ZULFAQAR Journal of Defence Science, Engineering & Technology
e-ISSN: 2773-5281
Vol. 6, Issue 2 (2023)
DOI: https://doi.org/10.58247/jdset-2023-0602-13
Journal homepage: https://zulfaqarjdset.upnm.edu.my

PERAMALAN RISIKO DIABETES MENGUNAKAN APLIKASI MYDIABETICRISK

Amir Afifuddin Ab Zayin[a], Zuraini Zainol[a*], Puteri Nor Ellyza Nohuddin[b], Hassan Mohamed[a]

[a] Department of Computer Science, Faculty of Science and Defence Technology, National Defence University of Malaysia, Sg. Besi Camp, 57000 Kuala Lumpur, Malaysia
[b] Faculty of Business, Higher College of Technology, Sharjah, United Arab Emirates

ARTICLE INFO

ARTICLE HISTORY
Received: 10-05-2023
Revised: 15-07-2023
Accepted: 01-09-2023
Published: 31-12-2023

KEYWORDS
Regresi logistic
Peramalan
Diabetes
Risiko
Perlombongan data

ABSTRACT

Diabetes merupakan salah satu penyakit pembunuh senyap yang mengancam nyawa di seluruh dunia. Pelbagai faktor risiko seperti umur, obesiti, diet pemakanan yang tidak seimbang, kurang bersenam, gaya hidup yang tidak sihat dan sebagainya telah dikenalpasti sebagai penyumbang utama kepada diabetes. MyDiabeticRisk merupakan aplikasi yang dapat digunakan untuk meramal peringkat awal diabetes di kalangan pesakit berdasarkan empat (4) faktor risiko iaitu glukos, tekanan darah, BMI dan umur menggunakan algoritma Regresi Logistik. Matlamat utama kajian ini adalah untuk membangunkan sebuah aplikasi berasaskan web yang dapat meramal diabetes mengunakan faktor risiko dan menghasilkan keputusan ramalan dalam bentuk grafik secara automatik. Dengan adanya aplikasi ini, seseorang individu dapat mengesan risiko penyakit diabetes lebih awal. Algoritma Regresi Logistik digunakan untuk membina model ramalan. Model yang telah dilatih digunakan untuk meramal risiko penyakit diabetes berdasarkan input pengguna.

Diabetes is one of the silent killer diseases that threaten lives all over the world. Various risk factors such as age, obesity, unbalanced diet, lack of exercise, unhealthy lifestyle and so on have been identified as major contributors to diabetes. MyDiabeticRisk is an application that uses logistic regression to predict early stage of Diabetes in patients based on four (4) primary factors such as glucose, blood pressure, BMI, and age. In this context, this web-based application is appropriate for individuals who want to determine whether they have or do not have Diabetes. The primary objective for this project is to create an application that can detect diabetes in patients using data and generate predictive findings in graphical style. The Logistic Regression algorithm was used in the methodology. From user input, the trained model is utilized to forecast diabetes illness outcomes.

1.0 INTRODUCTION

Diabetes atau penyakit kencing manis merupakan salah satu penyakit pembunuh senyap yang mengancam nyawa di seluruh dunia. Berdasarkan laporan [1] seramai 463 juta orang dewasa di seluruh dunia menghidap diabetes dan arah aliran (trend) ini terus meningkat terutamanya di negara-negara yang berpendapatan rendah. Pelbagai faktor risiko seperti umur, obesiti, diet pemakanan yang tidak seimbang, kurang bersenam, gaya hidup yang tidak sihat dan sebagainya telah dikenalpasti sebagai penyumbang utama kepada diabetes [2-4]. Diabetes di anggap sebagai salah satu penyakit kronik di mana paras gula dalam darah meningkat tinggi daripada paras normal [5-6] yang disebabkan oleh rembesan insulin yang rosak atau kesan biologinya yang terjejas, atau kedua-duanya [7]. Diabetes boleh mengakibatkan kegagalan organ berfungsi dalam badan seperti ginjal serta boleh mengakibatkan kematian. Diabetes juga merupakan penyakit jangka masa panjang yang disebabkan oleh kegagalan badan untuk menghasilkan insulin yang mencukupi, atau tidak dapat menggunakan insulin dengan betul. Insulin ialah hormon yang mengawal paras gula dalam darah dengan membenarkan gula masuk ke dalam sel-sel, dan dengan itu, menurunkan paras gula dalam darah. Kekurangan hormon insulin boleh

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my

menyebabkan serangan jantung, kegagalan buah pinggang, kemusnahan sel beta pankreas, penyakit vaskular perferal, koma dan sebagainya [8]. Terdapat dua (2) jenis diabetes iaitu jenis 1 dan 2. Hampir 90 % pesakit diabetes adalah jenis 2 di mana pankreas tidak dapat menghasilkan insulin sendiri, dan pesakit memerlukan suntikan insulin sepanjang hayat mereka [6].

Laporan Kementerian Kesihatan Malaysia (KKM) menunjukkan Malaysia mempunyai jumlah pesakit diabetes yang tertinggi di Asia Tenggara selepas Arab Saudi [9]. Kadar kematian yang semakin meningkat di kalangan masyarakat Malaysia berpunca daripada penyakit diabetes. Hal ini adalah kerana, masyarakat kurang mengambil berat tentang bagaimana diabetes ini boleh berada dalam tubuh badan. Oleh itu, masyarakat perlu tahu bagaimana penyakit ini boleh berlaku. Tanpa pengawasan yang berterusan, penyakit diabetes boleh mengakibatkan pertambahan gula di dalam darah dan seterusnya boleh membawa kepada komplikasi yang mengancam nyawa seperti strok dan penyakit berkaitan jantung [10]. Oleh itu, pengesanan awal risiko diabetes adalah penting bagi membantu masyarakat untuk hidup bebas daripada penyakit ini. Sekiranya diabetes dapat dikesan daripada awal, kesan bahaya penyakit ini dapat dielakkan dengan pengambilan ubat yang mencukupi [11]. Adalah diharapkan kajian ini dapat membantu pihak hospital dan pusat jagaan kesihatan untuk memantau tahap penyakit diabetes di kalangan rakyat Malaysia.

Teknik perlombongan data banyak digunakan dalam sektor kesihatan khususnya dalam penerokaan ramalan pelbagai penyakit [2, 12-21]. Pemodelan ramalan menggunakan teknik perlombongan data dan algoritma pembelajaran mesin untuk mengenal pasti corak dalam data dan pengesanan awal diabetes Sehingga kini, banyak kajian penyelidikan lebih tertumpu kepada pembinaan model ramalan diabetes dan tidak banyak kajian ramalan diabetes yang memfokuskan kepada pembangunan aplikasi atau sistem. Terdapat beberapa kajian yang dijalankan untuk membangunkan aplikasi peramalan diabetes. Sebagai contoh, penyelidik [2] telah membangunkan satu aplikasi pintar berasaskan web yang dapat meramal diabetes berdasarkan data klinikal. Kajian ini menggunakan pelbagai algoritma pembelajaran mesin seperti *Decision tree* (DT), *Naive Bayes* (NB), *k-nearest neighbor* (KNN), *Random Forest* (RF), *Gradient Boosting* (GB), *Logistic Regression* (LR) and *Support Vector Machine* (SVM). Dua set data iaitu *Pima Indians Diabetes* (set data 1) dan set data diabetes (set data 2) telah digunakan untuk kajian ini. Berdasarkan kajian mereka, pemodelan ramalan SVM telah memberikan peratus ketepatan ramalan tertinggi iaitu 80.26% untuk set data 1 manakala pemodelan ramalan DT dan RF memberikan ketepatan ramalan sebanyak 96.81% bagi set data 2. Untuk pembangunan aplikasi pintar, penyelidik telah mengintegrasi model ramalan SVM untuk meramal risiko diabetes mengunakan rangka kerja web mikro *Flask* berasaskan bahasa pengaturcaraan *Python*. Dalam aplikasi ini, pengguna perlu memasukkan 8 faktor risiko iaitu *Glucose, Pregnancy, BMI, Blood Pressure, Skin Thickness, Insulin, Age* dan *Diabetes Predigree.* Selepas memasukan input yang sah, aplikasi ini akan meramal pengesanan awal risiko penyakit diabetes sama ada pengguna tersebut menghidap diabetes atau tidak.

Dalam kajian lain, satu aplikasi berasaskan web dibangunkan yang mana dapat meramal sama ada seseorang pesakit menghidap diabetes atau tidak [20]. Pembangunan aplikasi ini menggunakan bahasa pengaturcaraan web PHP sebagai *back-end,* JavaScript sebagai *front-end* dan *Tensorflow.js* untuk perlaksanaan kod pemodelan ramalan *Artificial Neural Network* (ANN). Kajian ini menggunakan set data *Pima Indians Diabetes.* Sama seperti kajian [2], aplikasi ini memerlukan pengguna untuk menginput lapan (8) faktor risiko diabetes. Hasil keputusan experimen menunjukkan pemodelan ANN memberikan ketepatan ramalan sebanyak 82.35 %. Selain itu, satu aplikasi interaktif yang dapat meramal kebarangkalian normal (NGT), pra diabetes (IGT) dan diabetes (DM) telah dibangunkan oleh penyelidik [21]. Dalam kajian ini, model ramalan *Bayesian Network* (BN) telah dilatih menggunakan set data untuk 1531 subjek daripada *Ansung cohort of the Korean Genome and Epidemiological Study* (KoGES). Set data ini mengandungi faktor risiko tradisional (status semasa diabetes, jantina, umur, etc.), toleransi glukosa, biomarker EPC (AhRL, MIS-ATP, MIS-ROS). Biomarker serum EPC dapat mengukur tahap pendedahan individu kepada bahan kimia pencemar alam sekitar. Keputusan ramalan menunjukkan bahawa biomarker EPC mempunyai kesan interaktif terhadap perkembangan diabetes. Selain itu, penggunaan biomarker EPC ini telah menyumbang kepada peningkatan yang ketara dalam prestasi ramalan diabetes. Perisian yang digunakan untuk pembangunan aplikasi DiabetisBN tidak dibincangkan secara terperinci dalam kajian mereka.

Berdasarkan pemerhatian di atas, didapati hanya sebilangan kecil sahaja kajian yang memfokuskan kepada pembangunan sistem peramalan diabetes. Kajian ini akan membina model ramalan faktor risiko diabetes menggunakan algoritma Regresi Logistik. Pembangunan aplikasi ini menggunakan *Django*

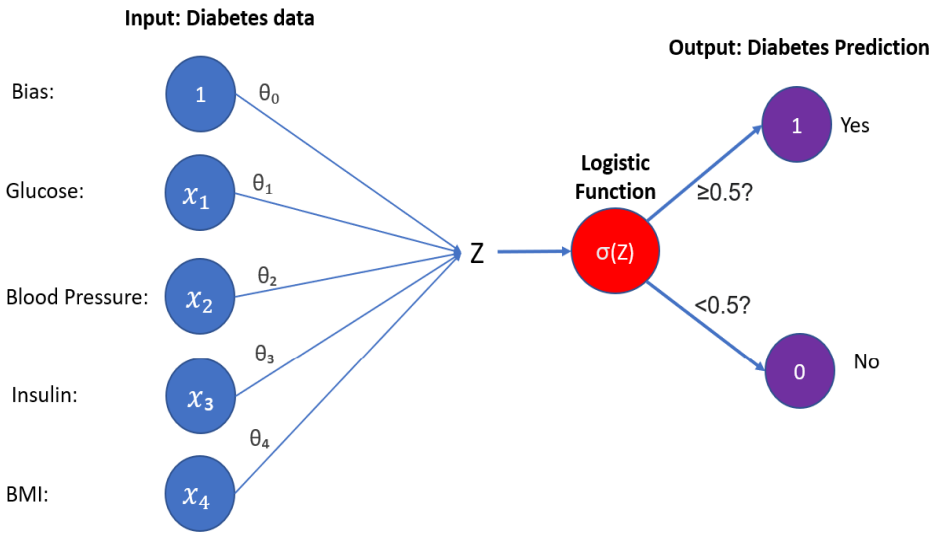
*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my

Python Web Framework dan *PyCharm* sebagai editor kod. *Django* menggunakan *HTML*, *CSS* dan *JavaScript* sebagai *front-end* dan bahasa program *Python* sebagai *back-end*. Set data yang akan digunakan dalam kajian ini ialah *PIMA Indian Diabetes* [24]. Set data ini mengandungi data pesakit wanita yang berumur sekurang-kurangnya 21 tahun ke atas. Oleh itu, matlamat kajian ini adalah untuk membangunkan satu aplikasi yang berasaskan web *MyDiabeticRisk* yang dapat meramal pengesanan awal risiko penyakit diabetes seseorang individu dengan lebih cepat dan efisien. Aplikasi ini menyediakan kemudahan kepada pengguna khususnya pihak hospital atau pusat penjagaan kesihatan untuk menganalisis data pesakit dengan lebih cekap dan menghasilkan keputusan ramalan risiko penyakit diabetes dalam bentuk yang mudah difahami oleh pengguna.

2.0 LATAR BELAKANG DAN KAJIAN LITERATUR

Perlombongan data, juga dikenali sebagai penerokaan data merupakan salah satu fasa yang terpenting dalam *Knowledge-Discovery in Databases* (KDD) [22]. Perlombongan data secara amnya merujuk kepada proses mencari korelasi atau corak yang menarik dalam sejumlah data besar dengan menggunakan pelbagai algoritma [23-27]. Perlombongan data merupakan gabungan kaedah dan peralatan (tools) daripada pelbagai bidang seperti pembelajaran mesin, statistik dan pangkalan data [28]. Pengarang [25] telah menyenaraikan tugasan (task) dan teknik perlombongan data. Antara contoh teknik perlombongan data yang popular ialah *DT*, *Bayesian Classification*, *NB*, *Fuzzy Logic*, *SVM*, *Hierarchical Clustering* dan sebagainya. Analisis peramalan merupakan kaedah yang mengintegrasikan pelbagai teknik perlombongan data, algoritma pembelajaran mesin dan statistik yang menggunakan set data yang terkumpul dan semasa untuk meramal risiko masa depan [29]. Analisis peramalan bergantung pada perhubungan antara pemboleh ubah bersandar (*dependent variable*) dan tidak bersandar (*indepedent variable*) daripada satu kejadian lalu dan mengeksplotasinya untuk meramalkan hasil yang boleh diketahui. Walau bagaimanapun, adalah penting untuk diambil perhatian bahawa aspek ketepatan akan bergantung pada tahap analisis data. Menurut [30], analisis ramalan merupakan kaedah umum untuk meramal ketepatan eksperimen kuantitatif.

Regresi Logistik ialah teknik pembelajaran mesin yang berasaskan konsep statistik dan banyak digunakan selain daripada regresi linear. Kedua-dua teknik ini adalah standing dalam pelbagai cara, tetapi perbezaan utama terletak pada cara algoritma ini digunakan [10]. Regresi Logistik selalu digunakan dalam kajian yang berkaitan dengan peramalan dan klasifikasi. Ia dinamakan sempena fungsi logistik yang juga dipanggil sebagai fungsi sigmoid atau fungsi logistik yang merupakan asas kepada kaedah ini. Rajah 1 menunjukkan model Regresi Logistik di mana 1, X_1, X_2, X_3 dan X_4 merupakan pemboleh ubah input. Manakala θ_0, θ_1, θ_2, θ_3, dan θ_4 merupakan koefisyen model dan Z (Rumus 1) merupakan output yang diramalkan. Regresi Logistik akan mengira jumlah pemberat bagi setiap pemboleh ubah input bagi menghasilkan output α(Z) melalui fungsi khas yang dikenali sebagai fungsi sigmoid (Rumus 2). Kemudian, nilai Z akan ditukarkan dalam bentuk kebarangkalian iaitu 0 atau 1. Sekiranya nilai kebarangkalian >= 0.5 maka model Regresi Logistik ini akan meramalkan *Yes* (1) dan *No* (0) sekiranya nilai kebarangkalian < 0.5.



Rajah 1. Model regresi logistik

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my

$$Z = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_4 \quad (1)$$

Formula untuk fungsi sigmoid ditakrifkan seperti Rumus 2:

$$sigmoid(Z) = \frac{1}{(1 + e^{-Z})} \quad (2)$$

3.0 PERAMALAN ANALISIS DIABETES MENGGUNAKAN REGRESI LOGISTIK

Banyak kajian telah dijalankan untuk membuat peramalan risiko penyakit diabetes menggunakan algoritma Regresi Logistik. Antaranya ialah penyelidik menggunakan algoritma Regresi Logistik untuk meramal penyakit diabetes jenis 2 [10]. Kajian ini menggunakan set data *Pima Indians Diabetes.* Hasil keputusan eksperimen menunjukkan pemboleh ubah-pemboleh ubah *pregnancies*, *glucose* dan BMI menunjukkan nilai kepentingan tertinggi dalam peramalan penyakit diabetes. Daripada analisis yang dijalankan ke atas set data, pemodelan Regresi Logistik memberikan peratus ketepatan ramalan sebanyak 75.32 %. Rajendra dan Latifi [14] menggunakan Regresi Logistik sebagai algoritma utama dalam pemodelan ramalan diabetes. Manakala kaedah *ensemble* (*Max Voting/Majority Voting and Stacking*) digunakan untuk membandingkan peningkatan prestasi asal kajian ini. Dua set data iaitu *Pima Indians Diabetes* (set data 1) dan set data diabetes *Vanderbilt* (set data 2) telah digunakan untuk kajian ini. Teknik pra pemprosesan data telah digunakan dalam kajian ini. Hasil kajian menunjukkan, pemodelan *Regresi Logistik* memberikan peratus ketepatan ramalan sebanyak 78 % bagi set data 1 manakala peratus ketepatan ramalan sebanyak 98% bagi pemodelan *ensemble* untuk set data 2.

Dalam kajian lain, Aishwarya dan Vaidehi [3] menggunakan pelbagai algoritma pembelajaran mesin seperti *Support Vector Machines (SVM)*, *Random Forest Classifier*, *Decision Tree Classifier*, *Extra Tree Classifier*, *Ada Boost* , *Perceptron*, *Linear Discriminant Analysis*, Regresi Logistik, KNN, *Gaussian Naïve Bayes*, *Bagging* dan *Gradient Boost Classifier* [3]. Kajian ini menggunakan dua set data berbeza iaitu set data PIMA Indian dan set data diabetes untuk menguji pelbagai model. Hasil kajian menunjukkan, pemodelan Regresi Logistik memberikan peratus ketepatan ramalan sebanyak 96 %. Dalam kajian yang dijalankan oleh Joshi dan Chawan [31], algoritma Regresi Logistik dan SVM diguna pakai untuk tujuan peramalan penyakit diabetes. Secara umumnya, semua set data telah melalui proses pembersihan bagi mendapatkan keputusan yang lebih baik. Berdasarkan kajian mereka, pemodelan SVM memberikan peratus ketepatan ramalan sebanyak 79 %. Berdasarkan permerhatian di atas, dapat disimpulkan bahawa algoritma Regresi Logistik telah terbukti sebagai salah satu algoritma yang cekap dalam pemodelan ramalan. Selain daripada pilihan algoritma, terdapat faktor lain yang boleh meningkatkan ketepatan dan masa larian model, seperti prapemprosesan data, penyingkiran nilai berlebihan dan normalisasi.

4.0 METODOLOGI

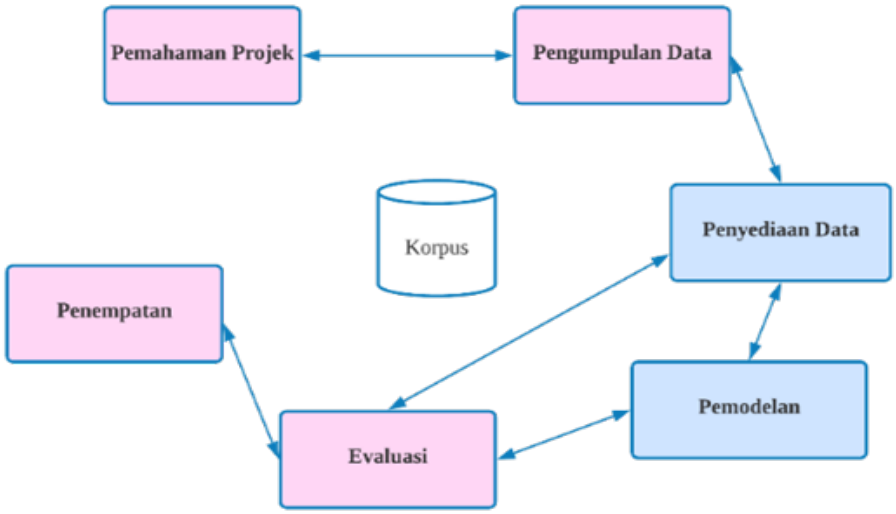
Kajian ini menggunakan metodologi CRISP-DM (*Cross Industry Standard Process for Data Mining*) untuk menerangkan pembangunan aplikasi *MyDiabeticRisk*. CRISP-DM merupakan model yang paling kerap digunakan oleh penyelidik dan pakar perlombongan data kerana ia menawarkan rangka kerja dan garis panduan lengkap untuk menyelesaikan masalah sedia ada [32]. Terdapat enam (6) fasa utama dalam Model CRISP-DM iaitu pemahaman projek, pengumpulan data, penyediaan data, pemodelan, penilaian dan penggunaan. Setiap fasa memainkan peranan penting dalam membangunkan aplikasi ini. Kelebihan model ini ialah, ia tidak perlu mengikut urutan fasa untuk membangunkan aplikasi. Model ini fleksibel di mana penyelidik sentiasa boleh kembali ke peringkat sebelumnya untuk melaraskan butiran. Rajah 2 menunjukkan fasa keseluruhan metodologi CRISP-DM dan aliran kerjanya.

4.1 Fasa 1: Pemahaman Projek

Sama seperti metodologi biasa yang lain, penyelidik perlu memahami objektif dan keperluan projek. Kajian ini tertumpu kepada membuat ramalan secara diagnostik samada seseorang individu menghidap penyakit diabetes atau tidak berdasarkan set atribut (faktor) yang dinyatakan dalam set data yang telah dimuat turun di laman sesawang *Kaggle*. Hasil keputusan ramalan ini dapat membantu seseorang individu meramal keputusan awal risiko diabetes seseorang individu dengan lebih efisien. Selain itu, ia dapat membantu pesakit diabetes untuk menjalani kehidupan yang lebih sihat. Objektif utama kajian ini

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my

ialah membuat analisis peramalan risiko penyakit diabetes dengan menggunakan set data yang dimuat turun di laman sesawang Kaggle. Secara keseluruhannya terdapat 768 data pesakit diabetes. Data ini mengandungi lapan (8) atribut (mewakili kriteria diagnosis perubatan) seperti *Glucose, Pregnancy, BMI, Blood Pressure, Skin Thickness, Insulin, Age, Diabetes Predigree* dan *satu target class* (mewakili status kesihatan individu yang telah diuji). Penerangan bagi set data mentah pesakit diabetes akan diterangkan di sub-bahagian seterusnya.



Rajah 2. Aliran kerja CRISP-DM diadaptasi daripada [33]

4.2 Fasa 2: Pengumpulan Data

Dalam kajian ini, set data diperolehi daripada *Pima Indians Diabetes Dataset* (PIDD) yang berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases* yang dimuat turun daripada laman sesawang *Kaggle*. Set data ini mengandungi maklumat tentang 768 sampel yang telah menjalani ujian diabetes. Jadual 1 menunjukkan lapan (8) atribut (faktor) yang berkaitan penyakit diabetes dan satu (1) atribut kelas (label). Atribut kelas ini mengandungi 2 nilai iaitu 0 dan 1. Nilai 0 ditafsirkan sebagai tidak menghidap diabetes manakala nilai 1 menunjukkan seseorang individu itu menghidap diabetes. Sebagai contoh, dalam 768 data, didapati 500 orang telah diuji dengan keputusan tidak menghidap diabetes manakala 268 orang lagi telah disahkan menghidap diabetes. Dalam kajian ini, lapan (8) atribut ini akan dikenali sebagai pemboleh ubah tidak bersandar (*independent variable*) manakala atribut kelas (label) dinamakan sebagai pemboleh ubah bersandar (*dependent variable*).

Jadual 1. Senarai atribut kajian

Bil	Nama Atribut	Perihalan pemboleh ubah Tidak Bersandar dan label
1.	*Pregnancy*	Bilangan kali kehamilan
2.	*Glucose*	Kepekatan glukosa plasma selama 2 jam dalam ujian toleransi glukosa oral
3.	*BloodPressure*	Tekanan darah diastolik (mm Hg)
4.	*Skin*	Ketebalan lipatan kulit trisep (mm)
5.	*Insulin*	Insulin serum 2 jam (mu U/ml)
6.	*Body Mass Index*	Indeks jisim badan (berat dalam Kg/Tinggi dalam m)2
7.	*Diabetes Pedigree function*	Fungsi keturunan diabetes
8.	*Age*	Umur (Tahun)
9.	*Class*	Pemboleh ubah kelas (0=Diuji Negatif atau 1=Diuji Positif)

Dalam kajian ini, hanya empat (4) atribut (faktor) seperti *Glucose, BMI, BloodPressure,* dan *Insulin* yang akan dipertimbangkan untuk analisis peramalan diabetes. Seterusnya set data yang terpilih ini akan

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my

dibersihkan dan penerangan terperinci tentang proses pembersihan akan dibicangkan dalam perenggan seterusnya.

4.3 Fasa 3: Penyediaan Data Kajian

Fasa penyediaan data juga dikenali sebagai proses pembersihan dan ianya sangat penting bagi memastikan data yang tidak lengkap dikeluarkan daripada set data [34]. Fasa ini akan menyediakan set data bersih yang akan digunakan dalam fasa pemodelan seterusnya. Ini termasuklah pemilihan atribut, pembersihan data dan perubahan (transformasi) data. Dalam fasa ini, masalah seperti data yang tidak bersih, mengandungi banyak ralat dan data yang tidak konsisten perlu menjalani prapemprosesan data. Ini kerana, data yang tidak bersih atau mempunyai nilai yang hilang menyebabkan keputusan akhir menjadi kurang tepat . Prapemprosesan merupakan proses pembersihan data bagi mendapatkan set data yang berkualiti. Berdasarkan Rajah 3, didapati set data ini mempunyai kehilangan nilai yang ditandakan dengan nilai 0 (warna kuning). Aturcara *Python* akan digunakan untuk mengenal pasti kehilangan nilai berdasarkan pemerhatian ke atas kehilangan data di lajur tertentu yang dilabelkan sebagai nilai 0. Teknik prapemprosesan data akan digunakan untuk membuang data mengikut keperluan kajian ini.

1	Pregnancie	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesPedigreeF	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1

Rajah 3. Contoh data mentah PIDD yang dipaparkan pada 10 baris pertama

Rajah 4 menunjukkan skor kehilangan data di 5 lajur utama iaitu *Glucose* (1), *BloodPressure* (2), *Skin* (3), *Insulin* (4) dan *BMI* (5). Berdasarkan keputusan eksperimen didapati lajur 4 (Insulin) menunjukkan skor kehilangan data yang tinggi di mana hampir separuh daripada set data kajian.

```
In [57]: # example of summarizing the number of missing values for each variable
         from pandas import read_csv
         # load the dataset
         dataset = read_csv('PIDD.csv', header=None)
         # count the number of missing values for each column
         num_missing = (dataset[[1,2,3,4,5]] == 0).sum()
         # report the results
         print(num_missing)
```

```
1      5
2     35
3    227
4    374
5     11
dtype: int64
```

Rajah 4. Contoh bilangan data yang hilang mengikut lajur

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my

```
In [59]: # example of review rows from the dataset with missing values marked
         from numpy import nan
         from pandas import read_csv
         # load the dataset
         dataset = read_csv('PIDD.csv', header=None)
         # replace '0' values with 'nan'
         dataset[[1,2,3,4,5]] = dataset[[1,2,3,4,5]].replace(0, nan)
         # print the first 10 rows of data
         print(dataset.head(10))
```

```
     0      1     2     3      4     5      6   7  8
0    0    1.0   2.0   3.0    4.0   5.0  6.000   7  8
1    6  148.0  72.0  35.0    NaN  33.6  0.627  50  1
2    1   85.0  66.0  29.0    NaN  26.6  0.351  31  0
3    8  183.0  64.0   NaN    NaN  23.3  0.672  32  1
4    1   89.0  66.0  23.0   94.0  28.1  0.167  21  0
5    0  137.0  40.0  35.0  168.0  43.1  2.288  33  1
6    5  116.0  74.0   NaN    NaN  25.6  0.201  30  0
7    3   78.0  50.0  32.0   88.0  31.0  0.248  26  1
8   10  115.0   NaN   NaN    NaN  35.3  0.134  29  0
9    2  197.0  70.0  45.0  543.0  30.5  0.158  53  1
```

Rajah 5. Data yang hilang ditandakan sebagai NaN

Berdasarkan Rajah 5, nilai NaN dapat dilihat dengan jelas dalam lajur 2, 3, 4 dan 5. Sebagai contoh, terdapat lima (5) data yang hilang dalam lajur 4. Nilai yang hilang ini akan memberi kesan kepada pengujian data. Strategi paling mudah untuk mengendalikan data yang hilang ialah dengan mengeluarkan rekod yang mengandungi nilai NULL. *Pandas* menyediakan fungsi *dropna()* yang boleh untuk membuang semua baris yang tiada data. Rajah 6 menunjukkan contoh set data mentah selepas fungsi *dropna()* digunakan dalam kajian ini.

```
In [63]: # example of removing rows that contain missing values
         from numpy import nan
         from pandas import read_csv
         # load the dataset
         dataset = read_csv('PIDD.csv', header=None)
         # summarize the shape of the raw data
         print(dataset.shape)
         # replace '0' values with 'nan'
         dataset[[1,2,3,4,5]] = dataset[[1,2,3,4,5]].replace(0, nan)
         # drop rows with missing values
         dataset.dropna(inplace=True)
         # summarize the shape of the data with missing rows removed
         print(dataset.shape)
```

```
(769, 9)
(393, 9)
```

Rajah 6. Semua baris yang tidak mempunyai data dialih keluar

Berdasarkan Rajah 6, 375 baris yang mengandungi NaN telah dialih keluar. Baki set data yang bersih adalah sebanyak 393 baris.

4.4 Fasa 4: Pemodelan

Dalam kajian ini, algoritma peramalan iaitu Regresi Logistik digunakan untuk membina model. Seperti yang dinyatakan dalam Fasa 3, set data yang telah dibersihkan akan dibahagikan kepada dua (2) bahagian iaitu set latihan dan set ujian. Pada peringkat ini, algoritma Regresi Logistik akan melatih set data latihan untuk menghasilkan model. Model yang terlatih ini akan menjadi pengelas dan data ujian akan digunakan untuk menilai model tersebut. Algoritma Regresi Logistik dipilih untuk membina model kerana ia memahami hubungan antara pemboleh ubah bersandar dan pemboleh ubah tidak bersandar dengan menganggar kebarangkalian. Jenis analisis ini boleh membantu penyelidik meramal kemungkinan

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my

peristiwa berlaku atau pilihan dibuat. Selepas itu, model yang telah dilatih ini akan digunakan untuk melakukan peramalan analisis ke atas data yang dikumpul.

4.5 Fasa 5: Evaluasi

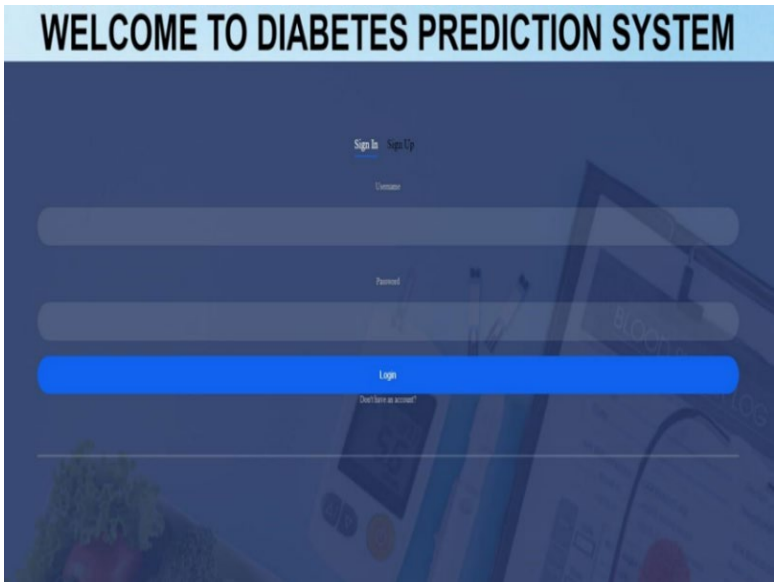
Dalam fasa ini, penyelidik boleh kembali ke peringkat penyediaan data dan pemodelan untuk menilai sama ada terdapat sebarang penambahbaikan atau sebarang butiran yang perlu ditambah pada kedua-dua peringkat. Tujuan fasa ini adalah untuk memastikan data dan model yang dibina disediakan dengan baik untuk menghasilkan keputusan yang baik. Set data ujian yang digunakan untuk menguji model yang dibina untuk mengetahui ketepatan model. Perlu ditekankan bahawa set data ujian mesti datang daripada set data yang sama seperti data latihan kerana ketepatan model akan dipengaruhi oleh perkara ini. Jika ketepatan tidak cukup ideal, adalah dinasihatkan untuk kembali dan melaraskan butiran pada fasa sebelumnya seperti penyediaan data dan pemodelan. Daripada analisis yang dijalankan ke atas set data latihan dan data ujian, peratus ketepatan ramalan pemodelan Regresi Logistik adalah 78 %.

4.6 Fasa 6: Penyerahan

Bergantung kepada keperluan projek, peringkat penggunaan boleh menghasilkan laporan dengan mudah bagi data yang dianalisis. Pada peringkat akhir, model yang dibina ini akan digunakan dalam aplikasi MyDiabeticRisk. Aplikasi ini direka bentuk untuk membuat ramalan penyakit diabetes secara masa nyata. Setiap data yang di analisa akan diklasifikasikan kepada dua tahap iaitu menghidap penyakit diabetes atau tidak menghidap penyakit diabetes. Semua keputusan akan divisualisasikan dalam bentuk graf.

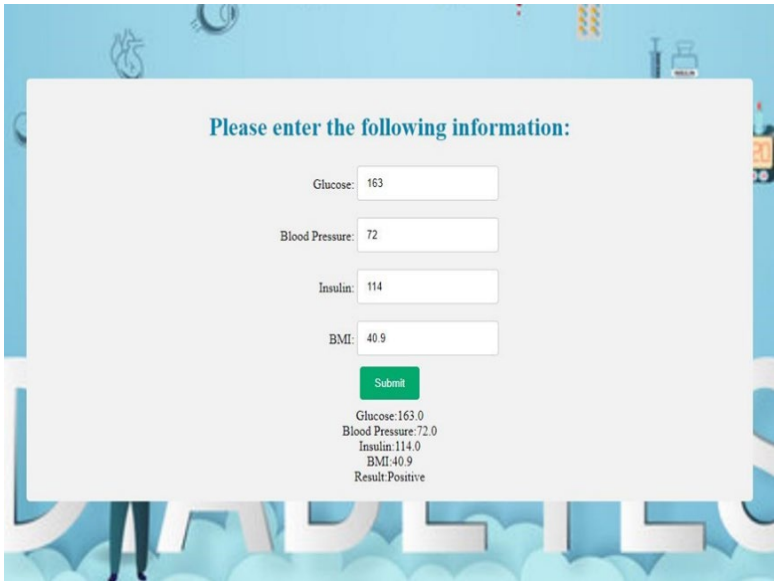
5.0 HASIL EXPERIMEN DAN PERBINCANGAN

Rajah 7 menunjukkan antara muka utama halaman log masuk bagi aplikasi *MyDiabeticRisk*. Pengguna yang berdaftar boleh melog masuk ke dalam aplikasi dengan memasukkan ID dan kata laluan yang tepat. Jika maklumat yang dimasukkan tidak tepat, pengguna tidak tidak dapat melog masuk ke dalam aplikasi dan satu mesej ralat akan dipaparkan kepada pengguna. Semua kata laluan yang disimpan dalam pangkalan data akan disulitkan untuk data privasi. Selepas log masuk ke aplikasi, pengguna akan dialihkan ke papan pemuka aplikasi ini. Pengguna perlu memasukkan data yang diperlukan iaitu *Glucose, BloodPressure Insulin* dan BMI. Semua data yang diinput tersebut mestilah dalam bentuk numerik. Model ramalan risiko diabetes yang dilatih meggunakan teknik Regresi Logistik akan meramal keputusan ramalan menggunakan input data. Contoh hasil keputusan ramalan risiko diabetes bagi seorang pengguna akan dipaparkan seperti di Rajah 8.



Rajah 7. Contoh Paparan log masuk aplikasi *MyDiabeticRisk*

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my



Rajah 8. Paparan hasil keputusan ramalan risiko diabetes berdasarkan input pengguna

Data yang diekstrak daripada pangkalan data akan diplotkan dalam bentuk graf. Rajah 9 memaparkan graf garis untuk semua data keputusan ramalan risiko penyakit diabetes bagi seseorang pesakit. Sebagai contoh, bermula dari tarikh 14-21 Januari, keputusan risiko diabetes pengguna adalah bersifat naik turun. Nilai 1 menunjukkan pengguna tersebut menghidap diabetes manakala nilai 0 adalah sebaliknya. Rajah 9 pula menunjukkan jujukan data untuk tahap *Glucose* bagi seseorang pesakit.

Rajah 9. Contoh paparan graf garis untuk ramalan risiko diabetes dan paras *glucose*

Rajah 10 memaparkan maklumat secara terperinci tentang keputusan ramalan risiko diabetes dan nilai data yang diinput di dalam aplikasi *MyDiabeticRisk*. Terdapat 22 baris yang mewakili 22 set data tersebut. Sebagai contoh, data yang bernombor 5, 6, 11, 15 dan 21 menujukkan pesakit tersebut diramalkan menghidap penyakit diabetes. Berdasarkan kepada fasa pengujian dan implementasi yang telah dilakukan, boleh disimpulkan bahawa aplikasi *MyDiabeticRisk* berjaya berfungsi dengan baik dan memenuhi objektif aplikasi. Walau bagaimanapun, penambahbaikan perlu dilakukan dari masa ke semasa supaya aplikasi ini menjadi lebih efektif dan efisien.

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my

No	Datetime	Glucose Level	Blood Pressure (mmHg)	Insulin	BMI (kg/m2)	Result
1	2022-01-15 00:16:46	6	148	72	35	0
2	2022-01-15 00:18:16	6	2	66	29	0
3	2022-01-15 00:21:50	32	2	2000	29	0
4	2022-01-15 00:33:17	6	148	72	35	0
5	2022-01-15 00:35:39	148	72	0	34	1
6	2022-01-15 02:09:35	183	64	0	23	1
7	2022-01-15 03:45:29	6	148	72	35	0
8	2022-01-15 15:49:11	6	148	72	35	0
9	2022-01-15 21:10:59	125	70	115	31	0
10	2022-01-15 21:11:46	125	70	115	31	0
11	2022-01-15 21:12:32	148	72	0	34	1
12	2022-01-15 21:25:33	125	70	115	31	0
13	2022-01-16 14:47:12	126	56	152	29	0
14	2022-01-17 23:51:12	139	80	0	27	0
15	2022-01-18 16:26:52	148	72	0	34	1
16	2022-01-19 01:17:00	100	66	90	33	0
17	2022-01-19 14:34:50	107	74	0	30	0
18	2022-01-19 22:37:52	158	76	245	32	0
19	2022-01-19 23:58:48	6	148	400	18	0
20	2022-01-19 23:59:07	88	188	400	35	0
21	2022-01-20 00:00:25	137	40	168	43	1
22	2022-01-20 23:14:37	118	84	230	46	0

Rajah 10. Contoh hasil eksperimen untuk set data yang diekstrak daripada aplikasi *MyDiabeticRisk*

6.0 KESIMPULAN

Penyakit diabetes seharusnya dicegah sebelum terlambat yang boleh mengakibatkan kematian. Di dalam kajian ini, penyelidik membangunkan aplikasi berasaskan web pada masa nyata (*MyDiabeticRisk*) yang boleh digunakan untuk menganalisis dan meramal penyakit Diabetes ini dihidap atau tidak dihidap oleh pesakit. Untuk tujuan ini, satu eksperimen telah disediakan dengan menggunakan teknik perlombongan data pada set data mentah. Ini adalah satu tugas yang sangat mencabar untuk menjalankan pra pemprosesan data kerana data mempunyai ralat nombor, nilai data hilang dan tidak lengkap. Algoritma Regresi Logistik telah digunakan pada data yang telah dibersihkan yang kemudiannya diklasifikasikan ke dalam empat (4) atribut utama untuk meramal keputusan kajian. Berdasarkan pada hasil kajian, pesakit telah menunjukkan beliau menghidap penyakit Diabetes di mana nilai *glucose, blood pressure*, insulin dan BMI mempunyai nilai yang tinggi berdasarkan daripada set data yang dilatih. Kesimpulannya, aplikasi *MyDiabeticRisk* boleh dimanfaatkan oleh pihak hospital atau mana-mana pusat kesihatan untuk meramal diabetes dalam kalangan pesakit berasaskan empat (4) atribut yang diperlukan dengan mudah dan sistematik. Selain itu, aplikasi ini menghasilkan keputusan ramalan diabetes tanpa perlu menunggu tempoh waktu yang lama. Seterusnya, applikasi ini dapat memudahkan proses menganalisis dan mengumpul data daripada orang ramai serta mempunyai pangkalan data yang mampu menyimpan data dan hasil keputusan dengan selamat dan sistematik. Kajian ini boleh dilanjutkan dengan mengguna pakai teknik pengelasan lain seperti *SVM, DT* dan termasuk perbandingan prestasi bagi setiap kaedah.

7.0 PENGHARGAAN

Penulis ingin mengucapkan ribuan terima kasih kepada semua pihak yang berkenaan dalam menyokong kerja-kerja penyelidikan ini terutamanya kepada Universiti Pertahanan Nasional Malaysia (UPNM) dan Higher College of Technology, Sharjah untuk penerbitan bersama ini.

Senarai Rujukan

[1] Mahmoud, M., & Mahmood, R. (2024). Differences in mental health status between individuals living with diabetes, and pre-diabetes in Qatar: A cross-sectional study. Heliyon, 10(1).

[2] Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., & Paul, B. K. (2021). Machine learning based diabetes prediction and development of smart web application. International Journal of Cognitive Computing in Engineering, 2, 229-241.

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my

[3] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292-299.
[4] Diksa, I. G. B. N., & Fithriasari, K. (2020). Analisis Faktor Resiko Penyebab Diabetes Mellitus dengan Regresi Logistik Biner. Inferensi, 4(1), 69-76.
[5] Vijayan, V. V., & Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 122-127). IEEE.
[6] Safitri, M., & Praba, A. D. (2024). PREDIKSI PENYAKIT DIABETES DENGAN MENGGUNAKAN ALGORITMA C4. 5. JIKA (Jurnal Informatika), 8(1), 74-81.
[7] Lonappan, A., Bindu, G., Thomas, V., Jacob, J., Rajasekaran, C., & Mathew, K. T. (2007). Diagnosis of diabetes mellitus using microwaves. Journal of Electromagnetic Waves and Applications, 21(10), 1393-1401.
[8] Vaishali, R., Sasikala, R., Ramasubbareddy, S., Remya, S., & Nalluri, S. (2017, October). Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset. In 2017 international conference on computing networking and informatics (ICCNI) (pp. 1-5). IEEE.
[9] dan Morbiditi, T. K. K. (2019). Penyakit tidak berjangkit, permintaan jagaan kesihatan dan literasi kesihatan. Diperoleh daripada: http://iku. moh. gov. my/images/IKU/Document/REPORT/NHMS2019/Fact_Sheet_NHMS_2019-BM. pdf.
[10] Tigga, N. P., & Garg, S. (2021). Predicting type 2 diabetes using logistic regression. In Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems: MCCS 2019 (pp. 491-500). Springer Singapore.
[11] Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC medical informatics and decision making, 19(1), 1-15.
[12] Resti, Y., Kresnawati, E. S., Dewi, N. R., & Eliyati, N. (2021). Diagnosis of diabetes mellitus in women of reproductive age using the prediction methods of naive bayes, discriminant analysis, and logistic regression. Science and Technology Indonesia, 6(2), 96-104.
[13] Zhou, Y. Y., Qiu, H. M., Yang, Y., & Han, Y. Y. (2020). Analysis of risk factors for carotid intima-media thickness in patients with type 2 diabetes mellitus in Western China assessed by logistic regression combined with a decision tree model. Diabetology & metabolic syndrome, 12, 1-13.
[14] Rajendra, P., & Latifi, S. (2021). Prediction of diabetes using logistic regression and ensemble techniques. Computer Methods and Programs in Biomedicine Update, 1, 100032.
[15] Abdollahi, J., & Nouri-Moghaddam, B. (2022). Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. Iran Journal of Computer Science, 5(3), 205-220.
[16] Palimkar, P., Shaw, R. N., & Ghosh, A. (2022). Machine learning technique to prognosis diabetes disease: Random forest classifier approach. In Advanced computing and intelligent technologies: proceedings of ICACIT 2021 (pp. 219-244). Springer Singapore.
[17] Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., ... & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. IEEE Access, 10, 8529-8538.
[18] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.
[19] Chan, N. K. W., Lee, A. S. H., & Zainol, Z. (2021, June). Predicting employee health risks using classification ensemble model. In 2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP) (pp. 52-58). IEEE.
[20] Dey, S. K., Hossain, A., & Rahman, M. M. (2018, December). Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm. In 2018 21st international conference of computer and information technology (ICCIT) (pp. 1-5). IEEE.
[21] Oh, R., Lee, H. K., Pak, Y. K., & Oh, M. S. (2022). An interactive online app for predicting diabetes via machine learning from environment-polluting chemical exposure data. International Journal of Environmental Research and Public Health, 19(10), 5800.
[22] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.
[23] Zainol, Z., Jaymes, M. T., & Nohuddin, P. N. (2018, May). Visualurtext: a text analytics tool for unstructured textual data. In Journal of Physics: Conference Series (Vol. 1018, No. 1, p. 012011). IOP Publishing.
[24] Dunham, M. H. (2006). Data mining: Introductory and advanced topics. Pearson Education India.

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my

[25] Nohuddin, P., Zainol, Z., Lee, A. S. H., Nordin, I., & Yusoff, Z. (2018). A case study in knowledge acquisition for logistic cargo distribution data mining framework. International Journal of Advanced and Applied Sciences, 5(1), 8-14.
[26] A. Rashid, R. A., Nohuddin, P. N., & Zainol, Z. (2017). Association rule mining using time series data for Malaysia climate variability prediction. In Advances in Visual Informatics: 5th International Visual Informatics Conference, IVIC 2017, Bangi, Malaysia, November 28–30, 2017, Proceedings 5 (pp. 120-130). Springer International Publishing.
[27] Nohuddin, P. N., Zainol, Z., & Hijazi, M. H. A. (2021). Study of B40 Schoolchildren Lifestyles and Academic Performance using Association Rule Mining. Annals of Emerging Technologies in Computing (AETiC), 5(5), 60-68.
[28] Mannila, H. (1996, June). Data mining: machine learning, statistics, and databases. In Proceedings of 8th International Conference on Scientific and Statistical Data Base Management (pp. 2-9). IEEE.
[29] Kalyankar, G. D., Poojara, S. R., & Dharwadkar, N. V. (2017, February). Predictive analysis of diabetic patient data using machine learning and Hadoop. In 2017 international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC) (pp. 619-624). IEEE.
[30] Wolberg, J. (2010). Designing quantitative experiments: prediction analysis. Springer Science & Business Media.
[31] Joshi, T. N., & Chawan, P. M. (2018). Logistic regression and svm based diabetes prediction system. International Journal For Technological Research In Engineering, 5, 4347-4350.
[32] Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), 217-222.
[33] Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29-39).
[34] Safitri, M., & Praba, A. D. (2024). PREDIKSI PENYAKIT DIABETES DENGAN MENGGUNAKAN ALGORITMA C4. 5. JIKA (Jurnal Informatika), 8(1), 74-81.

*Corresponding Author | Zainol, Z. | zuraini@upnm.edu.my