KAJIAN SILARA POKOK MENGGUNAKAN PEMBELAJARAN MESIN

Akmal Amien Talib
Nor Samsiah Sani

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Kajian silara pokok menggunakan kaedah pembelajaran mesin adalah satu kajian yang dibangunkan untuk menyelidik data LiDAR (Light Detection and Ranging) dataset yang disediakan oleh Pusat Angkasa Universiti Kebangsaan Malaysia. Dengan melakukan kajian terhadap data sebenar, banyak maklumat yang dapat diekstrak untuk kegunaan pada masa hadapan. Justeru itu, hasil kajian ini boleh dijadikan sebagai platform kepada penyelidik lain dalam menyelidik data sebenar yang berkaitan dengan dataset LiDAR. Metodologi yang digunakan dalam kajian ini ialah model Waterfall. Perisian yang digunakan untuk pembangunan kajian adalah LAStools dan Python. Maklumat ketinggian titik awan dalam data LiDAR dihasilkan berdasarkan lapisan pada dataset untuk menghasilkan kajian lanjut. Kajian silara pokok memfokuskan pencarian pengelompokan dengan menggunakan kaedah pembelajaran mesin k-means

1 PENGENALAN

Pengecaman Silara Pokok ditaktifkan sebagai pengecaman ciri-ciri pokok daripada kelompok pokok yang bertaburan di dalam hutan oleh pelbagai jenis atau spesis pokok yang berada di kawasan kajian. Dengan ketebalan hutan tropika di Malaysia ianya amatlah sukar untuk mengenal pasti Silara Pokok dengan cara tradisional dan sangat tidak efisen. Sejak beberapa tahun yang lepas penggunaan *Airborne Scanner System* yang menghasilkan tiga dimensi persekitaran maya titik awan kepada permukaan topologi hutan banyak digunakan dengan menggunakan LiDAR (*Light Detection and Ranging*) bagi memudahkan pengesanan pokok didalam hutan. LiDAR merupakan teknologi kawalan pengesanan yang menggunakan cahya dalam bentuk laser untuk mengukur variable jarak ke bumi dan ianya merupakan kawalan

pengesanan aktif. Namun, daripada LiDAR sahaja tidak dapat melakukan pengecaman silara pokok individu. Oleh itu, teknik algoritma pengelompokan *k-means*, *mean-shift* dan *DBSCAN* digunakan untuk membangunkan model pembelajaran mesin. Mengekstrak pokok individu secara terus melalui LiDAR akan membawa kepada kaedah pengelompokan. Kaedah ini khususnya memfokuskan pada pemerhatian *n* didalam kelompok *k* dimana setiap pemerhatian dipunyai kepada kelompok yang terdekat dengan mencuba mengecilkan jumlah keseluruhan jarak ruang titik *Euclidean* dengan kelompok sentroid-sentroid.

2 PENYATAAN MASALAH

LiDAR (*Light Detection and Ranging*) menghasilkan titik awan LiDAR untuk mengesan ciri individu pokok. Kaedah Silara Individu Pokok ialah algoritma untuk melihat silara pokok daripada imej *raster* ataupun titik awan densiti tinggi. Kaedah imej *raster* ini hanya mampu untuk mengesan pokok dibahagian atas permukaan sahaja dan tidak mengesan bahagian bawah pokok yang berada dihutan.

Seharusnya pokok dikesan sebagai individu pokok tetapi sekelompok pokok masih berada dalam kelompok yang sama bagi kaedah imej *raster* ini. Pengesanan sekelompok pokok mengakibatkan penyukaran untuk mendapatkan maklumat spesifik tentang ciri-ciri satu pokok dalam kawasan kajian. Kaedah yang lebih efektif perlu diaplikasikan bagi mengekstrak silara pokok individu didalam hutan dalam bentuk 3-dimensi daripada titik awan LiDAR.

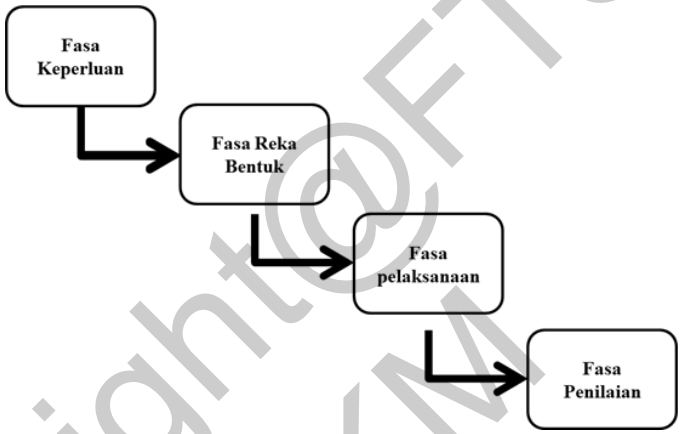
3 OBJEKTIF KAJIAN

Objektif projek ini adalah untuk membangunkan model pengelompokan k-means untuk mengenalpasti bilangan kelompok yang terhasil. Seterusnya mengenalpasti ciri-ciri setiap kelompok yang terhasil daripada kajian. Selain itu, dapat menghasilkan visualisasi kelompok yang terhasil dan akhir sekali membuat perbandingan dengan kaedah pengelompokan *mean-shift* dan *dbscan.*

4 METOD KAJIAN

Penggunaan model pembelajaran mesin yang berlainan dijalankan iaitu *k-means*, *DBSCAN* dan

mean-shift untuk memastikan perbezaan antara ketiga-tiga kaedah pengelompokan. Penggunaan model yang sesuai penting untuk memastikan perjalanan projek berjalan dengan lancar dan menajmin hasil kerja yang berkualiti. Penentuaan kelompok, visualisasi data dan analisa data melibatkan beberapa fasa pembangunan reka bentuk yang bersesuaian. Model pembangunan ini diadaptasi daripada model asas yang diperkenalkan oleh Van Den Broek (Riza dan Yuwaldi 2002). Fasa pembangunan reka bentuk termasuk fasa perancangan, reka bentuk, pelaksanaan dan penilaian. Model ini penting untuk memastikan perjalan projek berjalan dengan lancar. Rajah 1 menunjukkan model *waterfall* yang digunakan dalam kajian ini.



Rajah 1.1 Model *Waterfall*

4.1 Fasa Perancangan

Fasa ini melibatkan proses mengenalpasti masalah, objektif kajian, persoalan kajian dan penentuaan skop kajian dilakukan. Seterunya kajian susatera melibatkan pengumpulan dan pencarian pembacaan jurnal atau kajian yang lepas bagi menghasilkan pandangan yang jelas pada kajian yang akan dijalankan. Kajian berkaitan pengelompokan menggunakan kaedah pembelajaran mesin k-means, DBSCAN dan mean-shift dijalankan untuk mencetus idea dan inspirasi. Contoh topik yang berkaitan dikaji terutama berkaitan dengan kaedah pengelompokan menggunakan pembelajaran mesin. Penggunaan internet digunakan untuk mencari dan mengumpul maklumat kemudian dipersembahkan.

4.2 Fasa Reka Bentuk

Dalam fasa reka bentuk, kutipan awal data daripada LiDAR (*Light Detection and Ranging*)

dilaksanakan bagi memahami data dan biasakan dengan data yang diperoleh. Seterusnya algoritma *k-means* digunakan terhadap dataset LiDAR untuk menentukan bilangan kelompok yang terhasil daripada teknik pengelompokan *k-means.* Dalam kajian ini kaedah untuk mencari *centroid* awal perlu dicara dengan menggunakan dua kaedah iaitu kaedah *Elbow* dan juga kaedah *Silhouette.* Kedua-dua kaedah ini digunakan bagi menjalankan pengelompkan menggunakan kaedah pembelajaran mesin k-means yang digunakan. Seterusnya, analisa setiap kelompok daripada kelompok yang terhasil akan dijalankan bagi mengekstrak maklumat penting seperti ketinggian kelompok yang terhasil. Terdapat beberapa kaedah normalisasi yang digunakan bagi mendapatkan maklumat yang terdapat daripada pengelompokan tersebut.

4.3 Fasa Pelaksanaan

Dalam fasa pelaksanaan, ianya melibatkan pengubahan data mentah kepada bentuk yang boleh dijadikan model menggunakan *machine learning* algoritma. Seterusnya melibatkan ETL (*Extract, Transform, Load*) proses bagi membolehkan data menjadi data yang berkualiti tinggi. Dalam fasa ini juga pelaksanaan model pada data sebenar diadaptasikan pada model dan mengekstrak maklumat penting yang terhasil daripada kajian.

4.4 Fasa Penilaian

Pada fasa penilaian, perbandingan model pembelajaran mesin dibandingkan dan menilai daripada setiap algoritma yang menghasilkan analisa yang tepat pada data sebenar. Sekiranya keputusan tidak menepati kehendak, metodologi ini perlu disemak semua daripada langkah pertama untuk memahami bagaimana kesalahan ditimbulkan. Jadual Rancangan Pembangunan

5 HASIL KAJIAN

Bahagian ini membincangkan hasil daripada proses pembangunan kajian silara pokok menggunakan pembelajaran mesin. Penerangan yang mendalam tentang reka bentuk kajian diterangkan. Fasa reka bentuk adalah fasa yang penting dalam pembangunan kajian. Dalam kajian, penggunaan *Python* digunakan bagi membangunkan model pembelajaran mesin k-means. Dataset LiDAR dalam format LAS ditukarkan kepada CSV file dengan menggunakan LasTools kemudian divisualisasikan menggunakan *scatter* plot sebagai tujuan untuk

memberikan pandangan awal kepada dataset yang akan dikaji.

Seterusnya normalisasi data dilakukan terhadap dataset LiDAR dengan menggunakan normalisasi data yang berbeza iaitu *Z-Score*, *MinMaxScaler* dan *Quantile Transformer*. *Quantile Transformer* dipilih kerana memberikan nilai positif pada atribut Z dan signifikan berbanding teknik normalisasi yang lain.

Dataset yang telah dinormalisasi divisualisaikan bersama pengelompokan k-means untuk dianalisa pada fasa seterusnya. Seterusnya binned data pada atribut Z bagi penilaian ketinggian titik awan LiDAR dengan melabelkan data pada kluster bersama dengan lapisan satu, lapisan dua, lapisan tiga atau lapisan empat. Setiap lapisan akan dianalisa untuk menganalisa kelompok yang terhasil. Akhir sekali penilaian pada kaedah pengelompokan k-means, mean-shift dan DBSCAN dilakukan bagi menilai setiap teknik pengelompokan.

Jadual 1 Normalisasi Data *Quantile Transformer*

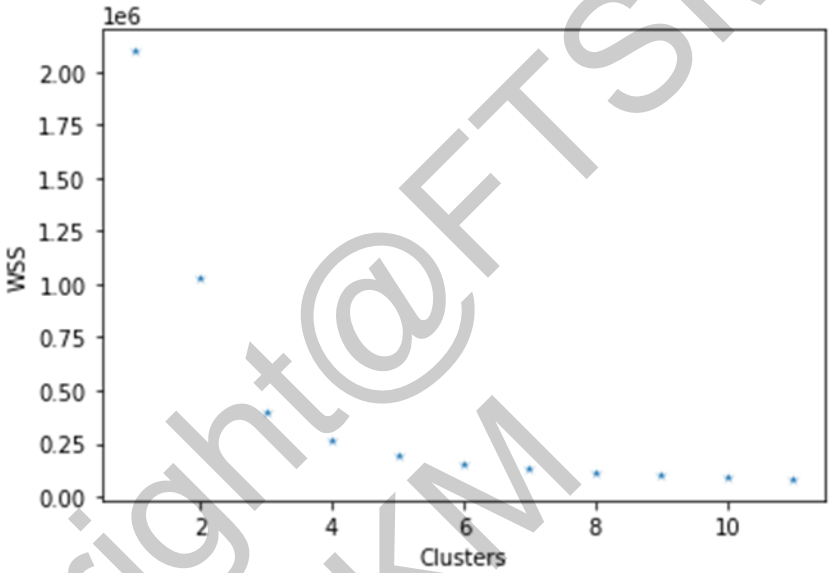
X	Y	Z
0.838367	0.912357	0.010511
0.909137	0.845899	0.014515
0.847284	0.829429	0.014515
0.847684	0.828122	0.016517
0.682775	0.790713	0.016517

Jadual 1 menunjukkan normalisasi data menggunakan kaedah *Quantile Transformer.* Nilai pada atribut Z iaitu ketinggian memberikan nilai positif. Ianya penting untuk memenuhi objektif kajian iaitu menilai ketinggian titik awan LiDAR.

Jadual 1.1 Kelompok dengan skor WSS

Clusters	WSS Skor
1	2.097150e+06

2	1.0322741e+06
3	3.9314220e+05
4	1.907348e+05
5	1.572271e+05
6	1.322721e+06
7	1.175232e+05
8	1.030480e+05
9	9.140575e+04
10	8.294498e+04



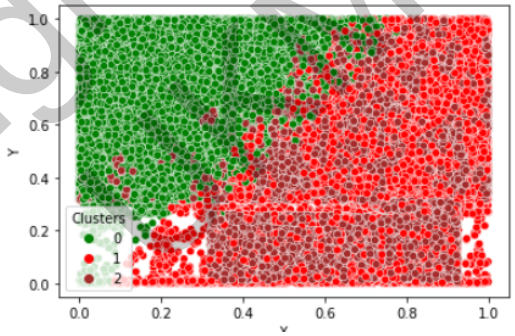
Rajah 1.2 *Elbow* Plot

Jadual 1.1 dan rajah 1.2 menunjukkan WSS(*Within Cluster-Sum of Squared*) skor bersama dengan Elbow plot bagi mencari nilai kelompok awal yang akan digunakan dalam kajian ini. Berdasarkan graf diatas, tiga merupakan kelompok awal yang terhasil namun kaedah ini agak samar dengan untuk menentukan nilai kelompok awal.

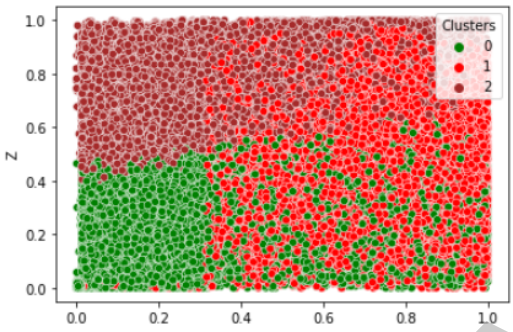
```
Silhouette score for k(clusters) = 2 is 0.5280798028775586
Silhouette score for k(clusters) = 3 is 0.5914617843927005
Silhouette score for k(clusters) = 4 is 0.5668357505404386
Silhouette score for k(clusters) = 5 is 0.5398923942314482
Silhouette score for k(clusters) = 6 is 0.5033446045817402
Silhouette score for k(clusters) = 7 is 0.503019562768979
Silhouette score for k(clusters) = 8 is 0.5008019536730351
Silhouette score for k(clusters) = 9 is 0.4823611233623312
Silhouette score for k(clusters) = 10 is 0.46906561021953974
Silhouette score for k(clusters) = 11 is 0.43614708143511804
Silhouette score for k(clusters) = 12 is 0.4287223633012518
```

Rajah 1.3 *Silhouette* Skor

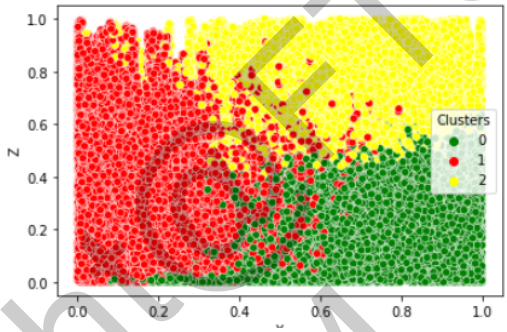
Rajah 1.3 menunjukkan *silhouette* skor bagi dataset LiDAR ini. Penentuaan kelompok awal berdasarkan *silhouette* skor yang tertinggi dipilih sebagai kelompok awal. Berdasarkan rajah 1.3 silhouette skor tertinggi ialah 0.5614 iaitu pada kelompok tiga. Kaedah Elbow dan Silhouette memberikan nilai kelompok awal yang sama , maka tiga merupakan kelompok awal yang terhasil.



(A)



(B)



(C)

Rajah 1.4 Visualisasi kelompok titik awan

Visualisai kelompok yang terhasil selepas pengelompokan k-means berjaya dijalankan. Visualisasi latitud bersama longitud (A) , visualisasi longitud bersama ketinggian (B) dan visualisasi logitud bersama ketinggian (C) ditunjukkan pada rajah 1.4. Terdapat tiga kelompok yang dibezakan dengan warna tersendiri.

Jadual 1.2 Rumusan data dalam kelompok 0, 1 dan 2

Lapisan	**Bilangan titik**	**Peratusan (&)**
Lapisan 1	14139	54
Lapisan 2	10487	41
Lapisan 3	1463	5
Lapisan 4	0	0

(A)

Lapisan	Bilangan titik	Peratusan (&)
Lapisan 1	12084	28
Lapisan 2	14803	34
Lapisan 3	11257	26
Lapisan 4	4599	10

(B)

Lapisan	Bilangan titik	Peratusan (&)
Lapisan 1	0	0
Lapisan 2	924	2
Lapisan 3	1340	37
Lapisan 4	21612	60

(C)

Berdasakan jadual 1.2, rumusan pada kelompok 0 (A) , kelompok 1 (B) dan kelompok 2 (C) yang terhasil terhadap setiap kelompok. Beberapa analisa dapat dirumuskan mengikut lapisan titik pada setiap kelompok. Kelompok 0 mempunyai banyak titik dikawasan rendah disebabkan pada lapisan 1 titik mempunyai 54% tinggi berbanding pada lapisan yang lain. Pada kelompok 1, ketinggian titik berada pada skala yang lebih kurang sama dan tiada perbezaan yang ketara . Akhir sekali pada kelompok 2, banyak titik tertumpu pada lapisan yang tinggi kerana data menunjukkan data sebanyak 60% berada pada lapisan 4.

Jadual 1.3 Keputusan *t-test* dan *p-value*

Kaedah	*t-test*	*p-value*
K-means & DBSCAN	0.1256	0.9115
Mean-Shift & DBSCAN	-0.1644	0.8844
K-means & Mean-Shift	-1.3052	0.3217

Jadual 1.3 menunjukkan kaedah k-means bersama means-shift dan mean-shift bersama DBSCAN memperoleh nilai negatif bagi ujian statistik iaitu t-value -1.3052 dan -0.1644 masing-masing. Nilai t negatif bermaksud kelompok dalam kedua-dua algoritma adalah hampir sama dan kaedah tidak signifikan. Dengan ini, kedua-dua pasangan kaedah ini adalah tidak berkaitan dan disahkan boleh dibangunkan tanpa sebarang pengaruhg antara satu sama lain. Namun

demikian, kaedah k-means dan DBSCAN memperoleh t-value yang positif iaitu 0.1256 bermaksud algoritma tersebut adalah tidak sama dan signifikan. Dengan ini, kedua-dua algoritma adalah berkaitan dan tidak disokong untuk membangunkan bersama disebabkan kemungkinan keputusan adalah hampir sama.

Bagi penilaian menggunakan p-value, kaedah mean-shif dan k-means merupakan kaedah yang paling bagus dan sesuai dibangunkan antara ketiga-tiga pasangan kaedah kerana algoritma memperoleh nilai p yang kecil iaitu 0.3217 dan kebarangkalian berlaku persamaan keputusan sampal data adalah kecil.

6 KESIMPULAN

Secara keseluruhannya, kajian silara pokok menggunakan pembelajaran telah berjaya dibangunkan dalam tempoh masa yang diberikan dengan kejayaan pencapaian maklumat, objektif, menepati skop yang didefinisikan serta mengikuti metodologi yang dicadangkan sehingga kajian diselesaikan. Pembangun kajian dan penyelidik dapat memperoleh maklumat dan kemudahan yang dibekalkan iaitu maklumat tentang dataset dan meghasilakn sebuah model pengelompokan. Banyak input yang dapat diperoleh serta banyak ilmu yang dipelajari sepanjang kajian pembangunan silara pokok menggunakan pembelajaran mesin dijalankan

7 RUJUKAN

Venter J, de Waal A., Willers C. 2019. Specializing CRISP-DM for Evidence Mining. Diakses dari: https://link.springer.com/chapter/10.1007%2F978-0-387-73742-3_21

Zhang, Jlin, X ,2013 Filtering airborne LiDAR data by embedding smoothness-constrained segmentation in progressive TIN densification. Diakses dari: https://www.sciencedirect.com/science/article/abs/pii/S0924271613001019

S.Saeedi, F. Samadzadegan b , N. El-Sheimy, 2009, Object extraction from LiDAR data using an artificial swarm bee colony clustering algorithm. Diakses dari: http://www.pf.bgu.tum.de/isprs/cmrt09/pub/CMRT09_Saeedi_et_al.pdf.

Denise Laes, Richard Warnick, Wendy Goetz, Paul Maus USDA Forest Service Remote Sensing tips, LiDAR Applications for forestry and geosciences Diakses dari: http://static1.squarespace.com/static/59c944de59cc68469d159b28/t/5a21e5b453450aa90cada3bd/1512170934265/lidar-overview.pdf

Gupta,Sandeep. 2013. Single tree delineation LiDAR. Diakses dari: https://www.researchgate.net/publication/298791738_Single_Tree_Delineation_Using_Airborne_LIDAR_Data.

XingboHu, Wei Chen. 2017.Adaptive mean shift-based identification of individual tree using Airborn LiDAR data, identifying individual trees and delineating Diakses dari: https://pdfs.semanticscholar.org/1f49/2950397a7d2d135f5f31e0d088bb3184fbee.pdf

R. Gaulton . 2010.LiDAR maping of canopy gaps in continuous cover forest. Diakses dari: https://www.tandfonline.com/doi/abs/10.1080/01431160903380565

Martin Ester. 2017.Density-based algorithm for discovering clusters. Diakses dari: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Konstantinos,G.Derpanis. 2005. Mean Shift Clustering. Diakses dari: http://www.cse.yorku.ca/~kosta/CompVis_Notes/mean_shift.pdf

Wilfred van Casteran, 2017. The waterfall model and agile. Diakses dari: https://www.researchgate.net/publication/313768860_The_Waterfall_Model_and_the_Agile_Methodologies_A_comparison_by_project_characteristics_-_short