Laporan UTS - IBDA3111

Calvin Institute of Technology

Semester Ganjil 2022/2023

Oleh

Jason Caleb Erwin Piay / 2020012340 / IT & Big Data Analytics

Jason Caleb Erwin Piay - 202001240

Indikasi Terbaik untuk Diabetes

Masalah:

Diabetes merupakan penyakit kronis yang disebabkan oleh tingginya kadar gula darah yang melewati batas normal di dalam tubuh seseorang. Penyakit ini sangatlah berbahaya karena salah satunya memiliki efek yang sangat besar di dalam meningkatkan resiko banyak masalah jantung. Hal tersebut bisa termasuk penyakit arteri koroner dengan nyeri dada (angina), serangan jantung, stroke, dan penyempitan arteri (aterosklerosis). Selain fakta tersebut, tercatat pada tahun 2021 bahwa sebanyak 537 juta orang di seluruh dunia (sekitar 25% dari populasi keseluruhan dan 10% dari orang yang berumur 20-79 tahun) terkena diabetes. Hal ini dianggap kebanyakan terjadi karena kurangnya, baik pengetahuan maupun kesadaran, di dalam mengindikasi kadar gula di dalam tubuh dan pola hidup yang sedang dijalani.

Oleh karena itu, dengan memanfaatkan data Diabetes Health Indicators Dataset yang didapatkan dari situs Kaggle, dihasilkan pemrosesan data dengan pendekatan secara analisis untuk mengetahui tingkat keterkaitan antara fitur-fitur input dengan outputnya, sebagaimana data yang digunakan merupakan jenis data yang *supervised*/terdapat target variabel, sehingga diketahui cara pengindikasian yang benar terhadap penyakit diabetes. Di dalam meraih *goal* atau tujuan tersebut terdapat beberapa *challenge* yang telah saya hadapi, seperti menentukan teknik prapemrosesan data mana yang terbaik atau yang paling tepat digunakan.

Kemudian, dataset mengenai diabetes ini juga akan digunakan hasil analisanya sebagai bahan bukti data serta riset/penelitian mandiri untuk memberikan beberapa solusi berkaitan dengan penyakit diabetes, yang topiknya akan diangkat pada proyek yang sedang dikerjakan pada mata kuliah Penatalayanan 1. Salah satu tujuan dari proyek pada mata kuliah tersebut adalah untuk dapat berpatisi secara aktif memberikan solusi terhadap masalah berkaitan dengan SDG (Sustainable Development Goals).

Data:

Data yang akan digunakan berasal dari situs website Kaggle dengan penyebar dari datanya adalah Alex Toboul. Di dalam data tersebut terdapat tiga buah data csv, tetapi yang digunakan hanya yang pertama dengan mempertimbangan kelengkapan dan tingkat kestabilan dari datanya dibandingkan ke dua data yang lainnya. Kemudian, data yang digunakan di dalam penganalisaan saya dapat diakses dengan mengklik di sini.

Berikut saya akan berikan penjelasan singkat mengenai setiap fitur pada data yang akan digunakan.

1. Diabetes_012 (*target variable*), berisikan data-data yang bersifat ketegori, yaitu 0 = tidak diabetes, 1 = pre-diabetes, dan 2 = diabetes,
2. HighBP, berisikan data-data yang bersifat ketegori, yaitu 0 =BP (*blood pressure*) tidak tinggi dan 1 = BP tinggi,
3. HighChol, berisikan data-data yang bersifat ketegori, yaitu 0 = kadar kolesterol rendah dan 1 = kadar kolesterol tinggi,
4. CholCheck, berisikan data-data yang bersifat ketegori, yaitu 0 = tidak pernah memeriksakan kadar kolesterol di dalam lima tahun terakhir dan 1 = pernah memeriksakan kadar kolesterol di dalam lima tahun terakhir,
5. BMI, berisikan data-data yang bersifat numerik, yaitu nilai berupa Body Mass Index,
6. Smoker, berisikan data-data yang bersifat ketegori, yaitu 0 = tidak pernah merokok lebih dari 100 kali dan 1 = pernah merokok lebih dari 100 kali,
7. Stroke, berisikan data-data yang bersifat ketegori, yaitu 0 = tidak pernah diisukan terkena penyakit stroke dan 1 = pernah diisukan terkena penyakit stroke,
8. HeartDiseaseorAttack, berisikan data-data yang bersifat ketegori, yaitu 0 = tidak pernah diisukan terkena CHD (*coronary heart disease*) atau MI (*myocardial infarction*) dan 1 = pernah diisukan terkena CHD (*coronary heart disease*) atau MI (*myocardial infarction*),
9. PhysActivity, berisikan data-data yang bersifat ketegori, yaitu 0 = tidak pernah melakukan kegiatan fisik atau berolahraga di dalam 30 hari terakhir dan 1 = pernah melakukan kegiatan fisik atau berolahraga di dalam 30 hari terakhir,

10. Fruits, berisikan data-data yang bersifat ketegori, yaitu 0 = tidak mengonsumsi buah-buahan setidaknya satu atau lebih per hari dan 1 = mengonsumsi buah-buahan setidaknya satu atau lebih per hari,
11. Veggies, berisikan data-data yang bersifat ketegori, yaitu 0 = tidak mengonsumsi sayur-sayuran setidaknya satu atau lebih per hari dan 1 = mengonsumsi sayur-sayuran setidaknya satu atau lebih per hari,
12. HvtAlcoholConsump, berisikan data-data yang bersifat ketegori, yaitu 0 = bukan merupakan seorang peminum alkohol dan 1 = merupakan seorang peminum alkohol,
13. AnyHealthcare, berisikan data-data yang bersifat ketegori, yaitu 0 = tidak memiliki asuransi kesehatan sama sekali dan 1 = memiliki asuransi kesehatan,
14. NoDocbcCost, berisikan data-data yang bersifat ketegori, yaitu 0 = tidak pernah tidak mengunjungi dokter karena alasan biaya dan 1 = pernah tidak mengunjungi dokter karena alasan biaya,
15. GenHlth, berisikan data-data yang bersifat ketegori, yaitu menilai kesehatan diri dari skala 1 sampai dengan 5,
16. MentHlth, berisikan data-data yang bersifat ketegori, yaitu berapa lama di dalam skala 30 hari kesehatan mental dapat sembuh dengan beberapa pertimbangan penyebabnya,
17. PhysHlth, berisikan data-data yang bersifat ketegori, yaitu berapa lama di dalam skala 30 hari kesehatan fisik dapat sembuh dengan beberapa pertimbangan penyebabnya,
18. DiffWalk, berisikan data-data yang bersifat ketegori, yaitu 0 = tidak memiliki masalah serius di dalam berjalan atau menaiki anak tangga dan 1 = memiliki masalah serius di dalam berjalan atau menaiki anak tangga,
19. Sex, berisikan data-data yang bersifat ketegori, yaitu 0 perempuan dan 1 = laki-laki,
20. Age, berisikan data-data yang bersifat ketegori, yaitu 1 = umur 18-24 tahun, 2 = umur 25-29 tahun, 3 = umur 30-34 tahun, 4 = umur 35-49 tahun, 5 = umur 50-54 tahun, 6 = umur 55-59 tahun, 7 = umur 60-64 tahun, 8 = umur 65-69 tahun, 9 = umur 70-74 tahun, …, 13= umur 80 atau lebih,
21. Education, berisikan data-data yang bersifat ketegori, yaitu 1 = tidak pernah sekolah atau hanya ketika sebelum sekolah dasar, 2 = kelas 1-8, 3, …, 6,

22. Income, berisikan data-data yang bersifat ketegori, yaitu 1 = gaji kurang dari $10000 per tahun, …, 5 = gaji kurang dari $35000 per tahun, …, dan 8 = gaji $75000 per tahun atau lebih.

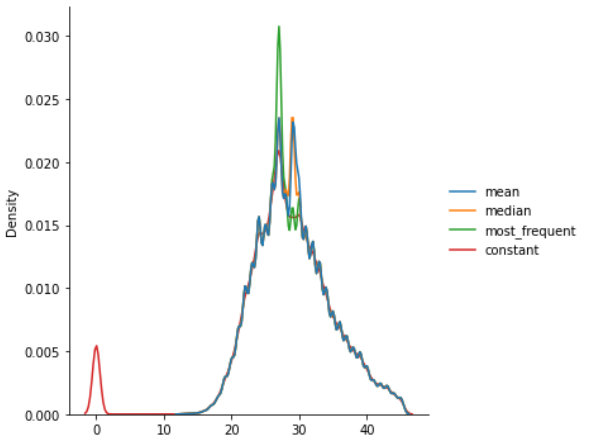
Diabetes Health Indicators Dataset ini merupakan data bersih yang diperoleh/dikumpulkan berdasarkan 253680 survei oleh CDC's BRFSS2015 (Center of Disease Control and Prevention, Behavioral Risk Factor Surveillance System 2015) pada tahun 2015. Data ini memiliki tiga buah kelas pengelompokkan pada variable outputnya, yaitu 0 yang mengindikasikan tidak terkena diabetes atau terkena pada periode hamil, 1 yang mengindikasikan pre-diabetes, dan 2 yang mengindikasikan terkena diabetes pada periode tidak hamil. Lalu, sisa kolom yang ada pada dataset ini, yaitu sebanyak 21 fitur, merupakan varibel inputnya.

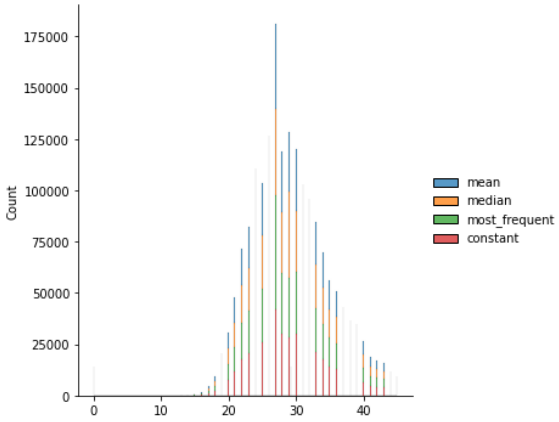
Di dalam melakukan persiapan sebelum data digunakan (prapemrosesan data), beberapa teknik telah diterapkan. Berikut saya akan berikan suatu daftar akan semua teknik tersebut dan rekayasa data yang telah dilakukan beserta penjelasannya.

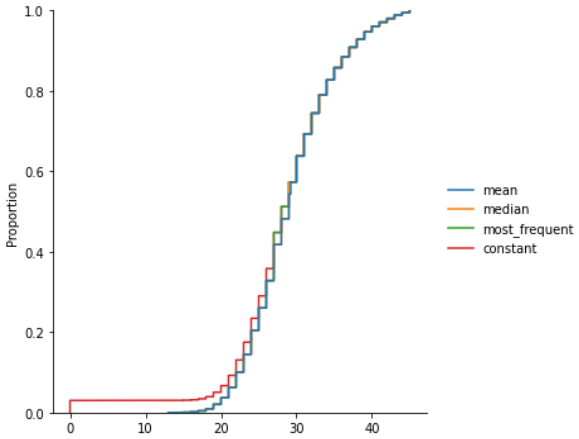
- *Cleaning Dataset*, sebagaimana data yang digunakan memiliki kelas yang kebanyakan merupakan bersifat kategori sehingga di dalam mengindari terhapusnya data-data tersebut beberapa teknik pembersihan data yang dilakukan berdasarkan tingkat keunikan dari data atau kolom tidak digunakan, seperti *few values,* dan *low variance*.
    - *Single Value*, pada proses awal, setelah data di-*import* dan dilihat keunikan dari datanya, ternyata tidak ada data yang memiliki *single value* sehingga teknik masih belum digunakan. Namun, setelah melakukan pengubahan data-data pencilan, barulah didapatkan ada dua buah kolom dari data yang memiliki *single value,* yaitu MentHlth dan PhysHlth sehingga dengan beberapa pertimbangan lainnya, saya memutuskan kedua buah fitur tersebut untuk dihapus.
    - *Duplicate Row*, pendeteksian terhadap baris-baris yang ada dilakukan secara otomatis oleh fungsi ".drop-duplicates", yang kemudian menghasilkan terdapat sejumlah 23899 buah baris dari data yang dihapus.
- *Resampling Data*, melakukan teknik prepemrosesan ini terhadap data merupakan tindakan yang tidak diekspetasikan untuk dilakukan, tetapi karena data yang saya gunakan ternyata memiliki *imbalance class* (kelas data yang tidak seimbang) yang

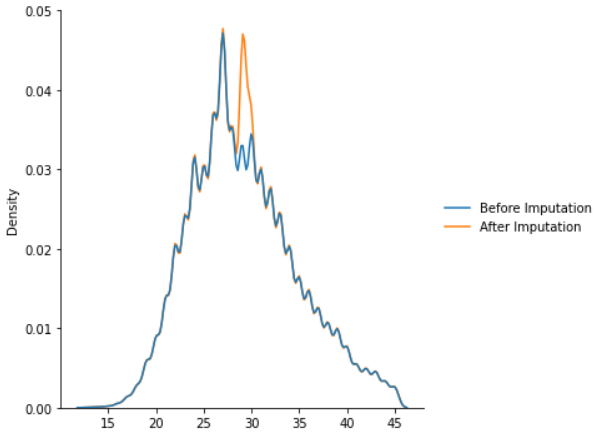
banyak sehingga untuk menghindari salah satunya akan ada banyak data yang dihapus karena terdeteksi menjadi suatu pencilan, dibutuhkan teknik *resampling data* ini, yaitu dengan mengubah jumlah data, baik dengan menambahkannya (menggunakan *random over-sampling*) maupun menguranginya (menggunakan *random under-sampling*).

- *Outliers Handling*, yaitu dengan mencari data-data yang merupakan pencilan dan lalu menggantinya dengan "nan" sebagai nilainya. Hal ini dilakukan sebagai penanganan terhadap data agar memiliki tingkat akurasi yang bertambah dan juga sebagai tindakan untuk menghindari adanya kerusakan pada model, seperti adanya sifat bias. Namun, di dalam melakukan hal ini, terdapat beberapa limitasi atau batasan yang saya berikan, seperti untuk kolom data atau fitur memiliki tingkat keunikan sebanyak sembilan atau kurang maka tidak akan diproses lebih lanjut di dalam pencarian data-data pencilannya. Hal ini saya lakukan dengan tujuan agar terhindarnya dari kekurangan jumlah data yang akan digunakan.
- *Data Imputation*, yaitu dengan menggunakan teknik-teknik khusus di dalam mengganti nilai dari, baik itu *missing data* (data yang hilang) maupun *outliers* atau pencilan, dengan nilai yang baru. Tujuan dari dilakukannya hal ini adalah agar tetap terjaganya tingkat akurasi dari data, mengurangi kemungkinan permodelan yang digunakan menjadi rusak, dan memperbaiki pendistribusian atau penyebaran dari data. Namun, di dalam menerapkan *data imputation* ini, saya dihadapi suatu masalah, yaitu karena banyaknya jumlah data yang saya gunakan mengakibatkan beberapa *advanced technique* (teknik yang bagus) menjadi tidak bisa saya gunakan. Oleh karena itu, di dalam menerapkan *data imputation* ini saya menggunakan Simpleimputer dengan juga melakukan evaluasi terlebih dahulu terhadap keempat jenis dari teknik ini, yaitu nilai rata-rata, nilai tengah, nilai modus, dan nilai konstan. Kemudian, setelah dilakukan pengevaluasian terhadap keempat jenis tersebut, saya memutuskan untuk menggunakan SimpleImputer yang memakai nilai rata-rata di dalam mengimputasi datanya.









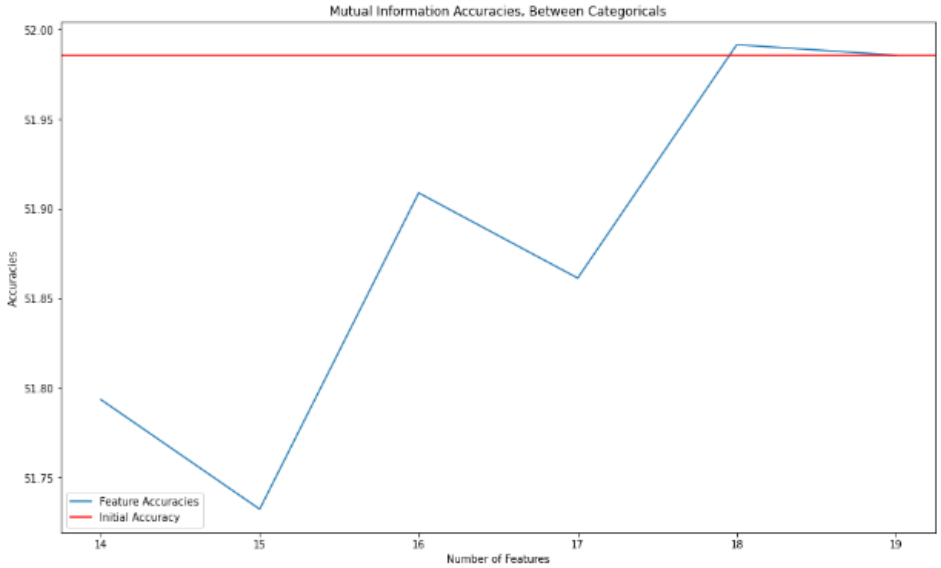
Hal ini karena setelah memperhatikan ketiga graph ini, saya menyadari bahwa perilaku yang ditunjukan dari nilai rata-rata atau *mean* tidak mengubah distribusi data secara ekstrem, sebagaimana salah satunya dapat dilihat dari yang ditunjukkan oleh garis warna merah, yaitu nilai konstan, pada graph. Terlihat bahwa garis tersebut membuat pendistribusian data menjadi berubah secara ekstrem.

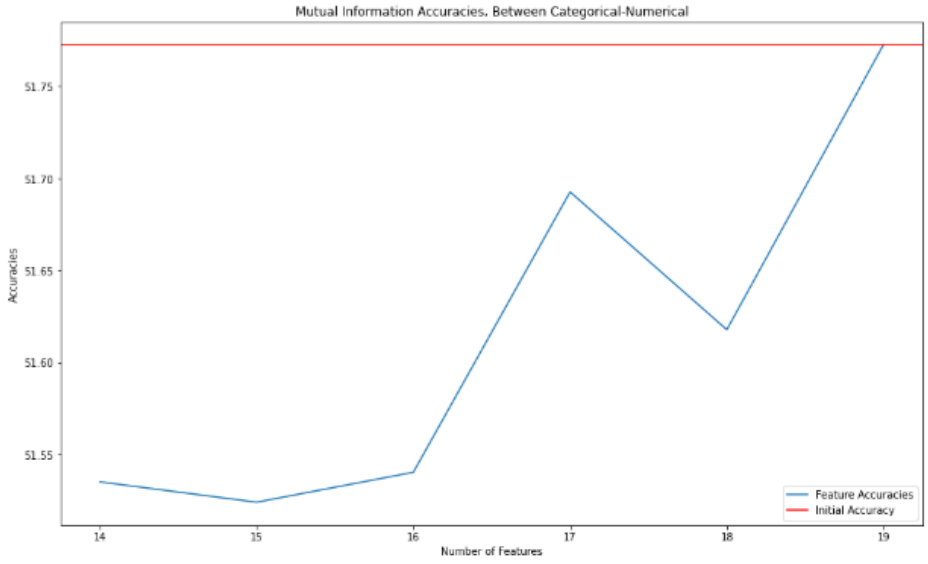
Kemudian, yang terakhir dari *data imputation*, dapat dilihat pada graph di samping merupakan visualisasi yang menunjukkan perubahan pada pendistribusian data, yaitu garis biru yang menunjukkan distribusi pada data sebelum dilakukan imputasi dan garis oren yang menunjukkan distribusi pada data sesudah dilakukan imputasi

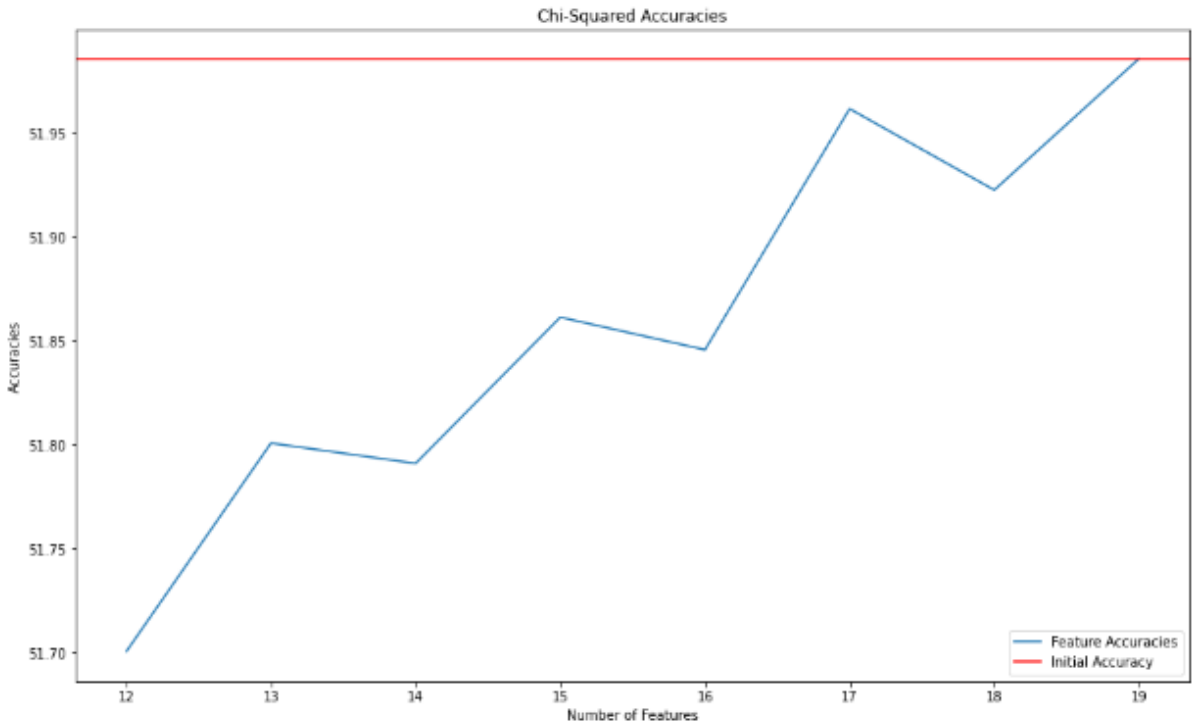
- *Feature Selection*, yaitu dengan menggunakan beberapa teknik penyeleksian fitur agar setiap kolom atau fitur di dalam data diberikan suatu nilai yang menunjukkan seberapa kuat relasi yang dibangun antara dua buah variabel, yaitu biasanya merupakan antara fitur dan target variabelnya, sehingga setelahnya dapat menjadi pertimbangan mengenai

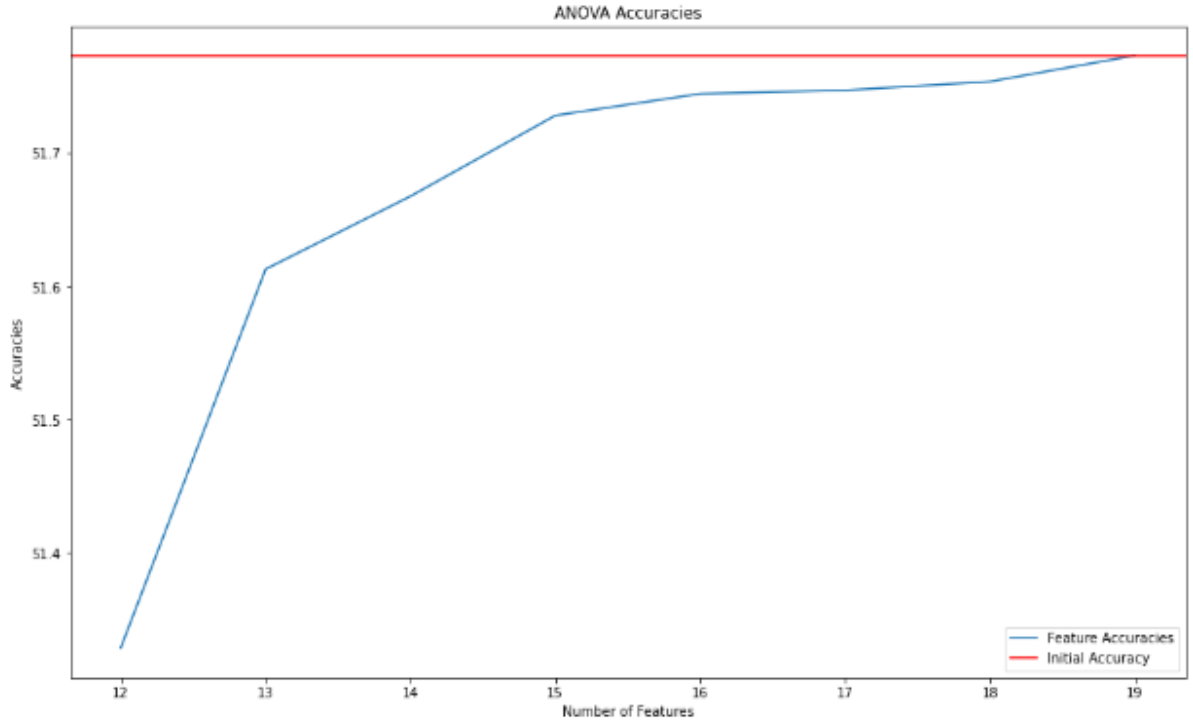
apakah suatu fitur harus dihapus dari data atau tidak. Selain itu, hal ini dilakukan agar dengan dihapusnya fitur-fitur yang tidak berguna maka harapannya data dapat memiliki tingkat akurasi yang lebih tinggi, mengurangi memori yang dipakai sehingga mengurangi juga waktu yang dihabiskan di dalam pemrosesan data seterusnya, dan juga agar mengurangi kemungkinan model yang akan digunakan menjadi rusak.

Di dalam menerapkan *feature selection* pada datanya, saya menggunakan tiga buah teknik, yaitu Mutual Information, Chi-Squared, dan ANOVA (Analysis of Variance). Ketiga teknik tersebut tidak saya pilih secara sembarangan saja, tetapi dengan memperhatikan kebutuhan dari data yang saya miliki, yaitu ada yang bersifat *categorical-numerical* dan *categorical- categorical*. Namun, untuk memastikan apakah ketiga teknik tersebut menghasilkan keuntungan bagi data, saya perlu melakukan suatu evaluasi terhadap ketiga teknik *feature selection* tersebut.







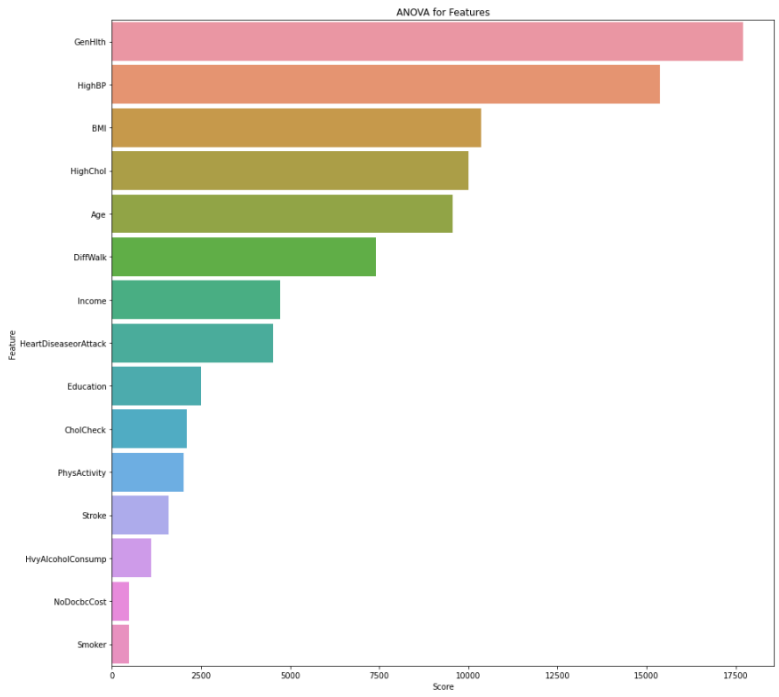


Dari keempat graph yang saya berikan, sayang sekali ternyata menunjukkan hasil yang kurang baik, yaitu bahwa garis merah, yang merupakan akurasi awal yang dihasilkan, berada pada posisi cukup paling atas ketimbang garis biru, yang merupakan nilai-nilai akurasi yang diberikan pada jumlah tertentu untuk fitur yang digunakan. Hal ini karena

dengan demikian maka pilihan bagi saya untuk mengurangi jumlah fitur, tetapi juga sekaligus mendapatkan nilai akurasi yang bertambah sangat sedikit.

Namun, terlepas dari keterbatasan hasil yang diberikan, hal tersebut tidak mengartikan bahwa saya tidak seharusnya mengurangi fitur jika akurasinya berkurang, tetapi masih ada beberapa pertimbangan lainnya, seperti jumlah memori dan model atau pun distribusi yang diberikan ketika fitur dikurangi. Kemudian, setelah juga memperhatikan dari graph-graph yang diberikan, saya memutuskan untuk menggunakan metode ANOVA (Analysis of Variance) sebagaimana terlihat bahwa hanya pada graph milik ANOVA sajalah yang memberikan pilihan terbanyak di dalam mengurangi jumlah fitur dengan akurasi yang tetap tidak terlalu jauh berbeda dengan akurasi awalnya. Lebih tepatnya, saya memilih ANOVA dengan jumlah fitur yang digunakan adalah menjadi 15 fitur input dan 1 buah output sehingga total kolom yang tersisa secara keseluruhan adalah 16.

Solusi:



Mengenai analisa dari datanya untuk menghasilkan solusi berdasarkan permasalahan yang saya angkat, yaitu untuk menentukan indikasi terbaik untuk penyakit diabetes, sebenarnya graph di samping sudah dapat cukup mewakilkannya atau menjawabnya. Hal ini karena graph menunjukkan tingkat relasi antara fitur input (yang berada pada graph) dan output atau target variabelnya (mengenai apakah suatu kasus dikatakan terkena diabetes atau tidaknya). Jika dilihat pada graph, terlihat bahwa fitur gen memiliki nilai yang tertinggi, yang mengartikan bahwa faktor gen ternyata merupakan yang paling berpengaruh di antara faktor-faktor lainnya, sedangkan

fitur *smoker* atau merokok memiliki tingkat yang terendah, yang mengartikan bahwa kemungkinan orang yang merokok akan terkena penyakit diabetes merupakan yang terkecil di antara faktor yang lainnya.

Kesimpulan:

Kesimpulan yang saya berikan dapat dibagi menjadi dua, yaitu berdasarkan pemrosesan data (prepemrosesan) dan masalah yang diangkat.

Berdasarkan pemrosesan data, terdapat beberapa poin yang dapat saya jadikan suatu kesimpulan. Pertama, prepemrosesan data memiliki sifat yang tidak bisa dianggap sebagai sesuatu yang berjalan lurus di dalam satu jalan, tetapi saya mendapati bahwa sifatnya lebih ditentukan kepada kondisi dari data yang digunakan, misalnya pada kasus saya yang mengharuskan adanya teknik prepemrosesan data tambahan, yaitu berupa *resampling data*. Selain itu, di dalam prepemrosesan data juga sangat penting untuk dapat memperhatikan, bukan hanya sifat dari datanya saja, melainkan juga jumlah fisik dari datanya, yang kemudian menjadi salah satu factor penting di dalam menentukan jenis teknik prepemrosesan data yang akan digunakan.

Kemudian, kesimpulan berdasarkan masalah yang diangkat, saya menyimpulkan bahwa solusi yang telah saya berikan masih belum dikatakan sempurna, bahkan sebenarnya masih sangat jauh. Hal ini karena beberapa pertimbangan berikut.

- Jumlah data yang digunakan masih belum cukup,
- Akurasi yang diberikan, meskipun sudah dilakukan banyak teknik prepemrosesan data, tetap memberikan nilai yang rendah,
- Menurut saya, *data imbalance* merupakan salah satu faktor utama yang menjadi alasan dari tidak terlalu baiknya, baik itu hasil dari model yang gunakan maupun akurasinya.

Sekian, kesimpulan yang dapat saya berikan atas proyek ini. Terima kasih.

“Di hadapan TUHAN yang hidup, saya menegaskan bahwa saya tidak memberikan maupun menerima bantuan apapun—baik lisan, tulisan, maupun elektronik—di dalam ujian ini selain daripada apa yang telah diizinkan oleh pengajar, dan tidak akan menyebarkan baik soal maupun jawaban ujian kepada pihak lain.”

Jason Caleb Erwin Piay