

ALGORITMA PENGELOMPOKAN BERASASKAN ANJAKAN-MIN UNTUK IDENTIFIKASI CIRI POKOK MENGGUNAKAN AIRBORNE LASER SCANNING (LIDAR)

Lim Qi Wen
Azuraliza Abu Bakar

ABSTRAK

Mengenal pasti pokok secara individu dan menggambarkan struktur kanopi pokok hutan daripada data awan titik yang diperoleh melalui Airborne Laser Scanning juga dikenali 'Light Detection and Ranging' (LiDAR) merupakan implikasi yang amat ketara dalam persediaan program perhutanan. Kajian ini mencadangkan algoritma anjakan-min untuk mengelompok data LiDAR termasuk tiga atribut utama iaitu x (titik latitud), y (titik longitud) dan z (titik ketinggian). Kemudian, pengekstrakan ciri-ciri pokok dalam tiga kelompok. Seterusnya, pengelasan lapisan pokok berdasarkan titik ketinggian iaitu tahap ketinggian pokok. Algoritma anjakan-min digunakan dalam kajian ini supaya dapat menganggarkan ciri-ciri pokok dengan lebih tepat. Algoritma ini dibangunkan atas cadangan pakar penyelidik dari Intitut Perubahan Iklim, UKM. Kajian ini telah menggunakan metodologi CRISP-DM yang terdiri daripada beberapa fasa iaitu pengenalpastian masalah, pengumpulan data, pra-pemprosesan data dan persediaan, pembangunan model ramalan, penilaian model dan penyebaran model. Antara proses utama yang digunakan adalah pengekstrakan ciri-ciri, pemilihan ciri-ciri yang penting daripada pemboleh ubah. Prestasi algoritma pengelompokan anjakan-min telah dibandingkan dengan dua algoritma pengelompokan iaitu *k-means* dan pengelompokan berasaskan ketumpatan. Keputusan mendapati algoritma pengelompokan berasaskan ketumpatan telah memberi prestasi yang lebih baik dari algoritma *k-means* dan algoritma anjakan-min menggunakan pengukuran *Silhouette Coefficient* dan perbandingan tiga kaedah penormalan data. Hasil kajian ini dapat digunakan untuk meramal ketinggian dan ketumpatan kanopi pokok dalam kawasan tertentu dan seterusnya dapat menganggarkan kadar perlepasan oksigen kawasan dan membantu pihak berkepentingan dalam perbuatan keputusan. Ia juga dijangka dapat membantu menilai pengukuran pertumbuhan pokok dalam kawasan tertentu. Selain itu, peta pokok mempunyai nilai praktikal untuk bidang ekologi dan sektor penghutanan dalam menguatkuasakan peraturan pengezonan di hutan riparian. Pendekatan identifikasi pokok individu adalah sangat padan dengan teknik berdasarkan bidang untuk mengukur karbon hutan dan mempunyai teori asas yang kuat dalam meminimumkan ralat dan ketidaktentuan.

1 PENGENALAN

Airborne Laser Scanning juga dikenali 'Light Detection and Ranging' (LiDAR) merupakan teknologi remote sensing aktif yang membekalkan pengukuran jarak secara terus antara alat pengimbas laser dengan permukaan bumi dan objek-objek di atasnya diterangi oleh pancaran laser dalam bentuk denyut laser. LiDAR merupakan teknologi yang direkabentuk bagi tujuan perolehan maklumat ketinggian samada ketinggian permukaan bumi ataupun ketinggian objek di permukaan bumi (Song et al., 2015). Mengikut Liu (2008), Raber et al. (2007) dan Tarolli (2014), teknologi ini adalah amat sesuai bagi mendapatkan data pokok yang berketepatan tinggi serta bagi kawasan yang luas seperti di hutan tropika. LiDAR sebagai remote sensing aktif bagi mendapatkan maklumat permukaan bumi tanpa menyentuh objek tersebut. Hal ini dapat dilakukan oleh LiDAR sendiri dengan mengesan dan merekod tenaga yang dipantulkan atau dipancarkan dan kemudian diproses, dianalisis dan diguna dalam program perhutanan. Denyutan cahaya kemudian akan digabungkan dengan data lain yang direkodkan oleh sistem airborne dan seterusnya menghasilkan maklumat individu pokok dari atas pokok hingga permukaan tanah dalam tiga-dimensi (3D) seperti saiz silara, diameter silara, ketinggian silara dan sebagainya dengan penyelesaian ruang dan ketepatan yang tinggi. Teknik Airborne LiDAR makin hari makin terkenal dan kerap digunakan dalam pengekstrakan dan perincian individu pokok kerana kecekapannya dalam memodelkan kanopi hutan dan menyelesaikan beberapa tugas inventori. Terdapat kaedah-kaedah atau algoritma yang dikembangkan untuk melaksanakan pengenalpastian individu pokok dengan menggunakan data airborne LiDAR. Pendekatan raster-based memanfaatkan model ketinggian kanopi (CHM) yang berasal daripada data mentah LiDAR kerap digunakan dalam pembahagian individu pokok. Pendekatan yang paling kerap digunakan untuk menjalankan pembahagian pokok dengan menggunakan data

awan titik secara tiga-dimensi (3D) yang dijana oleh LiDAR ialah kawasan pertumbuhan yang mana merupakan salah satu kategori kaedah pembahagian berdasarkan kawasan. Selain itu, pendekatan pengelompokan seperti algoritma *k-means* merupakan satu kaedah yang sangat popular untuk mengekstrakan individu pokok daripada data LiDAR secara langsung. Algoritma pengelompokan *k-means* memecahkan set data yang diberikan kepada nombor (k) yang tetap mengikut kelompok berdasarkan matlamat yang ingin dicapai. Pendekatan *raster-based* dapat melakukan penapisan yang cepat dengan data yang rumit dan serasi dengan gambar remote sensing. Walaubagaimanapun, CHM tidak sesuai untuk mencirikan pelbagai silara terutamanya hutan tropika. Hubungan antara ketinggian pokok dan lebar silara adalah lemah seterusnya mengakibatkan kesukaran dalam menentukan saiz penapis yang optimum untuk penapisan maksimum tempatan demi mengekalkan pelbagai skala silara di hutan yang padat seperti hutan tropika. Sebagai alternatif, pendekatan *point-based* dapat menjalankan secara langsung dengan data airborne laser scanning LiDAR yang asli dan juga pendekatan yang lebih sesuai digunakan. Pendekatan *point-based* dapat mengesan pokok-pokok kecil yang berada di permukaan tanah dengan ketepatan yang amat tinggi.

Sejak beberapa tahun yang lalu, teknik pengelompokan anjakan-min telah digunakan bersama dengan data LiDAR untuk menggambarkan pokok secara individu. Anjakan-min merupakan teknik analisis ruang-ciri yang serbaguna dan kuat dan didapati amat menarik dalam pemprosesan imej dan komuniti visi komputer (Cheng, Y., 1995; Comaniciu, D.; Meer, P, 2002). Hal ini disebabkan anjakan-min tidak bergantung pada andaian model geometri, aplikasinya dapat dipanjangkan sampai pemprosesan data awan titik yang tidak berstruktur. Contohnya, Melzer mendapati kegunaan anjakan-min dalam pemprosesan data LiDAR dan memanfaatkan prosedur anjakan-min sebagai alat untuk mengekstrak dan membina semula talian kuasa daripada ALS point cloud. (Melzer, T, 2007) Anjakan-min adalah berlainan dengan kaedah region growing dan algoritma *k-means*, anjakan-min tidak perlu titik benih, oleh itu

tidak sensitif terhadap permulaan. Selain itu, algoritma anjakan-min tidak perlu membuat andaian terhadap nombor kelompok dan tidak perlu mengandai bentuk yang ditetapkan berkaitan dengan data kelompok, satu-satunya parameter yang perlu dinyatakan ialah lebar jalur kernel untuk setiap atribut yang digunakan. Algoritma dapat memastikan algoritma pengelompokan anjakan-min boleh menunjukkan pokok-pokok yang kecil atau pokok yang di bahagian bawah, sementara ini tiada kaedah yang boleh capai sedemikian.

Projek ini bertujuan untuk mengelompok dan mengekstrak ciri serta mengelaskan lapisan pokok dengan menggunakan Airborne Laser Scanning (LiDAR) melalui proses pengekstrakan ciri dan pengelasan pokok. Ciri-ciri individu pokok yang telah diekstrak dapat digunakan dalam meramalkan dan mengelaskan pokok kepada lapisan.

2 PERNYATAAN MASALAH

Ketepatan dan kejituan model pembelajaran mesin yang sedia ada bagi pengekstrakan ciri-ciri pokok dan pengelasan pokok adalah kecil, tidak memuaskan dan perlu ditingkatkan lagi. Selain itu, pendekatan algoritma yang dibangunkan tidak lagi mencapai keputusan yang baik bagi proses pengekstrakan dan pengelasan pokok di hutan tropika yang padat dan seterusnya menyebabkan terdapat masalah berlakunya pokok rendah dan kecil telah dihalang oleh pokok besar dan tinggi akibat pokok kecil dan rendah tidak dapat dikesan. Kebanyakan kajian lepas yang telah dijalankan adalah berdasarkan imej spektrum dan aerial, kajian lepas kurang tertumpu kepada pengesanan pokok dan identifikasi spesies pokok dengan menggunakan titik pancaran dan kaedah pengelompokan.

Pemanasan global dan perubahan iklim makin hari makin serius berlaku di bumi kita pada masa kini. Hal ini perlu dipandang serius oleh masyarakat disebabkan kesan buruk akan mempengaruhi kehidupan masyarakat. Peningkatan suhu dunia akibat pemanasan global terjadi apabila atmosfera bumi dan permukaannya dipanaskan secara beransur-ansur oleh haba yang terperangkap serta tidak terlepas ke luar angkasa. Peningkatan suhu mendadak disebabkan pelepasan karbon dioksida yang tinggi dan pertambahan gas lain seperti klorofluorokarbon (CFC) oleh manusia seterusnya menggugat fungsi sebenar atmosfera bumi. Oleh itu, pengekstrakan dan perincian maklumat individu pokok adalah kaedah yang penting untuk dijalankan. Hal ini dapat mengekstrak data individu pokok yang mempunyai ketepatan yang tinggi supaya menyumbang dalam mendapatkan biomass, kadar karbon, isi padu, hasil pengeluaran dan struktur hutan untuk meramal iklim di bumi kita. Dengan ini, keadaan hutan dapat diramalkan dan kadar pelepasan karbon juga dapat sentiasa diperhatikan supaya langkah seterusnya boleh dirancang.

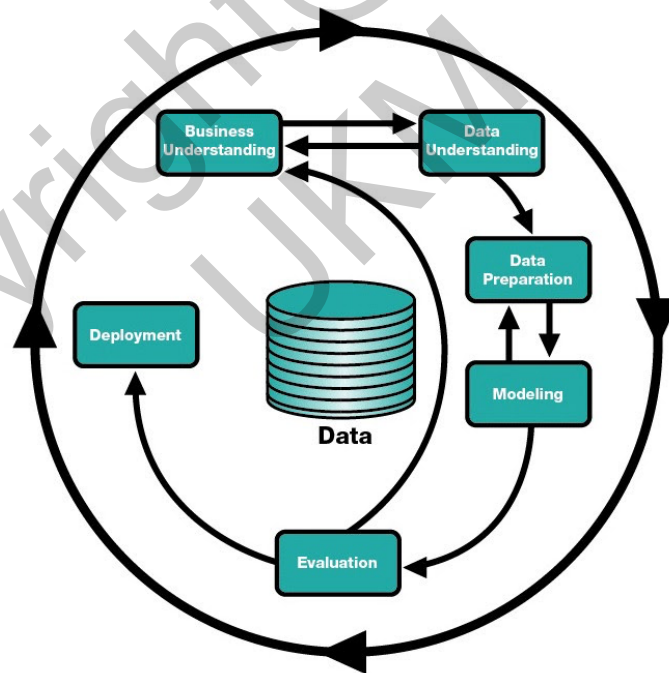
3 OBJEKTIF KAJIAN

Objektif kajian ini adalah:

1. Melaksanakan proses pra-pemprosesan dan penyediaan data awan titik LiDAR.
2. Mencadangkan pendekatan algoritma pengelompokan anjakan-min untuk pengekstrakan ciri-ciri dan pengelasan pokok dengan menggunakan data daripada Airborne Laser Scanning (LiDAR).
3. Menilai prestasi pengelompokan dengan perbandingan kaedah pengelompokan *K-Means* dan pengelompokan berasaskan ketumpatan (*Density Based Clustering*).

4 METOD KAJIAN

Projek ini akan dilaksanakan dengan menggunakan metodologi CRISP-DM (*The Cross Industry Standard Process for Data Mining*). Metodologi CRISP-DM ini merupakan model proses yang mempunyai enam fasa yang melabelkan kitaran hidup sains data secara semula jadinya. CRISP-DM adalah suatu konsorsium perusahaan yang didirikan oleh Komisi Eropa pada tahun 1996 dan telah ditetapkan sebagai proses standard dalam data mining yang dapat diaplikasikan di pelbagai sektor industri. Rajah 4.1 telah menunjukkan fasa-fasa metodologi CRISP-DM.



Rajah 4.1 Fasa-fasa Metodologi CRISP-DM

4.1 Fasa Pemahaman Perniagaan (*Business Understanding*)

Fasa pertama adalah untuk memahami tujuan dan kebutuhan dari sudut pandang perniagaan, kemudian menerjemahkan pengetahuan ini ke dalam pendefinisian masalah pada pemrosesan data. Pelbagai artikel, jurnal, buletin dan laporan kajian mengenai

domain ini perlu dikumpulkan untuk memahami situasi semasa. Algoritma anjakan-min dan model pembelajaran mesin yang bakal dibangunkan untuk menjalankan proses pengekstrakan pokok di hutan tropika juga dipastikan akan memenuhi keperluan sasaran pengguna iaitu dapat digunakan di kawasan yang besar secara luas demi mencapai matlamat pengurusan hutan dan inventori. Pengetahuan mengenai domain kajian adalah sangat penting untuk membangunkan algoritma, model dan sistem yang mempunyai nilai perniagaan yang diakui oleh pengguna.

4.2 Fasa Pemahaman Data (*Data Understanding*)

Fasa ini dimulai dengan pengumpulan data LiDAR. Data LiDAR (x, y, dan z) akan digunakan untuk menjalankan proses pengelompokan data LiDAR. Data x, y, dan z merupakan titik latitud, titik longitud dan titik ketinggian masing-masing. Selepas data x, y, dan z menjalankan proses pengelompokan, ciri dan maklumat pokok dapat diekstrak dengan menggunakan pendekatan algoritma anjakan-min supaya mendapatkan pemahaman yang mendalam tentang data.

4.3 Fasa Penyediaan Data (*Data Preparation*)

Fasa ini meliputi semua kegiatan untuk membangun set data akhir iaitu data yang akan diproses pada tahap pembangunan model pembelajaran mesin dari data mentah. Fasa ini dapat diulang beberapa kali. Pada tahap ini juga mencakup pemilihan tabel, rekod, dan atribut-atribut data x, y, dan z, termasuk proses pembersihan dan transformasi data untuk kemudian dijadikan masukan dalam fasa pembangunan model. Dalam kajian ini, tiga kaedah penormalan akan digunakan termasuk penormalan *min-max*, penormalan *z-score* dan penskalaan *quantile transformer*.

4.4 Fasa Pembangunan Model (*Modelling*)

Dalam fasa ini, pendekatan algoritma iaitu algoritma anjakan-min akan dipilih supaya membangunkan model untuk menjalankan proses pengelompokan, pengekstrakan

pokok serta pengelasan pokok. Algoritma anjakan-min merupakan teknik pengelompokan yang kebarangkalian ketumpatan bukan parametik untuk mengerat data awan titik hutan secara tiga-dimensi. Proses pengelompokan dengan menggunakan algoritma anjakan-min akan dijalankan kemudian mengekstrakan ciri dan mengelaskan lapisan pokok berdasarkan ciri yang telah diekstrak iaitu kelompok dan titik ketinggian. Algoritma anjakan-min akan digunakan dalam projek ini disebabkan algoritma anjakan-min merupakan teknik yang bagus dan sesuai untuk mencari anggaran ketumpatan yang lokal maxima secara optimal disebabkan kepadatan hutan tropika adalah tinggi.

4.5 Penilaian (*Evaluation*)

Pada tahap ini, algoritma dan model sudah terbentuk dan diharapkan memiliki kualiti yang baik jika dilihat dari sudut pandangan data analisis. Pada tahap ini akan dilakukan penilaian terhadap keberkesanan dan kualiti algoritma dan model sebelum digunakan dan menentukan algoritma dan model dapat mencapai tujuan yang ditetapkan pada fasa awal iaitu fasa kefahaman perniagaan. Latihan dan pengujian akan dijalankan dengan menggunakan data LiDAR untuk menguji ketepatan pengelompokan pokok secara individu. Kaedah yang akan digunakan ialah analisis perbandingan, *silhouette coefficient* dan plot serta *t-test* dan *p-value*. Keputusan akan berbanding dengan keputusan analisis menggunakan algoritma yang lain dan berbanding dengan menggunakan kaedah penskalaan ciri yang berlainan.

4.6 Fasa Penyebaran (*Deployment*)

Pada tahap ini, pengetahuan dan maklumat yang telah diperoleh akan diatur dan dibentangkan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Menurut catatan prestasi sistem pada fasa yang sebelumnya, fasa ini akan menentukan sama ada algoritma, model dan sistem yang dibangunkan sudah bersedia untuk disembarkan ke server untuk kegunaan awam. Algoritma juga akan diperlengkapkan dari segi sokongan

teknikal dan pengeluaran. Algoritma yang dibangunkan hanya akan disebarakan sekiranya semua komponen fungsi mencapai prestasi yang memuaskan.

4.7 Spesifikasi Keperluan

Perkakasan dan perisian yang diguna untuk membangun projek harus dipilih dengan teliti. Perkakasan dan perisian yang baik berfungsi dengan lancar serta menyokong pembangunan model pembelajaran mesin bagi identifikasi ciri pokok. Pemilihan perkakasan dan perisian yang tidak tepat boleh menjejaskan hasil projek. Jadual 4.1 merupakan senarai spesifikasi keperluan perkakasan yang digunakan dalam projek ini. Jadual 4.2 merupakan spesifikasi keperluan perisian untuk menyelesaikan pembangunan dan kajian projek ini.

Jadual 4.1 Spesifikasi Keperluan Perkakasan

Perkakasan	Komputer riba ACER SF314-54G
Ingatan Capaian Rawak (RAM)	4GB DDR4
Pemproses	Intel Core i5-8250U
Memori Dalaman	128GB SSD, 1TB HDD
Sistem Operasi	Windows 10 Home 64-bit
Kad Grafik	NVIDIA GeForce MX150

Jadual 4.2 Spesifikasi Keperluan Perisian

Perisian	Perincian
LAStool	Menukar .las file ke .text file.
Microsoft Word 2016	Menulis tesis projek tahun akhir.
Microsoft Excel 2016	Membuat data analisis bagi menunjukkan keputusan perjalanan pembangunan sistem.

Anaconda	Pengedaran sumber terbuka bagi bahasa pengaturcaraan Python untuk mengekalkan persekitaran Python.
Google Colab	Sebagai IDE untuk menulis algoritma python semasa membangunkan sistem ini.

5 HASIL KAJIAN

Dalam projek ini, algoritma anjakan-min akan dibangunkan bagi identifikasi ciri pokok individu. Data LiDAR yang telah menjalankan pra-pemprosesan akan dibahagikan kepada 90% data latihan dan 10% data ujian. Seterusnya, 10% data ujian iaitu 104858 data akan digunakan dalam memodelkan algoritma anjakan-min untuk mempercepatkan masa pemprosesan. Hal ini disebabkan data LiDAR yang besar menyebabkan proses pembangunan model menjadi perlahan. Terdapat tiga titik pusat kelompok berdasarkan x, y dan z telah dihasilkan melalui pengalihan dari titik ke titik data yang ketumpatan paling tinggi. Label algoritma anjakan-min telah digunakan untuk meramalkan bilangan kelompok.

Label algoritma anjakan-min telah dibahagikan kepada kelompok 0, kelompok 1 dan kelompok 2 berdasarkan titik pusat kelompok. Seterusnya, data x, y dan z akan dikelaskan mengikuti golongan kelompok. Kemudian, data LiDAR telah dikategorikan kepada empat lapisan dengan menggunakan fungsi '*binned*' berdasarkan z iaitu titik ketinggian secara individu. Lapisan 1 merupakan ketinggian dari -0.001 hingga 0.251. Lapisan 2 merupakan ketinggian dari 0.251 hingga 0.501. Lapisan 3 merupakan ketinggian dari 0.501 hingga 0.751 manakala lapisan 4 merupakan ketinggian dari 0.751 hingga 1.0. Hal ini juga dikatakan bahawa pokok dalam lapisan 1 lebih rendah dibandingkan dengan lapisan 2 seterusnya lapisan 3 lebih tinggi daripada lapisan 2. Lapisan 4 mempunyai pokok yang paling tinggi.

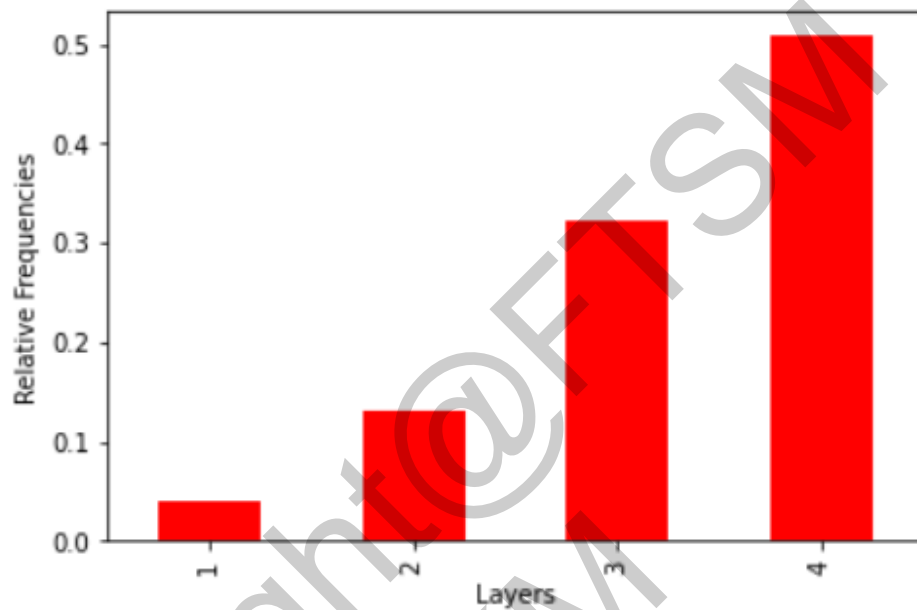
5.1 Kelompok 0

Dalam kelompok 0, data LiDAR telah dikategorikan berdasarkan lapisan 1 hingga lapisan 4. Jadual 5.1 telah menunjukkan jumlah data, peratusan dan kekerapan relatif berdasarkan lapisan dalam kelompok 0.

Jadual 5.1 Jumlah data dalam kelompok 1 berdasarkan lapisan

Lapisan	Jumlah	Peratusan (%)	Kekerapan Relatif
1	1509	3.94	0.03940
2	5004	13.06	0.1306
3	12339	32.22	0.3222
4	19450	50.78	0.5078

Rajah 5.1 telah menunjukkan carta turus secara menegak berdasarkan lapisan dan kekerapan relatif dalam kelompok 0.



Rajah 5.1 Carta turus secara menegak bagi kelompok 0

Berdasarkan jadual 5.1 dan rajah 5.1, lapisan 1 mempunyai bilangan data yang paling kecil manakala lapisan 4 mempunyai bilangan data yang paling banyak. Hal ini dapat dikatakan bahawa terdapat banyak pokok tinggi dan kurang pokok rendah dalam kelompok 0 disebabkan lapisan 4 mempunyai jumlah bilangan yang tinggi dan lapisan 1 mempunyai jumlah bilangan yang rendah.

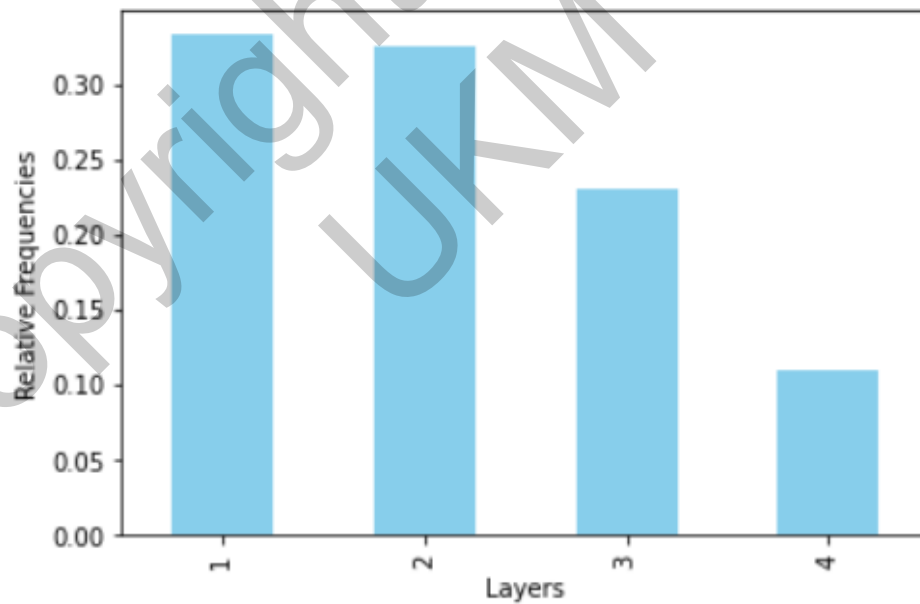
5.2 Kelompok 1

Dalam kelompok 1, data LiDAR juga dikategorikan berdasarkan lapisan 1 hingga lapisan 4. Jadual 5.2 telah menunjukkan jumlah data, peratusan dan kekerapan relatif berdasarkan lapisan dalam kelompok 1.

Jadual 5.2 Jumlah data dalam kelompok 1 berdasarkan lapisan

Lapisan	Jumlah	Peratusan (%)	Kekerapan Relatif
1	14951	33.33	0.3333
2	14622	32.59	0.3259
3	10344	23.06	0.2306
4	4947	11.03	0.1103

Rajah 5.2 telah menunjukkan carta turus secara menegak berdasarkan lapisan dan kekerapan relatif dalam kelompok 1.



Rajah 5.2 Carta turus secara menegak bagi kelompok 1

Berdasarkan jadual 5.2 dan rajah 5.2, lapisan 1 mempunyai bilangan data yang paling besar manakala lapisan 4 mempunyai bilangan data yang paling kecil. Hal ini dapat dikatakan bahawa terdapat banyak pokok rendah dan kurang pokok tinggi dalam

kelompok 1 disebabkan lapisan 1 mempunyai jumlah bilangan yang tinggi dan lapisan 4 mempunyai jumlah bilangan yang rendah.

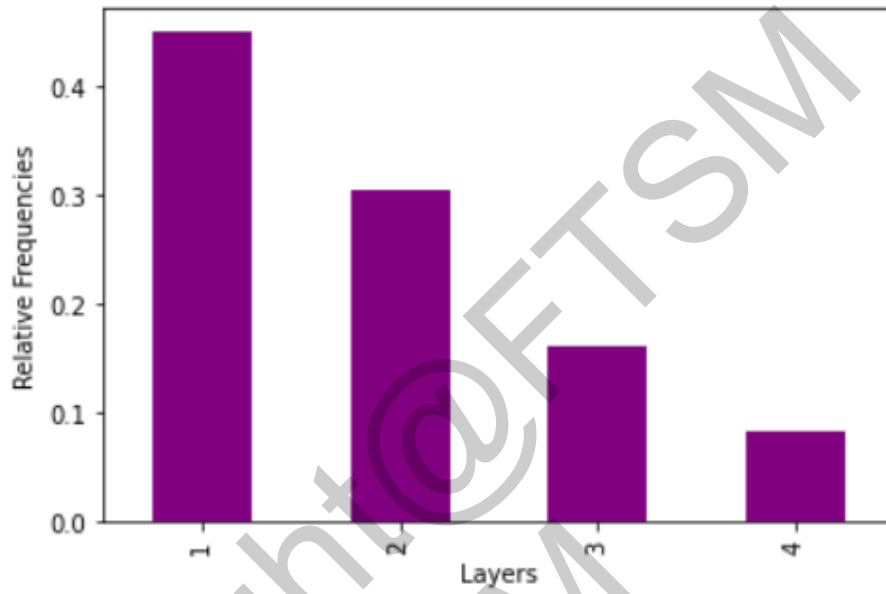
5.3 Kelompok 2

Dalam kelompok 2, data LiDAR juga dikategorikan berdasarkan lapisan 1 hingga lapisan 4. Jadual 5.3 telah menunjukkan jumlah data, peratusan dan kekerapan relatif berdasarkan lapisan dalam kelompok 2.

Jadual 5.3 Jumlah data dalam kelompok 2 berdasarkan lapisan

Lapisan	Jumlah	Peratusan (%)	Kekerapan Relatif
1	9763	45.01	0.4501
2	6601	30.43	0.3043
3	3529	16.27	0.1627
4	1799	8.29	0.0829

Rajah 5.3 telah menunjukkan carta turus secara menegak berdasarkan lapisan dan kekerapan relatif dalam kelompok 2.

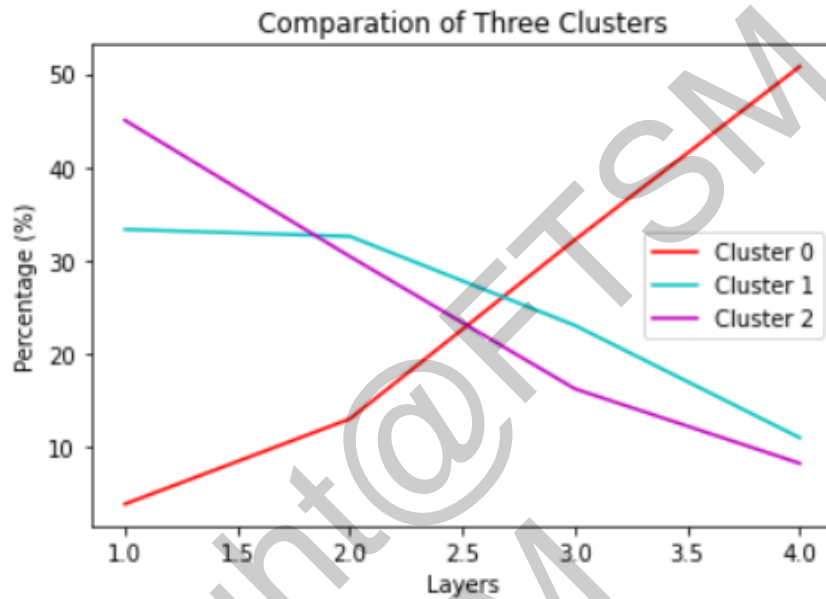


Rajah 5.3 Carta turus secara menegak bagi kelompok 2

Berdasarkan jadual 5.3 dan rajah 5.3, lapisan 1 mempunyai bilangan data yang paling besar manakala lapisan 4 mempunyai bilangan data yang paling kecil. Hal ini dapat dikatakan bahawa terdapat banyak pokok rendah dan kurang pokok tinggi dalam kelompok 1 disebabkan lapisan 1 mempunyai jumlah bilangan yang tinggi dan lapisan 4 mempunyai jumlah bilangan yang rendah. Kawasan kelompok 2 mempunyai persamaan dengan kawasan kelompok 1 di mana mempunyai lebih banyak pokok kecil.

5.4 Perbandingan Antara Tiga Kelompok

Rajah 5.4 telah menunjukkan plot garis perbandingan antara tiga kelompok iaitu kelompok 0, kelompok 1 dan kelompok 2.



Rajah 5.4 Plot Garis Perbandingan Antara Tiga Kelompok

Berdasarkan rajah 5.4, terdapat peratusan yang tinggi titik ketinggian pokok dalam kawasan lapisan 3 dan lapisan 4 di kelompok 0. Hal ini telah membuktikan kelompok 0 mempunyai pokok tinggi yang paling banyak. Di kelompok 1 dan kelompok 2, kedua-dua kelompok mempunyai bentuk plot garis yang hampir sama iaitu terdapat peratusan tinggi titik ketinggian pokok dalam kawasan lapisan 1 dan lapisan 2. Hal ini telah membuktikan kelompok 1 dan kelompok 2 mempunyai pokok rendah yang lebih banyak.

5.5 Pengujian

5.5.1 *Silhouette Coefficient*

Bagi perbandingan keputusan algoritma pengelompokan, algoritma berdasarkan ketumpatan mempunyai kejituan yang paling tinggi dan sesuai dipilih dan digunakan dalam identifikasi pokok secara individu melalui pengujian *Silhouette Coefficient*.

Silhouette coefficient juga dikenali sebagai *silhouette score* merupakan kaedah interpretasi dan pengesahan kekonsistenan antara kelompok data dan biasanya digunakan bagi algoritma pengelompokan. *Silhouette coefficient* dapat mengesahkan kesesuaian algoritma pengelompokan dengan menguji kesesuaian data dikelompok dengan data yang hampir sama dengan diri sendiri. Model pengelompokan yang mempunyai *Silhouette Coefficient* yang tinggi bermaksud model tersebut padat iaitu data dalam satu kelompok sama adalah hampir sama dengan diri sendiri dan data-data dikelompokkan secara baik dan seragam di mana data yang berada di kelompok lain tidak mempunyai sifat yang sama dengan kelompok lain. Nilai terbaik *Silhouette Coefficient* adalah 1 dan nilai paling buruk adalah -1. Jika nilainya adalah 0 bermaksud percanggahan kelompok-kelompok telah berlaku.

Algoritma anjakan-min, algoritma *k-means* dan algoritma berdasarkan ketumpatan telah dicatat dengan menggunakan ketiga-tiga kaedah penskalaan ciri yang berbeza iaitu penormalan *min-max*, penormalan *z-score* dan penskalaan *quantile transformer* bagi setiap algoritma. *Silhouette coefficient* telah dikira dan dicatatkan dalam jadual yang berikut. Jadual 5.4 telah menunjukkan *silhouette coefficient*.

Jadual 5.4 Silhouette Coefficient

	Algoritma Anjakan-Min	Algoritma <i>k-means</i>	Algoritma Berdasarkan Ketumpatan
Penormalan Min-Max	0.4225	0.5722	0.5774
Penormalan Z-Score	0.3675	0.3934	-
Penskalaan Quantile Transformer	0.3118	0.3147	0.6260

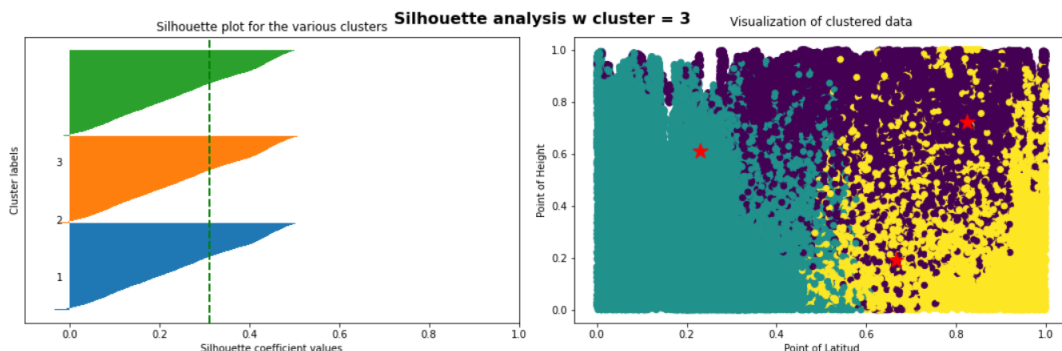
Berdasarkan Jadual 5.4, antara algoritma yang mempunyai *silhouette coefficient* yang paling tinggi adalah algoritma berdasarkan ketumpatan dengan menggunakan

penskalaan *quantile transformer* dengan nilai 0.6260. Dengan ini, algoritma berdasarkan ketumpatan merupakan algoritma yang paling padat dan seterusnya disahkan paling sesuai dalam identifikasi pokok secara individu.

5.5.2 *Silhouette Plot*

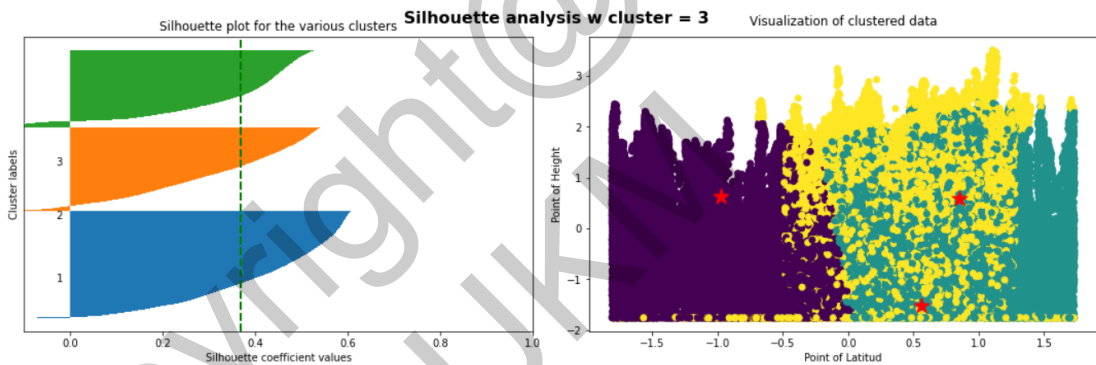
Silhouette plot boleh digunakan untuk menunjukkan jarak pengukuran antara setiap titik dalam satu kelompok dengan titik dalam kelompok yang lain seterusnya menjalankan analisis *silhouette* untuk pemilihan kaedah penskalaan ciri dan nombor kelompok yang sesuai. Aspek-aspek yang penting dengan menggunakan *silhouette* plot termasuk purata markah *silhouette*, turun naik saiz kelompok dan ketebalan *silhouette* plot bagi setiap kelompok.

Rajah 5.5 telah menunjukkan *Silhouette* plot bagi algoritma anjakan-min dengan menggunakan penskalaan *quantile transformer*. *Silhouette* coefficient yang dapat dilihat melalui rajah 5.5 adalah bawah purata iaitu bawah 0.4. Namun, ketebalan *silhouette* plot yang mewakili titik-titik penentu bagi ketiga-tiga kelompok adalah seragam. Hal ini dapat mengesahkan number kelompok bagi penskalaan *quantile transformer* adalah optimum.



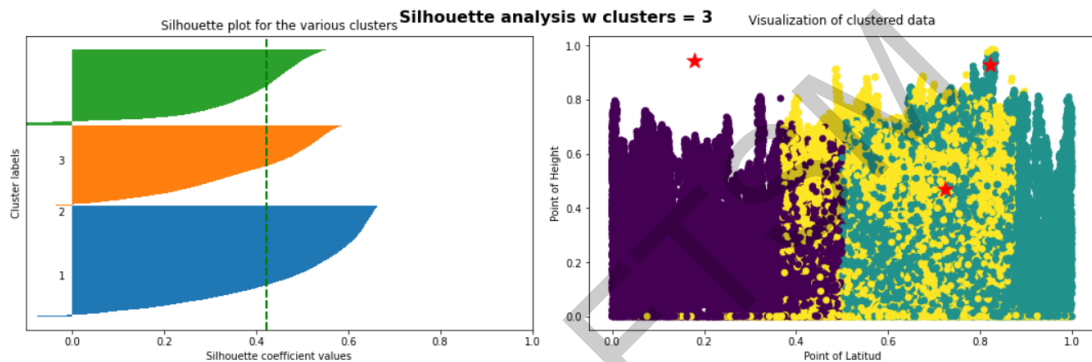
Rajah 5.5 Silhouette Plot (Penskalaan Quantile Transformer)

Rajah 5.6 telah menunjukkan *Silhouette* plot bagi algoritma anjakan-min dengan menggunakan penormalan *z-score*. *Silhouette* coefficient yang dapat dilihat melalui rajah 5.6 adalah bawah purata iaitu bawah 0.4 tetapi lebih tinggi daripada penskalaan *quantile transformer*. Namun, ketebalan *silhouette* plot yang mewakili titik-titik penentu adalah tidak seragam kerana kelompok 1 lebih tebal daripada kelompok 2 dan kelompok 3. Hal ini dapat mengesahkan number kelompok bagi penormalan *z-score* adalah kurang optimum.



Rajah 5.6 Silhouette Plot (Penormalan Z-Score)

Rajah 5.7 telah menunjukkan *Silhouette* plot bagi algoritma anjakan-min dengan menggunakan penormalan *min-max*. *Silhouette* coefficient yang dapat dilihat melalui rajah 5.7 adalah bawah purata iaitu bawah 0.5 tetapi lebih tinggi daripada penskalaan *quantile transformer* dan penormalan *z-score*. Namun, ketebalan *silhouette* plot yang mewakili titik-titik penentu adalah tidak seragam kerana kelompok 1 lebih tebal daripada kelompok 2 dan kelompok 3. Hal ini dapat mengesahkan number kelompok bagi penormalan *min-max* adalah kurang optimum.



Rajah 5.7 Silhouette Plot (Penormalan Min-Max)

Berdasarkan *silhouette plot*, penskalaan *quantile transformer* akan dipilih untuk dijadikan kaedah penskalaan ciri bagi algoritma anjakan-min dalam projek ini disebabkan nombor kelompok yang telah disahkan seragam seterusnya titik-titik penentu adalah seragam dalam setiap kelompok. Selain itu, penggunaan penskalaan *quantile transformer* dapat menunjukkan perbezaan data lapisan dengan paling ketara dan paling senang serta tepat untuk menganalisis peratusan bilangan pokok tinggi dan rendah di kawasan kelompok-kelompok.

5.5.3 T-Test

T-test merupakan satu teknik pengujian untuk membanding purata (min) dan menunjukkan perbezaan data-data antara satu sama lain, signifikan perbezaan tersebut serta kebarangkalian perbezaan tersebut akan berlaku. Nilai T merupakan ratio perbezaan antara dua kumpulan dan perbezaan dalam kumpulan masing-masing. Dengan ini, nilai t makin besar, perbezaan antara kumpulan makin besar, manakala, makin kecil nilai t, makin tinggi persamaan antara kumpulan. Nilai t makin besar bermaksud kumpulan-kumpulan tiada persamaan manakala nilai t makin kecil bermaksud kumpulan-kumpulan adalah hampir sama. Selain itu, nilai p ialah kebarangkalian keputusan sampel data berlaku persamaan secara rawak. Nilai p adalah

dari 0% hingga 100% tetapi nilai p ditulis dalam bentuk titik perpuluhan. Nilai p yang kecil adalah bagus disebabkan nilai p yang kecil bermaksud data tidak akan berlaku persamaan secara rawak.

Dalam kajian projek ini, t -test bagi sampel kebergantungan akan digunakan disebabkan ketiga-tiga algoritma iaitu algoritma anjakan-min, algoritma k -means dan algoritma berdasarkan ketumpatan menggunakan kaedah penskalaan ciri dan data yang sama untuk mendapatkan keputusan ramalan. Jadual 1.7 telah menunjukkan keputusan t -test.

Jadual 1.7 Keputusan T-Test

	T-Statistik	p-value	Peratusan
Anjakan-min & K-means	-0.7202	0.5112	51.12%
K-means & Berdasarkan Ketumpatan	0.1192	0.9108	91.08%
Anjakan-min & Berdasarkan Ketumpatan	-0.1664	0.8759	87.59%

Dalam t -test ini, aras signifikan yang digunakan iaitu α adalah 0.05 (95% sela keyakinan), nilai kritikal adalah 2.1318 dan darjah kebebasan (df) adalah 4. Terdapat tiga kes akan dijalankan termasuk antara algoritma anjakan-min dan k -means, algoritma k -means dan berdasarkan ketumpatan serta anjakan-min dan berdasarkan ketumpatan. Rajah 1.9 menunjukkan t -test bagi algoritma anjakan-min dan algoritma k -means.

□
 $t=-0.7202$, $df=4$, $cv=2.1318$, $p=0.5112$
 Accept null hypothesis that the mean shift and the k-means are not significant.
 Accept null hypothesis that the mean shift and the k-means are not significant.

Rajah 1.9 T-Test (Algoritma Anjakan-Min dan K-Means)

Bagi kes pertama, hipotesis sifar ialah algoritma anjakan-min dan algoritma *k-means* adalah signifikan manakala hipotesis alternatif ialah algoritma anjakan-min dan algoritma *k-means* adalah tidak signifikan. T-statistik adalah -0.7202 lebih kecil daripada nilai kritikal. Gagal dalam penolakan hipotesis sifar. Dengan ini, algoritma anjakan-min dan algoritma *k-means* adalah signifikan. Nilai p adalah 0.5112 lebih tinggi daripada aras signifikan. Pada 95% sela keyakinan, gagal dalam penolakan hipotesis sifar. Dengan ini, algoritma anjakan-min dan algoritma *k-means* adalah signifikan telah disahkan. Rajah 1.10 menunjukkan t-test bagi algoritma *k-means* dan algoritma berdasarkan ketumpatan.

□
 $t=0.1192$, $df=4$, $cv=2.1318$, $p=0.9108$
 Accept null hypothesis that the k-means and the density-based are not significant.
 Accept null hypothesis that the k-means and the density-based are not significant.

Rajah 1.10 T-Test (Algoritma K-Means dan Berdasarkan Ketumpatan)

Bagi kes kedua, hipotesis sifar ialah algoritma *k-means* dan algoritma berdasarkan ketumpatan adalah tidak signifikan manakala hipotesis alternatif ialah algoritma *k-means* dan algoritma berdasarkan ketumpatan adalah signifikan. T-statistik adalah 0.1192 lebih kecil daripada nilai kritikal. Gagal dalam penolakan hipotesis sifar, Dengan ini, algoritma *k-means* dan algoritma berdasarkan ketumpatan adalah tidak signifikan. Nilai p adalah 0.9108 lebih tinggi daripada aras signifikan. Pada 95% sela keyakinan, gagal dalam penolakan hipotesis sifar. Dengan ini, algoritma *k-means* dan algoritma berdasarkan ketumpatan adalah tidak signifikan telah disahkan. Rajah 1.11 menunjukkan t-test bagi algoritma anjakan-min dan algoritma berdasarkan ketumpatan.

□ $t=-0.1664$, $df=4$, $cv=2.1318$, $p=0.8759$
 Accept null hypothesis that the mean shift and the density-based are not significant.
 Accept null hypothesis that the mean shift and the density-based are not significant.

Rajah 1.11 T-Test (Algoritma Anjakan-Min dan Berdasarkan Ketumpatan)

Bagi kes ketiga, hipotesis sifar ialah algoritma anjakan-min dan algoritma berdasarkan ketumpatan adalah tidak signifikan manakala hipotesis alternatif ialah algoritma anjakan-min dan algoritma *k-means* adalah signifikan. T-statistik adalah -0.1664 lebih kecil daripada nilai kritikal. Gagal dalam penolakan hipotesis sifar. Dengan ini, algoritma anjakan-min dan algoritma berdasarkan ketumpatan adalah tidak signifikan. Nilai p adalah 0.8759 lebih tinggi daripada aras signifikan. Pada 95% sela keyakinan, gagal dalam penolakan hipotesis sifar. Dengan ini, algoritma anjakan-min dan algoritma berdasarkan ketumpatan adalah tidak signifikan telah disahkan.

Melalui *t-test*, ketiga-tiga algoritma adalah tidak signifikan. Dengan ini, ketiga-tiga algoritma adalah tiada perbezaan, tiada hubungan statistik dan tidak mempengaruhi antara satu sama lain. Oleh itu, ketiga-tiga algoritma boleh digunakan tanpa sebarang gangguan.

6 KESIMPULAN

Projek ini memperkenalkan algoritma anjakan-min bagi pengelompokan data LiDAR yang memperoleh dari sesuatu kawasan di hutan tropika dan membangunkan model pembelajaran mesin algoritma. Data awan titik secara tiga-dimensi (3D) yang dijana oleh LiDAR telah menjalankan pemprosesan awal kepada bentuk x , y , dan z . Seterusnya, data x , y , dan z akan menjalankan proses pengelompokan bagi pokok secara individu. Pengekstrakan ciri pokok dalam kelompok dan seterusnya pengelasan lapisan pokok berdasarkan data z iaitu titik ketinggian akan dijalankan. Keputusan terakhir yang dihasilkan akan diambil secara nilai kebarangkalian dan ketepatan hasil ramalan.

Kajian ini merupakan kajian yang dijalankan dengan pakar profesional dari Pusat Sains Angkasa, Institut Perubahan Iklim dengan menggunakan data LiDAR secara tiga-dimensi (3D). Lapisan pokok yang berdasarkan titik ketinggian dapat dicari dan seterusnya lapisan pokok tersebut dapat dianalisis. Melalui analisis yang dibuat terhadap data-data pokok sesuatu kawasan di hutan tropika, algoritma yang paling sesuai dan mempunyai kejituan yang tinggi untuk kajian identifikasi pokok dapat dikesan pada masa akan datang. Selain itu, pemilihan penskalaan ciri yang paling sesuai dan signifikan untuk membuat analisis dengan lebih tepat dan ketara juga dapat dijawab melalui kajian ini. Dalam projek ini, titik ketinggian telah berjaya diperolehi. Namun disebabkan formula untuk meramal kadar pelepasan karbon dioksida termasuk pengumpulan ketinggian pokok, berat bersih pokok, diameter pohon pokok, berat karbon pokok, berat karbon dioksida dalam pokok serta berat karbon dioksida tahunan dalam pokok. Oleh itu, kadar pelepasan karbon dioksida tidak dapat diramalkan kerana kekurangan beberapa parameter untuk dimasukkan ke formula ramalan kadar pelepasan karbon dioksida. Cadangan penambahbaikan kajian termasuk membangunkan kaedah untuk mengelas pokok individu berdasarkan ciri-ciri pokok individu, meramalkan sifat-sifat pokok secara individu yang lain seperti biomass, hasil pengeluaran dan struktur hutan. Selain itu, kadar pelepasan karbon dioksida, isi padu, hasil pengeluaran dan lain-lain dianggarkan daripada ciri pokok sesuatu kawasan dan meramalkan keadaan hutan terutamanya hutan tropika yang berkepadatan tinggi.

7 RUJUKAN

- Ajitesh Kumar. 2020. Kmeans Silhouette Score Explained With Python Example. <https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam>
- Amulya Aankul.2017. T-test using Python and Numpy. <https://towardsdatascience.com/inferential-statistics-series-t-test-using-numpy-2718f8f9bf2f>

- Analytics Vidhya. 2019. Statistics for Data Science: Introduction to t-test and its Different Types (with Implementation in R). <https://www.analyticsvidhya.com/blog/2019/05/statistics-t-test-introduction-r-implementation/>
- Andrea Jane Leys. 2019. How is carbon stored in trees and wood products. <https://forestlearning.edu.au/images/resources/How%20carbon%20is%20stored%20in%20trees%20and%20wood%20products.pdf>
- Ashutosh Bhardwaj. 2020. Silhouette Coefficient. <https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c>
- AskPython. 2021, RMSE-Root Mean Square Error in Python. <https://www.askpython.com/python/examples/rmse-root-mean-square-error>
- Arwan Badru Naim. 2009. Segmentasi Citra Dengan Menggunakan Metode Mean Shift Spectral Clustering. <http://digilib.its.ac.id/public/ITS-Undergraduate-7656-5105100169-SEGMENTASI%20CITRA%20DENGAN%20MENGUNAKAN%20METODE%20MEAN%20SHIFT%20SPECTRAL%20CLUSTERING.pdf>
- Bas Fransen. 2019. How to calculate CO2 sequestration. <https://www.ecomatcher.com/how-to-calculate-co2-sequestration/>
- Chew Jian Chieh. 2021. Making Sense of the Two-Sample T-Test. <https://www.isixsigma.com/tools-templates/hypothesis-testing/making-sense-two-sample-t-test/>
- Chris. 2020. How to perform Mean Shift clustering with Python in Scikit. <https://www.machinecurve.com/index.php/2020/04/23/how-to-perform-mean-shift-clustering-with-python-in-scikit/>
- Cory Maklin. 2019. Mean Shift Clustering Algorithm Example In Python. <https://towardsdatascience.com/machine-learning-algorithms-part-13-mean-shift-clustering-example-in-python-4d6452720b00>
- Damir Demirovic. 2019. An Implementation of the Mean Shift Algorithm. https://www.ipol.im/pub/art/2019/255/article_lr.pdf
- EDUCBA. 2020. Mean Shift Algorithm. <https://www.educba.com/mean-shift-algorithm/>

- Eva Lindberg & Johan Holmgren. 2017. Individual Tree Crown Methods for 3D Data from Remote Sensing. <https://link.springer.com/article/10.1007/s40725-017-0051-6>
- Edy Irwansyah. 2017. Clustering. <https://socs.binus.ac.id/2017/03/09/clustering/>
- GeeksforGeeks. 2019. ML|Mean-Shift Clustering. <https://www.geeksforgeeks.org/ml-mean-shift-clustering/>
- George Seif. 2018. The 5 Clustering Algorithms Data Scientists Need to Know. <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>
- Jason Brownlee. 2019. How to Code the Student's t-Test from Scratch in Python. <https://machinelearningmastery.com/how-to-code-the-students-t-test-from-scratch-in-python/>
- Jason Brownlee. 2020. How to Use Quantile Transforms for Machine Learning. <https://machinelearningmastery.com/quantile-transforms-for-machine-learning/>
- Jiawei Han, Jian Peo. 2012. Cluster Analysis. <https://www.sciencedirect.com/topics/computer-science/clustering-quality>
- Jim Frost. 2018. How t-Tests Work: t-Values, t-Distributions, and Probabilities. <https://statisticsbyjim.com/hypothesis-testing/t-tests-t-values-t-distributions-probabilities/>
- Jinglan Zhang. 2019. A Clustering-Based Approach for Improving the Accuracy of UWB Sensor-Based Indoor Positioning System. <https://www.hindawi.com/journals/misy/2019/6372073/>
- Jiping Liu, Jing Shen, Rong Zhao, Shenghua Xu. 2013. Extraction of individual tree crowns from airborne LiDAR data in human settlements. <https://www.sciencedirect.com/science/article/pii/S0895717711006765>
- K.Bogner, F.Pappenberger, H.L. Cloke. 2012. Technical Note: The normal quantile transformation and its application in a flood forecasting system. <https://hess.copernicus.org/articles/16/1085/2012/hess-16-1085-2012.pdf>
- Leon Hartwell Allen, Jr. 1991. Managing Water Resources in the West Under Conditions of Climate Uncertainty: A Proceedings. <https://www.nap.edu/read/1911/chapter/8>

[Melissa Latella, Fabio Sola, Carlo Camporaele. 2021 A Density-Based Algorithm for the Detection of Individual Trees from LiDAR Data. https://www.researchgate.net/publication/348603443_A_Density-Based_Algorithm_for_the_Detection_of_Individual_Trees_from_LiDAR_Data](https://www.researchgate.net/publication/348603443_A_Density-Based_Algorithm_for_the_Detection_of_Individual_Trees_from_LiDAR_Data)

Manimaran. 2019. Clustering Evaluation strategies. <https://towardsdatascience.com/clustering-evaluation-strategies-98a4006fcf>

Michael Brown. 1999. Parzen windows. <https://compbio.soe.ucsc.edu/genex/genexTR2html/node11.html#:~:text=Parzen%20windows%20classification%20is%20a,centered%20on%20the%20observed%20points>

Mohammad Amin Alizadeh Khameneh. 2013. Tree Detection and Species Identification using LiDAR Data. https://www.researchgate.net/publication/285593710_Tree_Detection_and_Species_Identification_using_LiDAR_Data

[Mohammed Senoussaoui, Patrick Kenny, Themis Stafylakis, Pierre Dumouchel. 2014. A Study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization. https://www.researchgate.net/publication/260661427_A_Study_of_the_Cosine_Distance-Based_Mean_Shift_for_Telephone_Speech_Diarization](https://www.researchgate.net/publication/260661427_A_Study_of_the_Cosine_Distance-Based_Mean_Shift_for_Telephone_Speech_Diarization)

Moto DEI. 2019. 11 Evaluation Metrics Data Scientists Should Be Familiar with – Lessons from A High-rank Kagglers' New Book. <https://towardsdatascience.com/11-evaluation-metrics-data-scientists-should-be-familiar-with-lessons-from-a-high-rank-kagglers-8596f75e58a7>

Mukesh Chaudhary. 2020. Silhouette Analysis in K-means Clustering. <https://medium.com/@cmukesh8688/silhouette-analysis-in-k-means-clustering-cefa9a7ad111>

N.Demir*, D. Poli, E. Baltsavias. Extraction of buildings and trees using images and LiDAR data. https://www.isprs.org/proceedings/XXXVII/congress/4_pdf/55.pdf

Nur Mohommad Farda, Pramaditya Wicaksono. 2015. Aplikasi Algoritma Klasifikasi Mean Shift untuk Pemetaan Habitat Bentik Studi Kasus Kepulauan Karimunjawa. https://www.researchgate.net/publication/283444119_Aplikasi_Algoritma_Klasifikasi_Mean_Shift_untuk_Pemetaan_Habitat_Bentik_Studi_Kasus_Kepulauan_Karimunjawa

- Qi Chen. 2007. Airborne LiDAR Data Processing and Information Extraction. <https://nature.berkeley.edu/biometlab/pdf/Chen%202007%20PERS%20highlight.pdf>
- Robert Kwiatkowski. 2021. Mall Customers Segmentation. <https://www.kaggle.com/datark1/customers-clustering-k-means-dbscan-and-ap>
- Sedef Cakir, Mikdat Kadioglu, Nihat Cubukcu. 2012. Multischeme ensemble forecasting of surface temperature using neural network over Turkey. https://www.researchgate.net/publication/257449459_Multischeme_ensemble_forecasting_of_surface_temperature_using_neural_network_over_Turkey
- Tan, P.N., Steinbach, M., Kumar, V. 2006. Introduction to Data Mining. Boston:Pearson Education. <https://www-users.cs.umn.edu/~kumar001/dmbook/sol.pdf>
- Tom Grigg. 2019. Concept Learning and Feature Spaces. Retrieved from: <https://towardsdatascience.com/concept-learning-and-feature-spaces-45cee19e49db>
- Tutorialspoint. 2020. ML-Clustering Mean Shift Algorithm. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_clustering_algorithms_mean_shift.htm
- Virginia Postrel. Clustering. <https://wiragotama.github.io/resources/ebook/parts/JWGP-intro-to-ml-chap10-secured.pdf>
- Wikipedia. 2019. Pembelajaran dalam. https://ms.wikipedia.org/wiki/Pembelajaran_dalam
- Xin Shen, Lin Cao. 2019. Tree-Species Classification in Subtropical Forests Using Airborne Hyperspectral and LiDAR Data. <https://www.mdpi.com/2072-4292/9/11/1180/htm>
- Ye Tian, Yasunari Yokota. 2019. Estimating the Major Cluster by Mean-Shift with Updating Kernel. https://www.researchgate.net/publication/335376286_Estimating_the_Major_Cluster_by_Mean-Shift_with_Updating_Kernel