Laporan Project Machine Learning

**Anggota kelompok:**

Albert Lilian Thamson / 2201754412

Daniel Santoso / 2201756506

Luwis Lim / 2201756506

Steven Odolf Yuwono / 2201758045

Tujuan:

* Mencari tahu apakah ada hubungan antara preferensi musik seseorang dengan gaya hidup kesehatannya, terutama merokok.

1. Metode

Kami mencoba semua metode di bawah ini untuk 2 dataset kami, yaitu dataset Slovakia dan dataset Indonesia.

1. Linear Regression
2. Polynomial Regression
3. Support Vector Machine (SVM)
4. Decision Tree
5. K-th Nearest Neigbour (KNN)
6. Random Forest
7. Experiment
8. Linear Regression

Regression berfungsi untuk mencari apakah ada hubungan antara 1 variabel dengan target. Kenapa memakai regression? Karena kita mau mencari hubungan antara preferensi music seseorang dengan gaya hidup kesehatannya (merokok atau tidak). Pertama kita mencoba menggunakan Multiple linear regression karena target kita tidak cuma 1 namun ada 3.

* 1. Slovakia Dataset

A picture containing table

Description automatically generated

Dari hasil yang kita dapat, R2 mendapatkan hasil 0.0229, yang berarti butuh penyesuaian yang sangat besar pada data karena minimal R2 harus lebih besar dari 0.05, dan juga intercept lebih besar dari 0.05 yang berarti tidak ada hubungan antara feature dengan target. Jadi kesimpulannya adalah tidak adanya hubungan antara feature dengan data.

* 1. Indonesia DatasetA picture containing bird

     Description automatically generated

Dari hasil yang kita dapat, R2 mendapatkan hasil -0.0277, yang berarti butuh penyesuain yang sangat besar pada data karena minimal R2 lebih besar dari 0.05, berarti data butuh disesuaikan agar bisa ke next step-nya.

1. Polynomial Regression
   1. Slovakia dataset

A screenshot of a cell phone

Description automatically generated

Dari hasil yang kita dapat, R2 mendapatkan hasil 0.03779, yang berarti butuh penyesuaian yang sangat besar pada data karena minimal R2 harus lebih besar dari 0.05, berarti data butuh disesuaikan agar bisa ke next step-nya.

* 1. Indonesia dataset

A screenshot of a cell phone

Description automatically generated

Dari hasil yang kita dapat, R2 mendapatkan hasil 1.0 yang berarti data yang dipakai merupakan perfect fit.



Tetapi hasil yang diberikan oleh interceptnya harusnya lebih besar dari 0.05 sedangkan dari data Indonesia didapatkan -0.0463, sehingga dapat disimpulkan bahwa tidak ada hubungan antar feature dengan target.

Kenapa tidak menggunakan Regression sebagai patokan kita? Karena data kita berbentuk discrete yang membuat Regression tidak dapat bekerja dengan semestinya. Jadi kita mencari algoritma lain untuk bisa mengolah data kita lebih baik.

1. Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah representasi dari data yang di training sebagai titik-titik yang dipisahkan ke dalam beberapa kategori dengan mencari celah antar titik selebar mungkin. SVM biasanya digunakan dalam ruang berdimensi tinggi dan menggunakan subset training points didalam decision function sehingga menggunakan memori secara efisien. Itu mengapa kami menggunakan SVM dalam projek Machine Learning kami.

1. Accuracy

A screen shot of a person

Description automatically generatedAccuracy yang kami dapatkan adalah 60% dan hasil tersebut masih kurang memuaskan.

1. Confusion Matrix

A picture containing drawing

Description automatically generatedDari confusion matrix yang kami peroleh, kami mendapatkan 0 nilai true positive dan 81 nilai true negative. Nilai true positive yang didapat sangat kecil jika dibandingkan dengan nilai true negative yang diperoleh.

1. Precision, Recall, F1-score, and Support

A picture containing drawing

Description automatically generatedPrecision yang kita dapatkan adalah 0.3 yang berarti kemampuan dalam melabel dari model kita masih kurang. Recall yang didapatkan adalah 0.5 yang berarti kemampuan model dalam mencari semua sample positif masih perlu ditingkatkan. F1-score yang didapatkan adalah 0.375 yang berarti sangat kurang karena masih sangat jauh untuk mencapai F1-score 1.

1. Decision Tree

Decision Tree adalah salah satu metode klasifikasi yang memiliki fungsi untuk menentukan suatu action yang akan diambil atau untuk menunjukan suatu statistik untuk probabilitas hal yang akan terjadi kedepannya. Decision Tree biasa dibuat dengan memanjang kebawah , namun juga terkadang dibuat kesamping. Cabang cabang dalam tree di decision tree mempresentasikan setiap kemungkinan yang akan terjadi.

A screenshot of a social media post

Description automatically generated

Dari data di atas dapat di lihat bahwa untuk dataset Slovakia memiliki akurasi sebesar 58,52% yang dimana cukup baik karena memiliki presentase di atas 50%, untuk confusion matrixnya memiliki True positive sebesar 65, false positive sebesar 16, false negative sebesar 40 dan true negative sebesar 14, dan juga di dapat precision dengan menggunakan rumus true positive / (true positive + false positive) dan di dapat angka sebesar 0.8024691358024691. Untuk bagian recall di dapat dengan rumus true positive / (true positive + false negative) dan di dapat recall sebesar 0.6190476190476191 dan juga di dapat nilai F1 dengan rumus (2 X precision X recall) / (precision + recall) dan didapat nilai f1 sebesar 0.6989247311827957.

A screenshot of a social media post

Description automatically generated

Dari data di atas dapat di lihat bahwa untuk dataset Indonesia memiliki akurasi sebesar 93,33% yang dimana sangat baik karena memiliki angka hampir sempurna. Untuk confusion matrixnya memiliki True positive sebesar 14, false positive sebesar 0, false negative sebesar 1 dan true negative sebesar 0, dan juga didapatkan precision dengan menggunakan rumus true Positive/ (true negative + false positive) dan di dapat angka sebesar 1. Untuk bagian recall di dapat dengan rumus true positive / (true positive + false negative) di dapat nilai recall sebesar 0.93333333 dan juga di dapat nilai F1 dengan rumus (2 X precision X recall) / (precision + recall) di dapat f1 sebesar 0.96551724

1. K-th Nearest Neigbour (KNN)

KNN adalah salah satu algoritma supervised machine learning yang dapat digunakan untuk menyelesaikan permasalahan classification maupun regression. Di sini kita menggunakan KNN lebih ke arah klasifikasi, karena data kita tidak continuous, data kita diskrit. KNN mengasumsikan bahwa titik-titik data yang berdekatan artinya ada dalam satu kelompok yang sama.

Berikut ini kami lampirkan beberapa hasil dari pemrosesan dataset Slovakia:

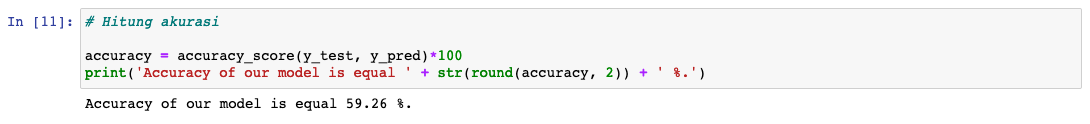
1. Correlation Matrix

A screenshot of a cell phone

Description automatically generated

Dari hasil correlation matrix tersebut, tidak terlihat adanya korelasi yang signifikan antara target kita yaitu “Smoking” dengan variable-variabel lainnya. Dari hasil correlation matrix ini saja kita sudah bisa mengkonklusikan bahwa sebenarnya agak sulit untuk membuat model learning menggunakan dataset ini, karena hubungan antara feature dan target tidak terlalu signifikan, namun tetap bisa.

1. Accuracy



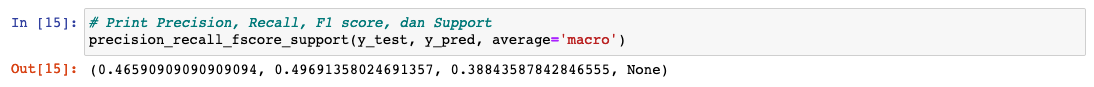
Accuracy yang kita dapatkan dari model ini yang adalah 59.26%, yang adalah akurasi yang tidak terlalu bagus dan masih kurang memuaskan.

1. Confusion matrix



Dari hasil confusion matrix tersebut, kita dapat mengetahui bahwa true positive dari model kita sangat kecil, namun true negative besar.

1. Precision, Recall, F1-score, dan Support



Precision yang kita dapatkan 0.465 yang berarti kemampuan melabel yang benar dari model kita masih kurang, terlihat dari jumlah true positive yang kecil. Recall yang kita dapatkan 0.49 yang berarti kemampuan model untuk mencari semua sample yang positif masih biasa-biasa saja, tidak bagus, tidak buruk juga. F1-score yang kita dapatkan adalah 0.388 berarti masih sangat kurang bagus, karena seharusnya yang paling bagus adalah 1 dan paling buruk adalah 0, dan kita cenderung ke arah 0.

Berikut ini kami lampirkan beberapa hasil dari pemrosesan dataset Indonesia:

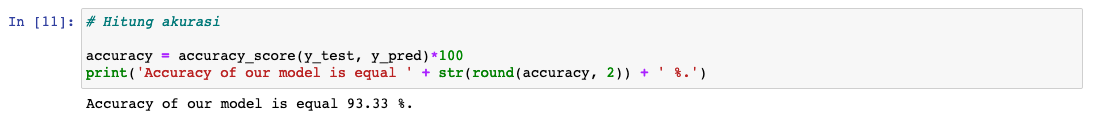
1. Correlation Matrix

A screenshot of a cell phone

Description automatically generated

Hasil correlation matrix yang kita dapatkan pada dataset Indonesia tidak jauh berbeda dari yang Slovakia, yaitu tidak adanya hubungan signifikan antara variable-variabel lain dengan target, kecuali ada agak sedikit hubungan yang ditunjukkan oleh variable “Rock n roll” sebesar 0.316.

1. Accuracy



Accuracy yang kita dapatkan dari model ini yang adalah 93.33%, yang adalah akurasi yang sangat bagus.

1. Confusion Matrix



True negative yang didapat yaitu 14 adalah nilai yang tinggi, mempertimbangkan nilai-nilai false yang hanya ada 1 yaitu false negative. Hal ini membuat model kita menjadi bagus, dan artinya akurasinya bagus juga. Namun bisa juga hal ini dikarenakan jumlah data Indonesia yang jauh lebih sedikit daripada data Slovakia.

1. Precision, Recall, F1-score, dan Support

A screenshot of a cell phone

Description automatically generated

Precision yang kita dapatkan 0.47 yang berarti kemampuan melabel yang benar dari model kita masih kurang, terlihat dari jumlah true positive yang kecil. Recall yang kita dapatkan 0.5 yang berarti kemampuan model untuk mencari semua sample yang positif masih kurang atau biasa-biasa saja, tidak bagus, tidak buruk juga. F1-score yang kita dapatkan adalah 0.482 berarti masih kurang bagus, karena seharusnya yang paling bagus adalah 1 dan paling buruk adalah 0, dan kita cenderung ke arah 0.

Kesimpulannya, metode KNN sebenarnya masih kurang cocok dengan dataset kita, dan masih kurang bisa mencapai harapan kita.

1. Random Forest

Random Forest adalah salah satu algoritma supervised machine learning yang dapat digunakan untuk menyelesaikan masalah classification dan regression. Random Forest membangun sebuah “hutan” yang terdiri dari banyak “pohon keputusan” atau “decision trees” yang mana nantinya model Random Forest akan mengambil nilai average yang di-“voting” oleh setiap pohon.

Berikut ini kami lampirkan beberapa hasil dari pemrosesan dataset Slovakia:

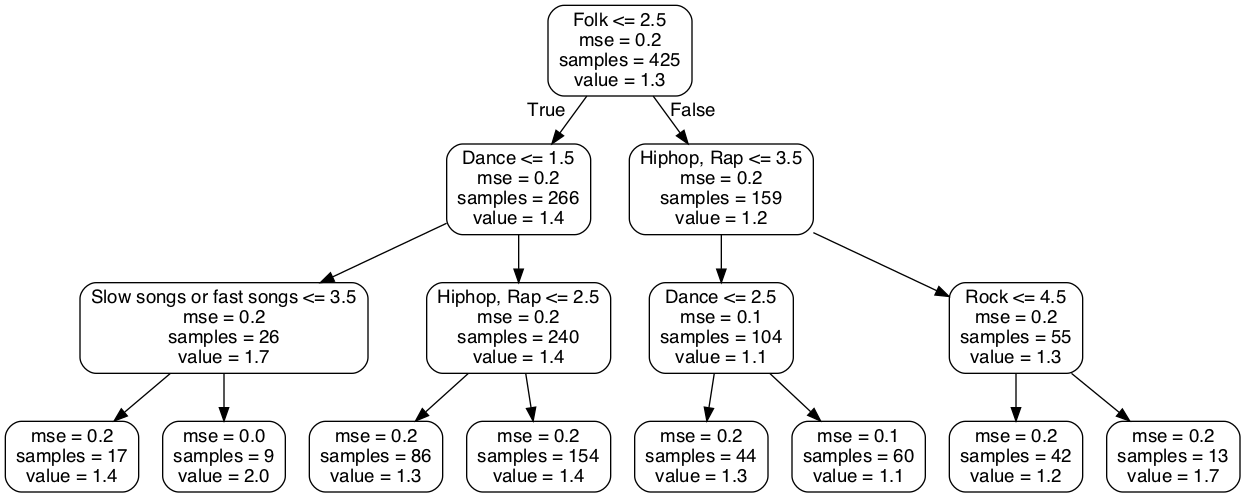
1. Akurasi dan Mean Absolute Error

A screenshot of a social media post

Description automatically generated

Error yang kita dapatkan adalah 0.45, yang mana itu adalah error yang cukup banyak dan akurasi yang kita dapatkan hanya 66.59%, yang mengindikasikan bahwa model kita masih belum cukup baik namun tidak buruk juga.

1. Penggambaran sederhana dari semua tree yang telah dibuat



Dari tree ini, kita bisa melihat beberapa fitur-fitur yang paling berpengaruh dan penting dalam memprediksi terhadap target. Fitur-fitur tersebut adalah folk; dance; hiphop, rap; dll.

1. Daftar kepentingan variable

A screenshot of a social media post

Description automatically generated

Dari daftar ini, kita bisa melihat bahwa sebenarnya kebanyakan dari fitur-fitur yang ada tidak ada yang dominan pengaruhnya terhadap targetnya, jadi sebenarnya tidak terlalu ada hubungan yang signifikan antara variable-variabel ini dengan targetnya.

1. Akurasi dan Mean Absolute Error menggunakan feature-feature yang terpenting saja

A screenshot of a social media post

Description automatically generated

Ternyata error yang diciptakan malah semakin besar, dan akurasinya berkurang. Hal ini dapat terjadi karena memang fitur-fitur yang ada tidak ada yang dominan, sehingga butuh semua fiturnya atau setidaknya sebagian besarnya untuk bisa menaikkan akurasinya sedikit saja.

Berikut ini kami lampirkan beberapa hasil dari pemrosesan dataset Indonesia:

1. Akurasi dan Mean Absolute Error

A screenshot of a social media post

Description automatically generated

Error yang kita dapatkan adalah 0.14, yang mana itu adalah error yang lumayan sedikit apabila dibandingkan dengan hasil dataset Slovakia. Akurasi yang kita dapatkan mencapai 91.01%, yang mengindikasikan bahwa model kita sudah cukup baik.

1. Penggambaran sederhana dari semua tree yang telah dibuat

A close up of text on a white background

Description automatically generated

Dari tree ini, kita bisa melihat beberapa fitur-fitur yang paling berpengaruh dan penting dalam memprediksi terhadap target. Fitur-fitur tersebut adalah metal or hardrock dan reggae, ska.

1. Daftar kepentingan variable

A screenshot of a social media post

Description automatically generated

Ternyata ada beberapa variable yang memang dominan dalam memprediksi target, seperti variable Opera, Folk, Rock n roll, dan juga Funk. Namun hasilnya berbeda dengan hasil yang didapat dari penggambaran tree tadi, yaitu Metal or hardrock dan Reggae, Ska. Hal ini dapat menjadi petunjuk bahwa variable Opera, Folk, Rock n roll, dan Punk memang menjadi genre musik yang paling bertolak belakang dalam menentukan seseorang punya kebiasaan merokok atau tidak.

1. Akurasi dan Mean Absolute Error menggunakan feature-feature yang terpenting saja

A screenshot of a social media post

Description automatically generated

Ternyata setelah menggunakan variable-variabel penting yang ditunjukkan pada daftar kepentingan variable tadi, nilai mean absolute error turun menjadi 0.12 dan akurasi meningkat menjadi 93%. Walaupun tidak signifikan, namun mengindikasikan bahwa kita sudah di jalan yang tepat.

Kesimpulannya, metode Random Forest kemungkinan adalah metode yang terbaik dari semua metode yang telah kita coba selama ini. Serta dari dataset Indonesia menunjukkan bahwa memang ada beberapa genre music yang sangat dominan dalam mempengaruhi target (merokok atau tidak), yang artinya orang yang memilih menilai tinggi variable-variabel ini sangat bertolak belakang dalam menilai targetnya. Genre-genre music tersebut adalah Opera, Folk, Rock n roll, dan juga Funk.

1. Thoughts about the result

Kami menyimpulkan bahwa metode Random Forest adalah metode terbaik dan paling cocok untuk permasalahan yang kami bahas di sini. Memang, hasilnya kurang memuaskan harapan kami untuk menemukan adanya hubungan yang kuat antara preferensi music seseorang dengan kesehatannya (merokok atau tidak). Yang kami dapatkan dari semua eksperimen ini adalah bahwa preferensi music seseorang, tidak terlalu berpengaruh atau menentukan apakah orang tersebut merokok atau tidak. Jadi stereotype bahwa orang yang suka music rock dan metal suka merokok dan gaya hidupnya kurang sehat tidak selalu benar. Hal tersebut kami buktikan lewat eksperimen ini. Namun stereotype tersebut terkadang juga bisa benar apabila ada di lingkungan negara Indonesia, seperti hasil dari percobaan Random Forest kami yang menunjukkan adanya pengaruh yang besar dari variable Rock n roll dan Punk terhadap variable Smoking.

1. Future works

Untuk kedepannya kami ingin mengumpulkan data yang lebih banyak dan lebih lengkap lagi agar algoritma kami kedepannya dapat lebih tepat untuk mencari apakah benar terdapat hubungan antara music dan gaya hidup. Dan juga kedepannya kami ingin mengeksperimen beberapa aspek-aspek lain yang dapat mempengaruhi gaya hidup seseorang seperti pekerjaan, hobi, dan lain sebagainya.