**Pantau penyimpangan data dengan Azure Machine Learning**

Mengubah tren data dari waktu ke waktu dapat mengurangi keakuratan prediksi yang dibuat oleh model. Pemantauan penyimpangan data ini adalah cara penting untuk memastikan model Anda terus memprediksi secara akurat.

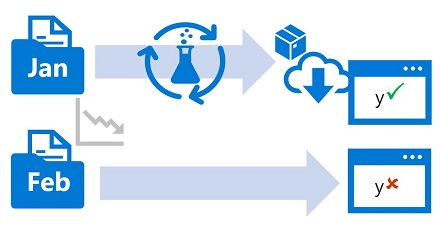
**Tujuan pembelajaran**

* Pelajari cara memantau penyimpangan data di Azure Machine Learning.

1. **Pengantar**

Anda biasanya melatih model machine learning menggunakan kumpulan data historis yang mewakili data baru yang akan diterima model Anda untuk inferensi. Namun, seiring waktu, mungkin ada tren yang mengubah profil data, sehingga membuat model Anda kurang akurat.

Misalnya, model dilatih untuk memprediksi jarak tempuh yang diharapkan dari sebuah mobil berdasarkan jumlah silinder, ukuran mesin, berat, dan fitur lainnya. Seiring berjalannya waktu, seiring dengan kemajuan teknologi mesin dan manufaktur mobil, efisiensi bahan bakar khas kendaraan mungkin meningkat secara dramatis; membuat prediksi yang dibuat oleh model yang dilatih pada data lama menjadi kurang akurat.



Perubahan profil data antara pelatihan dan inferensi ini dikenal sebagai **pergeseran data**, dan ini dapat menjadi masalah signifikan untuk model prediktif yang digunakan dalam produksi. Oleh karena itu penting untuk dapat memantau penyimpangan data dari waktu ke waktu, dan melatih kembali model yang diperlukan untuk menjaga akurasi prediksi.

1. **Tujuan pembelajaran**

Dalam modul ini, Anda akan mempelajari cara:

* Buat monitor penyimpangan data.
* Jadwalkan pemantauan penyimpangan data.
* Lihat hasil pemantauan penyimpangan data.

1. **Membuat monitor penyimpangan data**

Azure Machine Learning mendukung pemantauan penyimpangan data melalui penggunaan kumpulan data. Anda dapat menangkap data fitur baru dalam kumpulan data dan membandingkannya dengan kumpulan data yang digunakan untuk melatih model.

1. **Pantau penyimpangan data dengan membandingkan kumpulan data**

Sudah umum bagi organisasi untuk terus mengumpulkan data baru setelah model dilatih. Misalnya, klinik kesehatan mungkin menggunakan pengukuran diagnostik dari pasien sebelumnya untuk melatih model yang memprediksi kemungkinan diabetes, tetapi terus mengumpulkan pengukuran diagnostik yang sama dari semua pasien baru. Ilmuwan data klinik kemudian secara berkala dapat membandingkan kumpulan data baru yang terus bertambah dengan data pelatihan asli, dan mengidentifikasi tren data yang berubah yang mungkin memengaruhi keakuratan model.

Untuk memantau penyimpangan data menggunakan kumpulan data terdaftar, Anda perlu mendaftarkan dua kumpulan data:

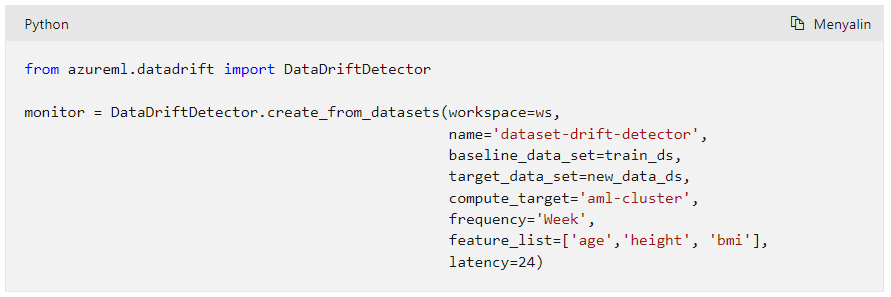
* Kumpulan data baseline - biasanya data pelatihan asli.
* Kumpulan data target yang akan dibandingkan dengan garis dasar berdasarkan interval waktu. Kumpulan data ini memerlukan kolom untuk setiap fitur yang ingin Anda bandingkan, dan kolom stempel waktu agar laju penyimpangan data dapat diukur.

**Catatan :**

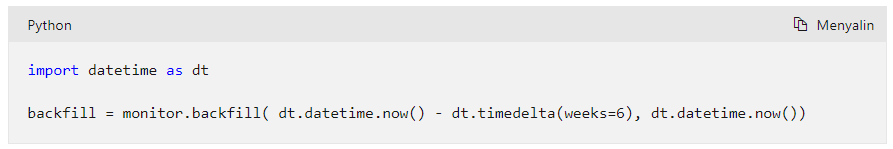
Anda dapat mengonfigurasi layanan yang disebarkan untuk mengumpulkan data baru yang dikirimkan ke model untuk inferensi, yang disimpan dalam penyimpanan gumpalan Azure dan dapat digunakan sebagai kumpulan data target untuk pemantauan penyimpangan data. Lihat **Mengumpulkan data dari model dalam produksi** di dokumentasi Pembelajaran Mesin Azure untuk informasi selengkapnya.

**Link** : <https://learn.microsoft.com/id-id/azure/machine-learning/v1/how-to-enable-data-collection>

Setelah membuat kumpulan data ini, Anda dapat menentukan monitor kumpulan data untuk mendeteksi penyimpangan data dan memicu peringatan jika tingkat penyimpangan melebihi ambang batas yang ditentukan. Anda dapat membuat monitor kumpulan data menggunakan antarmuka visual di studio Pembelajaran Mesin Azure, atau dengan menggunakan kelas **DataDriftDetector** di SDK Pembelajaran Mesin Azure seperti yang ditunjukkan dalam kode contoh berikut:



Setelah membuat monitor kumpulan data, Anda dapat mengisi ulang untuk segera membandingkan kumpulan data dasar dengan data yang ada di kumpulan data target, seperti yang ditunjukkan pada contoh berikut, yang mengisi ulang monitor berdasarkan perubahan mingguan dalam data selama enam bulan sebelumnya minggu:



1. **Peringatan penjadwalan**

Saat Anda menentukan monitor data, Anda menentukan jadwal yang harus dijalankan. Selain itu, Anda dapat menentukan ambang batas untuk laju penyimpangan data dan alamat email operator untuk pemberitahuan jika ambang batas ini terlampaui.

1. **Konfigurasikan jadwal monitor penyimpangan data**

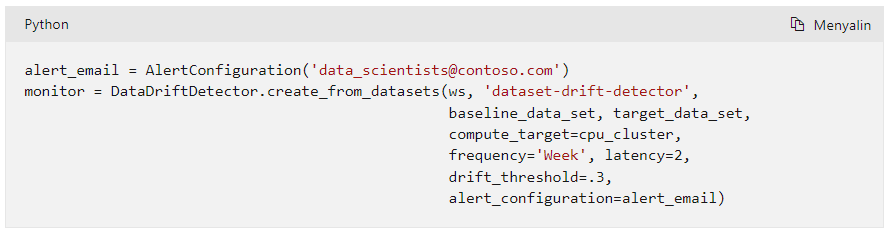
Pemantauan penyimpangan data bekerja dengan menjalankan perbandingan pada **frekuensi** terjadwal, dan menghitung metrik penyimpangan data untuk fitur dalam kumpulan data yang ingin Anda pantau. Anda dapat menentukan jadwal untuk dijalankan setiap H**ari, Minggu, atau Bulan.**

Untuk monitor kumpulan data, Anda dapat menentukan **latensi**, yang menunjukkan jumlah jam untuk memungkinkan data baru dikumpulkan dan ditambahkan ke kumpulan data target. Untuk monitor penyimpangan data model yang disebarkan, Anda dapat menentukan nilai waktu **schedule\_start** untuk menunjukkan kapan data drift run harus dimulai (jika dihilangkan, run akan dimulai pada waktu saat ini).

1. **Konfigurasikan peringatan**

Penyimpangan data diukur menggunakan besar perubahan dalam distribusi statistik nilai fitur dari waktu ke waktu. Anda dapat mengharapkan beberapa variasi acak alami antara dataset dasar dan target, tetapi Anda harus memantau perubahan besar yang mungkin mengindikasikan penyimpangan data yang signifikan.

Anda dapat menentukan **ambang** untuk besaran pergeseran data yang di atasnya Anda ingin diberi tahu, dan mengonfigurasi pemberitahuan lansiran melalui email. Kode berikut menunjukkan contoh penjadwalan monitor penyimpangan data untuk dijalankan setiap minggu, dan mengirim peringatan jika besarnya penyimpangan lebih besar dari 0,3:



1. **Latihan - Memantau penyimpangan data**

Sekarang adalah kesempatan Anda untuk memantau penyimpangan data menggunakan kumpulan data di Pembelajaran Mesin Azure.

Dalam latihan ini, Anda akan:

* Buat monitor penyimpangan data.
* Isi ulang monitor dan lihat penyimpangan data.

1. **Instruksi**

Ikuti petunjuk ini untuk menyelesaikan latihan.

1. Jika Anda belum memiliki langganan Azure, daftar untuk uji coba gratis di <https://azure.microsoft.com.>
2. Menampilkan repo latihan di <https://aka.ms/mslearn-dp100.> \ <https://microsoftlearning.github.io/mslearn-dp100/>
3. Jika Anda belum melakukannya, selesaikan latihan **Membuat ruang kerja Azure Machine Learning** untuk memprovisikan ruang kerja Azure Machine Learning, membuat instans komputasi, dan mengkloning file yang diperlukan.
4. Selesaikan latihan **Pantau penyimpangan data.**

**5. Ringkasan**

Dalam modul ini, Anda belajar cara:

* Buat monitor penyimpangan data.
* Jadwalkan pemantauan penyimpangan data.
* Lihat hasil pemantauan penyimpangan data.

Untuk mempelajari lebih lanjut tentang memantau penyimpangan data di Azure machine Learning, lihat **Mendeteksi penyimpangan data pada set data**

**Link :** <https://learn.microsoft.com/id-id/azure/machine-learning/v1/how-to-monitor-datasets?tabs=python>

dan **Mendeteksi penyimpangan data pada model yang disebarkan ke Azure Kubernetes Service (AKS)** di dokumentasi Azure Machine Learning .

**Link** : <https://learn.microsoft.com/id-id/azure/machine-learning/v1/how-to-enable-data-collection>