SISTEM PENCADANG PROGRAM LATIHAN DAN BIMBINGAN USAHAWAN INSKEN

NUR ATHIRAH BINTI RAMLAN[1]*

ZULAIHA ALI OTHMAN[2]

[1,2]*Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi,, Selangor Darul Ehsan, Malaysia*

Abstrak

Kajian ini bertajuk "Sistem Pencadang Latihan dan Bimbingan bagi Usahawan INSKEN" dengan objektif utama untuk membangunkan model latihan dan bimbingan usahawan berasaskan Teknik perlombongan data dan membangunkan sistem pencadang berasaskan model latihan dan bimbingan yang dibangunkan dalam objektif yang pertama. Melalui sistem ini, cadangan yang peribadi dan relevan akan disediakan kepada para usahawan berdasarkan ciri dan sejarah penglibatan mereka. Metodologi yang digunakan dalam kajian ini adalah CRISP-DM yang merupakan pendekatan berstruktur untuk perlombongan data yang berfokus kepada pemahaman, persediaan data, pemodelan dan penilaian. Dalam kajian ini, empat algoritma telah diaplikasikan ke dalam sistem: Naïve Bayes, Hutan Rawak, Pokok Keputusan dan Sokongan Mesin Vektor (SVM). Melalui analisis data sejarah penglibatan dan profil usahawan, empat algoritma telah memberikan hasil keputusan dan keberkesanan seperti berikut: Naïve Bayes (76%), Hutan Rawak (81%), Pokok Keputusan (72%) dan Sokongan Mesin Vektor (79%). Hasil keputusan ini memperlihatkan bahawa sistem cadangan yang dibangunkan dapat memberikan cadangan yang berkesan kepada usahawan dan membantu mereka memilih program latihan dan bimbingan yang sesuai. Secara kesuluruhannya, kajian ini telah Berjaya menghasilkan satu sistem pencadnag yang menggabungkan teknik perlomongan data dan algoritma terkini. Kajian ini merupakan langkah positif dalam membantu usahawan mencapai kejayaan dalam perniagaan mereka melalui program latihan dan bimbingan yang lebih tepat.

Kata kunci: [Perlombongan data, CRISP-DM, Sistem Pencadang]

Pengenalan

Pada era globalisasi ini, bilangan rakyat Malaysia yang menceburi bidang usahawan semakin meningkat seiring dengan peredaran pemodenan arus teknologi. Bidang perniagaan yang terdapat pelbagai cabang dan boleh dilakukan secara luar talian atau atas talian atau kedua-duanya sekali. Institut Keusahawanan Negara (INSKEN) mengambil berat dengan kadar peningkatan usahawan di Malaysia. Oleh yang demikian, mereka mengambil inisiatif bagi membantu para usahawan dalam menjayakan bidang perniagaan masing-masing dengan menyediakan perkhidmatan program latihan atau bimbingan bagi usahawan di Malaysia. Dengan peningkatan jumlah usahawan di Malaysia, bilangan permohonan bagi program INSKEN juga semakin meningkat. Oleh yang demikian, projek ini akan menghasilkan sebuah sistem pencadang yang boleh digunakan oleh pihak INSKEN dengan menggunakan Teknik perlombongan data.

Teknik perlombongan data sering digunakan untuk membuat data analisis oleh para ahli sains data di pelbagai sektor seperti dalam sektor pemasaran, pendidikan dan lain-lain (ALEXANDRA TWIN, 20 September 2020). Hasil daripada data analisis dapat membantu pengguna untuk membuat keputusan. Tetapi, hasil yang didapat melalui kaedah ini memerlukan pengetahuan pengatucaraan dan perlombongan data yang tinggi. Sistem yang dibina dapat memudahkan pengguna mendapatkan hasil dengan mudah.

Seiring dengan peningkatan jumlah usahawan di Malaysia pada masa kini, Institut Keusahawanan Negara (INSKEN) menerima jumlah permohonan daftar program latihan dan bimbingan yang tinggi bagi membantu para usahawan memajukan dan menaikkan bisnes mereka kepada satu tahap yang lebih tinggi. Penetapan program latihan atau bimbingan amat

penting dan memainkan peranan bagi usahawan menimba ilmu dan belajar dari INSKEN mengikut kriteria bisnes agar program yang diikuti oleh usahawan dapat memberikan kesan bagi meningkatkan bisnes para usahawan. Pihak INSKEN akan mengambil masa bagi menapis dan menentukan program latihan atau bimbingan yang sesuai mengikut kriteria bisnes bagi setiap permohonan oleh usahawan. Oleh yang demikian, INSKEN memerlukan satu sistem bagi membolehkan mereka untuk menetapkan usahawan tersebut perlu menjalani latihan atau bimbingan atau kedua-duanya yang sesuai bagi setiap permohonan daripada usahawan di Malaysia mengikut kriteria masing-masing.

Skop kajian ini adalah untuk meramal keputusan bagi program latihan dan bimbingan kepada usahawan mengikut kriteria. Jadi, untuk skop kajian ini hanya memfokuskan kepada kriteria usahawan. Sistem yang digunakan untuk melakukan perlombongan data dan menvisualisasikan keputusan-keputusan akan dibangunkan dengan menggunakan bahasa pengatucaraan Python.

Kajian "Sistem Pencadang Latihan dan Bimbingan bagi Usahawan INSKEN" sangat perlu dilakukan dan memiliki berbagai justifikasi dan kepentingan yang signifikan. Kajian "Sistem Pencadang Latihan dan Bimbingan bagi Usahawan INSKEN" memberikan manfaat yang signifikan dalam meningkatkan keberkesanan program, meningkatkan kepuasan peserta, meningkatkan efisiensi sumber daya, dan berkontribusi pada pengembangan bidang ilmu perlombongan data dan ekosistem keusahawanan. Diharapkan sistem ini akan menjadi aset berharga bagi INSKEN dan membantu mencapai misi mereka dalam membantu usahawan mencapai kejayaan dalam perniagaan mereka.

KAJIAN LEPAS SISTEM SOKONGAN KEPUTUSAN PERLOMBONGAN DATA

Kajian pertama oleh Lilly Sheeba (2021) yang bertujuan untuk mencadang acara untuk satu kumpulan individu. Satu sistem cadangan acara kumpulan berdasarkan teknik perlombongan data untuk pengkelasan telah dicadangkan dengan sistem sokongan keputusan. Pemprosesan data adalah proses yang digunakan oleh syarikat untuk mengubah maklumat menjadi maklumat yang berguna. Dengan menggunakan perisian untuk mencari corak dalam set data besar, syarikat boleh memahami lebih lanjut tentang pelanggan mereka untuk membangunkan strategi pemasaran yang lebih berkesan, meningkatkan jualan, dan mengurangkan kos. Pemprosesan data bergantung pada pengumpulan data yang berkesan, penyimpanan, dan pemprosesan komputer. Projek ini bertujuan untuk mencadang acara untuk satu kumpulan individu berdasarkan ciri pelanggan. Satu algoritma yang pantas yang dikenali sebagai algoritma pengkelasan teorem telah dibangunkan untuk mempelajari model pengkelasan untuk setiap kumpulan.

Kajian kedua oleh Monali Dey (2018) yang menyatakan teknologi perlombongan data menyediakan pendekatan berorientasikan pengguna untuk mendapatkan maklumat baru dan tersembunyi dalam data dan pengetahuan berharga dapat ditemui melalui aplikasi teknik perlombongan data dalam sistem kesihatan. Perlombongan data dalam bidang perubatan kesihatan membincangkan model pembelajaran untuk meramalkan penyakit pesakit.Monali Dey menjalankankan kajian dan analisis keunikan perlombongan data perubatan dan memberi gambaran tentang Sistem Sokongan Keputusan Kesihatan yang digunakan dalam bidang perubatan.

Sejumlah besar maklumat pelancongan disediakan kepada pengguna, tetapi pada masa yang sama, ia telah menyebabkan rambang mata kepada pengguna. Data yang besar telah mengatasi maklumat yang benar-benar diminati oleh pengguna. Dalam kajian yang ketiga ini oleh Zehao

Wang (2019), membuat satu sistem sokongan keputusan pelancongan membantu pengguna menyelesaikan masalah ini. Berdasarkan latar belakang di atas, kajian ketiga ini merancang dan melaksanakan sistem sokongan keputusan pelancongan berdasarkan teknik perlombongan data. Dari perspektif menambang persamaan antara pengguna, persamaan antara pengguna dikira melalui algoritma dan kemudian tarikan yang dikunjungi oleh pengguna dengan persamaan yang lebih tinggi disyorkan.

KAJIAN LEPAS PROGRAM LATIHAN DAN BIMBINGAN

Kajian ini secara keseluruhannya melibatkan dan berkait dengan program latihan dan bimbingan. Beberapa kajian mengenai program latihan dan bimbingan ini telah dilakukan. Antara kajian yang telah dilakukan ialah Ludavico Buratto (2022) bertajuk Cadagan Pengguna berdasarkan Prestasi yang Adil dalam Sistem e-Coaching. Dalam kajian ini, Ludavico (2022) memfokuskan pada scenario terakhir dengan mempertimbangkan platform e-Coaching untuk pelari. Kajian ini untuk menyediakan jurulatih dengan senarai pengguna yang diberi mengikut sokongan yang mereka perlukan. Lebih-lebih lagi, penjaminan pendedahan yang adil dalam peringkat, untuk memastikan bahawa pelari dari pelbagai kumpulan mempunyai peluang yang sama untuk disokong. Untuk melakukannya, kajian oleh Ludavico (2022) memodelkan prestasi dan tingkah laku berjalan mereka dan kemudian mempersembahkan algoritma untuk mengesyorkan pengguna kepada pelatih mengikut prestasi pelari pada sesi terakhir dan kualiti yang sebelumnya. Kajian ini telah menggunakan klasifikasi dan algoritma Pokok Keputusan (Decision Tree) dan Hutan Rawak (Random Forest) dengan ketepatan 83.34% dan 85.89% bagi menentukan ciri-ciri pengguna.

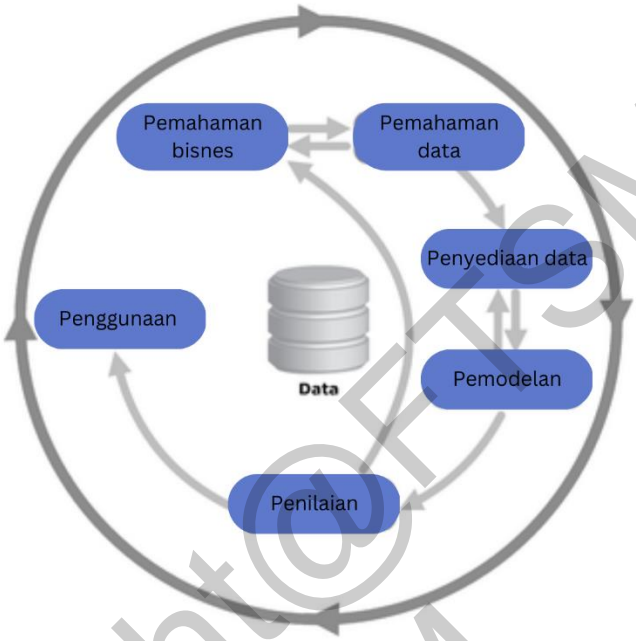
Seterusnya, kajian yang telah dilakukan oleh David Estrela (2017) iaitu Sistem Pencadang untuk Kursus Atas Talian. Kajian tersebut menyatakan pada masa kini terdapat banyak kursus yang tersedia untuk pelajar, dan beberapa kali sukar bagi pelajar untuk melihat maklumat yang berkaitan dengan kursus tersebut dengan memutuskan kursus mana yang harus

diambil. Kajian David Estrela (2017) bertujuan untuk membina sistem pencadang kursus atas talian kepada pengguna berdasarkan profil pengguna dan persamaan dengan pengguna yang lain. Tiga teknik digunakan bagi mengekstrak maklumat dan mencadangkan kursus atas talian iaitu Pokok Keputusan, Nearest Neighbour dan Sokongan Mesin Vektor (SVM). Dengan menggunakan tiga teknik tersebut, didapati bahawa Sokongan Mesin Vektor (SVM) sistem dapat memberikan cadangan yang lebih tepat dengan mempertimbangkan kepentingan setiap pengguna dengan ketepatan 88.82% manakala ketepatan bagi Pokok Keputusan ialah 85.25% dan Nearest Neighbour ialah 83.21%.

Akhir sekali, kajian dilakukan oleh Wen-Shung Tai (2008) yang bertajuk Sistem Pencadang e-Pembelajaran yang berkesan. Dua jenis algoritma yang digunakan untuk klasifikasi jenis e-pelajar.Weng Shung-Tai (2008) menggunakan dua model iaitu Decision Tree J48 dan Naïve Bayes dengan ketepatan 95.11% dan 87.26%. Berdasarkan kumpulan e - Learner ini, pengguna dapat memperoleh cadangan kursus dari pendapat kumpulan. Apabila kumpulan kepentingan yang berkaitan telah dibentuk, DM akan digunakan untuk mendapatkan peraturan jalan pembelajaran terbaik.

Metodologi yang akan digunakan dalam kajian ini ialah *Cross Industry Standard Process for Data Mining (CRISP-DM).* Metodologi ini adalah satu rangka kerja yang secara meluas digunakan untuk membimbing projek perlombongan data dan pembelajaran mesin. Ia merangkumi pendekatan menyeluruh,berulang dan terstruktur ke seluruh kitaran hidup perlombongan data, dari perancangan projek sehingga penggunaan. Metodologi CRISP-DM terdiri daripada enam fasa utama iaitu Pemahaman Perniagaan, Pemahaman Data, Penyediaan Data, Pemodelan, Penilaian, dan Penggunaan. Sepanjang fasa-fasa ini, penganalisis data dan pihak berkepentingan bekerjasama untuk menentukan objektif perniagaan, memahami sumber data, memproses dan mengubah data, memilih dan membina model, menilai prestasi, dan

akhirnya menyampaikan model ke dalam sistem operasi. CRISP-DM menyediakan proses sistematik dan berulang yang meningkatkan ketelusan projek, mengurangkan risiko, dan memastikan kejayaan usaha perlombongan data.



Rajah 1

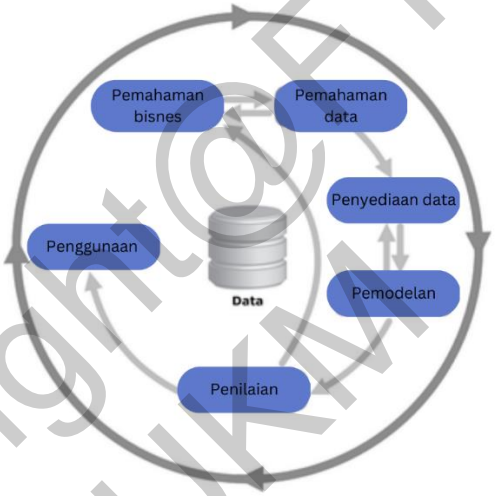
Laporan ini terdiri dari empat bahagian utama yang akan membahas sistem pencadang latihan dan bimbingan bagi usahawan INSKEN berdasarkan teknik perlombongan data. Pendahuluan akan menyediakan gambaran keseluruhan tentang projek ini, termasuk objektif laporan dan justifikasi mengapa projek ini perlu dilakukan. Pendahuluan juga akan memperkenalkan Institut Keusahawanan Negara (INSKEN) dan konteks di mana sistem pencadang ini akan diterapkan. Kemudian, bahagian metodologi akan menjelaskan pendekatan yang digunakan dalam projek ini, iaitu CRISP-DM (Cross-Industry Standard Process for Data Mining). Metodologi akan merangkumi langkah-langkah yang diambil dalam pemilihan dan pembersihan data, pemodelan, dan penilaian hasil algoritma yang diterapkan. Setelah itu, bahagian keputusan dan perbincangan akan menyajikan hasil dari implementasi algoritma-

algoritma seperti Naive Bayes, Random Forest, Decision Tree, dan Support Vector Machine (SVM). Ini akan mencakup skor ketepatan masing-masing algoritma dalam memberikan cadangan, serta analisis tentang bagaimana algoritma-algoritma tersebut berfungsi dalam mengenal pasti program latihan dan bimbingan yang sesuai.

Secara keseluruhan, laporan ini akan memberikan gambaran menyeluruh tentang projek "Sistem Pencadang Latihan dan Bimbingan bagi Usahawan INSKEN," mulai dari pendahuluan dan tujuan, metodologi yang digunakan, hasil dan perbincangan, hingga kesimpulan dan implikasi yang relevan. Struktur ini akan membantu pembaca memahami secara komprehensif bagaimana teknik perlombongan data telah digunakan untuk meningkatkan proses pemilihan program latihan dan bimbingan bagi usahawan INSKEN.

Metodologi Kajian

Objektif kajian ini adalah untuk membangunkan model latihan dan bimbingan usahawan berasaskan teknik perlombongan data. Di sini akan diperincikan metodologi yang digunakan dalam kajian ini iaitu Cross Industry Process for Data Mining (CRISP-DM) dan akan dipadankan mengikut set data kajian yang ingin dikaji. CRISP-DM adalah suatu pendekatan sistematik dan terstruktur untuk mengelola projek analisis data dan perlombongan data. Perlombongan data adalah salah satu cabang kepintaran buatan yang bertujuan untuk mendapatkan pengetahuan tertentu daripada maklumat yang ada.



Rajah 2

Metodologi CRISP-DM terbahagi kepada enam fasa iaitu fasa pemahaman bisnes, fasa pemahaman data, fasa penyediaan data, fasa pemodelan, fasa penilaian dan fasa pelaksanaan. CRISP-DM ini memastikan pengurangan risiko dan membantu memastikan matlamat pembangunan modul dikekalkan sepanjang fasa pembangunan modul dijalankan.

Fasa Pemahaman Bisnes

Fasa Pemahaman Bisnes merupakan proses pertama dalam memahami matlamat dan objektif kajian ini dilakukan. Data yang diperoleh adalah menggunakan set data yang diperolehi daripada Institut Keusahawanan Negara (INSKEN) dimana pihak INSKEN memberikan senarai

data peserta latihan dan bimbingan yang terdahulu yang mengandungi pelbagai atribut. Set data bagi senarai latihan dan bimbingan yang akan digunakan sebagai hasil sistem pencadang juga diberikan dan mempunyai jumlah latihan dan bimbingan sebanyak 58 program.

Fasa Pemahaman Data

Sumber data diperoleh daripada INSKEN mempunyai 4 fail data yang digunakan dalam kajian ini iaitu Maklumat Usahawan, Penglibatan Program, Senarai Usahawan dan Senarai Latihan Bimbingan. Data-data ini akan dianalisis terlebih dahulu sebelum beralih ke fasa yang seterusnya. Bagi set data mentah fail Maklumat Usahawan mempunyai 12,230 data serta 33 atribut yang menyatakan maklumat usahawan yang berdaftar dengan INSKEN. Fail data Pengilbatan Program menunjukkan senarai nama usahawan bersama sejarah program latihan dan bimbingan INSKEN yang pernah diikuti. Fail Penglibatan Program merangkumi 5 atribut bersama 19,119 jumlah data. Fail Senarai Usahawan mengandungi 28 atribut bersama 978 data. Fail data senarai latihan dan bimbingan mempunyai 3 atribut berserta 50 data.

Fasa Penyediaan Data

Fasa Penyediaan Data adalah fasa ketiga yang melibatkan pembersihan data sebelum diproses dan dianalisis. Sebelum menghasilkan set data akhir bagi kajian ini, atribut di dalam kesemua fail data mentah akan dilihat dan dikelaskan serta ditukarkan kepada atribut jenis numerik jika perlu bagi memudahkan proses penyediaan data sebelum ke fasa pemodelan. Atribut seperti ID dan Nombor IC tidak akan diubah kerana atribut tersebut mempunyai satu sahaja bagi setiap data dan hanya akan digunakan untuk menghubungkan setiap fail data mentah.

Seterusnya, satu fail set data ‘INSKEN Dataset’ diwujudkan bagi menggabungkan kesemua atribut dari semua fail data mentah. Dalam fail set data INSKEN Dataset, atribut Nama Program telah dipecahkan kepada 51 program itu sendiri. Ini bagi melihat data tentang usahawan tersebut

menghadiri setiap program tersebut. Berikut merupakan senarai sifat dengan jenis data serta butiran atribut serta atribut yang telah ditukarkan kepada numerik bagi set data INSKEN Dataset yang mengandungi 599 data dan 62 atribut.

i. Pembersihan Data

Pembersihan data adalah proses yang digunakan untuk mengenal pasti data yang tidak tepat, tidak lengkap, atau tidak munasabah, dan kemudian meningkatkan kualitinya melalui pembetulan kesalahan dan kekurangan yang dikesan (Juyiong Li, 2017). Berdasarkan pemerhatian terhadap sumber data, fail set data INSKEN Dataset merupakan fail akhir yang akan digunakan bagi prapemprosesan data. Fail data tersebut tidak mempunyai kehilangan data (missing value) dan set data telah bersturuktur maka pembersihan set data INSKEN Dataset telah selesai.

ii. Pengintegrasian Data

Data integrasi merupakan proses menggabungkan dan menggabungkan data dari pelbagai sumber. Matlamat data integrasi adalah untuk mencipta satu set data yang komprehensif dan boleh dipercayai yang membolehkan organisasi mendapatkan wawasan yang berharga, membuat keputusan yang berinformasi, dan melakukan analisis yang bermakna. Dengan menyelaraskan data dari pelbagai sumber, data integrasi memfasilitasi kerjasama yang lebih baik, mengurangkan redundansi, meningkatkan ketepatan data, dan meningkatkan kecekapan keseluruhan di seluruh organisasi, membawa kepada operasi yang lebih berkesan dan kelebihan bersaing di dunia yang dipacu oleh data pada masa kini.

a. Pewujudan label baru Latihan

Atribut ini diwujudkan dengan menggabungkan atribut nama program KAK, KAKSe, INBT dan IBT1 sehingga IBT41. Jenis atribut Latihan adalah numerik di mana mempunyai nilai 0 (Tidak

Hadir) dan 1 (Hadir). Tidak hadir di dalam konteks ini bermaksud usahawan tersebut tidak pernah menghadiri semua jenis latihan yang dianjurkan oleh INSKEN. Manakala, nilai 1 pula bermaksud usahawan tersebut pernah menghadiri satu atau lebih program latihan yang telah INSKEN anjurkan.

b. Pewujudan label baru Bimbingan

Atribut ini diwujudkan dengan menggabungkan atribut nama program PREMIUM, CIP, ME, BizClinic, APDP, E-Board, IBBC. . Jenis atribut Bimbingan adalah numerik di mana mempunyai nilai 0 (Tidak Hadir) dan 1 (Hadir). Tidak hadir di dalam konteks ini bermaksud usahawan tersebut tidak pernah menghadiri semua jenis program bimbingan yang dianjurkan oleh INSKEN. Manakala, nilai 1 pula bermaksud usahawan tersebut pernah menghadiri satu atau lebih program bimbingan yang telah INSKEN anjurkan.

c. Pewujudan label baru Program

Atribut ini diwujudkan dengan melihat atribut Latihan dan atribut Bimbingan. Data ini diperlihat satu demi satu untuk menentukan bahawa usahawan tersebut menghadiri Latihan sahaja atau Bimbingan sahaja atau menghadiri kedua-dua Latihan dan Bimbingan. Label program adalah numerik di mana mempunyai nilai 0 (Latihan), 1 (Bimbingan) dan 2 (Latihan & Bimbingan). Kewujudan atribut ini merupakan target keluaran (output) pada sistem cadangan yang akan dibangunkan dalam fasa seterusnya.

d. Pengurangan atribut

Pengurangan atribut ini melibatkan dua atribut dalam fail INSKEN Dataset iaitu atribut ID dan Nombor IC. Merujuk set data, atribut ID dan Nombor IC hanya digunakan bagi menghubungkan kesemua fail set data bagi menghasilkan sebuah fail set data akhir iaitu INSKEN Dataset. Atribut ID dan Nombor IC mempunyai nilai satu sahaja bagi semua data dan tidak memberikan makna dan

kesan kepada kajian ini. Oleh yang demikian, kedua-dua atribut ID dan Nombor IC akan dikeluarkan dan tidak akan digunakan bagi proses seterusnya.

Fasa Pemodelan

Pengujian Pertama Model

Pengujian pertama dilakukan dengan menggunakan set data INSKEN Dataset yang mengandungi 10 atribut dan 598 data. Pengujian dilakukan menggunakan empat model seperti yang dinyatakan di atas iaitu *Naïve Bayes,* Pokok Keputusan, Penyokong Mesin Vektor (SVM) dan Hutan Rawak. Hasil keputusan analisis bagi setiap model adalah seperti di jadual bawah.

Algoritma	**Ketepatan**	**Min Ralat Mutlak (MAE)**	**Min Ralat Kuasa Dua (MSE)**
Naïve Bayes	0.76	0.33	0.52
Pokok Keputusan	0.70	0.48	0.82
SVM	0.78	0.32	0.52
Random Forest	0.79	0.32	0.54

Jadual 1

Berdasarkan perbandingan analisis di atas, pengujian kedua dilakukan bagi meningkatkan ketepatan model dengan menggunakan set data INSKEN Dataset yang telah dikurangkan atribut. Pengurangan atribut dilakukan berdasarkan kod *feature_importances_* seperti yang ditunjukkan dalam rajah di bawah.

```
from sklearn.ensemble import ExtraTreesClassifier
# feature extraction
model = ExtraTreesClassifier(n_estimators=100)
model.fit(X, Y)
print(model.feature_importances_)
```

```
[0.073 0.167 0.074 0.052 0.437 0.068 0.025 0.074 0.031]
```

Rajah 3

Jadual 2 menunjukkan ringkasan nilai kepentingan ciri bagi setiap atribut di dalam INSKEN Dataset. Dua atribut yang mempunyai nilai terendah iaitu Jenis Perniagaan dan Bisnes Sektor telah dikeluarkan bagi membuat pengujian kedua model dan penilaian. Jangkaan bagi pengujian kedua model adalah model akan berfungsi dengan lebih baik dengan menunjukkan ketepatan yang lebih tinggi daripada pengujian pertama.

Atribut	**Nilai Kepentingan**
Purata Pendapatan Isi Rumah	0.073
Negeri Bisnes	0.167
Tahun Beroperasi	0.074
Jumlah Pekerja	0.052
Pekerjaan	0.437
Kedudukan	0.068
Jenis Perniagaan	0.025
Purata Pendapatan Tahunan	0.074
Bisnes Sektor	0.031

Jadual 2

Pengujian Kedua Model

Pengujian kedua model dilakukan dengan kaedah yang sama seperti pengujian pertama. Perbezaan di sini adalah nilai atribut yang terdapat di dalam set data INSKEN Dataset. Berikut merupakan rajah-rajah yang menunjukkan kod *Python* bagi setiap model yang dijalankan.

```python
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, mean_absolute_error, mean_squared_error, r2_score
from numpy import sqrt



# Create a Naive Bayes model
nb_model = GaussianNB()

# Perform 10-fold cross-validation on the model
cv_predictions = cross_val_predict(nb_model, X, Y, cv=10)

# Calculate the average accuracy from the cross-validation scores
average_accuracy = cross_val_score(nb_model, X, Y, cv=10, scoring='accuracy').mean()

# Calculate MAE, MSE, and R-squared using cross_val_predict
mae = mean_absolute_error(Y, cv_predictions)
mse = mean_squared_error(Y, cv_predictions)
rmse = sqrt(mse)
r2_value = r2_score(Y, cv_predictions)

# Print the results
print("Average Accuracy: {:.2f}%".format(average_accuracy * 100))
print('Mean Absolute Error of the model is: {}'.format(mae))
print('Mean Squared Error of the model is: {}'.format(mse))
print('Root Mean Squared Error of the model is: {}'.format(rmse))
print('R-squared value of the model is: {}'.format(r2_value))
```

Rajah 4

```python
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, mean_absolute_error, mean_squared_error, r2_score
from numpy import sqrt



# Create a Decision Tree model
dt_model = DecisionTreeClassifier()

# Perform 10-fold cross-validation on the model
cv_scores = cross_val_score(dt_model, X, Y, cv=10, scoring='accuracy')

# Calculate the average accuracy and other metrics from the cross-validation scores
average_accuracy = cv_scores.mean()
mae = round(cv_scores.mean(), 3)
mse = round(cv_scores.mean(), 3)
rmse = round(sqrt(mse), 3)
r2_value = round(cv_scores.mean(), 3)

print("Average Accuracy: {:.2f}%".format(average_accuracy * 100))
print('Mean Absolute Error  of the model is : {}'.format(mae))
print('Mean Squared Error of the model is : {}'.format(mse))
print('Root Mean Squared Error of the model is : {}'.format(rmse))
print('R-squared value of the model is : {}'.format(r2_value))
```

```
Average Accuracy: 71.96%
Mean Absolute Error  of the model is : 0.72
Mean Squared Error of the model is : 0.72
Root Mean Squared Error of the model is : 0.849
R-squared value of the model is : 0.72
```

Rajah 5

```
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, mean_absolute_error, mean_squared_error, r2_score
from numpy import sqrt


# Create an SVM model
svm_model = SVC()

# Perform 10-fold cross-validation on the model
cv_scores = cross_val_score(svm_model, X, Y, cv=10, scoring='accuracy')

# Calculate the average accuracy and other metrics from the cross-validation scores
average_accuracy = cv_scores.mean()
mae = round(cv_scores.mean(), 3)
mse = round(cv_scores.mean(), 3)
rmse = round(sqrt(mse), 3)
r2_value = round(cv_scores.mean(), 3)

print("Average Accuracy: {:.2f}%".format(average_accuracy * 100))
print('Mean Absolute Error  of the model is : {}'.format(mae))
print('Mean Squared Error of the model is : {}'.format(mse))
print('Root Mean Squared Error of the model is : {}'.format(rmse))
print('R-squared value of the model is : {}'.format(r2_value))
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A
  y = column_or_1d(y, warn=True)
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A
  y = column_or_1d(y, warn=True)
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A
  y = column_or_1d(y, warn=True)
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A
  y = column_or_1d(y, warn=True)
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A
  y = column_or_1d(y, warn=True)
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A
  y = column_or_1d(y, warn=True)
Average Accuracy: 78.97%
Mean Absolute Error  of the model is : 0.79
Mean Squared Error of the model is : 0.79
Root Mean Squared Error of the model is : 0.889
R-squared value of the model is : 0.79
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A
  y = column_or_1d(y, warn=True)
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A
  y = column_or_1d(y, warn=True)
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A
  y = column_or_1d(y, warn=True)
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:1143: DataConversionWarning: A
  y = column_or_1d(y, warn=True)
```

Rajah 6

```
from sklearn.model_selection import train_test_split, cross_val_score, cross_val_predict
from sklearn.metrics import accuracy_score, mean_absolute_error, mean_squared_error, r2_score, classification_report
from numpy import sqrt


# Create a Random Forest model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)

# Perform 10-fold cross-validation on the model
cv_predictions = cross_val_predict(rf_model, X, Y, cv=10)

# Calculate the average accuracy from the cross-validation scores
average_accuracy = cross_val_score(rf_model, X, Y, cv=10, scoring='accuracy').mean()

# Calculate MAE, MSE, and R-squared using cross_val_predict
mae = mean_absolute_error(Y, cv_predictions)
mse = mean_squared_error(Y, cv_predictions)
rmse = sqrt(mse)
r2_value = r2_score(Y, cv_predictions)

print("Average Accuracy: {:.2f}%".format(average_accuracy * 100))
print('Mean Absolute Error of the model is: {}'.format(mae))
print('Mean Squared Error of the model is: {}'.format(mse))
print('Root Mean Squared Error of the model is: {}'.format(rmse))
print('R-squared value of the model is: {}'.format(r2_value))
```

Rajah 7

Berdasarkan rajah-rajah di atas, bagi setiap model, data dibahagikan kepada set latihan dan set ujian menggunakan train_test_split. Fungsi *cross_val_score* digunakan dalam setiap model untuk menjalankan 10-*fold cross-validation*. Ia kemudian mengambil jenis model, ciri-ciri *input* (X), dan pemboleh ubah sasaran yang berkaitan (Y). Argumen cv=10 menentukan bilangan lipatan untuk pengesahan silang. Argumen *scoring='accuracy'* menunjukkan bahawa kita ingin menilai ketepatan model. Selepas pengesahan silang (cross-validation), kod ini mengira purata ketepatan, min ralat mutlak (MAE), min ralat kuasa (MSE), dan nilai R-kuasa. Metrik-metrik ini dinilai berdasarkan 10 lipatan pengesahan silang. Model ini dilatih dengan data latihan dan kemudian dinilai menggunakan data ujian untuk mengira ketepatan, min ralat mutlak, min ralat kuasa dan nilai R-kuasa. Laporan klasifikasi dicetak untuk semua model yang merangkumi ketepatan (precision), pengecaman semula (recall), dan skor F1 bagi setiap kelas.

Keputusan dan Perbincangan

Hasil daripada pengujian kedua model, model yang terkini menunjukkan peningkatan ketepatan berbanding dengan pengujian pertama setelah mengurangkan atribut. Penilaian setiap model akan diperincikan dalam subtopik seterusnya. Berikut merupakan perbandingan keputusan analisis bagi kesemua model yang digunakan dalam kajian ini setelah pengujian kedua dilakukan.

Algoritma	**Ketepatan**	**Min Ralat Mutlak (MAE)**	**Min Ralat Kuasa Dua (MSE)**
Naïve Bayes	0.75	0.34	0.53
Pokok Penetuan	0.72	0.43	0.74
SVM	0.79	0.31	0.51
Hutan Rawak	0.81	0.29	0.49

Jadual 2

Hasil akhir daripada Model Random Forest adalah keputusan kolektif atau ramalan purata semua pohon keputusan individu, menghasilkan ramalan yang lebih kukuh dan tepat berbanding sebarang pohon keputusan tunggal. Kod yang diberikan membuat iterasi ke atas semua pohon keputusan (*estimators_)* dalam model Hutan Rawak. Untuk setiap pokok keputusan, sebanyak 12 pokok keputusan pokok keputusan telah dikeluarkan bersama peraturan pada setiap pokok. Peraturan tersebut mewakili proses pembuatan keputusan bagi pokok keputusan yang tersebut. Rajah di bawah menunjukkan contoh hasil keluaran kod pengeluaran model yang telah dijalankan.

```
Decision Tree 100 Rules:
|--- Purata Pendapatan Tahunan <= 1.50
|   |--- Negeri Bisnes <= 2.50
|   |   |--- Purata Pendapatan Isi Rumah <= 0.50
|   |   |   |--- Negeri Bisnes <= 1.50
|   |   |   |   |--- Pekerjaan <= 0.50
|   |   |   |   |   |--- Kedudukan <= 1.50
|   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |--- Kedudukan >  1.50
|   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |--- Pekerjaan >  0.50
|   |   |   |   |   |--- Kedudukan <= 1.50
|   |   |   |   |   |   |--- Negeri Bisnes <= 0.50
|   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |--- Negeri Bisnes >  0.50
|   |   |   |   |   |   |   |--- Purata Pendapatan Tahunan <= 0.50
|   |   |   |   |   |   |   |   |--- Jumlah Pekerja <= 0.50
|   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |   |   |--- Jumlah Pekerja >  0.50
|   |   |   |   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |   |   |   |--- Purata Pendapatan Tahunan >  0.50
|   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |--- Kedudukan >  1.50
|   |   |   |   |   |   |--- Negeri Bisnes <= 0.50
|   |   |   |   |   |   |   |--- Jumlah Pekerja <= 0.50
|   |   |   |   |   |   |   |   |--- Purata Pendapatan Tahunan <= 0.50
|   |   |   |   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |   |   |   |   |--- Purata Pendapatan Tahunan >  0.50
|   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |   |--- Jumlah Pekerja >  0.50
|   |   |   |   |   |   |   |   |--- Purata Pendapatan Tahunan <= 0.50
|   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |   |   |--- Purata Pendapatan Tahunan >  0.50
|   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |--- Negeri Bisnes >  0.50
|   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |--- Negeri Bisnes >  1.50
|   |   |   |   |   |--- Kedudukan <= 0.50
|   |   |   |   |   |   |--- Tahun Beroperasi <= 5.50
|   |   |   |   |   |   |   |--- Jumlah Pekerja <= 1.50
|   |   |   |   |   |   |   |   |--- Tahun Beroperasi <= 3.50
|   |   |   |   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |   |   |   |   |--- Tahun Beroperasi >  3.50
|   |   |   |   |   |   |   |   |   |--- Tahun Beroperasi <= 4.50
|   |   |   |   |   |   |   |   |   |   |--- Jumlah Pekerja <= 0.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |   |   |   |   |   |   |--- Jumlah Pekerja >  0.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |   |   |   |   |--- Tahun Beroperasi >  4.50
|   |   |   |   |   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |   |   |   |--- Jumlah Pekerja >  1.50
|   |   |   |   |   |   |   |   |--- Purata Pendapatan Tahunan <= 0.50
|   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |   |   |--- Purata Pendapatan Tahunan >  0.50
|   |   |   |   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |   |   |--- Tahun Beroperasi >  5.50
|   |   |   |   |   |   |   |--- class: 0.0
```

Rajah 8

Kesimpulan

Projek kajian ini adalah menggunakan algoritma-algoritma pembelajaran mesin untuk menghasilkan sistem cadangan bagi program latihan dan bimbingan INSKEN kepada usahawan yang berdaftar. Dua objektif kajian yang disebut pada bab awal, ingin membangunkan model latihan dan bimbingan usahawan berasaskan teknik perlombongan data telah berjaya dicapai. Dalam erti kata, pernyataan masalah bisnes kajian ini juga dapat diselesaikan. Set data yang mengandungi sejarah penglibatan usahawan dalam program latihan dan bimbingan didapati daripada pihak Institusi Keusahawanan Negara (INSKEN) telah melalui pemprosesan data pada fasa penyediaan data untuk menjamin kualiti data supaya tidak mempengaruhi keputusan ramalan.

Kekangan yang dihadapi dalam sepanjang projek ini dijalankan adalah pada waktu fasa penyediaan data. Hal yang demikian kerana set data mentah yang diperolehi dari pihak INSKEN terlalu banyak juga tidak berada dalam keadaan yang tersusun dan terpaksa melalui pembersihan data berulang kali serta beberapa ciri perlu dikeluarkan. Cabaran dilalui dalam tempoh yang panjang bagi menghasilkan satu set fail data yang cantik dan tersusun bagi ke fasa seterusnya.

CADANGAN PENAMBAHBAIKAN KAJIAN

Cadangan penambahbaikan kajian ini adalah dengan melihat semula semasa fasa penyediaan data iaitu ketika atribut menjalani fasa pengelasan. Pemilihan ciri-ciri usahawan dan bisnes juga perlu disemak semula bagi menghasilkan satu sistem yang boleh menampung pelbagai ciri dan memberikan kesan serta impak kepada usahawan INSKEN. Selain daripada itu, sistem ini hanya mencadangkan usahawan tersebut perlu menjalani latihan sahaja atau bimbingan sahaja atau kedua-duanya. Penambahbaikan yang boleh dilakukan pada sistem ini adalah mungkin boleh mencadangkan program latihan dan bimbingan secara terpeinci mengikut jenis

program latihan dan bimbingan yang disediakan oleh INSKEN. Secara kesuluruhannya, penambahbaikan boleh dilakukan kepada sistem ini supaya dapat menjadi lebih efisien dan berguna kepada pihak berkepentingan.

KEPUTUSAN

Kesimpulannya, bab ini telah menerangkan keseluruhan projek secara ringkas dan kekangan yang dihadapi sepanjang kajian sistem pencadang ini dijalankan. Cadangan penambahbaikan model dan sistem juga dinyatakan bagi masa akan datang. Projek ini telah klasifikasikan 7 kategori ciri usahawan dan satu sistem dibangunkan untuk melihat cadangan program latihan dan bimbingan bagi usahawan INSKEN. Hasil keputusan boleh dilihat pada satu pembangunan sistem yang dibangunkan.

Penghargaan

Alhamdulillah, syukur ke hadrat Ilahi kerana dengan izinNya saya dapat menyiapkan laporan projek tahun akhir ini. Setinggi-tinggi penghargaan saya merakamkan kepada penyeliasa saya iaitu Prof. Madya Dr Zulaiha Ali Othman diatas segala tunjuk ajar dan bimbingan yang diberikan kepada saya sepanjang saya menyiapkan kajian ini. Segala jasa, ilmu dan tunjuk ajar beliau saya akan kenang kerana ianya menjadi salah satu asbab dalam kejayaan projek saya.

Terima kasih saya ucapkan kepada ahli keluarga saya terutama ibu dan bapa, Ramlan Ibrahim dan Sarina Nong yang sentiasa menjadi tulang belakang dan menyokong dari segi doa, fizikal, mental dan kewangan. Segala jasa ahli keluarga saya yang sentiasa bersabar melihat air mata saya sepanjang menyiapkan projek ini akan dikenang. Jasa kalian tidak akan dilupakan dan akan sentiasa menjadi pembakar semangat dalam perjuangan kehidupan di university mahupun perjuangan alam dewasa setelah tamat di universiti.

Terima kasih saya ucapkan kepada rakan-rakan saya yang sangat banyak membantu saya dari segi tunjuk ajar dan sokongan yang sentiasa diberikan kepada saya. Segala pengorbanan kalian dari segi masa dan tenaga tidak akan dilupakan dan semoga kebaikan tersebut menjadi asbab untuk segala urusan kalian dipermudahkan.

Seterusnya, terima kasih saya kepada semua kucing-kucing peliharaan saya kerasa sentiasa menghiburkan saya tatkala saya sedang bertungkus-lumus dalam menyiapkan tugasan.

Saya juga berterima kasih kepada diri saya sendiri, Nur Athirah Ramlan kerana tidak berputus asa untuk menyiapkan projek tahun akhir ini. Anda sudah buat yang terbaik, Athirah. Terima kasih untuk tidak berhenti di tengah jalan.

Akhir kata, saya berterima kasih kepada semua pihak yang terlibat secara langsung atau tidak langsung dalam menghasilkan laporan projek tahun akhir saya ini. Saya juga ingin memohon maaf kerana terdapat kekurangan sepanjang menjalankan projek tahun akhir ini.

Sekian, terima kasih

RUJUKAN

Boratto, L., Carta, S., Iguider, W., Mulas, F., & Pilloni, P. (2022, August 5). *Fair performance-based user recommendation in eCoaching systems - user modeling and user-adapted interaction*. SpringerLink. Retrieved Dicember 29, 2022, from https://link.springer.com/article/10.1007/s11257-022-09339-6

Brand, M., Fast online svd revisions for lightweight recommender systems. In *SIAM International Conference on Data Mining (SDM)*, 2003.

Burke, R., Hybrid web recommender systems. pages 377–408. 2007.

Chimieski, B. F., & Fagundes, R. D. R. (n.d.). *Association and Classification Data Mining algorithms comparison over medical datasets*. Journal of Health Informatics. https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/226

Classification and prediction based data mining ... - IEEE xplore. (n.d.). https://ieeexplore.ieee.org/abstract/document/8365371/

Data Mining System and Applications: A review - researchgate. (n.d.-b). https://www.researchgate.net/profile/V-M-Thakare/publication/264841908_Data_Mining_System_and_Applications_A_Review/links/57404bf408ae9f741b32d72a/Data-Mining-System-and-Applications-A-Review.pdf

Data Mining Techniques - Javatpoint. www.javatpoint.com. (n.d.). Retrieved November 17, 2022, from https://www.javatpoint.com/data-mining-techniques

Dwivedi, R. (n.d.). *What are recommendation systems in machine learning?* Analytics Steps. Retrieved December 22, 2023, from https://www.analyticssteps.com/blogs/what-are-recommendation-systems-machine-learning

Estrela, D. (n.d.). *(PDF) a recommendation system for online courses - researchgate*. Retrieved December 29, 2022, from https://www.researchgate.net/publication/31588045_A_Recommendation_System_for_Online_Courses

Lilly Sheeba S, et. al. (n.d.). *Group event recommendations framework based on Data Mining*. Turkish Journal of Computer and Mathematics Education (TURCOMAT). https://turcomat.org/index.php/turkbilmat/article/view/832

Najafabadi, M. K., Mohamed, A. Hj., & Mahrin, M. N. (2017, November 7). *A survey on data mining techniques in recommender systems - soft computing*. SpringerLink. https://link.springer.com/article/10.1007/s00500-017-218-7

Natarajan, K., Li, J., & Koronios, A. (1970, January 1). *Data mining techniques for data cleaning*. SpringerLink. https://link.springer.com/chapter/10.1007/978-0-85729-320-6_91

Tai, D. W. S., Wu, H. J., & Li, P. H. (2008, June 6). *Effective e-learning recommendation system based on Data Mining*. The Electronic Library. Retrieved January 12, 2023, from https://doi.org/10.1108/0240470810879482

Nur Athirah Binti Ramlan (A179646)

Prof. Dr Zulaiha Ali Othman

Fakulti Teknologi & Sains Maklumat,

Universiti Kebangsaan Malaysia