**Mendeteksi dan mengurangi ketidakadilan dalam model dengan Azure Machine Learning**

Model pembelajaran mesin sering kali dapat merangkum bias yang tidak disengaja yang mengakibatkan ketidakadilan. Dengan Fairlearn dan Azure Machine Learning, Anda dapat mendeteksi dan mengurangi ketidakadilan dalam model Anda.

**Tujuan pembelajaran**

Dalam modul ini, Anda akan mempelajari:

* + Cara mengevaluasi model pembelajaran mesin untuk keadilan.
  + Cara mengurangi disparitas prediktif dalam model pembelajaran mesin.

1. **Pengantar**

Model pembelajaran mesin semakin banyak digunakan untuk menginformasikan keputusan yang mempengaruhi kehidupan masyarakat. Misalnya, prediksi yang dibuat oleh model pembelajaran mesin mungkin memengaruhi:

* Persetujuan untuk pinjaman, asuransi, atau layanan keuangan lainnya.
* Penerimaan ke sekolah atau perguruan tinggi.
* Kelayakan untuk uji coba medis atau pengobatan eksperimental.
* Inklusi dalam promosi pemasaran.
* Seleksi untuk pekerjaan atau promosi.

Dengan keputusan kritis seperti itu dalam keseimbangan, keyakinan bahwa model pembelajaran mesin yang kami andalkan memprediksi, dan tidak membeda-bedakan atau terhadap subset populasi berdasarkan etnis, jenis kelamin, usia, atau faktor lainnya.



1. **Tujuan pembelajaran**

Dalam modul ini, Anda akan mempelajari cara:

* Mengevaluasi model pembelajaran mesin untuk keadilan.
* Mitigasi disparitas prediktif dalam model pembelajaran mesin.

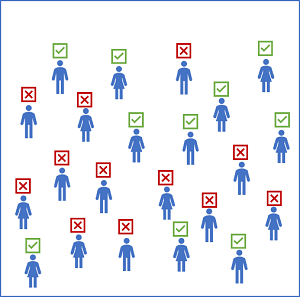
1. **Pertimbangkan keadilan model**

Ketika kita mempertimbangkan konsep keadilan mengenai prediksi yang dibuat oleh model pembelajaran mesin, itu membantu untuk menjadi jelas tentang apa yang kita maksud dengan "adil".

Misalnya, model klasifikasi digunakan untuk memprediksi probabilitas keberhasilan pembayaran pinjaman dan karenanya mempengaruhi apakah pinjaman disetujui atau tidak. Model ini kemungkinan akan dilatih menggunakan fitur yang mencerminkan karakteristik pemohon, seperti:

* Usia
* Status ketenagakerjaan
* Pendapatan
* Penghematan
* Utang saat ini

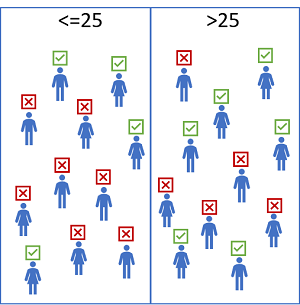
Fitur-fitur ini digunakan untuk melatih model klasifikasi biner yang memprediksi apakah pemohon akan membayar pinjaman.



Misalkan model memprediksi bahwa sekitar 45% pelamar akan berhasil membayar kembali pinjaman mereka. Namun, pada meninjau catatan persetujuan pinjaman, Anda mulai menduga bahwa lebih sedikit pinjaman yang disetujui untuk pelamar berusia 25 atau lebih muda daripada untuk pelamar yang berusia di atas 25 tahun. Bagaimana Anda bisa yakin model ini adil bagi pelamar di kedua kelompok usia?

1. **Mengukur disparitas dalam prediksi**

Salah satu cara untuk mulai mengevaluasi keadilan model adalah dengan membandingkan prediksi untuk setiap grup dalam fitur sensitif. Untuk model persetujuan pinjaman, Usia adalah fitur sensitif yang kami pedulikan, sehingga kami dapat membagi data menjadi subset untuk setiap kelompok usia dan membandingkan tingkat seleksi (proporsi prediksi positif) untuk setiap kelompok.



Katakanlah kami menemukan bahwa model memprediksi bahwa 36% pelamar berusia 25 atau lebih muda akan membayar pinjaman, tetapi memprediksi keberhasilan pembayaran untuk 54% pelamar berusia di atas 25 tahun. Ada perbedaan prediksi 18%.

Pada pandangan pertama, perbandingan ini tampaknya mengkonfirmasi bahwa ada bias dalam model yang mendiskriminasi pelamar yang lebih muda. Namun, ketika Anda mempertimbangkan populasi secara keseluruhan, mungkin orang yang lebih muda umumnya berpenghasilan kurang dari orang yang lebih mapan dalam karier mereka, memiliki tingkat tabungan dan aset yang lebih rendah, dan memiliki tingkat gagal bayar pinjaman yang lebih tinggi.

Poin penting yang perlu dipertimbangkan di sini adalah bahwa hanya karena kami ingin memastikan keadilan mengenai usia, itu tidak selalu mengikuti usia itu bukan faktor dalam probabilitas pembayaran pinjaman. Ada kemungkinan bahwa secara umum, orang yang lebih muda lebih kecil kemungkinannya untuk membayar pinjaman daripada orang yang lebih tua. Untuk mendapatkan gambaran lengkapnya, kita perlu melihat sedikit lebih dalam ke dalam performa prediktif model untuk setiap subset populasi.

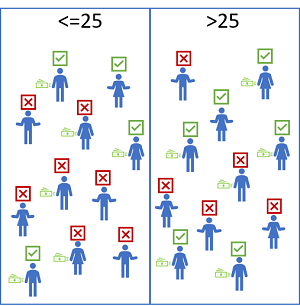
1. **Mengukur disparitas dalam performa prediksi**

Saat Anda melatih model pembelajaran mesin menggunakan teknik yang diawasi, seperti regresi atau klasifikasi, Anda menggunakan metrik yang dicapai terhadap data validasi hold-out untuk mengevaluasi performa prediktif keseluruhan model. Misalnya, Anda dapat mengevaluasi model klasifikasi berdasarkan akurasi, presisi, atau pengenalan.

Untuk mengevaluasi kewajaran model, Anda dapat menerapkan metrik performa prediktif yang sama ke subset data, berdasarkan fitur sensitif tempat populasi Anda dikelompokkan, dan mengukur perbedaan dalam metrik tersebut di seluruh subkelompok.

Misalnya, model persetujuan pinjaman menunjukkan metrik pengenalan keseluruhan 0,67 - dengan kata lain, itu dengan benar mengidentifikasi 67% kasus di mana pemohon membayar kembali pinjaman. Pertanyaannya adalah apakah model menyediakan tingkat prediksi yang sama untuk kelompok usia yang berbeda atau tidak.

Untuk mengetahuinya, kami mengelompokkan data berdasarkan fitur sensitif (Usia) dan mengukur metrik performa prediktif (pengenalan) untuk grup tersebut. Kemudian kita dapat membandingkan skor metrik untuk menentukan perbedaan di antara mereka.



Katakanlah kita menemukan bahwa pemanggilan kembali untuk kasus validasi di mana pemohon adalah 25 atau lebih muda adalah 0,50, dan ingat untuk kasus di mana pemohon di atas 25 adalah 0,83. Dengan kata lain, model ini dengan benar mengidentifikasi 50% orang dalam kelompok usia 25 atau lebih muda yang berhasil membayar pinjaman (dan karena itu salah diklasifikasikan 50% dari mereka sebagai defaulter pinjaman), tetapi menemukan 83% pembayar pinjaman di kelompok usia yang lebih tua (hanya salah mengklasifikasikan 17% dari mereka). Disparitas dalam performa prediksi antara kelompok adalah 33%, dengan model memprediksi negatif yang jauh lebih palsu untuk kelompok usia yang lebih muda.

1. **Potensi penyebab disparitas**

Saat Anda menemukan perbedaan antara tingkat prediksi atau metrik performa prediksi di seluruh grup fitur sensitif, ada baiknya mempertimbangkan potensi penyebabnya. Tugas tersebut mungkin termasuk:

* Ketidakseimbangan data. Beberapa grup mungkin direpresentasikan secara berlebihan dalam data pelatihan, atau data mungkin miring sehingga kasus dalam kelompok tertentu tidak mewakili populasi keseluruhan.
* Korelasi tidak langsung. Fitur sensitif itu sendiri mungkin tidak prediktif dari label, tetapi mungkin ada korelasi tersembunyi antara fitur sensitif dan beberapa fitur lain yang mempengaruhi prediksi. Misalnya, kemungkinan ada korelasi antara usia dan riwayat kredit, dan kemungkinan ada korelasi antara riwayat kredit dan gagal bayar pinjaman. Jika fitur riwayat kredit tidak disertakan dalam data pelatihan, algoritma pelatihan dapat menetapkan bobot prediktif ke usia tanpa memperhitungkan riwayat kredit, yang mungkin membuat perbedaan dengan probabilitas pembayaran pinjaman.
* Bias sosial. Bias bawah sadar dalam proses pengumpulan, persiapan, atau pemodelan data mungkin telah memengaruhi pemilihan fitur atau aspek desain model lainnya.

1. **Memitigasi bias**

Mengoptimalkan keadilan dalam model pembelajaran mesin adalah tantangan sosioteknik. Dengan kata lain, itu tidak selalu sesuatu yang dapat Anda capai murni dengan menerapkan koreksi teknis pada algoritma pelatihan. Namun, ada beberapa strategi yang dapat Anda adopsi untuk mengurangi bias, termasuk:

* Menyeimbangkan data pelatihan dan validasi. Anda dapat menerapkan teknik over-sampling atau under-sampling untuk menyeimbangkan data dan menggunakan algoritme pemisahan bertingkat untuk mempertahankan proporsi yang representatif untuk pelatihan dan validasi.
* Lakukan pemilihan fitur yang ekstensif dan analisis teknik. Pastikan Anda sepenuhnya menjelajahi korelasi yang saling terhubung dalam data Anda untuk mencoba membedakan fitur yang secara langsung prediktif dari fitur yang merangkum hubungan yang lebih kompleks dan bernuansa. Anda dapat menggunakan dukungan **interpretabilitas model di Azure Machine Learning : modul 20 - Menjelaskan model pembelajaran mesin dengan Azure Machine Learning** untuk memahami bagaimana fitur individual memengaruhi prediksi.
* Mengevaluasi model untuk disparitas berdasarkan fitur yang signifikan. Anda tidak dapat dengan mudah mengatasi bias dalam model jika Anda tidak dapat mengukurnya.
* Performa prediktif keseluruhan trade-off untuk disparitas yang lebih rendah dalam performa prediktif antara grup fitur sensitif. Model yang 99,5% akurat dengan performa yang sebanding di semua kelompok sering lebih diinginkan daripada model yang 99,9% akurat tetapi mendiskriminasi subset kasus tertentu.

Sisa modul ini mengeksplorasi paket **Fairlearn** - paket Python yang dapat Anda gunakan untuk mengevaluasi dan mengurangi ketidakadilan dalam model pembelajaran mesin.

1. **Menganalisis kewajaran model dengan Fairlearn**

**Fairlearn** adalah paket Python yang dapat Anda gunakan untuk menganalisis model dan mengevaluasi perbedaan antara prediksi dan performa prediksi untuk satu atau beberapa fitur sensitif.

Ini bekerja dengan menghitung metrik grup untuk fitur sensitif yang Anda tentukan. Metrik itu sendiri didasarkan pada metrik evaluasi model **scikit-learn standar**, seperti akurasi, presisi, atau pengenalan untuk model klasifikasi.

Fairlearn API sangat luas, menawarkan beberapa cara untuk mengeksplorasi perbedaan dalam metrik di seluruh pengelompokan fitur sensitif. Untuk model klasifikasi biner, Anda mungkin mulai dengan membandingkan tingkat pilihan (jumlah prediksi positif untuk setiap grup) dengan menggunakan fungsi **selection\_rate** sama. Fungsi ini mengembalikan tingkat seleksi keseluruhan untuk himpunan data pengujian. Anda juga dapat menggunakan fungsi **sklearn.metrics** standar (seperti **accuracy\_score, precision\_score, atau recall\_score**) untuk mendapatkan tampilan keseluruhan tentang performa model.

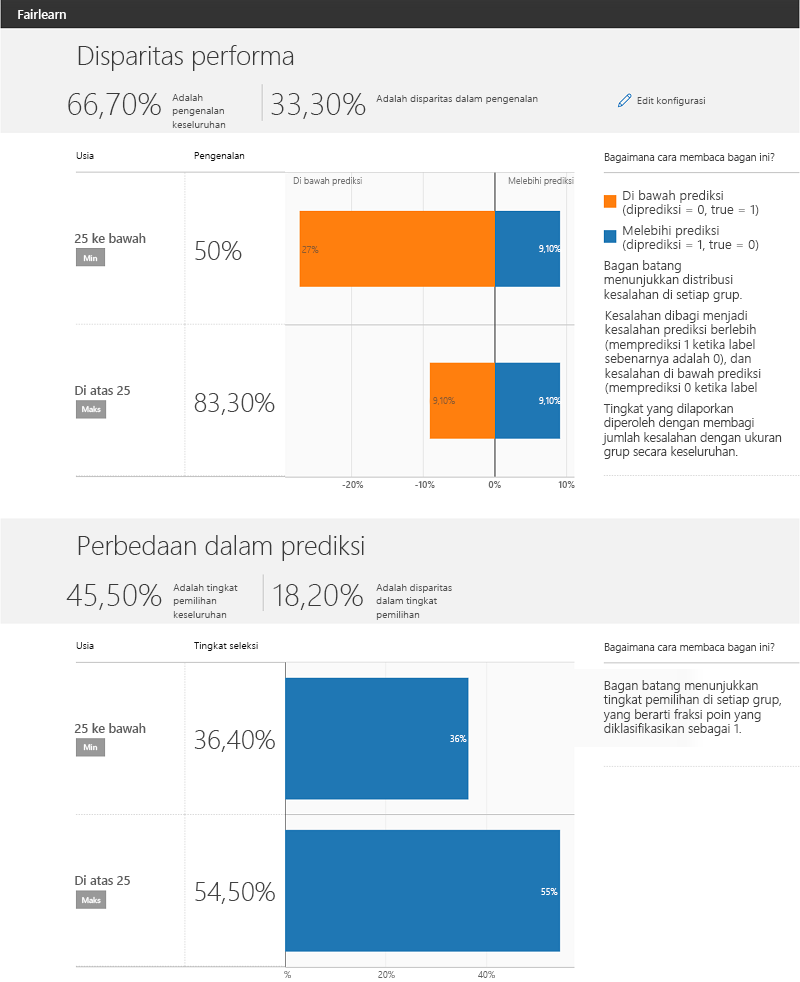
Kemudian, Anda dapat menentukan satu atau beberapa fitur sensitif dalam himpunan data yang ingin Anda kelompokkan subset populasi dan membandingkan tingkat seleksi dan performa prediktif. Fairlearn menyertakan fungsi **MetricFrame** yang memungkinkan Anda membuat kerangka data beberapa metrik menurut grup.

Misalnya, dalam model klasifikasi biner untuk prediksi pembayaran pinjaman, di mana fitur sensitif Usia terdiri dari dua nilai kategoris yang mungkin (**25-dan-bawah dan lebih-25**), MetricFrame untuk grup-grup ini mungkin mirip dengan tabel berikut:



1. **Memvisualisasikan metrik di dasbor**

Sering kali lebih mudah untuk membandingkan metrik secara visual, sehingga Fairlearn menyediakan widget dasbor interaktif yang dapat Anda gunakan dalam buku catatan untuk menampilkan metrik grup untuk model. Widget memungkinkan Anda memilih fitur sensitif dan metrik performa untuk dibandingkan, lalu menghitung dan memvisualisasikan metrik dan perbedaan, seperti ini:



1. **Integrasikan dengan Azure Machine Learning**

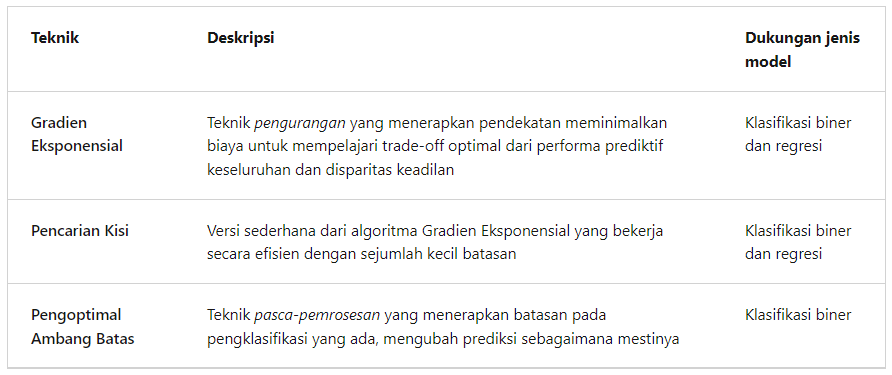
Fairlearn terintegrasi dengan Azure Machine Learning dengan memungkinkan Anda menjalankan eksperimen di mana metrik dasbor diunggah ke ruang kerja Azure Machine Learning Anda. Ini memungkinkan Anda untuk berbagi dasbor di studio Azure Machine Learning sehingga tim ilmu data Anda dapat melacak dan membandingkan metrik disparitas untuk model yang terdaftar di ruang kerja.

1. **Mitigasi ketidakadilan dengan Fairlearn**

Selain memungkinkan Anda menganalisis disparitas dalam tingkat seleksi dan performa prediktif di seluruh fitur sensitif, Fairlearn menyediakan dukungan untuk mengurangi ketidakadilan dalam model.

1. **Algoritma mitigasi dan batasan paritas**

Dukungan mitigasi di Fairlearn didasarkan pada penggunaan algoritma untuk membuat model alternatif yang menerapkan batasan paritas untuk menghasilkan metrik yang sebanding di seluruh grup fitur sensitif. Fairlearn mendukung teknik mitigasi berikut.



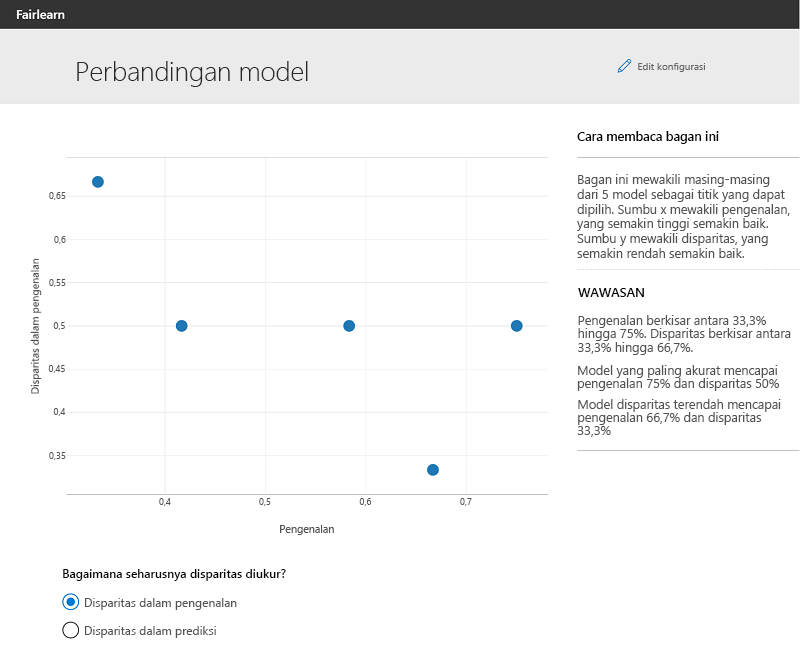
Pilihan batasan paritas tergantung pada teknik yang digunakan dan kriteria kewajaran spesifik yang ingin Anda terapkan. Kendala di Fairlearn meliputi:

* **Paritas demografis:** Gunakan batasan ini dengan algoritme mitigasi untuk meminimalkan disparitas dalam tingkat pemilihan di seluruh grup fitur sensitif. Misalnya, dalam skenario klasifikasi biner, batasan ini mencoba memastikan bahwa jumlah prediksi positif yang sama dibuat di setiap grup.
* **Kesamaan rasio positif sejati:** Gunakan batasan ini dengan salah satu algoritme mitigasi untuk meminimalkan perbedaan dalam rasio positif sejati di seluruh grup fitur sensitif. Misalnya, dalam skenario klasifikasi biner, batasan ini mencoba memastikan bahwa setiap grup berisi rasio prediksi positif yang sebanding.
* **Paritas tingkat positif palsu:** Gunakan batasan ini dengan salah satu algoritme mitigasi untuk meminimalkan perbedaan dalam tingkat\_positif palsu di seluruh grup fitur sensitif. Misalnya, dalam skenario klasifikasi biner, batasan ini mencoba memastikan bahwa setiap grup berisi rasio prediksi positif palsu yang sebanding.
* **Peluang yang sama:** Gunakan batasan ini dengan salah satu algoritme mitigasi untuk meminimalkan perbedaan dalam gabungan tingkat positif benar dan tingkat\_positif palsu di seluruh grup fitur sensitif. Misalnya, dalam skenario klasifikasi biner, batasan ini mencoba memastikan bahwa setiap grup berisi rasio yang sebanding dari prediksi positif dan positif palsu yang sebenarnya.
* **Kesamaan tingkat kesalahan:** Gunakan batasan ini dengan salah satu algoritme mitigasi berbasis pengurangan (**Gradien Eksponen** dan **Penelusuran Kotak**) untuk memastikan bahwa kesalahan untuk setiap fitur sensitif kelompok tidak menyimpang dari tingkat kesalahan keseluruhan lebih dari jumlah yang ditentukan.
* **Kerugian grup terikat:** Gunakan batasan ini dengan algoritma mitigasi berbasis pengurangan untuk membatasi kerugian untuk setiap grup fitur sensitif dalam model regresi.

1. **Pelatihan dan evaluasi model yang dimitigasi**

Pendekatan umum untuk mitigasi adalah menggunakan salah satu algoritme dan batasan untuk melatih beberapa model, dan kemudian membandingkan performa, tingkat seleksi, dan metrik disparitas mereka untuk menemukan model optimal untuk kebutuhan Anda. Seringkali, pilihan model melibatkan trade-off antara performa prediktif mentah dan keadilan - berdasarkan definisi Anda tentang keadilan untuk skenario tertentu. Umumnya, keadilan diukur dengan pengurangan disparitas pemilihan fitur (misalnya, memastikan bahwa proporsi anggota yang sama dari setiap kelompok gender disetujui untuk pinjaman bank) atau dengan pengurangan disparitas metrik performa (misalnya, memastikan bahwa model sama akuratnya dalam mengidentifikasi pembayar dan defaulter di setiap kelompok usia).

Fairlearn memungkinkan Anda untuk melatih model yang dimitigasi dan memvisualisasikannya menggunakan dasbor, seperti ini.



Anda dapat memilih model individual di scatterplot untuk melihat detailnya, memungkinkan Anda menjelajahi opsi dan memilih model terbaik untuk persyaratan keadilan Anda.

1. **Integrasikan dengan Azure Machine Learning**

Sama seperti saat menganalisis model individual, Anda dapat mendaftarkan semua model yang ditemukan selama pengujian mitigasi Anda dan mengunggah metrik dasbor ke Azure Machine Learning.

1. **Latihan - Gunakan Fairlearn dengan Azure Machine Learning**

Sekarang adalah kesempatan Anda untuk mendeteksi dan mengurangi ketidakadilan dalam model.

Dalam latihan ini, Anda akan:

* Gunakan paket Fairlearn dengan Azure Machine Learning untuk mengevaluasi disparitas performa prediksi.
* Gunakan paket Fairlearn dengan Azure Machine Learning untuk mengurangi ketidakadilan.

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 menyediakan ruang kerja Azure Machine Learning, membuat instans komputasi, dan mengkloning file yang diperlukan.
4. Selesaikan latihan **Deteksi dan mitigasi ketidakadilan.**

**6. Ringkasan**

Dalam modul ini, Anda belajar cara:

* Mengevaluasi model pembelajaran mesin untuk keadilan.
* Mitigasi disparitas prediktif dalam model pembelajaran mesin.

Untuk mempelajari lebih lanjut tentang Fairlearn, lihat dokumentasi **Fairlearn di GitHub.**

**Link :** https://github.com/fairlearn/fairlearn