

PERAMALAN HUJAN MENGGUNAKAN KAEDAH PERLOMBONGAN DATA

NORLIYANA ROSMI
SUHAILA ZAINUDIN

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Kajian ini dijalankan bertujuan untuk membangunkan sebuah model ramalan hujan menggunakan kaedah perlombongan data. Kajian ini menggunakan teknik pengelasan iaitu Pohon Keputusan (DT), Rangkaian Neural Buatan (ANN), Mesin Vektor Sokongan (SVM) dan Pendekatan Bayesian (NB). Pembangunan model ini memfokuskan kepada data yang mempengaruhi hujan di Selangor yang terdiri daripada 5173 jumlah data siri masa bagi tempoh 15 tahun (2000-2014). Sampel diperoleh dari Jabatan Meteorologi Malaysia dan Jabatan Pengaliran dan Saliran Malaysia. Lima atribut yang digunakan dalam kajian ini ialah taburan hujan, paras air, kelembapan relatif, suhu, arus sungai dan taburan hujan. Set data siri masa ini telah dibahagikan kepada dua jenis data iaitu data 10-Pengesahan Bersilang dan Pembahagian Peratusan serta set data latihan dan data ujian. Kajian ini terbahagi kepada dua jenis set eksperimen iaitu set Eksperimen I (ujian pilihan 10- pengesahan bersilang dan pembahagian peratusan 66%) dan set Eksperimen II (set data latihan dan data ujian akan digunakan). Hasil dapatan kajian akan dinilai mengikut peratusan model yang paling tinggi di antara keempat – empat teknik berdasarkan Eksperimen I dan Eksperimen II. Ini bermaksud kajian ini memfokuskan kepada ketinggian peratusan nilai *Correctly Classified* (CC) di samping turut mengambil kira nilai *Root Mean Square Error* (RMSE) dan *Receiver Operating Characteristics* (ROC) bagi setiap set ujian yang dijalankan. Selepas proses perlombongan data dijalankan serta berdasarkan model yang dihasilkan, peratus ketepatan ramalan yang diperoleh akan dianalisa. Kesimpulannya, bagi Eksperimen I, peratus ketepatan paling tinggi bagi ujian pilihan pengesahan bersilang adalah teknik Pohon Keputusan (J48) iaitu sebanyak 65.13% dan peratus ketepatan ujian pilihan *percentage split 66%* adalah teknik pengelasan Rangkaian Neural Buatan (ANN) iaitu sebanyak 65.32%. Manakala peratus ketepatan paling tinggi bagi ujian pilihan *percentage split* yang dilakukan pada Eksperimen II adalah Pohon Keputusan (J48) iaitu sebanyak 64.95% pada pembahagian peratusan 70% data latihan-30% data.

1 PENGENALAN

Ramalan cuaca telah menjadi masalah yang paling mencabar secara saintifik dan teknologinya di dunia dalam abad yang lalu. Hal ini adalah disebabkan oleh dua faktor iaitu yang pertama, ia digunakan untuk pelbagai aktiviti manusia dan kedua ialah, peluang yang dicipta oleh pelbagai kemajuan teknologi yang secara langsung berkait dengan bidang kajian ini seperti evolusi dalam pengiraan dan peningkatan dalam sistem pengukuran (Dutta & Tahbilder 2014). Maka, peramalan hujan adalah penting untuk merancang aktiviti harian seperti pertanian yang merupakan salah satu ekonomi utama Malaysia. Hujan adalah penting untuk pengurusan hasil tanaman, pengurusan sumber air dan perancangan untuk aktiviti lainnya. Keadaan tempoh yang kering atau hujan secara berpanjangan boleh mengganggu sistem pertumbuhan tanaman serta mengakibatkan kepada pengurangan hasil tanaman

(M.Kannan et al. 2010). Selain itu, peramalan hujan ini juga amat penting sebagai amaran agar orang awam mengambil langkah berjaga-jaga untuk merangka aktiviti yang melibatkan kawasan seperti laut, sungai dan air terjun pada waktu-waktu kritikal monsun. Proses peramalan hujan ini agak sukar kerana parameter yang diperlukan sangat kompleks maka ketidakpastian adalah sangat besar (Onwubolu 2007). Teknik perlombongan data telah dipilih untuk menyelesaikan masalah ini. Di dalam teknik ini, proses prapemprosesan telah dilakukan kepada data cuaca mentah. Model ini dilatih menggunakan set data latihan dan diuji ketepatannya melalui data ujian (Nikam & Meshram 2013).

2 PENYATAAN MASALAH

Ramalan hujan amat penting di negara ini kerana faktor taburan hujan yang banyak sepanjang tahun. Maklumat ini penting kepada penduduk terutamanya yang terlibat dengan industri pertanian dan pembinaan. Hal ini penting agar dapat merancang pengurusan dengan berkesan. Hujan cenderung untuk mengganggu atau melambatkan pembinaan luar di samping kebanyakan bahan – bahan pembinaan tidak boleh terdedah kepada air (Bordoli 2010). Selain itu, maklumat berkenaan ramalan hujan diperlukan terutamanya pada musim monsun yang kebiasaannya akan hujan lebat yang berpanjangan sehingga banjir besar seperti yang berlaku pada tahun 2014 di negeri-negeri pantai timur. Sistem ramalan hujan yang lebih berkesan serta lebih tepat dapat membantu penduduk membuat persediaan menghadapi musim tengkujuh. Hal ini dapat mengurangkan kerugian harta benda serta mengelakkan kehilangan nyawa.

3 OBJEKTIF KAJIAN

Projek ini bertujuan menguji keberkesanan model ramalan hujan dengan membuat kajian perbandingan perlombongan data terhadap teknik pengelasan Pohon Keputusan (DT), Rangkaian Neural Buatan (ANN), Mesin Vektor Sokongan (SVM) dan Pendekatan Bayesian (NB) dengan menfokuskan kepada kejituan nilai yang terhasil. Membangunkan antara muka sistem ramalan hujan menggunakan model yang dipilih.

4 METOD KAJIAN

Metodologi yang digunakan adalah CRISP - DM (*cross-industry process for data mining*). CRISP – DM adalah metodologi perlombongan data yang komprehensif yang boleh digunakan sama ada orang yang baru terlibat dengan perlombongan data sehinggalah kepada pakar – pakar

dengan satu pelan lengkap untuk menjalankan projek perlombongan data. CRISP – DM bertujuan membuat projek perlombongan data yang lebih besar, jimat kos, lebih dipercayai, lebih terkawal dan lebih cepat (Wirth 2000). Metodologi ini dibahagikan kepada 6 bahagian iaitu pemahaman bisnes, pemahaman data, penyediaan data, pemodelan, pengujian dan penggunaan. Turutan antara fasa – fasa ini adalah tidak begitu ketat. Bergerak di antara satu fasa kepada fasa yang seterusnya atau yang sebelumnya adalah diperlukan bergantung kepada hasil fasa tersebut, fasa atau tugas fasa yang tertentu untuk dilakukan seterusnya.

Bulatan luar di rajah 1.1 menunjukkan proses ini adalah berterusan dan tidak berakhir walaupun setelah solusi diaplikasi. Ia belajar semasa proses tersebut dan daripada penyelesaian tersebut boleh mencetuskan soalan bisnes yang baru dan lebih berfokus (Chapman et al. 1999).

1.7.1 Pemahaman Bisnes

Fasa pertama di dalam metodologi ini memfokus kepada pemahaman objektif serta keperluan projek daripada perspektif bisnes seterusnya menukarkan pengetahuan ini kepada permasalahan perlombongan data. Plan awal dirangka untuk mencapai objektif tersebut.

1.7.2 Pemahaman Data

Fasa pemahaman data bermula dengan pengumpulan data seterusnya meneroka data – data yang bermasalah atau pun mengesan subset menarik untuk membentuk hipotesis daripada maklumat yang tersembunyi.

1.7.3 Penyediaan Data

Ini adalah fasa di mana pemilihan data dilakukan untuk digunakan semasa melombong data. Penyediaan data merangkumi keseluruhan aktiviti untuk membina set data akhir yang akan digunakan semasa proses pemodelan (Chapman et al. 1999). Pemilihan adalah berdasarkan kualiti data, perkaitan antara data dengan matlamat perlombongan data serta kekangan-kekangan yang akan mempengaruhi proses perlombongan data. Fasa ini kebiasaannya dilakukan beberapa kali dan tidak ditetapkan turutan.

1.7.4 Pemodelan

Semasa fasa ini, pelbagai teknik pemodelan dipilih untuk diaplikasikan. Kebiasaannya, terdapat beberapa teknik untuk permasalahan perlombongan data yang sama. Beberapa teknik

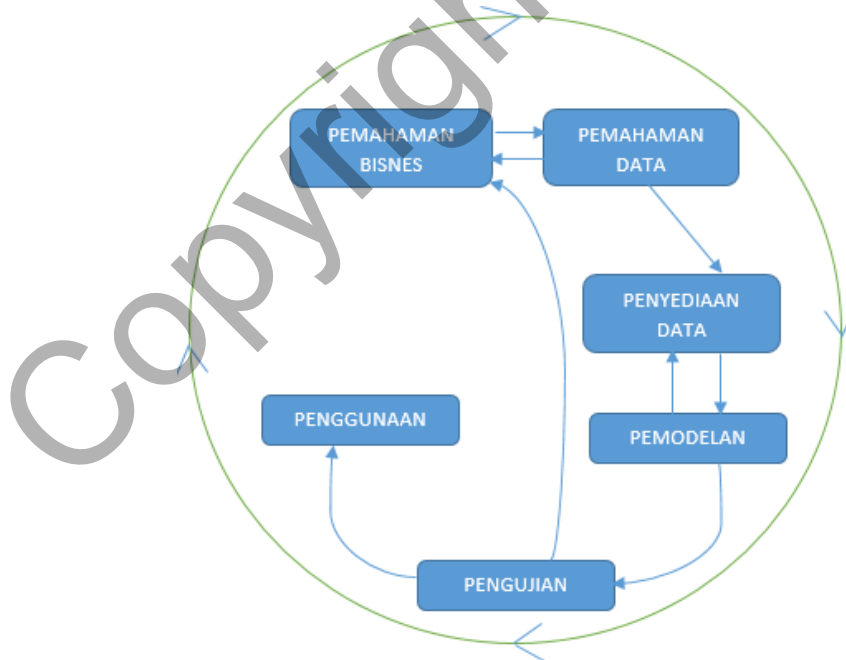
memerlukan data dalam bentuk yang khusus. Hal ini perlu diambil perhatian semasa proses penyediaan data (Chapman et al. 1999).

1.7.5 Penilaian

Semasa di dalam empat fasa sebelum ini, segala data telah dikaji dan trend di sebalik data telah diperolehi serta model telah dibina berdasarkan analisis data yang dilakukan. Maka, untuk fasa ini, hasil daripada kajian sebelum ini akan dinilai. Penilaian dibuat terhadap model yang telah dicipta, proses mencipta model, tahap seberapa praktikal model tersebut serta menilai sama ada telah mencapai objektif atau tidak.

1.7.6 Penggunaan

Ini adalah fasa terakhir dalam CRISP-DM. Bergantung kepada keperluan, fasa ini boleh menjadi ringkas membua report atau menjadi kompleks seperti melakukan proses perlombongan data berulang kali. Kebiasaannya pelanggan yang melakukan fasa ini dan bukannya penganalisis data. Namun, adalah penting untuk pelanggan memahami langkah seterusnya yang perlu dilakukan.



Rajah 1.1 Metodologi Pembangunan Model CRISP-DM

5 HASIL KAJIAN

Bahagian ini membincangkan hasil daripada kajian ramalan hujan.

5.1 Perbincangan Keputusan Eksperimen I

Bahagian ini membincangkan hasil keputusan Eksperimen I yang menggunakan pilihan ujian 10 kali lipatan pengesahan bersilang dan pembahagian peratusan (66%) terhadap empat teknik pengelasan iaitu Pohon Keputusan (DT), Mesin Vektor Sokongan (SVM), Rangkaian Neural Buatan (ANN) dan Pendekatan Bayesien (NV).

Jadual 5.1 : Keputusan ujian 10-pengesahan bersilang dan pembahagian peratusan (66%) bagi setiap teknik pengelasan yang dipilih

Berdasarkan keputusan Jadual 5.1 (model ramalan) yang menggunakan data dari Eksperimen I iaitu ujian 10-pengesahan bersilang, teknik pengelasan Pohon Keputusan (DT) memberikan nilai kejituan yang paling tinggi berbanding teknik lain berdasarkan nilai CC iaitu 65.13%. Nilai RMSE bagi teknik pengelasan DT juga merupakan yang paling rendah (nilai paling rendah adalah nilai yang paling baik iaitu menghampiri nilai 0) di antara keempat-empat teknik iaitu 0.4685. Namun, teknik pengelasan yang mempunyai nilai ROC yang paling tinggi iaitu yang menghampiri 1 (nilai yang menghampiri nilai 1 adalah nilai yang paling tepat) ialah teknik pengelasan NB iaitu 0.688 berbanding nilai ROC bagi teknik pengelasan DT yang hanya kurang 0.003 iaitu 0.685.

Keputusan bagi perbandingan pembahagian peratus (66%) pula, NN mencatatkan nilai bacaan CC yang paling tinggi iaitu 65.32. Selain itu, teknik pengelasan yang memberikan nilai bacaan RMSE yang paling rendah juga merupakan teknik NN iaitu 0.4689 dan nilai bacaan ROC yang paling tinggi bagi pembahagian peratus (66%) adalah 0.701 iaitu teknik NN.

Pilihan Ujian	Atribut (6)	Teknik	CC (%)	RMSE	ROC
10 – Lipatan Pengesahan bersilang	Arus sungai, Suhu, Kelembapan, Paras air, Kelajuan angin, Kelas	DT	65.13	0.4685	0.685
		NB	63.89	0.4751	0.688
		NN	64.26	0.4722	0.685
		SVM	63.57	0.6036	0.626
Pembahagian peratusan (66%)	Arus sungai, Suhu, Kelembapan, Paras air, Kelajuan angin, Kelas	DT	64.13	0.4696	0.689
		NB	63.73	0.4731	0.697
		NN	65.32	0.4689	0.701
		SVM	63.82	0.6015	0.626

5.2 Perbincangan Keputusan Eksperimen II

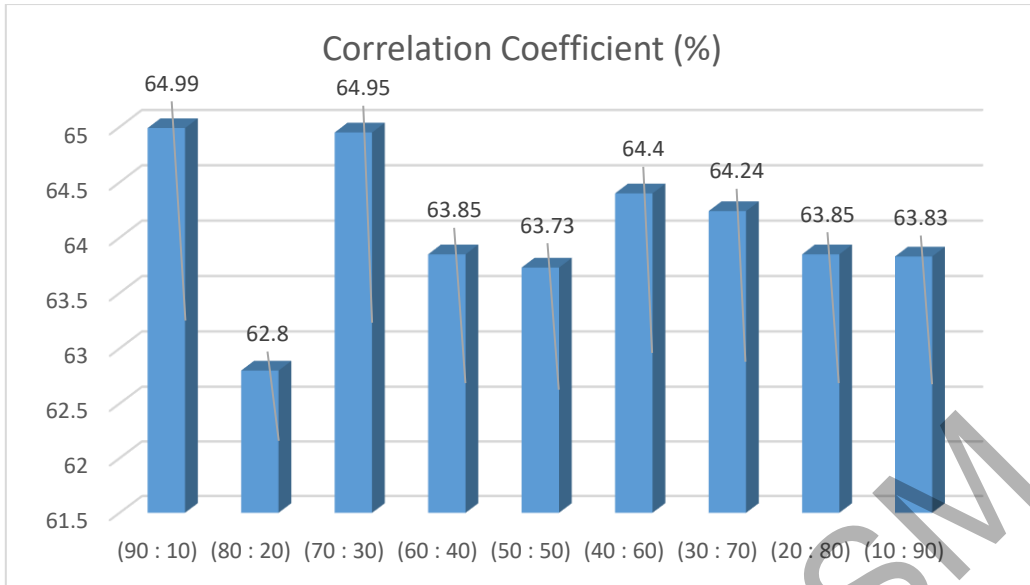
Bahagian ini membincangkan hasil keputusan Eksperimen II yang telah menggunakan set data latihan dan data ujian. Set data latihan dan data ujian tersebut dibahagikan kepada beberapa pembahagian peratusan seperti 90% data latihan - 10% data ujian, 80% data latihan – 20% data ujian, 70% data latihan – 30% data ujian, 60% data latihan – 40% data ujian, 50% data latihan – 50% data ujian, 40% data latihan – 60% data ujian, 30% data latihan – 70% data ujian, 20% data latihan – 80% data ujian dan 10% data latihan – 90% data ujian.

5.2.1 Pohon Keputusan (J48)

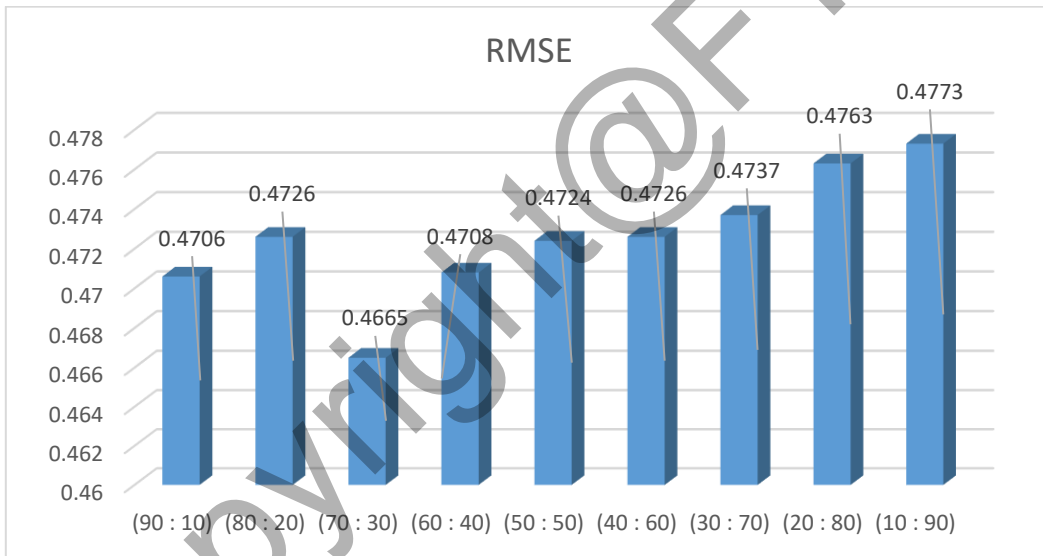
Pohon keputusan (DT) adalah teknik pengelasan yang ringkas dan digunakan secara meluas dalam kerja perlombongan data. Berikut merupakan hasil keputusan ujian pembahagian peratusan data latihan dan data ujian bagi teknik pengelasan Pohon Keputusan (DT) yang diringkaskan ke dalam bentuk Jadual 5.2 dan Rajah 5.1, Rajah 5.2 serta Rajah 5.3. Hasil dari ringkasan keputusan tersebut telah dibincangkan dan dianalisa dengan lebih terperinci pada bahagian ini.

Pilihan Ujian	Atribut (6)	Percentage Split	CC (%)	RMSE	ROC
Percentage Split	Arus sungai, Suhu, Kelembapan, Paras air, Kelajuan angin, Kelas	(90 : 10)	64.99	0.4706	0.696
		(80 : 20)	62.80	0.4726	0.687
		(70 : 30)	64.95	0.4665	0.703
		(60 : 40)	63.85	0.4708	0.689
		(50 : 50)	63.73	0.4724	0.682
		(40 : 60)	64.40	0.4726	0.681
		(30 : 70)	64.24	0.4737	0.680
		(20 : 80)	63.85	0.4763	0.650
(10 : 90)	63.83	0.4773	0.650		

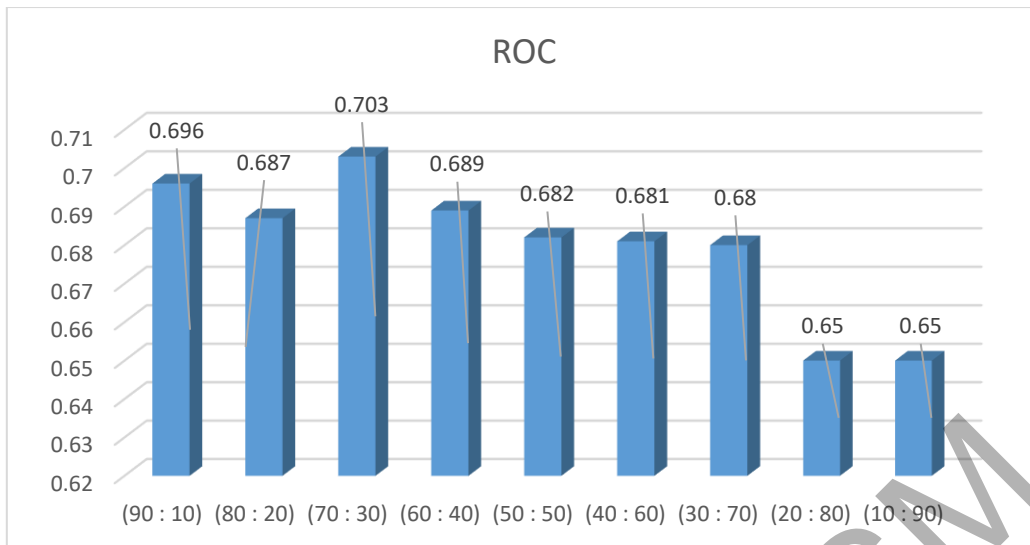
Jadual 5.2 : Keputusan ujian *Percentage Split* (DT)



Rajah 5.1 : Graf bacaan *Correlation Coefficient* (DT)



Rajah 5.2 : Graf bacaan RMSE (DT)



Rajah 5.3: Graf bacaan ROC (DT)

Berdasarkan Jadual 5.2 di atas, pembahagian peratusan data latihan dan data ujian telah menggunakan teknik pengelasan Pohon Keputusan (J48). Hasil dari keputusan ujian tersebut menunjukkan pembahagian peratusan pada 90% data latihan-10% data ujian mencatatkan bacaan CC tertinggi berbanding pembahagian peratusan yang lain iaitu sebanyak 64.99% manakala pembahagian peratusan pada 70% data latihan-30% data ujian memberikan nilai kedua tertinggi iaitu 64.95%. Bagi nilai bacaan RMSE dan ROC, pembahagian peratusan pada 70% data latihan-30% data ujian mencatatkan nilai ketepatan yang paling baik di mana masing-masing mencatat nilai bacaan 0.4665 (menghampiri nilai 0) bagi RMSE dan 0.703 (menghampiri nilai 1) bagi ROC.

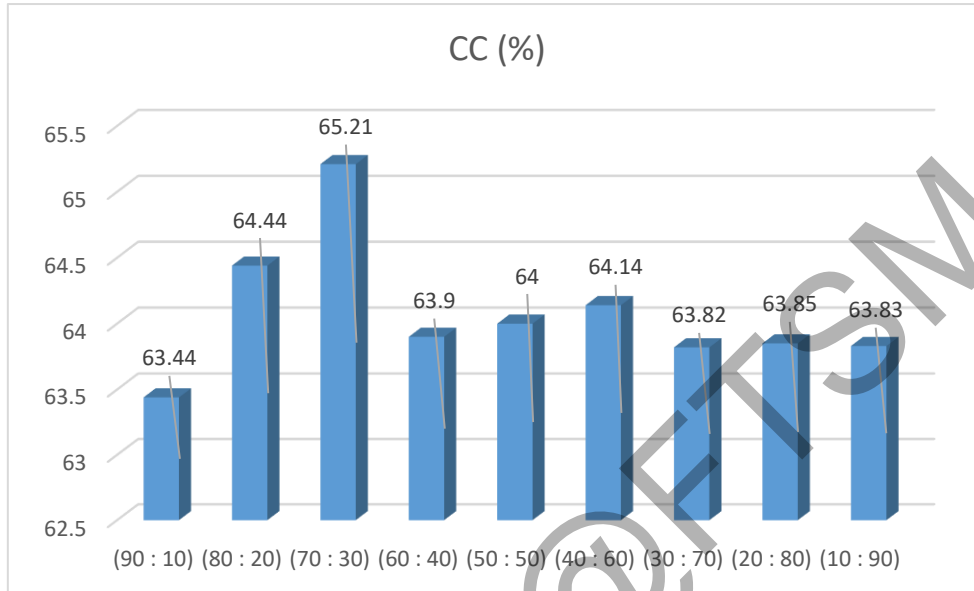
5.2.2 Mesin Vektor Sokongan (*Library Support Vector Machine*)

Berikut merupakan hasil keputusan ujian pembahagian peratusan data latihan dan data ujian bagi teknik pengelasan Mesin Vektor Sokongan (LibSVM) yang diringkaskan ke dalam bentuk Jadual 5.3 dan Rajah 5.4, Rajah 5.5 serta Rajah 5.6. Hasil dari ringkasan keputusan tersebut telah dibincangkan dan dianalisa dengan lebih terperinci pada bahagian ini.

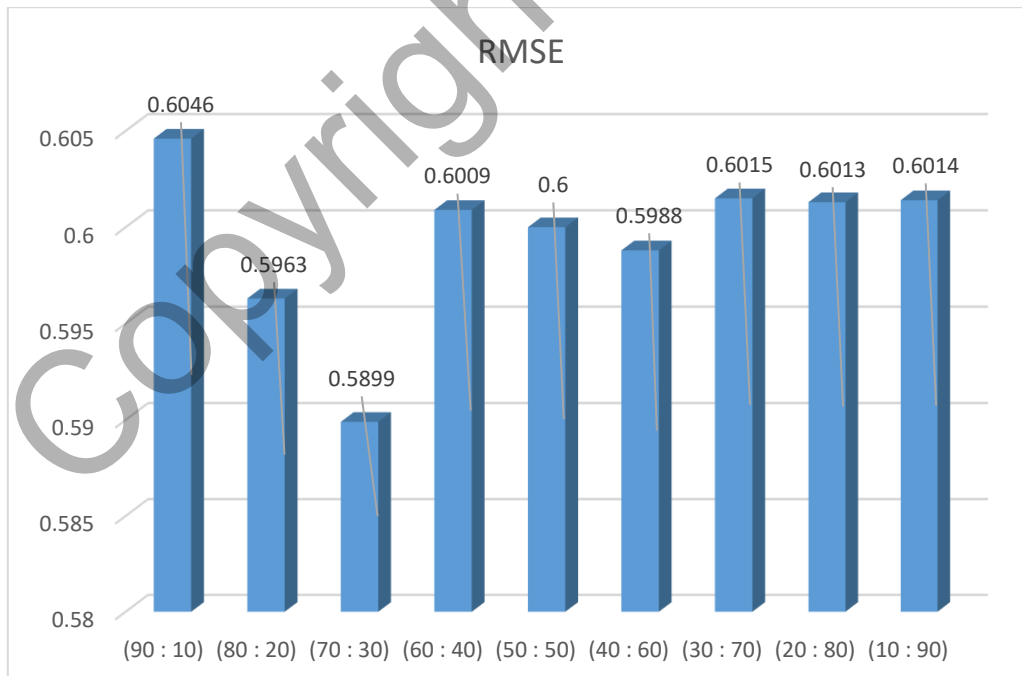
Pilihan Ujian	Atribut (6)	Percentage Split	CC (%)	RMSE	ROC
<i>Percentage Split</i>	Arus sungai, Suhu, Kelembapan, Paras air, Kelajuan angin, Kelas	(90 : 10)	63.44	0.6046	0.632
		(80 : 20)	64.44	0.5963	0.646
		(70 : 30)	65.21	0.5899	0.656
		(60 : 40)	63.90	0.6009	0.644
		(50 : 50)	64.00	0.6000	0.645
		(40 : 60)	64.14	0.5988	0.646

		(30 : 70)	63.82	0.6015	0.643
		(20 : 80)	63.85	0.6013	0.646
		(10 : 90)	63.83	0.6014	0.646

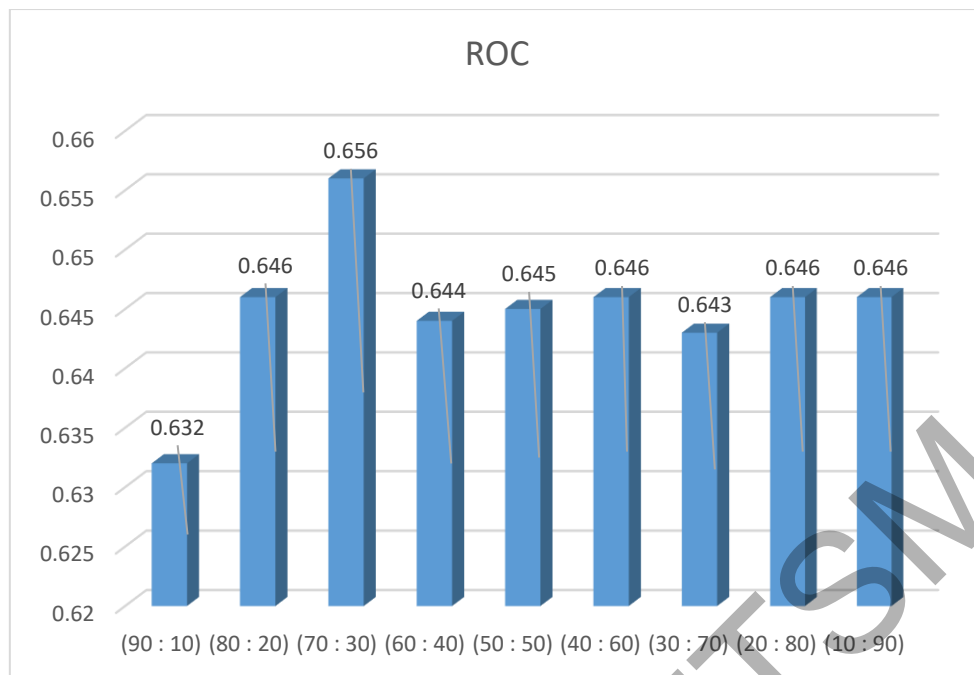
Jadual 5.3 : Keputusan ujian *Percentage Split* (SVM)



Rajah 5.4 : Graf bacaan *Correlation Coefficient* (SVM)



Rajah 5.5 : Graf bacaan RMSE (SVM)



Rajah 5.6 : Graf bacaan ROC (SVM)

Berdasarkan Jadual 5.3 di atas, pembahagian peratusan 70% data latihan-30% data ujian bagi teknik pengelasan SVM mencatatkan nilai bacaan CC yang tertinggi bagi keputusan pembahagian peratusan iaitu dengan nilai bacaan 65.21% berbanding pembahagian peratusan yang lain. Seterusnya bagi nilai bacaan RMSE dan ROC, pembahagian peratusan pada 70% data latihan-30% data ujian mencatatkan nilai ketepatan yang paling baik di mana masing-masing mencatat nilai bacaan 0.5899 (menghampiri nilai 0) bagi RMSE dan 0.656 (menghampiri nilai 1) bagi ROC.

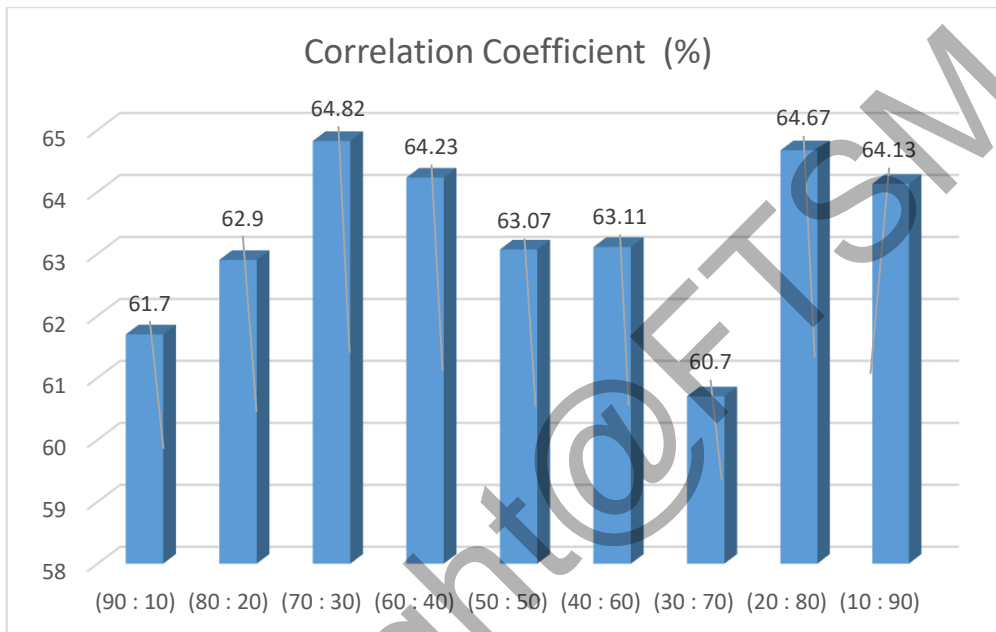
5.2.3 Rangkaian Neural Buatan (ANN)

Berikut merupakan hasil keputusan ujian pembahagian peratusan data latihan dan data ujian bagi teknik pengelasan Rangkaian Neural Buatan (ANN) yang diringkaskan ke dalam bentuk Jadual 5.4 dan Rajah 5.7, Rajah 5.8 serta Rajah 5.9. Bagi teknik pengelasan ini, algoritma *Multi Layer Perceptron* (MLP) telah digunakan. Hasil dari ringkasan keputusan tersebut telah dibincangkan dan dianalisa dengan lebih terperinci pada bahagian ini.

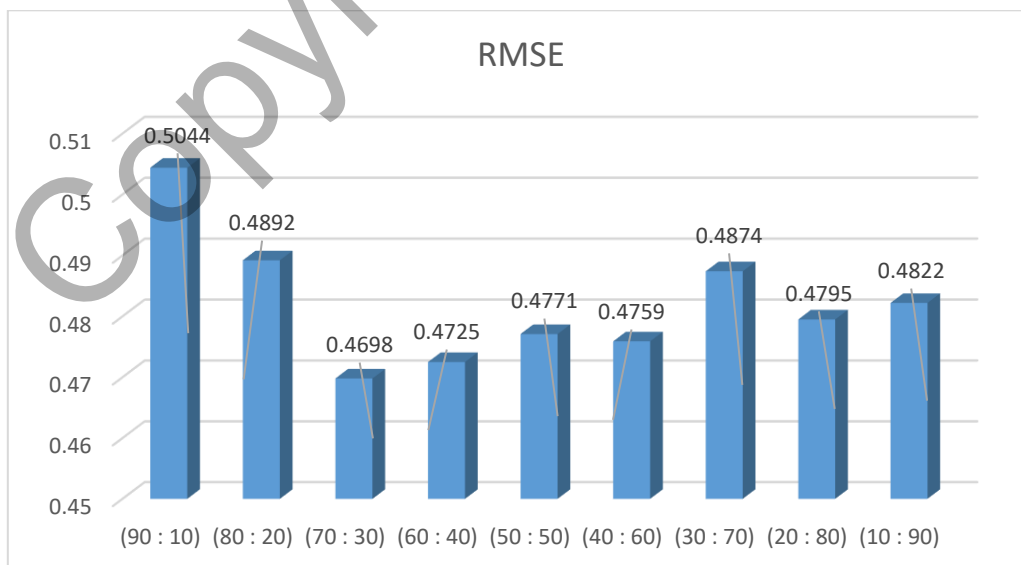
Pilihan Ujian	Atribut (6)	Percentage Split	CC (%)	RMSE	ROC
<i>Percentage Split</i>	Arus sungai, Suhu, Kelembapan,	(90 : 10)	61.70	0.5044	0.696
		(80 : 20)	62.90	0.4892	0.695
		(70 : 30)	64.82	0.4698	0.705

Paras air, Kelajuan angin, Kelas	(60 : 40)	64.23	0.4725	0.683
	(50 : 50)	63.07	0.4771	0.684
	(40 : 60)	63.11	0.4759	0.681
	(30 : 70)	60.70	0.4874	0.688
	(20 : 80)	64.67	0.4795	0.681
	(10 : 90)	64.13	0.4822	0.680

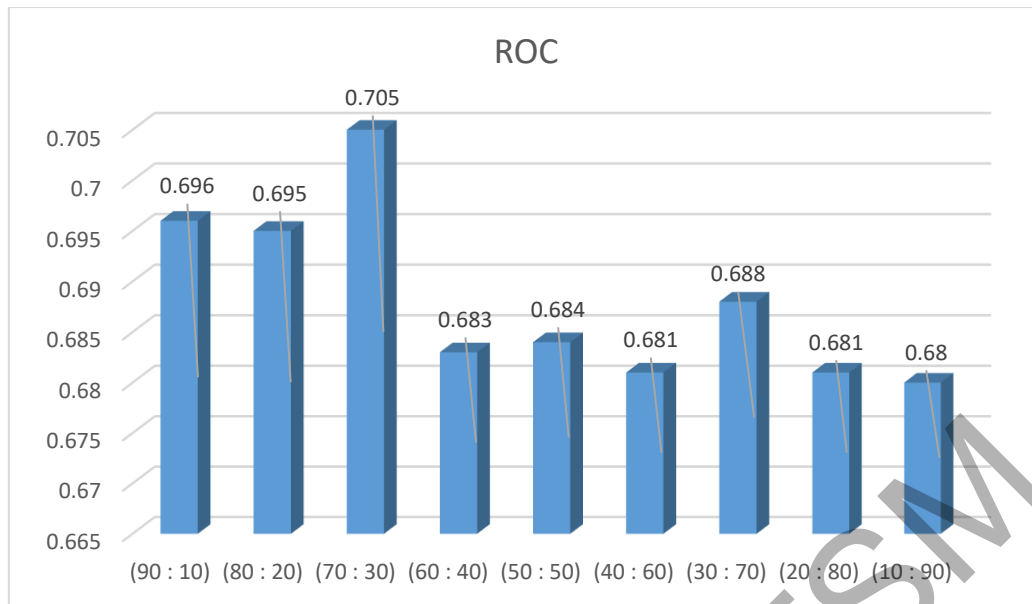
Jadual 5.4 : Keputusan ujian *Percentage Split* (ANN)



Rajah 5.7 : Graf bacaan *Correlation Coefficient* (ANN)



Rajah 5.8 : Graf bacaan RMSE (ANN)



Rajah 5.9 : Graf bacaan ROC (ANN)

Berdasarkan Jadual 5.4 di atas, pembahagian peratusan data latihan dan data ujian telah menggunakan teknik pengelasan Rangkaian Neural Buatan (ANN). Hasil dari keputusan ujian tersebut menunjukkan pembahagian peratusan pada 70% data latihan-30% data ujian mencatatkan bacaan CC tertinggi berbanding pembahagian peratusan yang lain iaitu sebanyak 64.82%. Bagi nilai bacaan RMSE dan ROC, pembahagian peratusan pada 70% data latihan-30% data ujian mencatatkan nilai ketepatan yang paling baik di mana masing-masing mencatat nilai bacaan 0.4698 (menghampiri nilai 0) bagi RMSE dan 0.705 (menghampiri nilai 1) bagi ROC berbanding teknik pengelasan yang lain.

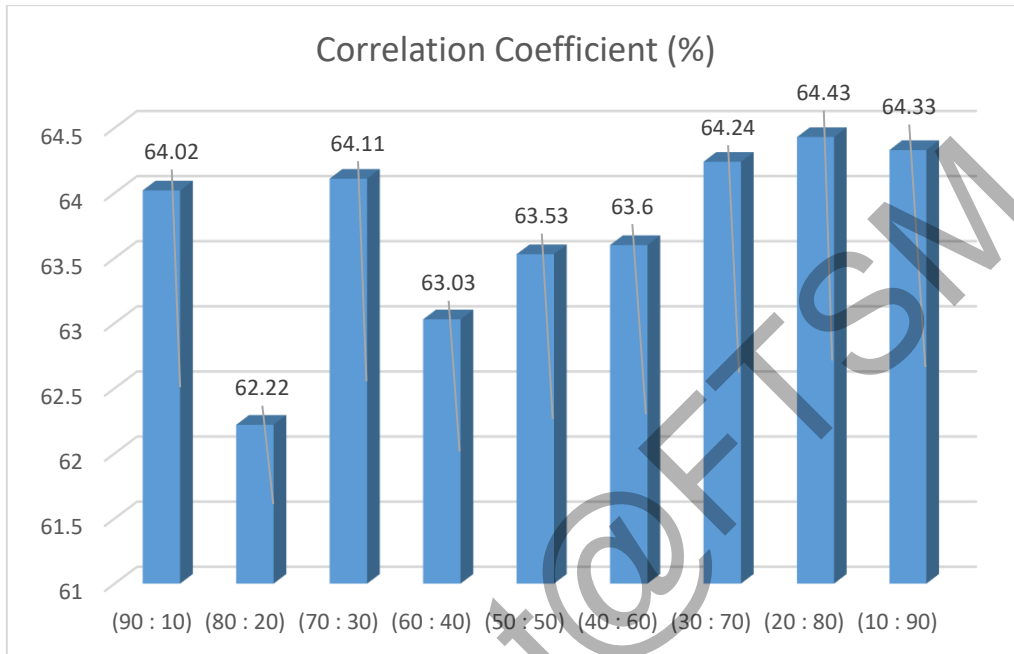
5.2.4 Pendekatan Bayesian (NB)

Berikut merupakan hasil keputusan ujian pembahagian peratusan data latihan dan data ujian bagi teknik pengelasan Pendekatan Bayesian (NB) yang diringkaskan ke dalam bentuk Jadual 5.5 dan Rajah 5.10, Rajah 5.11 serta Rajah 5.12. Hasil dari ringkasan keputusan tersebut telah dibincangkan dan dianalisa dengan lebih terperinci pada bahagian ini.

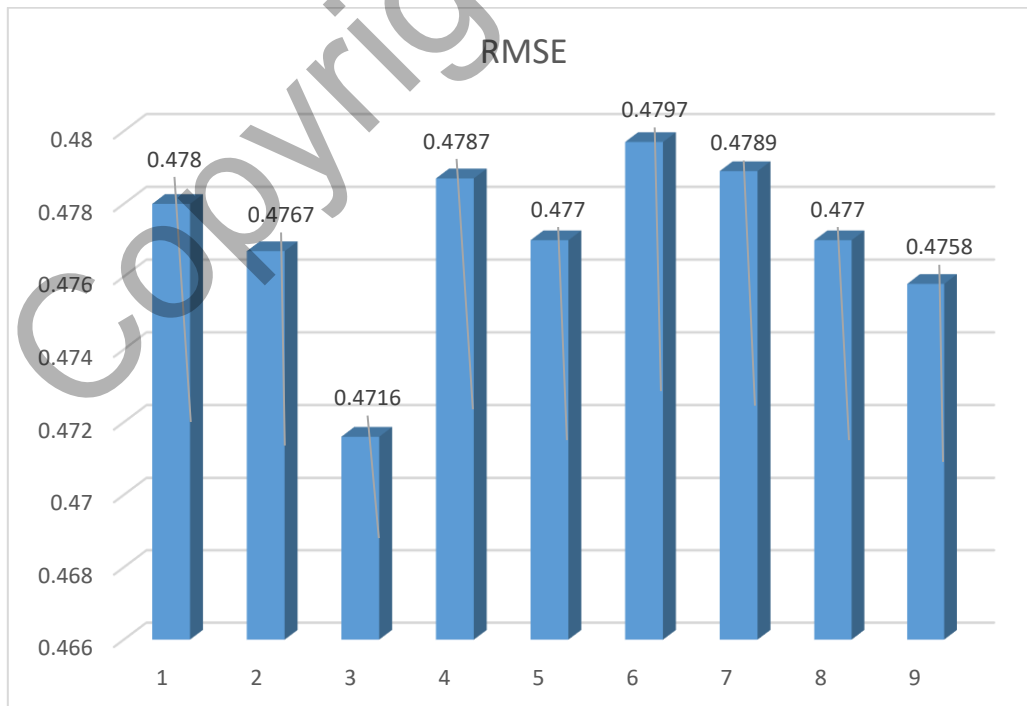
Pilihan Ujian	Atribut (6)	Percentage Split	CC (%)	RMSE	ROC
Percentage Split	Arus sungai, Suhu, Kelembapan, Paras air, Kelajuan angin, Kelas	(90 : 10)	64.02	0.4780	0.686
		(80 : 20)	62.22	0.4767	0.692
		(70 : 30)	64.11	0.4716	0.702
		(60 : 40)	63.03	0.4787	0.685
		(50 : 50)	63.53	0.4770	0.690
		(40 : 60)	63.60	0.4797	0.686

		(30 : 70)	64.24	0.4789	0.687
		(20 : 80)	64.43	0.4770	0.689
		(10 : 90)	64.33	0.4758	0.692

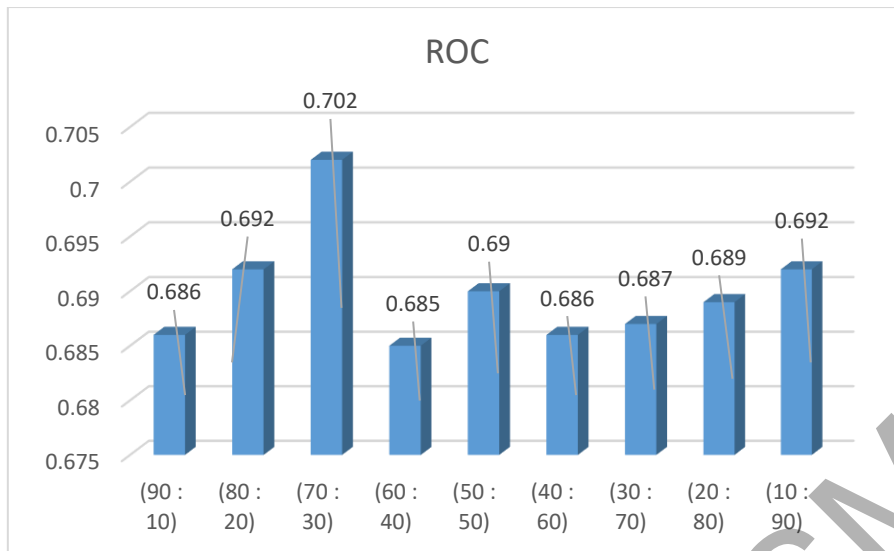
Jadual 5.5 : Keputusan ujian *Percentage Split* (NB)



Rajah 5.10 : Graf bacaan *Correlation Coefficient* (NB)



Rajah 5.11 : Graf bacaan *RMSE* (NB)



Rajah 5.12 : Graf bacaan ROC (NB)

Berdasarkan Jadual 5.5 di atas, pembahagian peratusan data latihan dan data ujian telah menggunakan teknik pengelasan Pendekatan Bayesian (NB). Hasil dari keputusan ujian tersebut menunjukkan pembahagian peratusan pada 20% data latihan-80% data ujian mencatatkan bacaan CC tertinggi berbanding pembahagian peratusan yang lain iaitu sebanyak 64.43% manakala pembahagian peratusan pada 70% data latihan-30% data ujian memberikan nilai kedua tertinggi iaitu 64.95%. Namun bagi nilai bacaan RMSE dan ROC, pembahagian peratusan pada 70% data latihan-30% data ujian mencatatkan nilai ketepatan yang paling baik di mana masing-masing mencatat nilai bacaan 0.4716 (menghampiri nilai 0) bagi RMSE dan 0.702 (menghampiri nilai 1) bagi ROC berbanding yang lain.

5.3 Antara Muka Sistem Ramalan Hujan



Rajah 5.12 : Antara Muka Sistem Ramalan Hujan

6 KESIMPULAN

Daripada keputusan analisis yang telah diperoleh, dapat disimpulkan bahawa peratus ketepatan paling tinggi bagi ujian Eksperimen I iaitu ujian pilihan 10 kali lipatan pengesahan bersilang adalah teknik Pohon Keputusan (DT) dan peratus ketepatan ujian pilihan pembahagian peratusan 66% adalah teknik pengelasan Rangkain Neural Buatan (ANN). Manakala peratus ketepatan paling tinggi bagi ujian pilihan pembahagian peratusan yang dilakukan pada Eksperimen II adalah Pohon Keputusan (DT) iaitu pada pembahagian peratusan 70% data latihan-30% data.

7 RUJUKAN

- Iklim Malaysia - Malaysian Meteorological Department. (n.d.). <http://www.met.gov.my/web/metmalaysia/education/climate/generalclimateofmalaysia> [13 May 2017].
- Bordoli, D. 2010. Weather Claims in United Kingdom Construction Industry. *Construction Law Journal* 26: 19–32. Retrieved from <https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjRw5Gi0LbQAhXBJMAKHJjC0AQFggbMAA&url=http%3A%2F%2Fcs.ananas.chaoxing.com%2Fdownload%2F56741057498ed8a4295d2106&usq=AFQjCNGrUdAGbPRcxwdLsQwYHwYbPRwg-Q&sig2=OndR85Y>
- Chapman, P., Clinton, J., Khabaza, T., Reinartz, T. & Wirth, R. 1999. the Crisp-Dm Process Model. *The CRISP-DM Consortium* 310(C).
- Dutta, P. S. & Tahbilder, H. 2014. Prediction of Rainfall Using Datamining Technique Over Assam. *Indian Journal of Computer Science and Engineering* 5(2): 85–90. Retrieved from <http://www.ijcse.com/docs/INDJCSE14-05-02-081.pdf>
- M.Kannan, S.Prabhakaran & P.Ramachandran. 2010. Rainfall Forecasting Using Data Mining Technique. Engg Journals Publications.
- Nikam, V. B. & Meshram, B. B. 2013. Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach. *2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation*, hlm.132–136. IEEE. doi:10.1109/CIMSim.2013.29
- Onwubolu, G. 2007. Self-organizing data mining for weather Forecasting. ... *Conference Data ...*, 81–88. Retrieved from http://wstst05.softcomputing.net/ecdm07_1.pdf
- Wirth, R. 2000. CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (24959): 29–39. doi:10.1.1.198.5133