**Nama : Jihan Lailatul A**

**NIM : 1301170455**

**Kelas : IF 41 03**

**LAPORAN TUGAS BESAR MACHINE LEARNING**

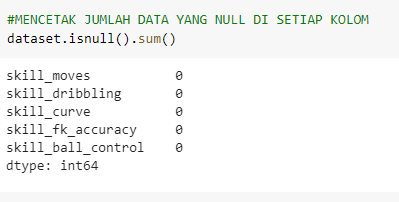
**Classification**

1. **Formulasi Masalah**

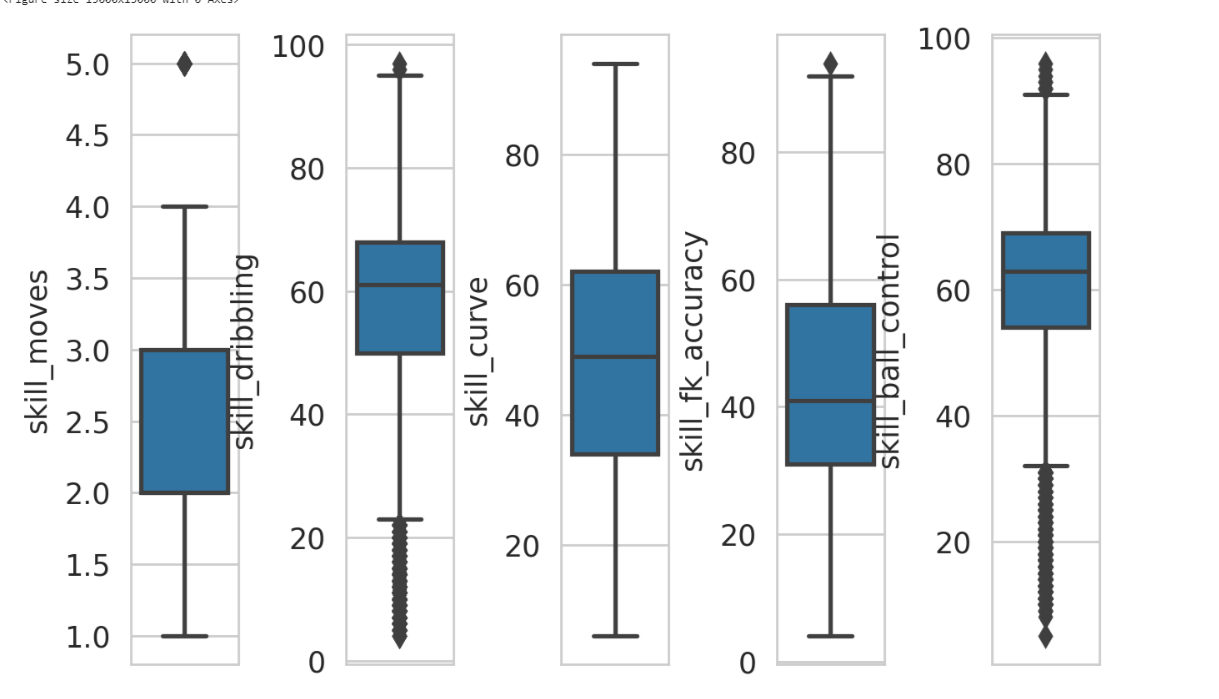
* Problem yang akan diselesaikan adalah memprediksi player fifa20 termasuk ke dalam ke dalam kelas Skill Moves rate yang mana – { 1,2,3,4,5} berdasarkan fitur Skill Dribbling, Skill Curve, Skill fk Accuracy, dan Skill Ball Control
* Dataset yang tersedia memiliki banyak kolom namun tipe data pada setiap kolom mayoritas adalah integer dan nilai keterkaitannyapun lumayan tinggi sehingga pemilihan kolom maupun label lebih mudah
* Permasalahan yang ditemukan ada banyaknya outlier
* Terdapat potensial bias apabila menggunakan fitur height

1. **Eksplorasi dan Persiapan Data**

* Pada dataset tidak terdapat nilai null atau NaN sehingga tidak diperlukan penghapusan maupun penggantian data (Missing Value)



* Namun, terdapat outlier pada beberapa fitur sehingga data-data tersebut perlu di diganti dengan nilai meannya. Outlier pada Skill Dribbling merupakan data yang nilainya dibawah 22 diganti dengan 55. Outlier pada Skill Ball Control adalah data yang nilainya dibawah 35 diganti dengan 58.



1. **Pemodelan**

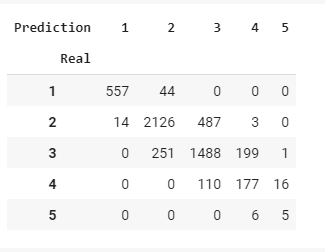
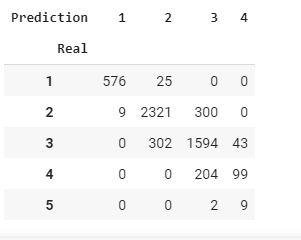
Pemodelan dilakukan petama-tama dengan memisahkan data fitur dengan labelnya. Kemudian data dibagi menjadi 70% data train dan 30% data test.

Saya memilih metode naïve Bayes karena merupakan salah satu algoritma /metode terbaik untuk klasifikasi dan jika ada nilai yang hilang maka dapat diabaikan dalam perhitungan.

1. **Eksperimen**

Eksperimen dilakukan dengan mengubah metode pemodelan dengan metode SVM. Saya memilih SVM karena SVM memiliki konsep yang lebih matang dan lebih jelas secara matematis dibandingkan dengan teknik-teknik klasifikasi lainnya.

1. **Evaluasi**

Confusion Table Naïve Bayes Confusion Table SVM

Evaluasi menggunakan Confusion table mempermudah melihat banyaknya data yang meleset pada setiap kelas. Confusion Table atau Confusion Matrix tersebut menunjukan banyaknya data prediksi yang sesuai dan tidak sesuai dengan kelas yang sebenarnya. Pada table confusion tersebut menunjukkan table SVM banyak data yang tidak sesuai yang kelas seharusnya lebih sedikit daripada table Naïve Bayes.

Hasil Akurasi dari Naïve Bayes : 79.88%

Hasil Akurasi dari SVM : 84%

Untuk data yang telah disediakan metode yang memiliki akurasi yang lebih baik adalah metode SVM.

1. **Kesimpulan**

Pembuatan classification ini termasuk mudah karena diuntungkan oleh dataset yang bentuk datanya mempermudah pembuatan program. Mempermudah pembuatan program dalam hal, data pada setiap fitur tidak perlu di encoding terlebih dahulu, tidak memiliki banyak Null, nilai korelasi antar fiturnya cukup banyak yang tinggi, selain itu mudah menentukan label kelas yang variasinya tidak banyak. Sehingga nilai akurasi yang didapatkan juga bisa lebih tinggi.

**Clustering**

**Formulasi Masalah**

Pada clustering ini fitur Y adalah Height dan fitur X adalah Weight karena 2 fitur tersebut secara umum saling bergantung.

**Eksplorasi dan Persiapan Data**

Persiapan data yang dilakukan adalah dengan melihat apakah terdapat data yang NaN atau kosong agar tidak mengganggu plotting dan perhitungan.

**Pemodelan**

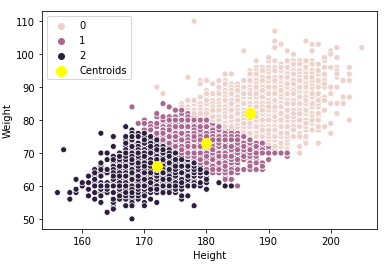
Model yang dipilih adalah K Means karena menurut penulis K Means memiliki tingkat kerumitan yang rendah untuk perhitungan dan pembuatan programnya. Pembuatan program algoritma K Means dilakukan dengan cara:

Pertama-tama, penginisialisasian K dan titip centroid sebanyak K.

Kemudian, lakukan looping untuk menghitung nilai Euclidian distance dan mengupdate nilai centroidnya setiap round loop.

Loop berakhir ketika perhitungan nilai titik centroid tidak berubah.

Didapatkan model berikut:



SSE = 490321

**Kesimpulan**

Meskipun pembuatan k-means merupakan clustering paling sederhana dan mudah dipahami, pembuatan algoritma atau programnya bisa sedikit rumit karena banyaknya variable yang dibutuhkan terutama untuk menyimpan nilai sementara yang digunakan pada looping.