupyter, R, Python, and Spark are popular among data scientists, the platform can be leveraged to enable data scientists to streamline their work and facilitate communication between teams via a flexible and highly interactive platform that supports multiple programming languages.

![Metadata and governance](image1)

```
Metadata and governance
Data discovery | Data profiling | Metadata cataloging
```

![Intelligent processing](image2)

```
Intelligent processing
Notebook | Content | Machine learning deployment | AutoML | BYOL
```

![Data orchestration and monitoring](image3)

```
Data orchestration and monitoring
Connection management | Workflows | Scheduling | Machine learning tracking | Metrics
```

![Data pipelining](image4)

```
Data pipelining
Data ingestion | Data processing | Data enrichment | Operators
```

Figure 5.5 SAP Data Intelligence: Components
Data orchestration and monitoring

SAP Data Intelligence empowers enterprise users to discover, refine, govern, and orchestrate numerous types of data from various enterprise sources and ensure business transformation through managed interfaces, open source integration, cloud-based and on-premise deployments, and third-party interfaces. Out of the data orchestration and monitoring portfolio, machine learning tracking can take care of managing machine learning experiments while also preserving the results of machine learning experiments as performed by the data scientist team. Thus, machine learning tracking can be an essential unit in any machine learning project by ultimately helping data scientists track code, metadata, inputs, and outputs; easily reproduce results when needed; and finally determine which model should be promoted to production. One the other hand, the metrics explorer enables the visualization of machine learning runs, including comparison metrics between different runs belonging to one or more scenarios and the parameters captured as part of the experimentation and training phase either using the tracking SDK or the submit metric operator. In this context, runs are instances of a machine learning experiments, consisting of metrics and parameters from one iteration, along with the associated tags and labels.

Data pipelining

The core of the SAP Data Intelligence platform, data pipelining enables you to build modular data pipelines connecting disparate sources and apply a wide gamut of logical operators to dive more deeply into and extract insights out of enterprise data in ways never imagined before. You can execute data orchestration processes with modular data pipelines to scale up and down according to demand and deliver trustworthy data to the right users with the right context at the right time.

Bring Your Own License Model

SAP Data Intelligence is available with the Bring Your Own License (BYOL) model and can be deployed on-premise; in a public cloud (for instance, from Amazon, Google, or Microsoft); or in a private cloud running on Kubernetes compute clusters. From a release point of view, SAP Data Intelligence follows a frequent release cycle in an “as-a-service” version specifically for new machine learning-specific features.

SAP Data Intelligence applications

SAP Data Intelligence applications encapsulate various tools and applications that are accessible through the SAP Data Intelligence launchpad and offers a comprehensive set of features to execute an end-to-end data science execution cycle. The main entry point to SAP Data Intelligence is the SAP Data Intelligence launchpad, which allows you to launch numerous SAP Data Intelligence applications including the following:

- **SAP Data Intelligence connection management** to configure connections to remote destinations
- **SAP Data Intelligence database tools** for accessing the SAP Vora database
- **SAP Data Intelligence diagnostics** for troubleshooting along with access to logs, traces, and metrics
- **SAP Data Intelligence metadata explorer** useful for browsing, profiling, crawling, and organizing the technical metadata of connected systems
- **SAP Data Intelligence modeling**, the modeling environment for data pipelines using SAP Data Intelligence pipeline engine
- **SAP Data Intelligence monitoring** for monitoring SAP Data Intelligence pipelines
- **SAP Data Intelligence system management** for managing the SAP Data Intelligence system

A simple, browser-based, and easy-to-access launchpad enables you to connect all the necessary data sources in your landscape, extract the metadata from your connected sources, build a catalog of metadata, explore and search metadata in your catalog, perform machine learning experiments with tracking, and much more. Additionally, the launchpad can be personalized based on roles so that users can be assigned applications as favorites or be presented with applications displayed in groups.

Phew! This list is quite comprehensive, isn’t it? Overall, SAP Data Intelligence is the “umbrella” that encompasses the integration of all the solutions, data, and processes involved in SAP’s overall approach to big data.

5.3.3 Values of SAP Data Intelligence

Before dive more deeply into the individual features available with SAP Data Intelligence, let’s quickly go through a list of values offered by SAP Data Intelligence from an enterprise AI perspective, which includes the following:

Connecting data silos

SAP Data Intelligence is designed to connect and bring together the world of structured and unstructured data. SAP Data Intelligence’s unique integration with various enterprise data sources like SAP BW4/HANA and SAP Cloud Analytics can become a game-changer for your enterprise by helping you realize the potential value of your integrated data sources. Out of the box support for the widest range of connectors includes SAP HANA, SAP Vora, Object Stores (S3, GCS, etc.), HDFS servers, SAP Business Warehouse (SAP BW), Oracle, Microsoft Windows Azure Storage Blobs, SAP Cloud Platform Integration, Google Cloud BigQuery, SAP IQ databases, and many more.

Scaling AI across the enterprise

By moving AI projects from the proof of concept (POC) stage to production and creating an AI assembly line to better manage the data science, machine learning,
AI processes, bringing together IT team members and data scientists is now a reality with SAP Data Intelligence.

- **Extracting value from distributed data**
  Crossing every system blockage, even for data spread across a complex landscape, you can orchestrate and process this data using the most suited engine for each variation of data, thus minimizing the amount of data being moved while also connecting data that may be held in data silos.

- **Embracing open source technologies**
  SAP Data Intelligence is laid on a foundation of open source and data science principles and lets you monetize your enterprise data using open source technologies including Docker containers; orchestration by Kubernetes; and native support for SAP, open source, and third-party machine learning frameworks, such as TensorFlow, R, Python, Spark, etc.

- **Scenario-driven model building**
  A unique approach that uses the concepts of machine learning scenarios to unite several artifacts belonging together into a virtual container—scenario and holds machine learning experiment data, documentations, experimental runs, model key performance indicators (KPIs) and API deployments in one place.

- **Ensuring MLOps**
  A unified platform enabling MLOps facilitates processes for operationalizing data science and machine learning solutions using code and best practices that promote efficiency, speed, and robustness while establishing seamless collaboration between data scientists, data engineers, and cloud engineers and building upon iterative feedback by stakeholders.

- **Leveraging the experimental phase**
  Harness the usage of Jupyter Notebook in data-driven applications by benefitting from the seamless integration of the JupyterLab environment and the Pipeline Modeler application.

- **Managing everything at once**
  A single interface to view and orchestrate all data-driven applications and data flows across the organization with tools to help manage governance, auditability, and transparency.

### 5.4 Key Capabilities

Now that you have an idea of the overall features available with SAP Data Intelligence, you should see its value as a complete data science platform. In this section, we’ll focus on the specific capabilities offered by SAP Data Intelligence and look at how key data science-oriented capabilities can empower you to bring enterprise intelligence into action.

#### 5.4.1 Metadata and Governance

A deep understanding of metadata and its efficient use is key towards building intelligent enterprise solutions, and SAP Data Intelligence eases the process of metadata exploration and processing. Your metadata assets originate from the related source systems and can be made available in a catalog called the SAP Data Intelligence Metadata Explorer, which includes functionalities for labeling data, searching data, and tracking data lineages. Using the Metadata Explorer, you can gather information about the various aspects of your data including locations, attributes, qualities, and the sensitivity of the data, which ultimately can help you make informed decisions about publishing datasets, including accessibility. A self-learning metadata and governance tactic can take care of metadata consistency, accuracy, and completeness by applying unified metadata cataloging and also provides better visibility of data assets across the landscape along with pervasive data consumption and improved user experiences. The metadata management functionalities enable you to discover, understand, and consume information about your data with the ability to synchronize, share, and perform version lineages and impact analysis. In these ways, you can get answers to related information requests without browsing through multiple systems or repositories or touching various data models. Figure 5.6 shows a snapshot of the SAP Data Intelligence Metadata Explorer, which is equipped with the following capabilities:

- Features to connect to different enterprise data sources. You can also automatically crawl through the metadata structures and create references on your data to store them in the Metadata Catalog.

- Along with a set of services, the Metadata Explorer includes browse, fact sheet, preview, profile, publish, rules, and data preparation functions for sources like IBM DB2, Oracle MySQL, Oracle, MS SQL Server, and MS Azure Cloud SQL.

- Rule-based data quality management to better understand the quality of data and improve data through data preparation and modeling. This kind of parameterized, dynamic rule lets you define data quality thresholds and monitor incoming data using scorecards driven by key data quality KPIs. Also, data quality rules can be easily reused: Simply bind these rules to new data coming from various other sources.

- Self-service and data-driven data preparation is another key feature offered as part of data management and is primarily targeted for business users willing to transform, shape, harmonize, curate, and enrich the datasets based on different business scenarios and project requirements, sometimes with a single click.

- A sophisticated business glossary and catalog editor to define terms and associate these terms with related Metadata Explorer objects, which allows you to search for terms and view their relationships.

- Ability to handle large-scale enterprise deployments with holistic lifecycle management, elasticity, low total cost of ownership (TCO), and resilience.
5.4.2 Machine Learning Scenario Manager

After tailoring and making the dataset ready for the business scenario, now it’s time for data scientists to focus on the business problem they are trying to solve, and SAP Data Intelligence’s machine learning scenario manager is a central place to deal with versioning, connecting datasets, building models, deployments, monitoring model performance and lifecycle management, ultimately freeing up more time for the IT team and data scientists to focus on more productive work. Besides serving as a unified platform for modeling machine learning scenarios, SAP Data Intelligence also fulfills a key demand from data scientists seeking a computational notebook for practicing machine learning experiments, sharing code, and quick prototyping. The SAP Data Intelligence platform leverages Jupyter Notebook, which is a free, open source, interactive web-based computational notebook. Even though various computational notebooks have been around for several years, Jupyter has exploded in popularity recently, enthusiastically adopted by the data science community, and is now the default environment for research.

As a unified workspace, the machine learning scenario manager is capable of dealing with all machine learning models across the organization including the management of datasets, notebooks, and pipelines in a single interface. According to SAP’s official definition, a machine learning scenario is a collection of design-time elements, such as pipelines and notebooks in Jupyter Notebook, and runtime elements such as training runs, models, and model deployments.

Built around the key concepts of machine learning, the machine learning scenario manager, as shown in Figure 5.7, lets you deal with data science artifacts and manage related tasks with the following features:

- The machine language scenario manager in SAP Data Intelligence can serve as a central place to access the necessary tools and services for data scientists and machine learning experts to play with data and execute machine learning experiments.
- A single cockpit, covering a wide range of data science capabilities including dataset binding, performing experiments in Jupyter Notebook, and automating model training and deployment via templated pipelines, ultimately allowing for seamless transitions from prototyping phases into production.
- The multifaceted machine learning scenario manager unites various artifacts into a virtual container that is built on a foundational concept—the machine learning scenario, which ultimately holds all experiment data, scenario versions, model performance metrics, deployment history documents, experimental runs, model KPIs, and API deployments in a single place. A machine learning scenario is the ultimate entry point for kickstarting machine learning projects using the SAP Data Intelligence platform.
- Embedded inside machine learning scenario manager, a modeler tool based on the SAP Data Intelligence pipeline engine uses a flow-driven programming paradigm. Thus, users can freehand model data processing pipelines as computation graphs to achieve the required data ingestion and transformation capabilities. Inside a computation graph, nodes represent operations on the data and edges represent data flows.
- A pipeline represents a concrete and complex data flow and helps transform data between elements connected in a series. Executing a data pipeline can help process raw data from multiple sources and make this data available for different use cases.

Figure 5.6 SAP Data Intelligence: Metadata Explorer

Figure 5.7 SAP Data Intelligence: Machine Learning Scenario Manager
5.4.3 Machine Learning Data Manager

Data scientists devote a substantial amount of effort and time preparing the datasets they need to experiment with different machine learning scenarios. In the process of determining which class of models to use and what types of features to include, these experiments produce a number of different artifacts. Without a standardized way to manage the resulting artifacts, data scientists can find reproducing their analyses and comparing the results of their experiments difficult. Moreover, since each experiment cycle requires a considerable amount of time, the data science team will want to preserve the experiment’s result for future reference. The machine learning data manager in SAP Data Intelligence helps you store the necessary datasets into data workspaces and collections, so that they can be easily accessed whenever needed. According to the official definition, a data workspace groups together any number of data collections, which are logical groups of data consumed or produced by a notebook or a pipeline during the execution of a machine learning scenario. Figure 5.8 shows a quick snapshot of machine learning data manager. The key features of machine learning data manager include the following capabilities:

- A unique set of features lets you organize machine learning datasets into a folder-based hierarchy combining data workspaces and data collections, along with relevant metadata like features and lineage. These data collections can be easily consumed as part of an AutoML experiment or programmatically in a Jupyter Notebook.
- As a unique framework, the machine learning data manager enhances the traceability and reproducibility of each pipeline correlated to machine learning use cases by organizing the relevant data into workspaces and data collections.
- To preserve the outcome of different machine learning experiments, the machine learning data manager offers provisions for transforming your data into clean and well-structured datasets that can be written to another data collection within the workspace and supports relationship identification through hierarchy. Unique machine learning functionalities like feature modeling and splitting also allow the segregation of data into train, test, and validation datasets.
- The machine learning data manager also supports lineage analysis, where you can trace your data from a dataset to its source easily. You can also navigate through various dependencies between objects, for instance, whether the data has been transformed or where the data originated and how the dataset may have been modified throughout.

Now that you know how to manage your machine learning experiment datasets, let’s now focus on one of the most interesting and offbeat features offered by SAP Data Intelligence, called AutoML. Every day, this feature is becoming more and more popular, even among data scientists.

5.4.4 AutoML

For enterprises to infuse intelligence into their business process, AutoML is a critical step to lowering the barriers of entry into data science and to opening the way for data-driven automation at scale. In recent years, interest in adopting AutoML tools to automate several tasks in the data science workflow has spiked. In contrast to the typical automated machine learning features found in various tools, as shown in Figure 5.9, AutoML categorically automates various tasks in the machine learning workflow, including acquiring and preparing data; engineering features from the data; picking the best algorithms; fine-tuning algorithms; and deploying and monitoring production models. With a typical machine learning application, development involves many steps, such as data preparation, feature selection, model choice, and model parameter optimization, but with AutoML, all these steps are automated intelligently, which reduces the manual effort required of users and the risk of error. If you don’t want to govern or control data science experiments that you typically perform with the machine learning scenario manager, you can easily opt for AutoML to generate machine learning pipelines automatically.
5.4.5 Machine Learning Tracking

Why is machine learning experiment tracking vital? Data science processes are generally research-centric work; often, multiple parallel experiments are performed, many of which might not ever make it into production. While working on a machine learning project, getting good results from a single model training run is rare that keeping all your machine learning experiments organized can be a real challenge. Having a process that lets you draw valid conclusions out of various experiments is quite vital, which is where machine learning experiment management becomes a key factor in delivering successful outcomes. Day-to-day effort with machine learning models repeatedly boils down to a sequence of experiments before picking a modeling approach, which combines network topology, training data, and various optimizations or tweaks to the models. Also, enterprise users, like machine learning engineers and data scientists, tend to work alone, which gives the false illusion that keeping track of the experimentation process is not important so long as a final model is delivered. Problems arise when you need to go back to an earlier idea, re-run a model that you may have experimented with only months ago, or simply compare and visualize the differences between runs. With these cases, the need for a system or tool for tracking machine learning experiments is (painfully) apparent. Machine learning tracking allows data scientist teams to see what others...
are doing; share ideas and insights; and store experiment metadata, which can be retrieved and analyzed at any time. Teamwork can be much more efficient, situations where several people work on the same task can be prevented, and the onboarding of new members can be made much more smooth.

The tracking service on the SAP Data Intelligence platform offers a complete framework for data scientists to configure, capture, organize, reproduce, and visualize machine learning experiments. With provisions for capturing multiple aspects of tracking, data scientists can stipulate what information must be captured as part of the experiment (e.g., model parameters, model metrics, model hyperparameters, etc.) and can later view this information through the Metric Explorer within the machine learning scenario manager. Machine learning tracking on the SAP Data Intelligence platform is primarily driven by the tracking SDK, and the required modules, methods, and scripts can be incorporated into experiments in Jupyter Notebook. But before we discuss experiment tracking in more detail, you must first understand some key entities related to tracking, as shown in Figure 5.11, which we’ll describe next.

Let’s review the key entities you’ll need understand to track your machine learning models, such as the following:

- **Run**
  A run is an instance of a machine learning experiment. Each run consists of metrics and parameters from one iteration of the machine learning experiment, along with the associated tags and labels. Metrics are used to measure the performance of your machine learning models. Different metrics are used to evaluate different machine learning models. For example, log loss or average accuracy could be used to evaluate a classification task, whereas mean absolute error or root mean square error (RMSE) could be used to evaluate a model that provides a numeric output.

- **Run collection**
  A run collection allows you to logically group a number of runs. You can generate run collections by using the tracking functionality of the SAP Data Intelligence Python SDK. When you log runs using the tracking functionality, you can specify a name for your run collection. If you don’t specify a name, the system creates a default run collection for each source (that is, notebook or pipeline). The name of the default run collection is the same as that of the source. These default run collections are shown in the metric explorer with the suffix (`Default`).

- **Visual board**
  The visual board is the central place for comparing different runs sourced from one or more run collections and is generated using runs that are called source runs. Data from source runs is shown in tabular format under the Data tab of the visual board.

- **Canvas and charts**
  The canvas is the part of the visual board that lets you add charts to compare source runs graphically. For example, to plot a metric against a run ID, parameter, or timestamp, you can create charts from the metrics and parameters of source runs and compare this information visually using the canvas.

Figure 5.12 shows a glimpse of SAP Data Intelligence’s monitoring capabilities. Now that you have a pretty good understanding of the measurement units behind machine learning tracking, let’s learn how you can implement the machine learning tracking capability.
Getting started with experiment tracking using Jupyter Notebook is simple. The following code snippet will automatically include all the necessary modules required for experiment tracking:

```python
from sapdi import tracking

After the module-importing statement, you’ll use the following code snippets to mark the beginning and the end of the run:

```
tracking.start_run(run_collection_name='runcollect1')
tracking.end_run()
```

In this case, `start_run()` initializes the tracking run under the provided run collection called ‘runcollect1’ of type string, which basically allows you to group together different runs of your experiment. Furthermore, when the run is initialized via `start_run()`, the metadata information, such as the Scenario Details, Source Details, Start Time, etc., is automatically recorded.

To log the execution-related metrics, generate list of metrics, or log execution-related parameters, respectively, you can use the following methods:

- `log_metric`
- `log_metrics`
- `log_parameters`

The following code snippet shows you comprehensively how to embed machine learning tracking scripts inside a pipeline:

```python
from sapdi import tracking
run = tracking.start_run()
# tracking.start_run(run_collection_name='rc1')
# Use this to group different runs under a collection
params = {
    'input_size': 784,
    'hidden_size': 500,
    'num_classes': 10,
    'num_epochs': 2,
    'batch_size': 100,
    'learning_rate': 0.001,
    'model_file': 'model.ckpt'
}
tracking.log_parameters(params)
```

Now that we’ve covered machine learning tracking, in the next section, we’ll introduce you to the SAP Data Intelligence modeler and explore its operator-driven capabilities for building graphs for your business processes.

### 5.4.6 SAP Data Intelligence Modeler and Built-in Operators

Designed on the SAP Data Intelligence pipeline engine, the SAP Data Intelligence modeler delivers a flow-based programming paradigm for data scientists to create data processing pipelines modeled as computation graphs and executed in a containerized environment that runs on Kubernetes. The modeler application helps you model graphs using a graphical interface and achieve the required data ingestion and transformation capabilities. The modeler comes with 250+ built-in operators that can be directly used to build graphs for your business processes or can be used as the basis for a custom operator to design powerful data pipelines. Using the modeler, you can easily design your data processing pipelines using existing graphs or the available operators.

As shown in Figure 5.13, on the left side of the application, you can access all the objects you’ll need to interact with to create your pipelines, while the main screen stores the pipelines and allows you to execute and debrief the results.

Inside the modeler, as shown in Figure 5.14, a pipeline or graph is a network of operators connected to each other using typed input ports and output ports for exchanging data. In this context, an operator is a reactive component representing a vertex of a graph and reacts only to events or messages delivered to the operator through its input ports.
Since the operator is totally unaware of the graph in which it is defined and unfamiliar with the source and target of its incoming and outgoing connections, the operator can only interact with the environment through its output ports. Figure 5.15 shows a sample operator with its input ports and output ports. Each port is associated with a port type, and the tool uses color coding to identify compatible port types.

![Figure 5.15 Operators](image)

An operator is associated with a port of any data type, such as string, blob, int64, float64, byte, message stream, and so on. Operators require a certain runtime environment for their execution, and the SAP Data Intelligence Data modeler provides predefined environments for operators. These environments are made available to users as a library of Docker files. When you execute a pipeline (a graph), the tool translates each operator in the graph into a process. The modeler then searches the Docker files for an environment suitable for the operator’s execution and instantiates a Docker image of that environment. This Docker image and the operator process are executed on a Kubernetes cluster. Predefined operators within the SAP Data Intelligence modeler can be classified into various categories, which we’ll describe next.

**Converter**

A handful of converters are grouped under this category, as shown in Figure 5.16, and these operators support the different types of conversions often required to transform your business data in various ways, for instance, StreamToString, StringToStream, ToBlob, ToStringType, ToString, ToNumber, and ToMessage.

![Figure 5.16 Converters](image)

**Data Quality**

This key set of operators, as shown in Figure 5.17, is focused on managing data quality, including masking data, cleansing data, gaining statistical insights, applying data validation rules, and more.
Data Workflows
These are dedicated operators for transforming and transferring data; managing workflows including merging, splitting, and triggering workflows; improving data quality using jobs in SAP Data Services; and sending notifications at certain points in the workflow.

Files
As shown in Figure 5.18, file management is easy with this set of file handling operators, including the ability to monitor, read, write, remove, and list files.

Hadoop and Spark
Figure 5.19 shows a set of dedicated operators for managing Hadoop jobs, especially for submitting jobs to Hadoop and clusters using the Apache Livy API, a Representational State Transfer (REST) API provided by different cloud providers. Apache Livy is a service that enables easy interaction with a Spark cluster over a REST interface. With Apache Livy, you can easily submit Spark jobs or snippets of Spark code and retrieve results synchronously or asynchronously. In this set, you can also access Spark Context management, all via a simple REST interface or a remote procedure call (RPC) client library.

Processing
This unique set of operators handles common requirements within graphs, like command execution; constant generation; random data generation at certain intervals; inclusion of Go, JavaScript, Python, and R code snippets; and more. Figure 5.22 shows a quick snapshot of processing-related operators.
5 Machine Learning with SAP Data Intelligence

SAP Integration
As shown in Figure 5.23, this set of operators are focused on data integration, taking care of the sending, receiving, and formatting of data to/from SAP Analytics Cloud and SAP Cloud Platform Integration.

![SAP Integration Operators](figure5_23.png)

Figure 5.23  SAP Integration Operators

Machine Learning Core Operators
These dedicated operators are only for creating machine learning pipelines and supporting capabilities, such as model training and serving, and integrates with the scenario management features of the platform. Figure 5.24 shows a quick snapshot of processing-related operators.

![Machine Learning Core Operators](figure5_24.png)

Figure 5.24  Machine Learning Core Operators

We’d like to focus on a particular core operator called functional services. With SAP Leonardo Machine Learning Foundation now being dissolved into SAP Data Intelligence, Functional Services are longer available separately and instead is included as an operator as part of the pipeline modeler, as shown in Figure 5.25. Currently, the Functional Services operator offers a rather limited set of pretrained image and text processing-oriented functions, including image classification, feature extraction, OCR, similarity scoring, text classifier, and topic detection.

![Functional Services Operators](figure5_25.png)

Figure 5.25  Functional Services Operators

SAP Vora
As shown in Figure 5.26, these dedicated operators handle SAP Vora database-related operations, including the reading, ingestion, and loading of data and includes a message decoder.

![SAP Vora Operators](figure5_26.png)

Figure 5.26  SAP Vora Operators

Structured Data Operators
These operators, shown in Figure 5.27, handle structured data coming from various data sources like Azure SQL DB, DB2, Google BigQuery SQL, Microsoft SQL Server, MySQL, Oracle, Redshift, and SAP IQ databases.

![Structured Data Operators](figure5_27.png)

Figure 5.27  Structured Data Operators

Connectivity
This set of operators primarily takes care of connectivity through HTTP requests and providing WAMP record access and includes the ability to consume or send records from Kafka clusters. You can also send emails; replicate tables; configure MQTT-based connectivity; configure connections Microsoft Azure, Google Cloud Platform, or Amazon Web Services (AWS); and more, as shown in Figure 5.28.

![Connectivity Operators](figure5_28.png)

Figure 5.28  Connectivity Operators
Tables
This set of operators, shown in Figure 5.29, handles the generation of table messages based on a user-defined schema, the generation of random data, and the decoding of tabular data.

Figure 5.29 Tables Operators

Text Analysis
This unique set of operators, shown in Figure 5.30, handles natural language processing (NLP) for unstructured textual data by using the text analysis capabilities of SAP Vora.

Figure 5.30 Text Analysis Operators

Utilities
As shown in Figure 5.31, these operators deal with the conversion, aggregation, defining scripts, serializing Python data, application logging, and compiling flowgraphs along with many more helpful operators.

Figure 5.31 Utility Operators

Figure 5.32 shows the overall category of operators supported by the SAP Data Intelligence platform.

5.5 Migrating to SAP Data Intelligence from SAP Data Hub
With the release of SAP Data Intelligence 3.0 and the lack of new functionalities planned for release after SAP Data Hub 2.7, perhaps migrating to SAP Data Intelligence from SAP Data Hub may seem unnecessary. According to SAP, with the release of SAP Data Intelligence 3.0, SAP Data Hub customers will automatically receive SAP Data Intelligence, and no additional migration effort is needed. Existing customers will now have access to all cloud-based machine learning tools as part of support packages without any additional fee or migration effort required. You’ll still be at liberty to select the deployment model of your choice—whether on-premise, in the cloud, or a combination of the two. In addition to SAP Data Hub, SAP Leonardo Machine Learning Foundation has also been dissolved into the SAP Data Intelligence platform since that solution...
will no longer be available as a standalone service for new subscriptions and can only be consumed via SAP Data Intelligence. As described throughout this chapter, SAP Data Intelligence offers all the integration, orchestration, metadata management, connectivity, and other rich services of SAP Data Hub with the services of SAP Leonardo Machine Learning in the cloud, along with a lot more new features.

In case you still don’t have access to an SAP Data Intelligence tenant and are really eager to get your hands dirty, you can start immediately by signing up for an on-premise trial edition of SAP Data Intelligence 3.0, which is available for Google Cloud Platform, Amazon Web Services (AWS), and Microsoft Azure, through the SAP Cloud Appliance Library, as shown in Figure 5.33.

Figure 5.33 SAP Data Intelligence Trial Edition 3.0

5.6 Summary

Finally, we’ve concluded our chapter on SAP Data Intelligence, SAP’s data science platform that offers a coherent machine learning and data science foundation that can lead your business processes to evolve into an intelligent enterprise. The name SAP Data Intelligence highlights the solution’s focus on intelligent data processing and the operationalization of machine learning processes in enterprise landscapes. This combination allows you to get the most out of your disparate data assets, which can be easily integrated, intelligently refined, and seamlessly used in your business processes. Data scientists can continue to work with the tools they know and love while IT focuses on prototyping machine learning models to put into production and ensures that innovation and machine learning projects can scale and be properly managed. Accelerating the efficient delivery of your intelligent data with innovative data pipelines can enable smart customer experiences. Knowing your data and improving user experiences with metadata management strategies can improve data visibility and enable pervasive data consumption, thus bringing people, processes, and technologies together to drive enterprise-level AI and more importantly gain business insight. A key point of SAP Data Intelligence is providing you with the freedom to choose the deployment model of your choice—whether on-premise, in the cloud, or a combination of the two.
Contents

Preface ........................................................................................................................ 15

PART I Introduction

1 Machine Learning and Intelligent Enterprise ......................................................... 23
  1.1 What Is Machine Learning? ................................................................................... 25
  1.2 Transition from the Digital Era to the Intelligent Era .................................. 25
  1.3 Intelligent Enterprise Use Cases ......................................................................... 26
    1.3.1 Intelligent Invoice Matching ................................................................. 26
    1.3.2 Smart Recruiting .................................................................................... 27
    1.3.3 Service Ticket Intelligence ................................................................. 27
    1.3.4 Product Similarity Matching ................................................................. 27
    1.3.5 Brand Impact ......................................................................................... 28
    1.3.6 Robotic Process Automation ............................................................... 28
    1.3.7 Virtual Assistance .................................................................................. 29
  1.4 SAP's Intelligent Enterprise Strategy ................................................................. 29
    1.4.1 Intelligent Suite ....................................................................................... 30
    1.4.2 Digital Platform ....................................................................................... 31
    1.4.3 Intelligent Technologies .......................................................................... 31
  1.5 SAP's Machine Learning Technologies and Applications ............................. 32
    1.5.1 Embedded Machine Learning Applications ........................................ 32
    1.5.2 SAP Intelligent Robotic Process Automation .................................... 33
    1.5.3 SAP Conversational Artificial Intelligence .......................................... 34
    1.5.4 SAP Data Intelligence .......................................................................... 34
    1.5.5 Embedded Machine Learning within SAP HANA ................................ 35
  1.6 Summary ........................................................................................................... 36

2 Machine Learning Fundamentals .......................................................................... 37
  2.1 Basic Probability Concepts ................................................................................ 37
    2.1.1 What is Probability? ................................................................................ 38
    2.1.2 Conditional Probability and Independence ........................................... 47
## 2 Random Variables
- Covariance and Correlation 62
- Variance and Standard Deviation 61
- Expectation 60
- Density and Distribution Function 54
- Random Variables 49

## 2.1 Basic Machine Learning Concepts
- Different Types of Machine Learning Scenarios 63
- Types of Learning 65

## 2.2 Machine Learning Algorithms
- Linear Regression Analysis 66
- Classification 85
- Cluster Analysis 113
- Other Machine Learning Techniques 125

## 2.3 Machine Learning Concepts
- Density and Distribution Function 54
- Variance and Standard Deviation 61
- Expectation 60
- Covariance and Correlation 62

## 3 Implementation Lifecycle
- Understanding the Implementation Lifecycle 140
- Define, Measure, Analyze, Improve, Control 141
- Analytics Solutions Unified Method for Data Mining/ Predictive Analytics 142
- Key Components in a Machine Learning Project Lifecycle 142

## 3.1 Understanding the Implementation Lifecycle
- Cross Industry Standard Process for Data Mining 140
- Analytics Solutions Unified Method for Data Mining/ Predictive Analytics 142

## 3.2 Knowing the Business
- Design, Measure, Analyze, Improve, Control 143

## 3.3 Understanding and Exploring Data
- Checking Data Quality 147
- Checking Summary Statistics 151
- Visualizing Data 153

## 3.4 Preparing Data
- Data Conversion 157
- Dimensionality Reduction 160
- Variable Transformation 161

## 3.5 Developing the Model
- Model Overfitting 165
- Model Performance 166

## 3.6 Evaluating and Fine-Tuning Model
- Model Fine-Tuning 168
- Diagnostic Checks 168
- Validation Set Approach 170
- K-Fold Cross Validation 171

## 3.7 Deploying the Model
- Summary 173

## 4 Machine Learning on SAP HANA
- SAP HANA Machine Learning Components 175
- SAP HANA Predictive Analysis Library 177
- SAP HANA Automated Predictive Library 181
- SAP HANA Extended Machine Learning Library 184
- R Integration 188

## 5 Machine Learning with SAP Data Intelligence
- Data Science Project Lifecycle 207
- Managing the Data Science Project Lifecycle 209
- SAP Data Intelligence 210
- SAP Data Intelligence Versus SAP Data Hub 211
- Architecture 212
- Values of SAP Data Intelligence 215

## 5.1 Data Science Project Lifecycle
- Metadata and Governance 216
- Machine Learning Scenario Manager 218
- Machine Learning Data Manager 220
- AutoML 221
- Machine Learning Tracking 223
- SAP Data Intelligence Modeler and Built-in Operators 227

## 5.2 Managing the Data Science Project Lifecycle
- Migrating to SAP Data Intelligence from SAP Data Hub 235

## 5.6 Summary
- Model Fine-Tuning 168
- Diagnostic Checks 168
- Validation Set Approach 170
- K-Fold Cross Validation 171
- Summary 173
- SAP HANA Machine Learning Components 175
- SAP HANA Predictive Analysis Library 177
- SAP HANA Automated Predictive Library 181
- SAP HANA Extended Machine Learning Library 184
- R Integration 188
- Data Science Project Lifecycle 207
- Managing the Data Science Project Lifecycle 209
- SAP Data Intelligence 210
- SAP Data Intelligence Versus SAP Data Hub 211
- Architecture 212
- Values of SAP Data Intelligence 215
- Migrating to SAP Data Intelligence from SAP Data Hub 235
- Summary 236
PART II  Building Machine Learning Applications

6  SAP HANA Predictive Analysis Library and R Integration  

6.1  SAP HANA Predictive Analysis Library  
6.1.1  Prerequisites and Installation  
6.1.2  SAP HANA PAL Procedures  
6.1.3  Calling SAP HANA PAL Procedures  
6.1.4  Model Evaluation  

6.2  R Integration  
6.2.1  Prerequisites and Installation  
6.2.2  R Packages  
6.2.3  Calling R Functions from SAP HANA  
6.2.4  Model Evaluation  

6.3  Summary  

7  Developing Applications with SAP HANA Predictive Analysis Library  

7.1  Introduction to the Use Case  

7.2  Building a Predictive Analytics Application Using SAP HANA PAL  
7.2.1  Understanding the Business Problem and the Data  
7.2.2  Data Preparation  
7.2.3  Model Building  
7.2.4  Model Validation and Refinement  
7.2.5  Model Deployment  
7.2.6  Interpretation and Displaying Results  

7.3  Summary  

8  SAP AI Business Services  

8.1  Overview  

8.2  Document Classification  
8.2.1  Test Run of the Document Classification Service  

8.3  Document Information Extraction  
8.3.1  Test Run of the Document Information Extraction Service  

8.4  Business Entity Recognition  

8.5  Data Attribute Recommendation  
8.5.1  Test Run of the Data Attribute Recommendation Service  

8.6  Invoice Object Recommendation  

8.7  SAP Service Ticket Intelligence  
8.7.1  Test Run of the Service Ticket Intelligence Service  

8.8  Summary  

9  Building Scenarios Using Jupyter Notebook  

9.1  Adding a Notebook  

9.2  SAP Data Intelligence Python SDK  
9.2.1  Setting Up and Leveraging the SAP Data Intelligence SDK  
9.2.2  SAP Data Intelligence Python SDK  
9.2.3  Popular Python Libraries  
9.2.4  Various Library Options  

9.3  Use Case  
9.3.1  Setting Up the Scenario  
9.3.2  Scripting  
9.3.3  Data Plotting  
9.3.4  Code Execution  

9.4  Summary  

10  Automated Machine Learning Data Science Automation  

10.1  AutoML on SAP Data Intelligence  

10.2  Features of AutoML  

10.3  AutoML Step-by-Step  
10.3.1  Configuring the Training Data Collection  
10.3.2  Configuring the Test Data Collection  
10.3.3  Setting Up AutoML  
10.3.4  Testing AutoML  

10.4  Summary  
11 Conversational Artificial Intelligence 399

11.1 Introduction to SAP Conversational Artificial Intelligence 399

11.2 SAP Conversational AI 401
   11.2.1 Natural Language Processing Engine 402
   11.2.2 Intents 404
   11.2.3 Expressions 405
   11.2.4 Entities 406
   11.2.5 Languages 409
   11.2.6 Channels 412

11.3 Bot Building Techniques 412
   11.3.1 Bot Building Process 413
   11.3.2 Bot Connector 417
   11.3.3 Bot Monitoring and Analytics 418
   11.3.4 SAP Conversational AI APIs 419

11.4 Building a Chatbot Using SAP Conversational AI 421
   11.4.1 Let’s Build a Bot! 423
   11.4.2 Setting Up the Bot Project 424
   11.4.3 Creating Intents 425
   11.4.4 Setting Up Entities 427
   11.4.5 Setting Up Skills 429
   11.4.6 Enabling Voice 435

11.5 Summary 436

PART III Use Cases and Roadmaps

12 Integrating Machine Learning with the Internet of Things and Blockchain 439

12.1 Technology-Driven Transformation 441

12.2 Data—The Common Theme 442

12.3 Use Cases 445
   12.3.1 Supply Chain 445
   12.3.2 Healthcare 450

12.4 Summary 456

13 Industry Use Cases for Machine Learning Applications 457

13.1 Acceptance of Machine Learning across Different Industries 457
   13.1.1 Assessing the Impact of Machine Learning Applications 458
   13.1.2 Challenges of Machine Learning Applications 460

13.2 Machine Learning Ecosystem 461

13.3 Identifying Industry Use Cases 464
   13.3.1 Manufacturing Industry 465
   13.3.2 Customer Service Industry 469
   13.3.3 Healthcare Service Industry 473
   13.3.4 Finance Industry 476

13.4 Summary 480

14 Conclusion and Roadmap 481

14.1 Recap 481

14.2 Best Practices 483

14.3 Roadmap 484

14.4 Summary 486

The Authors 487

Index 489
Index

A
Action ........................................... 415
Activation function ......................... 108
Agglomerative clustering ................. 119
proportion measure ....................... 121
Aggregation .................................. 157
Akaike Information Criteria ............ 169, 295
Analysis Solutions Unified Method for
Data Mining/Predictive Analytics ...... 142
Application function library ............. 241
Application function modeler .......... 36
Artificial intelligence .................... 25, 399
Artificial neural network ................. 105
hidden layers ................................ 109
hidden nodes ................................ 109
model design ................................ 110
multilayer .................................. 109
multilayer training ....................... 111
Association analysis .................... 128
ASUM-DM .................................. 142
Attributes .................................. 146
categorical ................................ 147
continuous ................................. 147
discrete .................................... 147
interval .................................... 147
nominal ................................... 146
ordinal .................................... 147
ratio ....................................... 147
Autogenerated pipeline ................. 394
Automate quality inspection .......... 467
Automated machine learning .......... 375
Automated model training ............. 353
Automated predictive library .......... 358
AutoML ................................... 221, 375
application ................................ 385
authorization .............................. 385
cockpit .................................. 377
data science automation ............. 375
experiment target ....................... 388
features .................................. 376
metric .................................... 388
performance checkpoint ............... 390
prediction scenario ..................... 388
process cycle ............................ 376
AutoML (Cont.)
scenario ................................... 392
scenario popup ........................... 392
scenario section ......................... 394
scenarios list ............................. 386
setup ..................................... 385
task ....................................... 388
testing .................................... 395
AutoML → Automated machine learning
Auto-sklearn ................................ 375
Auto-WEKA ................................ 375
Azure Machine Learning .............. 25
B
Back-propagation ......................... 112
Bayes classifier pipeline ............... 392
Bayesian information criteria ........ 169
Best practices ............................ 483
Big data .................................. 442
Binarization ................................ 158
Blockchain .................................. 32, 439
data flow .................................. 444
use cases .................................. 445
Bot analytics ............................. 418–419
BOT API .................................. 419
Bot Builder ................................ 412–413
architecture .............................. 413
Bot Builder API ......................... 420
Bot building ............................. 417
custom code .............................. 417
process ................................... 413–414
user authentication ..................... 417
Bot Connector ........................... 416
architecture ............................. 417
Bot Connector API ..................... 421
channels .................................. 421
connectors .............................. 421
conversations .......................... 421
messages ................................ 421
participants ............................. 421
Bot monitoring ........................ 418
Brand impact ........................... 28
Bring Your Own License ............. 214
Budget ................................... 389
Business entity recognition .......... 339
Business entity recognition service .. 318
Laboni Bhowmik is an SAP manufacturing consultant working at IBM since 2007. She has more than 15 years of experience with SAP solutions and technologies, including SAP Manufacturing Integration and Intelligence, SAP Manufacturing Execution, SAP Overall Equipment Effectiveness Management, SAPUI5, and SAP Cloud Platform. Laboni has led several large-scale SAP manufacturing implementation projects for multiple global clients. She is also extensively involved in progressing IBM’s SAP Digital Manufacturing and SAP Cloud Platform competency, including designing cutting-edge industry solutions, development, publications, and mentoring.

Avijit Dhar is a senior consultant with IBM India. He has more than 15 years of experience in leading the design and implementation of multiple large-scale projects, particularly in the SAP manufacturing domain. He is also responsible for growing IBM’s SAP Internet of Things and machine learning practice.

Ranajay Mukherjee is a senior solution architect of SAP advanced analytics in IBM India. He has more than 21 years of experience in solutioning, leading, and delivering complex analytics projects across industries. Ranajay is a subject matter expert in data science, SAP HANA, SAP BW, and has extensively worked on the research, design, and development of machine learning and AI solutions using SAS, SPSS, R, and SAP HANA. He has held numerous leadership roles in IBM India, including SAP HANA Enterprise Practice Lead of IBM India and SAP BW powered by SAP HANA Conversion Factory Lead of IBM India.