

ALAT PRAPEMPROSESAN DATA

UNTUK SISTEM SOKONGAN EKSEKUTIF UNIVERSITI

Muhammad Faiz Bin Abdullah
Prof. Madya Dr. Mohd Zakree Ahmad Nazri

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Projek ini dijalankan bertujuan untuk membangunkan alat prapemprosesan untuk membersihkan data yang dimuat naik oleh pengguna. Sains data merupakan suatu bidang yang berkembang pesat selari dengan Revolusi Perindustrian 4.0. Perkara teras yang perlu dilakukan oleh saintis data sebelum melakukan tugas analitik adalah memproses data tersebut untuk menjadi data yang bebas daripada sebarang kecacatan dan kesilapan. Proses prapemprosesan data adalah teknik perlombongan data yang melibatkan perubahan data mentah kepada format yang mudah difahami. Data dunia sebenar selalunya tidak lengkap dan mungkin mengandungi banyak kesilapan. Prapemprosesan data adalah kaedah yang terbukti dapat menyelesaikan masalah tersebut. Alat perlombongan data seperti RapidMiner atau Weka menyediakan alat untuk prapemprosesan data tetapi tidak menyediakan bimbingan kepada pengguna tentang apa yang perlu dilakukan untuk membersihkan data. Metodologi yang digunakan bagi pembangunan sistem ini adalah *iterative*. Untuk tujuan ujian, data pekerja dan pelajar universiti akan digunakan. Reka bentuk aplikasi ini adalah berdasarkan seni bina web dengan reka bentuk antara muka yang responsif. Aplikasi ini akan dipasang di pelayan dan pengguna boleh mencapainya di Internet. Selepas proses prapemprosesan berlaku, data-data tersebut boleh digunakan untuk tujuan analitik dan visualisasi. Modul analitik dan visualisasi adalah luar daripada skop projek ini. Bahasa pengaturcaraan yang digunakan untuk membangunkan sistem ini ialah Python. Adalah diharapkan, dengan terbangunnya alat ini, saintis data dapat menjalankan kerja-kerja prapemprosesan dengan lebih efisien.

1 PENGENALAN

Kepentingan penggunaan perlombongan data oleh dunia industri pada Zaman Revolusi Industri ke-4 ini tidak boleh disangkal lagi. Setiap organisasi mempunyai simpanan data mereka yang tersendiri. Set data yang besar kadangkala adalah terlalu kompleks untuk difahami. Namun, organisasi yang berjaya menganalisis data tersebut akan mempunyai kelebihan yang agak signifikan.

Sains data amat membantu dalam membangunkan dan meningkatkan kompetensi sesebuah organisasi. Ini kerana sains data dapat membantu sesebuah organisasi untuk membuat keputusan dengan lebih baik, mengenalpasti peluang dan membuat ramalan dengan mengenali corak tersembunyi dalam sesebuah set data.

Sains data merangkumi pelbagai tugas seperti klasifikasi, peramalan, penggabungan data dan sebagainya dalam menghasilkan sesebuah keputusan yang baik. Sebelum menganalisis

sesebuah set data, data tersebut perlulah melalui prapemprosesan untuk membersihkan data. Prapemprosesan data adalah proses yang merangkumi pembersihan, integrasi, pendiskritan data dan sebagainya.

2 PENYATAAN MASALAH

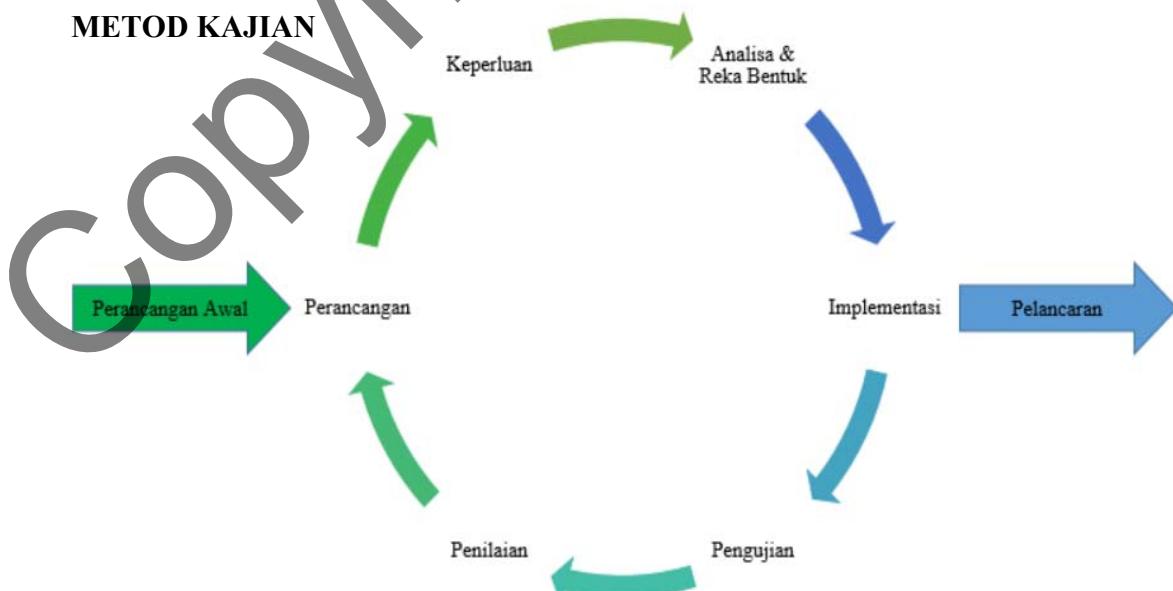
Aktiviti pengumpulan set data yang besar seringkali akan berhadapan dengan masalah. Antara masalah yang biasa dihadapi ialah data yang terkeluar daripada julat, penggabungan data yang mustahil dan kehilangan nilai. Analisis data yang dilakukan dari set data yang bermasalah akan berkemungkinan menghasilkan keputusan yang tidak tepat. Justeru, perkara yang paling penting sebelum melakukan analisis data ialah mengesahkan kualiti data tersebut sebelum menggunakan.

3 OBJEKTIF KAJIAN

Objektif utama kajian ini adalah seperti berikut:

- i. Membangunkan alat prapemprosesan data yang akan dijadikan sebagai platform untuk melakukan tugas prapemprosesan data yang memfokuskan kepada data universiti.
- ii. Membangunkan modul pembersihan data berdasarkan seni bina web

4 METOD KAJIAN



Rajah 1: Metodologi *iterative*.

Metodologi yang digunakan sepanjang pembangunan sistem ini ialah metodologi *iterative*. Metodologi ini digunakan secara meluas untuk kerja-kerja pembangunan yang besar. Sepanjang pembangunan sistem, kitaran hidup pembangunan sistem boleh dilakukan berulang kali. Metodologi ini dibahagikan kepada 6 fasa utama iaitu:

4.1 Fasa Perancangan

Tujuan fasa ini dilakukan adalah untuk mencari skop masalah dan mengenalpasti penyelesaian masalah.

4.2 Fasa Keperluan

Fasa ini dilakukan untuk mengenal pasti dan menyediakan kesemua keperluan pengguna dan sistem yang diperlukan untuk membangunkan sistem ini.

4.3 Fasa Analisa & Reka Bentuk

Fasa ini dilakukan untuk menganalisis dan merangka seni bina sistem ini. Spesifikasi, ciri dan operasi sistem akan dibincangkan secara lebih mendalam dalam sistem ini.

4.4 Fasa Implementasi

Pada fasa ini, pembangunan sebenar sistem akan dimulakan di mana pengaturcaraan perisian dilakukan.

4.5 Fasa Pengujian

Pada fasa ini, modul yang telah disiapkan akan diuji untuk memastikan kod tidak bermasalah dan fungsi sistem menepati keperluan pengguna.

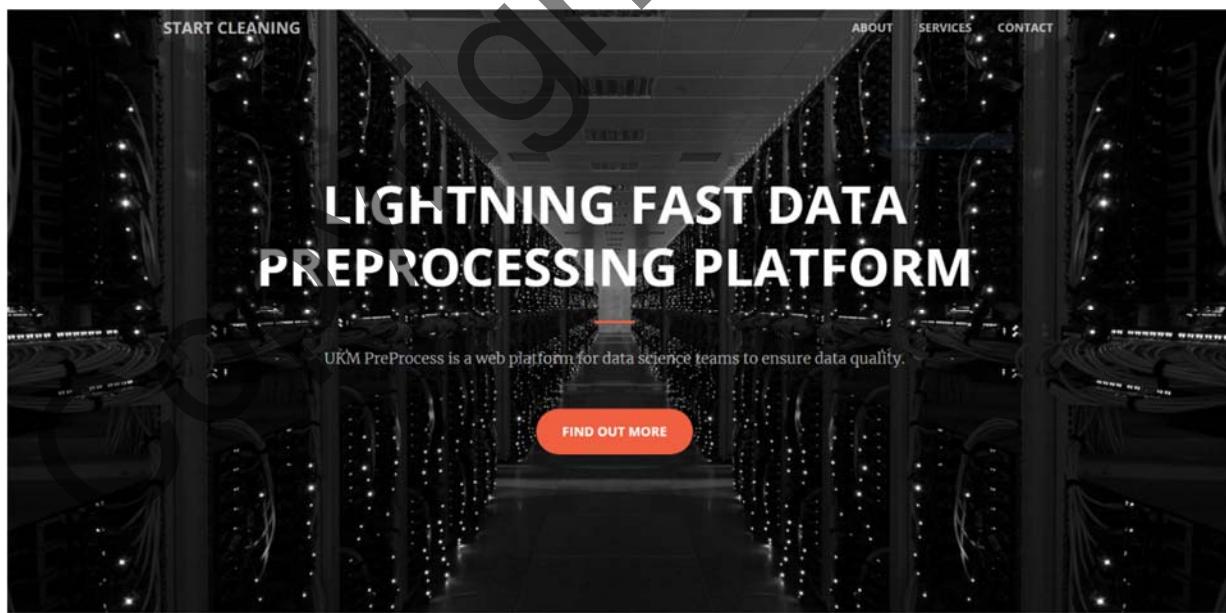
4.6 Fasa Penilaian

Fasa ini akan menilai hasil daripada dapatan kajian empat fasa sebelum ini. Keefisienan sistem akan diukur dan sebarang cadangan penambahbaikan akan dinilai.

5 HASIL KAJIAN

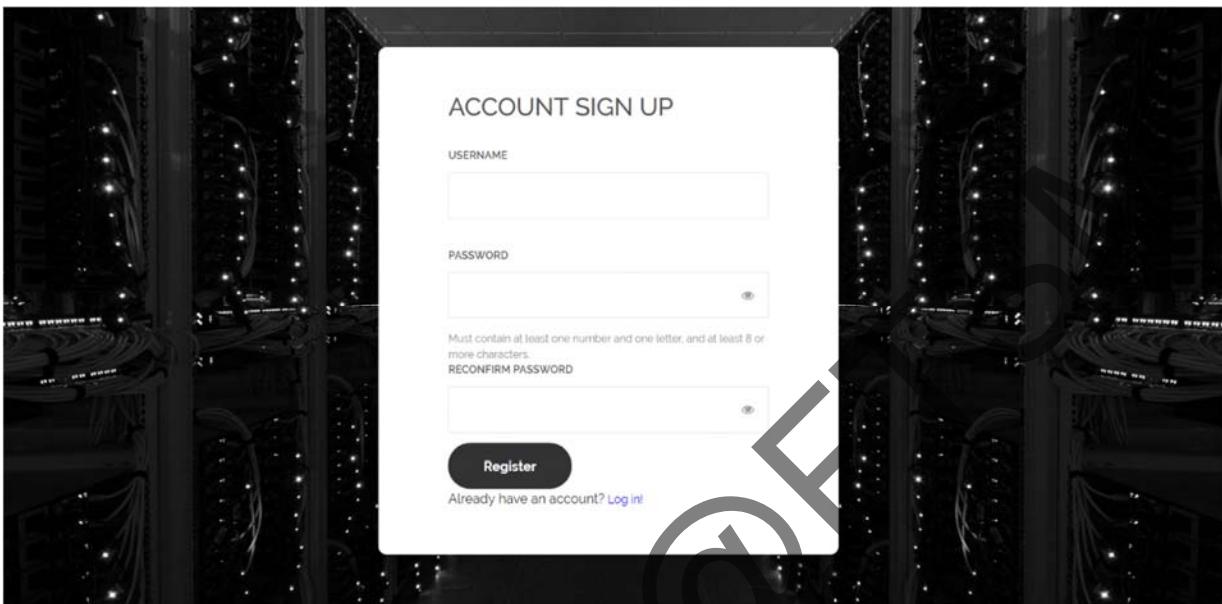
Bahagian ini membincangkan hasil daripada proses pembangunan Alat Prapemprosesan Data untuk Sistem Sokongan Eksekutif Universiti. Dalam projek ini, bahasa pengaturcaraan Python versi 3.6 dan rangka kerja Django versi 2.0 telah digunakan untuk membangunkan bahagian belakang web ini. Sementara itu, bahagian hadapan pula dibangunkan menggunakan HTML5, Cascading Style Sheet (CSS) dan Bootstrap untuk menjadikan antara muka lebih responsif. Antara muka dan fungsi sistem akan diterangkan lebih lanjut dalam bahagian ini.

Rajah 2 menunjukkan antara muka laman bagi alat prapemprosesan data ini. Ciri-ciri dan informasi tentang sistem boleh dibaca oleh pengguna pada laman ini.



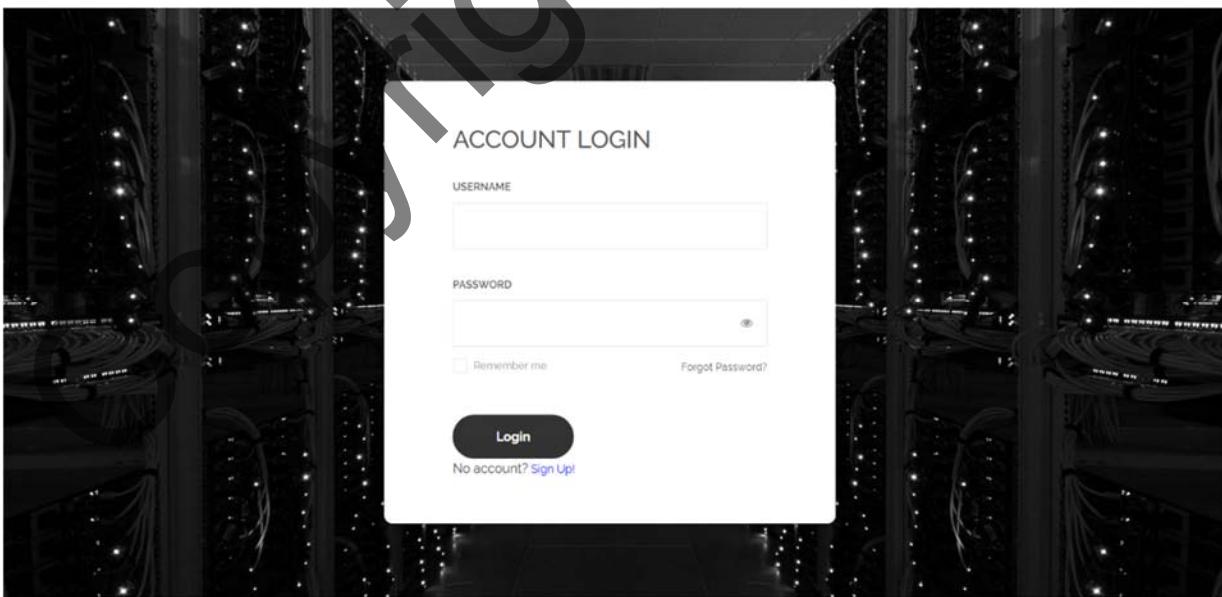
Rajah 2: Laman utama bagi Alat Prapemprosesan Data.

Rajah 3 menunjukkan antara muka bagi pengguna untuk mendaftar masuk ke dalam sistem ini. Pengguna hanya perlu memasukkan nama pengguna dan kata laluan untuk mencipta akaun baharu.



Rajah 3: Antara muka daftar pengguna.

Rajah 4 menunjukkan antara muka untuk pengguna log masuk ke dalam sistem ini. Dengan memasukkan nama pengguna dan kata laluan yang betul, pengguna akan dapat memasuki sistem ini.



Rajah 4: Antara muka log masuk.

Rajah 5 menunjukkan bar navigasi pengguna setelah log masuk ke dalam sistem.

Pengguna boleh memilih untuk ke laman muat naik data, paparan senarai data, tetapan akaun dan log keluar.



Rajah 5: Antara muka bar navigasi.

Rajah 6 menunjukkan antara muka untuk pengguna melakukan tetapan akaun. Pengguna boleh mengubah nama pengguna dan kata laluan baharu pada laman ini.

 A screenshot of the "Manage Account" form. The form has a light gray background with a green header bar at the top containing the title "Manage Account". Below the header, there are three input fields: "New Username" with placeholder "New Username", "New Password" with placeholder "New Password" and a note "Must contain at least one number and one letter, and at least 8 or more characters", and "Confirm Password" with placeholder "Retype New Password". At the bottom of the form, there is a blue button labeled "Create Form" and a green button labeled "Update Account".

Rajah 6: Antara muka tetapan akaun pengguna.

Rajah 7 menunjukkan antara muka untuk pengguna memuat naik fail berformat *comma-separated values* (CSV). Pengguna boleh memilih sama ada untuk memuat naik fail dari peranti mereka atau memasukkan *uniform resource locator* (URL) untuk menyimpan fail di dalam pelayan fail. Pengguna boleh memasukkan nama dan deskripsi fail.

The screenshot shows a web-based upload interface titled 'Upload'. It includes fields for 'Name' and 'Description', and a section for 'Upload via' with options for 'File' (selected) and 'URL'. A 'Browse' button is available for selecting files. Below these is a 'URL' input field. At the bottom of the form is a green horizontal bar containing two buttons: 'Clear Form' and 'Upload'.

Rajah 7: Antara muka muat naik fail CSV.

Rajah 8 menunjukkan antara muka untuk pengguna melihat fail CSV yang disimpan dalam pelayan fail. Pengguna boleh memilih sama ada untuk melakukan prapemprosesan data, muat turun fail berformat CSV atau membuang fail dari pelayan fail.

The screenshot displays a table of uploaded data entries. The columns are labeled 'Name', 'Description', 'Time Uploaded', and 'Action'. The table contains four rows:

Name	Description	Time Uploaded	Action
Data FKAB	Data ini tidak lengkap.	May 23, 2018, 11:57 a.m.	Clean Download Delete
Data FSSK	Format tanak, perlu dibetulkan	May 23, 2018, 12:44 p.m.	Clean Download Delete
Data FTSM	Data ini mempunyai banyak atribut tidak berguna.	May 23, 2018, 12:33 p.m.	Clean Download Delete
Data Pensyarah	Nama universiti berlindan.	May 23, 2018, 10:47 a.m.	Clean Download Delete

At the bottom left, it says 'Showing 1 to 4 of 4 entries'. At the bottom right, there are 'Previous' and 'Next' buttons.

Rajah 8: Antara muka senarai data.

Rajah 9 menunjukkan antara muka untuk pengguna melakukan prapemprosesan data. Pengguna boleh melihat kandungan fail data yang dimuat naik. Pengguna boleh menetapkan bilangan entri yang ingin dilihat dan melakukan carian perkataan di kotak pencarian. Kolumn

bernilai ‘nan’ yang berwarna kuning menandakan data dalam fail tersebut hilang atau tidak mempunyai nilai.

PreProcess   List of Data 												
	Show <input type="button" value="10"/> entries	Search:										
Bil. UKM(Per)	Nama	Status Semasa	Status Cuti	Jabatan	Fakulti	Jawatan	Tarikh Lantikan	Tarikh Tamat Jawatan	Tarikh Mula Cuti	Tarikh Tamat Cuti	Jam	
0 1 K559699	HABABAH PATANG	Berhenti (Bersara Wajib - 26/06/2015)	nan	jabatan DEKAN FAKULTI TEKNOLOGI & SAINS MAKLUMAT	FAKULTI TEKNOLOGI & SAINS MAKLUMAT	SETIAUSAHA PEJABAT N28 (KUP)	1975-07-14	26/6/2015	nan	nan	52.00	
1 2 K559668	NEGAT RAZAK HANBAN	Aktif	nan	PUSAT PENYELIDIKAN TEKNOLOGI KECERDASAN BUAUTAN (CAIT)	FAKULTI TEKNOLOGI & SAINS MAKLUMAT	PROFESOR VK7 (DS)	1975-04-09	4/2/2018	nan	nan	160.50	
2 2 K566668	NEGAT RAZAK HANBAN	Aktif	nan	PUSAT PENYELIDIKAN TEKNOLOGI KECERDASAN BUAUTAN (CAIT)	FAKULTI TEKNOLOGI & SAINS MAKLUMAT	PROFESOR VK7 (DS)	1975-04-09	4/2/2018	nan	nan	160.50	
3 4 K557696	ABI ATAN	Aktif	nan	jabatan DEKAN FAKULTI TEKNOLOGI & SAINS MAKLUMAT	FAKULTI TEKNOLOGI & SAINS MAKLUMAT	PEMBANTU TADBIR (P/O) N22 (KUP)	1979-04-06	11/12/2015 14/12/1995 13/6/1999	11/12/2015	14/12/1995 13/6/1999	107.80	
4 5 K558386	ZIHANA SAKAN	Aktif	nan	PUSAT PENYELIDIKAN TEKNOLOGI KECERDASAN BUAUTAN (CAIT)	FAKULTI TEKNOLOGI & SAINS MAKLUMAT	PENGAJAR KHAS SARJANA TAMU	1982-06-26	2/5/2017	28/10/1999	23/4/2004	69.00	
5 6 K558339	ZIRAAHAB NEGATKAH	Berhenti (Bersara Wajib - 09/04/2016)	nan	PUSAT PENYELIDIKAN TEKNOLOGI & PENGURUSAN PERISIAN (SOFTAM)	FAKULTI TEKNOLOGI & SAINS MAKLUMAT	PENSYARAH UNIVERSITI DS52	1982-01-08	9/4/2016	31/12/1999	30/9/2003	75.00	
6 7 K558678	NIRAATA BT. NIKHTAR	Aktif	nan	PUSAT PENYELIDIKAN TEKNOLOGI & PENGURUSAN PERISIAN (SOFTAM)	FAKULTI TEKNOLOGI & PENDYARAH UNIVERSITI DS54	TEKNOLOGI & PENDYARAH UNIVERSITI DS54	1982-06-09	11/12/2015	nan	nan	45.00	

Rajah 9: Antara muka paparan kandungan data.

Rajah 10 menunjukkan antara muka pengguna untuk melihat statistik data. Statistik data yang dipaparkan adalah bilangan entri, bilangan data unik, mod, frekuensi, purata, sisihan piawai, nilai minimum dan nilai maksimum.

PreProcess   List of Data 											
Data Statistics											
Bil. UKM(Per)	Nama	Status Semasa	Status Cuti	Jabatan	Fakulti	Jawatan	Tarikh Lantikan	Tarikh Tamat Jawatan	Tarikh Mula Cuti	Tarikh Tamat Cuti	Jam
count 192.000000	192	192	9	192	192	192	192	192	109	109	168
uniqueNaN	157	162	22	3	3	30	146	27	86	97	100
top NaN	K566668	NAAR FAEZAH NAHB. YATAN	InAktif	PUSAT PENYELIDIKAN BELAJAR PENGURUSAN PERISIAN (SOFTAM)	FAKULTI TEKNOLOGI & SAINS MAKLUMAT	PENSYARAH UNIVERSITI DS52	2000-06-11	11/12/2015 27/10/1997 26/1/1999 12/2000			
freq	NaN	3	2	164	6	72	192	36	4	163	7
mean	96.494792	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	55.578631	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	48.750000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	98.500000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	144.250000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	192.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Rajah 10: Antara muka statistik data.

Rajah 11 menunjukkan antara muka bagi penstrukturkan semula data. Pengguna boleh menggugurkan kolumn pilihan atau menamakan semula kolumn pilihan.

The screenshot shows the 'PreProcess' software interface with the following details:

- Top Bar:** PreProcess, Upload New Data, List of Data, and a user icon.
- Data Preview:** A table showing various statistics for a dataset. The columns include: top, freq, mean, std, min, 25%, 50%, 75%, and max. The rows include: NAAR, FAEZAH, NAHB., YATAN, K566866, Cuti, PUSAT PENYELIDIKAN TEKNOLOGI & Belajar PENGURUSAN PERISIAN (SOFTAM), FAKULTI TEKNOLOGI & SAINS MAKLUMAT, PENSYARAH UNIVERSITI DS52, and dates/times.
- Left Panel:**
 - Restructure Table:** Contains buttons for Drop Column, Rename Column, Normalize Data Type, Incomplete Row, and Modify Data.
 - Column:** Shows the current column name 'Bil.' and a 'Drop Column' button.
 - Old Header Name:** Shows the current header 'Bil.'
 - New Header Name:** An empty input field for renaming the column.
- Bottom Buttons:** Accept and Cancel.

Rajah 11: Antara muka penstrukturkan semula data.

Rajah 12 menunjukkan antara muka bagi normalisasi jenis data. Pengguna boleh menukar jenis data kepada objek, *float*, *integer* atau tarikh mengikut kolumn pilihan.

The screenshot shows the 'PreProcess' software interface with the following details:

- Top Bar:** PreProcess, Upload New Data, List of Data, and a user icon.
- Left Panel:** Normalization settings for 'Normalize Data Type'.
- Table:** A grid for changing data types. It lists columns: Bil., UKM(Per), Nama, Status Semasa, Status Cuti, Jabatan, Fakulti, Jawatan, Tarikh Lantikan, Tarikh Tamat Jawatan, Tarikh Mula Cuti, Tarikh Tamat Cuti, and Jam. For each column, it shows the current 'Data Type' (e.g., float64, object) and a dropdown menu for 'Change To' with options like Object, Float-64, Integer-64, Datetime-64, and Object. The 'Integer-64' option is highlighted in blue.
- Bottom Buttons:** Accept and Cancel.

Rajah 12: Antara muka normalisasi jenis data.

Rajah 13 menunjukkan antara muka bagi penggantian data tidak lengkap. Pengguna boleh menggugurkan baris data yang tidak mempunyai sebarang nilai, menggugurkan baris data yang

tidak menepati syarat yang ditetapkan, menggantikan nilai hilang dengan purata dan menggantikan nilai hilang dengan mod.

The screenshot shows the 'Incomplete Row' section of the PreProcess software. It includes buttons for 'Drop Empty Row', 'Drop Row with Missing Attribute', 'Drop Row without this Column', and 'Fill with Mean' or 'Fill with Mode' for specific columns. A large watermark 'Copy Right © FTSM' is overlaid across the interface.

Rajah 13: Antara muka penggantian data tidak lengkap.

Rajah 14 menunjukkan antara muka bagi modifikasi nilai data. Pengguna boleh menggugurkan baris data yang berulang, membuang whitespace, dan menggantikan karakter mengikut kolumn pilihan.

The screenshot shows the 'Modify Data' section of the PreProcess software. It includes buttons for 'Drop Duplicate Data', 'Remove Left Whitespace', 'Remove Right Whitespace', and 'Remove Both Side Whitespace'. Below these are sections for 'Select Column' and 'Char to be replaced' with a 'Replace with' field and a 'Modify Char' button. A large watermark 'Copy Right © FTSM' is overlaid across the interface.

Rajah 14: Antara muka modifikasi nilai data.

6 KESIMPULAN

Sebagai seorang pelajar yang telah mempelajari teori berkenaan kejuruteraan perisian, projek ini telah berjaya memupuk diri dalam menterjemah teori kepada praktikal pembangunan perisian. Pendedahan seperti ini telah mempersiapkan diri saya dalam menghadapi dunia realiti pembangunan perisian yang semakin hari semakin mencabar. Sepanjang perjalanan projek ini, saya telah menguasai bahasa baharu iaitu Python dan rangka kerja Django untuk membangunkan laman web. Akhir sekali, diharap bahawa cadangan penambahbaikan dapat diimplementasikan supaya sistem ini menjadi lebih sempurna.

7 RUJUKAN

Pyle, D., Editor, S., & Cerra, D. D. 1999. Data Preparation for Data Mining. Jil. 1. San Diego: Morgan Kauffman Publishers.

RapidMiner. 2017. Data Science Platform | RapidMiner. <https://rapidminer.com>. [21 Oktober 2017].

University of Waikato. 2017. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. <https://www.cs.waikato.ac.nz/ml/weka/>. [21 Oktober 2017].

KNIME. 2017. KNIME Product Matrix | KNIME. <https://www.knime.com/products/product-matrix>. [21 Oktober 2017].