**Analisa Dota 2 Hero Clustering**

**Menggunakan Metode K-Means++**

**Muhammad Imanudin Saputra (1706047100)**

**LATAR BELAKANG**

**Dota 2** adalah sebuah permainan [Arena pertarungan daring multipemain](https://id.wikipedia.org/wiki/Arena_pertarungan_daring_multipemain), dan merupakan sekuel dari [*Defense of the Ancients*](https://id.wikipedia.org/wiki/Defense_of_the_Ancients) mod pada *Warcraft 3: Reign of Chaos* dan *Warcraft 3: The Frozen Throne*. DotA 2 dikembangkan oleh [Valve Corporation](https://id.wikipedia.org/wiki/Valve_Corporation), terbit juli 2013 dota 2 dapat dimainkan secara gratis pada [sistem operasi](https://id.wikipedia.org/wiki/Sistem_operasi) [Microsoft Windows](https://id.wikipedia.org/wiki/Microsoft_Windows), [OS X](https://id.wikipedia.org/wiki/OS_X) and [Linux](https://id.wikipedia.org/wiki/Linux). Dota 2 dapat dimainkan secara eksklusif melalui distributor resmi valve, [Steam](https://id.wikipedia.org/wiki/Steam). (Wikipedia)

Dota 2 dimainkan oleh 2 tim yang beranggota 5 orang pemain, setiap tim memiliki markas yang berada dipojok peta, setiap markas memiliki satu bangunan bernama "Ancient", Di mana tim harus berusaha menghancurkan "Ancient" tim lawan agar dapat memenangkan pertandingan. Setiap pemain hanya dapat mengontrol satu karakter "Hero" yang berfokus pada menaikan level, mengumpulkan gold, membeli item dan melawan tim lawan untuk menang. (Wikipedia)

Dalam game online pada umumnya memiliki istilah-istilah tersendiri yang digunakan pada game tersebut. Berikut beberapa istilah umum pada game dota 2 yang akan digunakan:

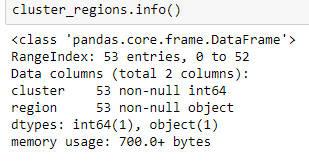
* Unit: seluruh bagian dari permainan yang termasuk sebagai tim sendiri atau tim lawan.
* Hero: Karakter yang dapat digunakan setiap pemain. Dalam 1 permainan, hero hanya dapat dipilih 1 kali (khusus pada data permainan yang digunakan pada analisis data).
* Creeps: Monster yang termasuk tim sendiri, tim lawan, ataupun tidak keduanya. Ada beberapa hero yang dapat mengendalikan creep, sehingga creep tersebut menjadi bagian dari tim sendiri.
* Tower: Bangunan yang ada pada setiap tim yang dapat menyerang yang bukan bagian dari timnya. Untuk menghancurkan ancient lawan (memenangkan permainan) harus menghancurkan beberapa tower lawan terlebih dahulu
* Ancient: Bangunan tim lawan yang harus dihancurkan agar menang.
* Gold: Mata uang yang digunakan untuk membeli item. Setiap saat selalu bertambah dan bisa didapatkan dengan membunuh creep atau hero lawan, dan menghancurkan bangunan tim lawan (tower).
* Xp: experience point, untuk menambahkan level. Didapatkan dengan berada di sekitar creeps atau hero lawan saat mereka terbunuh, atau membantu memberikan serangan untuk membunuh hero lawan, dan berada di sekitar bangunan (tower) tim lawan yang hancur.
* Kills: terhitung sebanyak banyaknya serangan terakhir sehingga membuat hero lawan mati
* Last hits: banyaknyanya serangan terakhir yang diberikan kepada creeps sehingga menyebabkan creep mati
* Deny: membunuh unit tim sendiri (creeps, bangunan, hero). Menyebabkan lawan tidak mendapatkan gold (jika berada di sekitar unit yang hancur atau terbunuh) dan hanya mendapatkan sebagian xp.
* Assist: Membantu memberikan serangan kepada hero lawan sehingga hero lawan mati.
* Level: besarnya tingkatan hero yang dimainkan.
* HP: health point, menyatakan besarnya poin sampai unit hancur atau mati.
* Healing: menyembuhkan HP suatu unit.

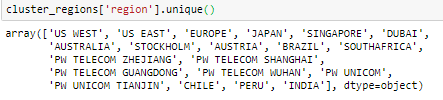
1. Data

Data yang digunakan adalah data mengenai 50.000 permainan dota 2 yang diambil dari <https://www.kaggle.com/devinanzelmo/dota-2-matches>, yaitu data *cluster regions,* *match*, *hero names*, dan *players*. Berikut rincian data set tersebut:

1. Cluster Regions

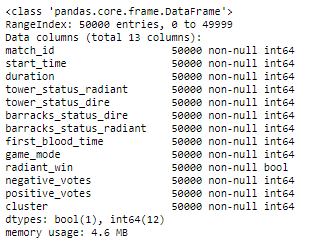
Cluster regions berisi data mengenai kode cluster dan region berdasarkan cluster tersebut. Pada data tersebut terdapat 53 kode cluster dan terdapat 20 region dengan rincian seperti berikut:





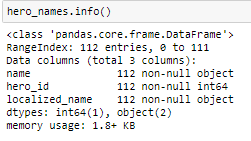
1. Match

Data match berisi 50.000 data permainan yang berbeda. Berikut seluruh variabel pada data *match*:



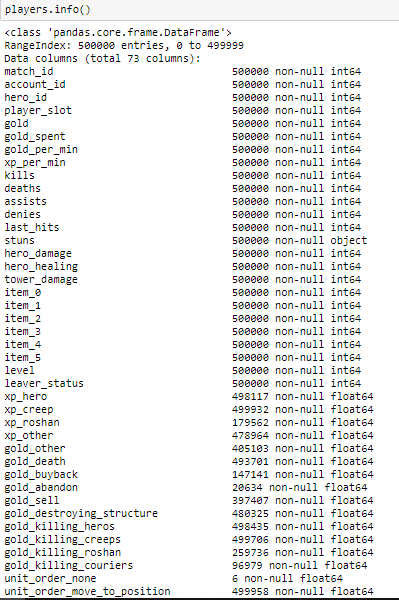
1. Hero Names

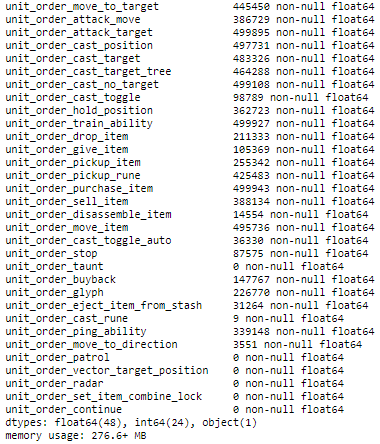
Data *hero names* berisi mengenai nama hero, hero\_id, dan nama hero yang umum digunakan. Pada data tersebut terdapat 112 *hero* berbeda yang dapat dimainkan oleh pemain. Berikut rincian data yang digunakan:



1. Players

Data *players* berisi mengenai permainan sebanyak 50.000 dan 500.000 baris. Setiap baris berisi detail mengenai statistik setiap pemain/*player*. Data tersebut terdiri dari 73 variabel. Berikut rincian data tersebut:





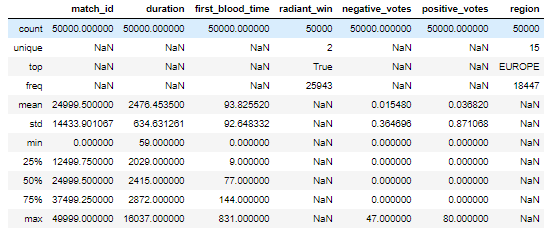
Analisis data akan menggunakan metode K-mean++ dengan bantuan pyhton 3.7.

K-mean clustering adalah metode unsupervised learning yang digunakan untuk mengelompokkan data berdasarkan karakteristik-karakteristik tertentu. Pada metode tersebut akan dipilih berapa banyak cluster yang akan dibentuk. Selanjutnya dipilih centroid pertama secara acak sebagai titik awal setiap cluster dan kemudia dilakukan perhitungan berulang-ulang sampai centroid stabil dengan menggunakan nilai mean. K-mean++ adalah perbaikan pada metode K-mean dimana inisialisasi centroin pertama dilakukan dengan tidak acak.

1. Tujuan Analisis Data
2. Mencari tahu karakteristik dari permainan dota 2
3. Membuat suatu pengelompokkan *hero* yang digunakan dalam permainan dota 2
4. Analisis Data
5. Pre-processing
6. Pada data *match* terdapat 13 variabel termasuk variabel *cluster*. Sehingga akan ditambahkan variabel *region* juga dengan melihat data *cluster regions*, sehingga sekarang terdapat 14 variabel. Selanjutnya dari 14 variabel hanya akan digunakan 7 variabel yaitu:

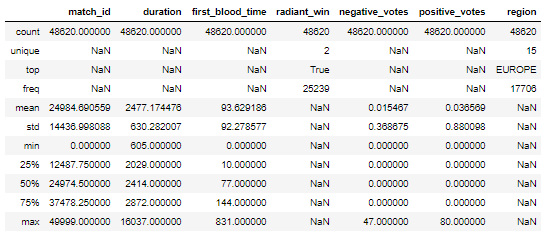
* match\_id: id permainan
* duration: durasi permainan (dalam detik)
* first\_blood\_time: waktu pertama kali *hero* dibunuh oleh *hero* lawan (dalam detik)
* radiant\_win: indikator apakah tim radian menang atau tidak (1: radian menang, 0: dire menang)
* negative\_votes: banyaknya votes negatif yang diberikan pada permainan tersebut. Votes dapat diberikan oleh pemain permainan tersebut ataupun orang lain yang menonton permainan tersebut.
* positive\_votes: banyaknya votes positif yang diberikan pada permainan tersebut. Votes dapat diberikan oleh pemain permainan tersebut ataupun orang lain yang menonton permainan tersebut.
* region: wilayah permainan tersebut dimainkan.

Berikut rincian data:

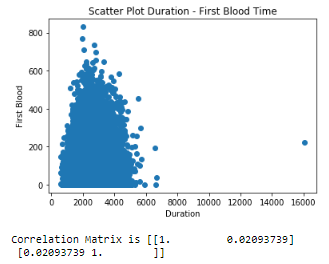


Jika dilihat dari durasi permainan tercepat yaitu 59 detik atau kurang dari 1 menit. Hal ini bisa terjadi jika semua pemain dalam 1 tim meninggalkan permainan. Sehingga jika digunakan data dari permainan yang hanya sebentar dimana pemain meninggalkan permainan, statistik dari pemain tersebut (*hero* yang dimainkan) tidak dapat digunakan karena diinginkan hasil statistik yang tidak menggambarkan permainan dari *hero* tersebut sehingga pengelompokkan yang dibuat memiliki hasil yang tidak optimal. Dalam hal ini digunakan permainan dengan durasi minimal selama 10 menit atau 600 detik, karena pada umumnya dalam waktu tersebut setidaknya masih dapat memungkinkan untuk suatu tim memenangkan permainan, yaitu mengahancurkan *ancient* tim lawan.

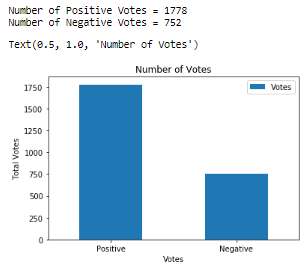
Sehingga rincian data yang digunakan adalah sebagai berikut:



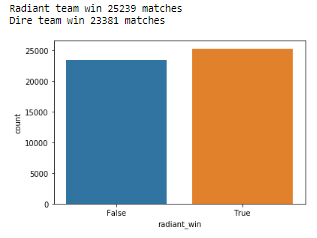
Pada data, durasi tercepat permainan adalah 605 detik atau sekitar 10 menit dan waktu tercepat *hero* dibunuh pertama kali dalam permainan oleh *hero* lain adalah sekitar 0 detik setelah waktu permainan dimulai. Hal ini karena setelah pemain memasukin *map* permainan terdapat waktu yang dihitung mundur sampai waktu permainan berjalan, pada saat tersebut bisa terjadi pembuhan *hero*. Sedangkan durasi terlama dari permaian adalah selama 16037 detik atau sekitar 268 menit dan waktu terlama *hero* dibunuh adalah 831 detik setelah waktu permainan dimulai. Selain itu akan dilihat apakah terdapat korelasi antara kedua waktu tersebut, berikut *scatter plot* dan matriks korelasi:



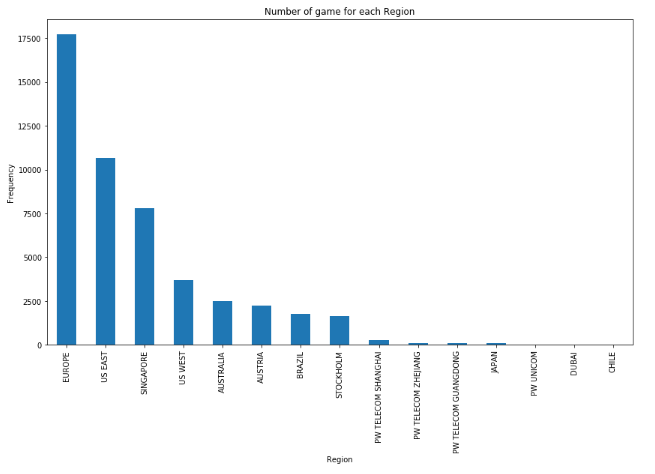
Jika dilihat terdapat *outlier* pada data durasi permainan, yaitu nilai maksimumnya. Untuk negative votes dan positive votes berikut chart perbandingannya:



Dapat dilihat jauh lebih banyak jumlah total positive votes (1778) lebih banyak dibandingkan negative votes (752), sekilas hal ini dapat berarti bahwa orang-orang lebih sering memberikan respon positif terhadap suatu permainan dibandingkan negatif. Berikut adalah perbandingan jumlah tim yang memenangkan permainan:

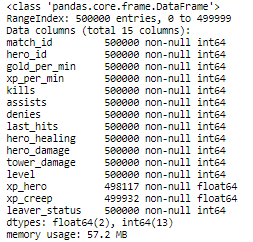


Jika dilihat tidak terdapat perbedaan yang jauh untuk perbandingan jumlah kemenangan tim radiant dengan tim dire. Sekilas dapat dikatakan bahwa tidak terlihat pengaruh tim mana yang lebih sering menang, untuk menentukannya lebih pasti diperlukan analisis lebih lanjut. Berikut adalah wilayah dimana permainan dimainkan:



Jika dilihat dari wilayah permainan yang dimainkan terbanyak adalah EUROPE, diikuti dengan US EAST, SINGAPORE dan seterusnya. Hal ini dapat berarti bahwa jumlah pemain dota 2 terbanyak berdasarkan data yang digunakan adalah EUROPE, karena region menggambarkan server daerah yang digunakan untuk bermain sedangkan orang-orang akan cenderung memilih server yang dekat dengan wilayah mereka bermain karena jaringan koneksi dengan permainan yang lebih tidak terganggu.

1. Pada data players terdapat 73 variabel yang menyatakan karakteristik dari permainan seoranng pemain untuk setiap permainan yang ada. Dari 73 variabel tersebut hanya akan digunakan sebanyak 15 variabel yang dipilih sehingga memungkinkan untuk dapat menggambarkan karakteristik *hero* yang dimainkan seorang pemain pada analisis yang akan dilakukan. Berikut ringkasan data *players* dan penjelasan setiap variabelnya:



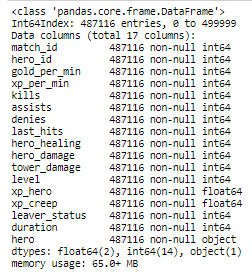
* match\_id: id permainan
* hero\_id: id setiap hero
* gold\_per\_min: rata-rata banyaknya gold yang didapat setiap menit
* xp\_per\_min: rata-rata banyaknya xp yang didapat setiap menit
* kills: banyaknya kill yang dilakukan
* assists: banyaknya assist yang dilakukan
* denies: banyaknya deny yang dilakukan
* last\_hits: banyaknya last hit yang dilakukan
* hero\_healing: besarnya helaing yang diberikan terhadap hero
* hero\_damage: besarnya kerusakan yang diberikan kepada hero
* tower\_damage: besarnya kerusakan yang diberikan kepada tower tim lawan
* level: besarnya level saat permainan berakhir
* xp\_hero: besarnya xp yang didapat dari kill atau assist hero tim lawan
* xp\_creep: besarnya xp yang didapat dari creep
* leaver\_status: indikator apakah pemain tersebut meninggalkan permainan (abandon). 1: iya, 0: tidak.

Berdasarkan variabel leaver\_status, maka tidak diinginkan mereka/pemain yang memiliki nilai leaver\_status 0 karena statistik dari pemain tersebut tidak dapat menggambarkan dengan baik statistik dari hero yang dimainkan. Maka untuk pemain yang meninggalkan permainan akan dibuang.

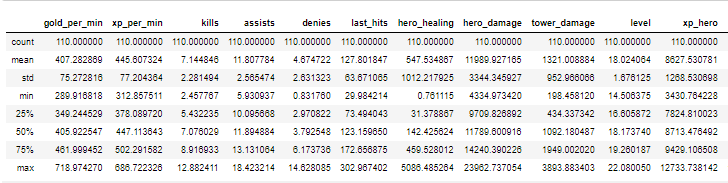
Pada variabel xp\_hero dan xp\_creep terdapat beberapa data yang kosong, sehingga seluruh data pada pemain tersebut akan dibuang.

Sebelumnya telah dipilih permainan yang akan digunakan datanya berdasarkan durasi permainan tersebut. Pada data *players* akan diseleksi permainan mana yang datanya akan digunakan berdasarkan analisis sebelumnya.

Selanjutnya akan ditambahkan nama hero yang umumnya digunakan (variabel hero pada data *hero\_names*) berdasarkan hero\_id yang ada. Berikut ringkasan data *players* dengan data berjumlah 487,116:

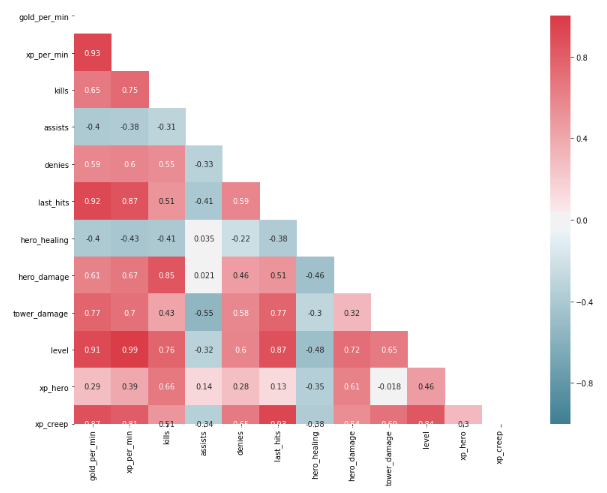


1. Untuk melakukan pengelompokkan berdasarkan hero maka data tersebut akan disusun berdasarkan hero-nya dengan menghitung mean dari seluruh data hero tersebut. Sebelumnya dibuang terlebih dahulu variabel yang tidak tidak akan digunakan kembali yaitu hero\_id, match\_id, leaver\_status, dan duration. Berikut ringkasan data yang didapatkan untuk beberapa variabel:



1. K-Means++ Clustering

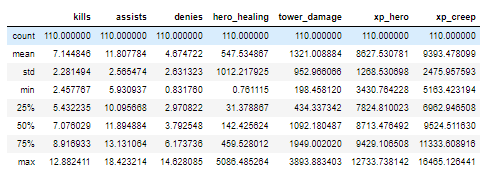
Untuk melakukan pengelompokkan hero akan digunakan metode K-Means++ clustering. Sebelumnya akan dilihat apakah ada korelasi antara setiap variabel yang akan digunakan. Berikut heat map yang didapat:



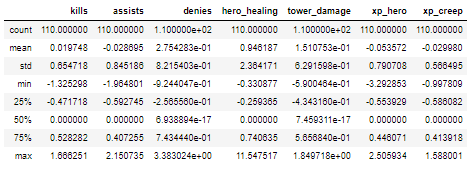
Berdasarkan gambar di atas dapat dilihat bahwa terdapat beberapa variabel yang saling berkorelasi dengan nilai koefisien korelasi lebih dari 0.8. Berdasarkan paper Cluster Analysis Gets Complicated oleh Rajan Sambadan, dikatakan bahwa adanya kolinearitas antara variabel yang digunakan dapat mempengaruhi hasil dari clustering terutama k-means clustering sehingga diperlukan penanganan-penanganan tertentu, salah satunya (dan yang dipilih) yaitu membuang salah satu variabel yang berkorelasi tersebut. Sehingga dari variabel-variabel tersebut yang akan dibuang adalah variabel gold\_spent, xp\_per\_min, gold\_per\_min, last\_hits, dan level. Sementara variabel yang tetap digunakan adalah kills, assists, denies, hero\_healing, tower\_damage, xp\_hero, dan xp\_creep. Pemilihan variabel dilakukan dengan pertimbangan karakteristik dari setiap variabel yang dipilih sebelumnya dapat dijelaskan dengan cukup jelas, berikut alasan pemilihan variabel-variabel tersebut:

1. gold\_per\_min dapat diwakilkan oleh kills dan last\_hits, karena semakin banyak kills dan last\_hits yang didapat maka semakin banyak gold yang didapatkan juga
2. level dan xp\_per\_min dapat diterangkan oleh xp\_hero dan xp\_creep, karena xp\_per\_min dipengaruhi secara langsung oleh xp\_hero dan xp\_creep, sedangkan xp\_per\_min semakin besar maka level semakin besar.
3. last\_hits memiliki korelasi yang cukup besar dengan xp\_creep. Hal ini menandakan bahwa umumnya hero mendapatkan xp dari creep (membunuh creep) dan bukan hero lain (membunuh hero lain). Tetapi dengan melihat jumlah last\_hits dapat diketahui informasi lebih banyak karena dengan mendapatkan last\_hits pemain tersebut mendapatkan gold tambahan
4. kills dipilih menggantikan hero\_damage karena dengan banyaknya kills yang didapat dapat memberikan informasi lebih banyak mengenai xp dan gold yang didapat. Untuk hero\_damage lebih menjelaskan mengenai assists (yang belum tentu didapatkan), xp dan gold yang didapat, sedangkan variabel assists sendiri akan digunakan sehingga tidak perlu pengganti.
5. denies, hero\_healing, dan tower\_damage tidak dapat dijelaskan variabel lainnya, dan hal ini juga dapat dilihat dari korelasi dengan variabel lain yang tidak terlalu besar.

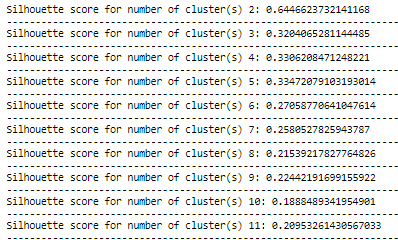
Berikut ringkasan data yang didapat:

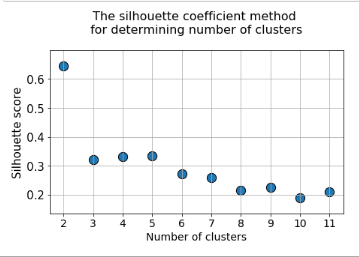


Selanjutnya akan dilakukan standardisasi menggunakan RobustScaler pada data untuk dapat digunakan pada metode k-means++ clustering karena data sebelumnya memiliki ukuran yang berbeda-beda. Berikut ringkasan data yang didapat setelah distandardisasi:



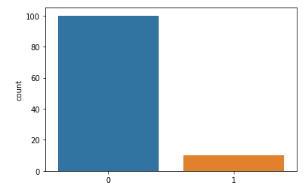
Akan ditentukan jumlah cluster optimum dengan menggunakan metode silhouette. Berikut output yang didapat:

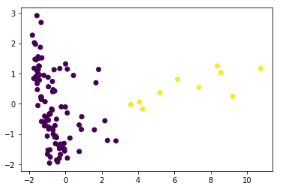




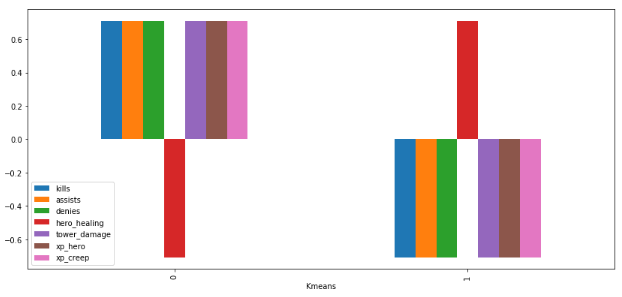
Didapatkan nilai cluster k optimum sebesar 2.

Berikut penggunakan metode k-means++ dan hasilnya:

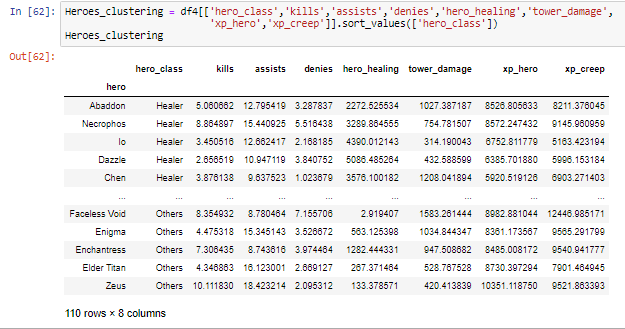




Jika dilihat dari hasilnya, jumlah pada kedua kelompok tidak seimbang dan berdasarkan scatter plot terlihat persebaran data yang tidak rata. Hal ini bisa terjadi karena adanya outlier. Hasil pengelompokkan datanya adalah sebagai berikut:



Jika dianalisa lebih jauh, kelompok kedua telihat memiliki nilai hero\_healing di atas rata-rata hero lainnya. Maka akan digunakan penamaan kelompok pertama sebagai ‘Others’ dan kelompok kedua sebagai ‘Healers’. Berikut hasil akhir data yang didapat.



1. Kesimpulan dan Saran

DOTA 2 termasuk salah satu game yang memiliki fitur yang sangat banyak dan rumit untuk dilakukan analisa. Hero pada DOTA 2 memiliki berbagai macam karakteristik masing-masing. Walaupun pada umumnya sudah terdapat standar role hero yang digunakan tetapi banyak pemain yang mengabaikannya.

Dari hasil analisis didapatkan 2 kelompok yang cukup berbeda jauh karakteristiknya, kelompok yang memiliki kill, assist, deny, tower damage dan xp yang tinggi dengan kelompok dengan healing\_hero yang tinggi. Hasil yang didapatkan tersebut cukup bagus karena pada kenyataannya hero healer pada hasil analisis memang merupakan hero yang memiliki skill untuk healing. Namun pada kelompok pertama terlalu banyak hero yang menempati kelompok tersebut dan terlihat tidak seimbang.

Hal ini dapat dikarenakan adanya outlier pada data, dapat dilihat juga pada scatter plot hasil yang didapatkan setelah melakukan k-means++. Karena k-means++ sensitif terhadap outlier. Oleh karena itu akan lebih baik jika dilakukan pengecekan outliers terlebih dahulu sebelumnya untuk dapat menggunakan metode tersebut. Selain itu masih banyak fitur game DOTA 2 yang tidak dimasukkan dalam analisis, sehingga akan lebih baik jika fitur-fitur tersebut ditambahkan sebagai variabel-variabel yang akan dianalisis

1. Referensi

Sambandan, Rajan. Cluster Analysis Get Complicated. <https://www.trchome.com/docs/5-cluster-analysis-gets-complicated/file>, diakses pada tanggal 5 Januari 2020, pukul 17.00.

Dota 2. <https://id.wikipedia.org/wiki/Dota_2>, diakses pada tanggal 6 Januari 2020, pukul 18.00.

Ramadhani, RizkY Desi. Memahami K-Mean Clustering Pada Machine Learning Dengan Python. <https://medium.com/@16611129/memahami-k-mean-clustering-pada-machine-learning-dengan-phyton-430323d80868>, diakses pada tanggal 6 Januari 2020, pukul 20.00.

Dataset. <https://www.kaggle.com/devinanzelmo/dota-2-matches>