MERAMAL FAKTOR KOS TUNTUTAN PERTUBUHAN KESELAMATAN SOSIAL(PERKESO) MENGGUNAKAN PEMBELAJARAN MESIN

NOR SYAHIDA BINTI CHE PA @ MUSTAPHA

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPROLEHI
IJAZAH SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

16 Februari 2024

NOR SYAHIDA BINTI CHE PA
@ MUSTAPHA
P107647

PENGHARGAAN

Dengan nama Allah Yang Maha Pemurah lagi Maha Mengasihani. Dipanjatkan setinggi kesyukuran kehadrat Ilahi di atas bimbingan hidayah-Nya dalam usaha menuntut ilmu ini. Tanpa inayat dari-Nya, usaha dan perjuangan menuntut ilmu ini pasti akan menjadi sia-sia.

Setinggi perhargaan buat penyelia projek akhir saya, Prof Madya Dr. Suhaila Zainudin di atas bimbingan dan kepakaran beliau yang sangat berharga. Komitmen dan kesabaran beliau dalam menelusuri perkembangan dan selok-belok projek ini amat saya hargai. Terima kasih juga buat semua pensyarah-pensyarah Program Sarjana Sains Data di atas didikan dan curahan ilmu yang diberikan sepanjang tempoh pengajian saya di UKM.

Terima kasih khas ditujukan buat ibu, suami dan keluarga di atas sokongan yang tidak berbelah bagi, doa yang tidak putus dan toleransi yang sentiasa ada. Anda merupakan tonggak kekuatan saya untuk meneruskan dan menghabiskan pengajian ini.

Ucapan terima kasih tidak terhingga buat majikan saya atas pemahaman dan sokongan mereka semasa saya menyambung pelajaran ke peringkat sarjana. Fleksibiliti dan dorongan membolehkan saya untuk mengimbangi tanggungjawab kerja dengan tuntutan akademik.

Kepada semua yang namanya tidak tersenarai di sini tetapi mereka memberi kesan yang mendalam dalam mujahadah menuntut ilmu ini, saya ucapkan bebanyak terima kasih terutama sekali kawan rapat dan rakan sekerja. Segala ucapan, doa yang baik serta kata-kata semangat mendorong saya untuk mengharungi cabaran ini dengan penuh ketahanan.

ABSTRAK

Penyelidikan ini mengkaji penggunaan teknik pembelajaran mesin untuk meramal faktor penyumbang kepada kos tuntutan bencana kerja di PERKESO. Enam algoritma pembelajaran mesin seperti *Linear Regression, Ridge Regression, Support Vector Machine, Decision Tree*, *Random Forest* dan XGBoost telah diaplikasikan dan kesemua algoritma dinilai prestasi ramalannya. Penggunaan teknik pembelajaran mesin terutamanya algoritma XGBoost menunjukkan prestasi paling unggul dalam kajian ini di mana algoritma ini berfungsi sebagai ramalan praktikal dalam memacu PERKESO sebagai organisasi yang mampu untuk membuat ramalan berasaskan data. Dengan memanfaatkan model ramalan ini, PERKESO sebagai penyedia pekhidmatan keselamatan sosial boleh memperolehi pandangan yang bernilai tentang faktor-faktor yang mempengaruhi kos tuntutan bencana kerja di samping meningkatkan keupayaan untuk menilai risiko dengan tepat. Sehubungan dengan itu tiga ciri teratas yang dikenalpasti (tempoh cuti sakit, zon dan umur) telah muncul sebagai faktor penting bagi penentu kos insurans. Penemuan ini boleh digunakan secara langsung oleh PERKESO untuk memperhalusi strategi pencegahan kemalangan dan meningkatkan proses membuat keputusan. Akhirnya penyelidikan ini bukan sahaja menyumbang kepada kemajuan teknik permodelan ramalan tetapi juga dapat memberikan garis panduan yang nyata bagi menghadapi kerumitan dalam membuat anggaran faktor kos tuntutan bencana kerja. Dengan menerima pembelajaran mesin sebagai suatu alat ramalan, PERKESO boleh mendapat manfaat daripada ketepatan dalam membuat perancangan dan keputusan, pengurusan risiko yang lebih baik dan akhirnya dapat memacu ke arah industri keselamatan sosial atau insurans yang lebih cekap dan adaptif.

PREDICTING THE CONTRIBUTING FACTORS FOR PERKESO CLAIMS USING A MACHINE LEARNING APPROACH

ABSTRACT

This research investigates the applicability of machine learning approaches for predicting factors contributing to insurance cost. Six machine learning algorithms were developed and evaluated for their predictive performance, namely Linear Regression, Ridge Regression, Support Vector Machine, Decision Tree, Random Forest and XGBoost. The utilisation of machine learning techniques, notably the XGBoost algorithm, which demostrated superior performance in this study, can be a practical dan data-driven predictive model for PERKESO or insurance companies. By leveraging these predictive models, PERKESO as social security provider, can gain valuable insights into the factors influencing insurance costs, thereby enhancing its ability to assess risks accurately. The identified top three features (duration of medical leaves, zones and age) has emerged as crucial predictors of insurance costs. PERKESO can directly apply this insight to refine prevention accident strategies and inform decision making processes. The predictive capabilities of machine learning enable more accurate cost predictions and provide proactive approach for PERKESO to strategise and plan effectively. As a result, this research contributes to the advancement of predictive modeling techniques and offers a tangible guidelines for PERKESO to navigate the complexities of insurance cost factor estimation. By embracing machine learning as a predictive tool, insurance companies can benefit from enhanced precision, improved risk management and the ability to make informed decisions, ultimately leading to a more efficient and adaptive insurance industry.

KANDUNGAN

Halaman

SENARAI JADUAL

No. Jadual **Halaman**

SENARAI ILUSTRASI

SENARAI SINGKATAN

ANN	*Artificial Neural Network*
DT	*Decision Tree*
KNN	*K-Nearest Neighbours*
LR	*Linear Regression*
MC	*Medical Certificate*
ML	*Machine Learning*/Pembelajaran Mesin
MSE	*Mean Squared Error*
OB	Orang Berinsurans
PKS	Pertubuhan Keselamatan Sosial
RF	*Random Forest*
RMSE	*Root Mean Squared Error*
UKM	Universiti Kebangsaan Malaysia
SVM	*Support Vector Machine*

BAB I

PENGENALAN

1.1 PENDAHULUAN

Deklarasi Hak Asasi Manusia (UDHR,1948) di dalam artikel 25 menyatakan setiap orang berhak mendapat taraf hidup yang mencukupi untuk kesihatan dan kesejahteraan diri dan keluarganya, termasuk makanan, pakaian, tempat tinggal dan rawatan perubatan serta perkhidmatan sosial yang diperlukan dan hak untuk mendapatkan keselamatan sekiranya berlaku pengangguran, sakit, hilang upaya, perceraian, usia yang atau kekurangan mata pencarian yang di luar kawalannya (UN,1948).

Oleh itu, satu dasar dan program yang direkabentuk bagi mengurangkan jurang kemiskinan serta bertujuan untuk melindungi dari sebarang ketidaktentuan yang boleh mengugat kesejahteraan hidup manusia sekaligus dapat memenuhi keperluan hak asasi manusia yang dinyatakan dalam artikel 25. Dasar atau program ini dinamakan sebagai Perlindungan Sosial.

Perlindungan sosial secara umumnya merujuk kepada rangkaian dasar, program, dan langkah-langkah yang dilaksanakan oleh kerajaan dan masyarakat untuk menangani dan mengurangkan isu kemiskinan, ketidakseimbangan jurang sosial masyarakat. Perlindungan sosial merupakan komponen penting dalam dasar sosial yang bertujuan untuk mempromosikan kesejahteraan dan keselamatan sosial individu dan komuniti (DOSM,2021).

Tujuan utama perlindungan sosial adalah untuk memastikan semua ahli masyarakat mempunyai akses kepada barangan, perkhidmatan, dan sistem sokongan asas, terutamanya pada waktu rentan atau dalam keadaan terdesak. Perlindungan sosial merangkumi beberapa dimensi, termasuk keselamatan sosial, bantuan sosial, insurans

sosial, dan dasar pasaran buruh. Berikut adalah beberapa elemen penting perlindungan sosial daripada maklumat Bank Negara Malaysia , https://www.bnm.gov.my/documents/20124/3026377/emr2020_bm_box1_socialprotection.pdf :

1. **Keselamatan Sosial**: Keselamatan sosial merujuk kepada set program dan langkah-langkah yang menyediakan sokongan pendapatan dan perlindungan terhadap risiko yang berkaitan dengan usia tua, kecacatan, pengangguran, dan penyakit. Ini termasuk pencen, faedah kecacatan, faedah pengangguran, dan perkhidmatan penjagaan kesihatan.

2. **Bantuan Sosial**: Program bantuan sosial direka untuk memberikan sokongan segera dan sasaran kepada individu dan keluarga yang menghadapi kemiskinan atau kesukaran dalam hidup. Program ini dilaksanakan dalam bentuk pemberian wang tunai, bantuan makanan, subsidi perumahan, atau baucar untuk memenuhi keperluan asas.

3. **Insurans Sosial**: Skim insurans sosial biasanya merupakan program kontributif yang dibiayai melalui sumbangan wajib daripada individu dan majikan. Program ini memberikan perlindungan dan faedah kepada individu dalam keadaan kecemasan seperti pengangguran, penyakit, kematangan, atau kecederaan berkaitan dengan kerja.

4. **Dasar Pasaran Buruh**: Dasar pasaran buruh bertujuan untuk mempromosikan pekerjaan yang baik, memastikan upah yang adil, dan melindungi hak pekerja. Ini termasuk langkah-langkah seperti undang-undang upah minimum, perlindungan pekerjaan, peraturan keselamatan dan kesihatan pekerjaan, dan program pembangunan kemahiran dan latihan pekerjaan.

5. **Pengurangan Kemiskinan**: Perlindungan sosial memainkan peranan penting dalam mengurangkan kemiskinan dan ketidakseimbangan dengan menyediakan sokongan pendapatan, akses kepada perkhidmatan asas, dan peluang pembangunan manusia. Ia membantu individu dan kumpulan rentan untuk mengatasi halangan dan meningkatkan kesejahteraan keseluruhan.

6. **Perlindungan Universal**: Konsep perlindungan sosial universal berusaha untuk memastikan semua orang dalam masyarakat, tanpa mengira status sosioekonomi atau jenis pekerjaan mereka, mempunyai akses kepada program dan perkhidmatan perlindungan sosial asas. Ia menekankan inklusiviti dan selari dengan frasa *no one left behind* (Datin Azlaily Binti Abd Rahman et.al,2020).

Program perlindungan sosial dapat dibiayai melalui pelbagai sumber, termasuk bajet kerajaan, sumbangan keselamatan sosial, bantuan antarabangsa, dan perkongsian awam-swasta. Reka bentuk dan pelaksanaan sistem perlindungan sosial tertentu berbeza antara negara, bergantung kepada konteks sosioekonomi, keutamaan politik, dan sumber yang ada.

Di Malaysia, perlindungan sosial merangkumi pelbagai dasar dan program yang bertujuan untuk memastikan kesejahteraan dan keselamatan sosial penduduknya. Sistem perlindungan sosial di Malaysia meliputi komponen-komponen penting berikut:

1. **Pertubuhan Keselamatan Sosial (PERKESO) :** PERKESO berdasarkan maklumat yang diperolehi daripada https://www.perkeso.gov.my/ adalah badan berkanun yang bertanggungjawab mengendalikan Akta Keselamatan Sosial Pekerja 1969 pada asalnya. Ia menyediakan perlindungan keselamatan sosial kepada pekerja dalam kes kemalangan berkaitan dengan pekerjaan, penyakit pekerjaan, ketidakupayaan, dan kematian. Ia menawarkan faedah seperti kos perubatan, faedah kecacatan, dan faedah tanggungan.

2. **Insurans Kesihatan Nasional** : Malaysia mempunyai skim insurans kesihatan nasional yang dikenali sebagai skim MySalam berdasarkan maklumat yang didapati daripada https://www.bnm.gov.my/documents/20124///dddb1e16-b0b0-2495-0f57-9ac6a59eec91/. Skim ini menyediakan bantuan kewangan dan perlindungan untuk penyakit kritikal kepada individu dan isi rumah berpendapatan rendah. Skim ini bertujuan untuk mengurangkan beban kewangan yang berkaitan dengan kos penjagaan kesihatan bagi mereka yang layak.

3. **Program Bantuan Awam** : Kerajaan Malaysia menyediakan pelbagai bentuk bantuan sosial kepada individu dan keluarga berpendapatan rendah. Program utama adalah Bantuan Sara Hidup (BSH), sebelum ini dikenali sebagai BR1M (Bantuan Rakyat 1Malaysia) mengikut informasi yang diperolehi daripada https://www.br1m.info/bantuan-sara-hidup-bujang-2020-bsh/, dimana BR1M menyediakan pemindahan wang tunai kepada isi rumah yang layak untuk mengurangkan kemiskinan dan meningkatkan taraf hidup mereka.

4. **Faedah Pencen** : Kumpulan Wang Simpanan Pekerja (KWSP) adalah skim simpanan wajib untuk pekerja di Malaysia. Ia berfungsi sebagai faedah pencen dan menyediakan kestabilan kewangan kepada individu pada usia persaraan. Pekerja dan majikan menyumbang sebahagian daripada gaji pekerja ke KWSP, yang berkumpul sepanjang masa dan boleh dikeluarkan pada masa persaraan (DOSM,2021).

5. **Bantuan Pendidikan** : Kerajaan Malaysia menawarkan pelbagai bentuk bantuan kewangan untuk menyokong pendidikan pada pelbagai peringkat. Ini termasuk biasiswa, pinjaman pengajian, dan geran untuk memastikan akses kepada pendidikan bagi pelajar dari latar belakang berpendapatan rendah menurut maklumat yang didapati pada https://www.malaysia.gov.my/portal/content/29611?language=my.

6. **Sistem Insurans Pekerjaan (SIP):** Malaysia melaksanakan Sistem Insurans Pekerjaan pada tahun 2018, yang menyediakan sokongan pendapatan sementara kepada pekerja yang kehilangan pekerjaan. Ia menawarkan bantuan kewangan dan perkhidmatan penempatan semula untuk membantu individu semasa tempoh peralihan antara pekerjaan. SIP ini mula diperkenalkan oleh PERKESO pada tahun 2018(Datin Azlaily Binti Abd Rahman et.al,2020).

7. **Perumahan Subsidi** : Kerajaan Malaysia menyediakan skim perumahan subsidi, seperti Program Perumahan Rakyat, untuk memastikan perumahan yang terjangkau bagi isi rumah berpendapatan rendah menurut maklumat yang didapati pada https://www.pmo.gov.my/wp-

content/uploads/2019/07/BUKU_Dasar_Perumahan_Mampu_Milik_Negara_11052019.pdf

Ini adalah beberapa langkah perlindungan sosial utama di Malaysia. Program dan dasar khusus boleh berubah dari semasa ke semasa apabila kerajaan memperkenalkan inisiatif baru atau mengubah yang sedia ada untuk menangani keperluan yang berubah dalam populasi dan keadaan sosioekonomi.

Secara keseluruhannya, perlindungan sosial adalah instrumen penting untuk mempromosikan keadilan sosial, kestabilan ekonomi, dan hak asasi manusia. Dengan menangani kemiskinan dan risiko sosial, ia bertujuan untuk mencipta masyarakat yang lebih adil dan inklusif di mana individu memperoleh sokongan yang diperlukan untuk menjalani kehidupan yang bermaruah dan selamat.

1.2 LATAR BELAKANG KAJIAN

PERKESO merupakan sebuah badan berkanun di Malaysia yang mentadbir program keselamatan sosial untuk pekerja. Ia menjalankan beberapa fungsi untuk memastikan kesejahteraan dan perlindungan sosial pekerja. Berikut adalah fungsi utama PERKESO (PERKESO,2023) :

1. **Pampasan Bencana Kerja** : PERKESO menyediakan pampasan dan faedah kepada pekerja yang mengalami kemalangan berkaitan dengan pekerjaan atau penyakit pekerjaan. Ini termasuk kos perubatan, perkhidmatan rehabilitasi, faedah kecacatan sementara atau kekal, dan bantuan kewangan kepada tanggungan dalam kes kematian(Datin Azlaily Binti Abd Rahman et.al,2020).

2. **Program Kembali Bekerja** : PERKESO melaksanakan program dan inisiatif untuk memudahkan pekerja yang cedera untuk kembali bekerja. Ini melibatkan penyediaan perkhidmatan rehabilitasi vokasional, bantuan penempatan pekerjaan, dan program latihan untuk meningkatkan kebolehpasaran dan mengintegrasikan individu ke dalam tenaga kerja(Datin Azlaily Binti Abd Rahman et.al,2020).

3. **Insurans Bencana Pekerjaan** : PERKESO mengendalikan skim insurans bencana pekerjaan yang memberikan perlindungan kepada pekerja dalam kes kemalangan berkaitan dengan pekerjaan atau penyakit pekerjaan. Majikan dikehendaki untuk menyumbang kepada dana insurans ini, dan pekerja menerima faedah daripada dana tersebut dalam kes kecederaan atau kecacatan(Datin Azlaily Binti Abd Rahman et.al,2020).

4. **Skim Pencen Keilatan** : PERKESO menguruskan Skim Pencen Keilatan, yang memberikan sokongan kewangan kepada pekerja yang tidak dapat bekerja secara kekal disebabkan kecacatan atau keilatan. Individu yang layak menerima pencen bulanan untuk menggantikan pendapatan mereka(Datin Azlaily Binti Abd Rahman et.al,2020).

5. **Skim Pencen Untuk Tanggungan** : PERKESO mentadbir urus Skim Pencen ini yang memberikan sokongan kewangan kepada tanggungan pekerja yang meninggal dunia. Ini termasuk pencen bulanan,dan faedah pendidikan untuk kanak-kanak(Datin Azlaily Binti Abd Rahman et.al,2020).

6. **Pendidikan dan Latihan** : PERKESO melaksanakan program pendidikan dan latihan untuk meningkatkan kesedaran tentang keselamatan dan kesihatan pekerjaan di kalangan majikan dan pekerja. Ia bertujuan untuk mempromosikan persekitaran kerja yang selamat dan mencegah kemalangan dan penyakit berkaitan dengan pekerjaan(Datin Azlaily Binti Abd Rahman et.al,2020).

7. **Penyelidikan dan Advokasi** : PERKESO menjalankan penyelidikan dan kajian mengenai keselamatan sosial, keselamatan dan kesihatan pekerjaan. Ia memainkan peranan advokasi dalam mempromosikan perlindungan sosial dan meningkatkan dasar dan program keselamatan sosial di Malaysia(Datin Azlaily Binti Abd Rahman et.al,2020).

Dalam memenuhi fungsi-fungsi utama tesebut, PERKESO bertanggungjawab mentadbir dan menguatkuasakan empat (4) iaitu Akta Keselamatan Sosial Pekerja(Akta 4), Akta Keselamatan Sosial Pekerjaan Sendiri 2017 (Akta 789), Akta Sistem Insurans

Pekerjaan (Akta 800) dan Akta Skim Keselamatan Sosial Suri Rumah 2022 (Akta 838). Berikut adalah maklumat mengenai akta-akta tersebut :

1. **Akta Keselamatan Sosial Pekerja 1969** : terdapat dua skim perlindungan yang ditawarkan iaitu Skim Bencana Kerja dan Skim Keilatan. Skim Bencana Kerja memberi perlindungan kepada para pekerja daripada bencana pekerjaan termasuk penyakit khidmat dan kemalangan semasa berkaitan dengan pekerjaan. Manakala Skim Keilatan memberi pelindungan 24 jam kepada pekerja terhadap keilatan atau kematian akibat sebarang sebab yang berlaku di luar waktu kerja. Kedua-dua skim adalah untuk menyediakan faedah tunai kepada pekerja dan tanggungannya apabila berlaku kejadian di luar jangka di samping menyediakan rawatan perubatan, pemulihan jasmani atau latihan vokasional. Pada mulanya akta ini memberi perlindungan kepada pekerja bermajikan sahaja dan kemudian perlindungan akta diperluaskan kepada pekerja asing dan pekerja domestik (Pertubuhan Keselamatan Sosial,2021).

2. **Akta Keselamatan Sosial Pekerjaan Sendiri 2017 (Akta 789)** : memberi perlindungan di bawah Skim Bencana Pekerjaan kepada pemandu teksi yang bekerja sendiri dan individu yang menjalankan perkhidmatan pemandu seperti *Uber* dan *Grab Car*. Mulai 1 Januari 2020, Akta Keselamatan Sosial Pekerjaan Sendiri 2017 (Akta 789) telah diperluaskan kepada 19 sektor informal pekerjaan sendiri yang lain seperti nelayan, petani, penjaja, penggiat seni,dan sebagainya secara berperingkat. Perluasan skim ini secara tidak langsung dapat memantapkan lagi jaringan keselamatan sosial yang progresif dan komprehensif yang memberi manfaat kepada lebih 2.5 juta pekerja dalam sektor informal di Malaysia. Pelaksanaanya bermula 1 Januari 2020 (Pertubuhan Keselamatan Sosial,2021).

3. **Akta Sistem Insurans Pekerjaan (Akta 800)** : diperkenalkan bertujuan untuk melindungi dan membantu pekerja yang kehilangan pekerjaan melalui dua komponen utama iaitu Insurans Pekerjaan dan Perkhidmatan Pekerjaan untuk menggalakkan pasaran buruh aktif. Sistem Insurans Pekerjaan(SIP) merupakan skim baharu perlindungan tambahan kepada pekerja-pekerja yang kehilangan

pekerjaan bagi menggantikan pendapatan yang hilang, memberi latihan *reskilling* dan *upskilling* untuk mendapatkan pekerjaan baharu serta menyediakan perkhidmatan carian pekerjaan supaya mereka yang kehilangan pekerjaan mendapat pekerjaan yang sesuai dengan lebih cepat. Pelaksanaan SIP menjadi platform yang signifikan mendepani situasi luar jangka pemberhentian pekerja dalam pasaran tenaga buruh di Malaysia berikutan cabaran serta kesan buruk kepada ekonomi akibat pandemik COVID-19 yang telah memberi kesan mendalam kepada pasaran pekerjaan. Menerusi PERKESO, pekerja yang kehilangan pekerjaan akan diberi bantuan kewangan segera untuk menampung perbelanjaan semasa mencari pekerjaan baharu. Mereka juga akan dibantu dalam penilaian kerjaya, kaunselling, carian kerja serta pemadanan dan penempatan pekerjaan. Pelaksanaan Dasar Pasaran Buruh Aktif seperti ini dapat mengelakkan pengganggguran jangka panjang dan menjadi penstabil ekonomi semasa krisis ekonomi terutama semasa negara dilanda pandemik COVID-19 (SIP,2022).

4. **Akta Skim Keselamatan Sosial Suri Rumah 2022 (Akta 838)** : Skim Keselamatan Sosial Suri Rumah (SKSSR) yang berkuat kuasa pada 1 Disember 2022, diperkenalkan di bawah Akta Keselamatan Sosial Suri Rumah 2022. Skim ini bertujuan untuk memberi perlindungan bencana domestik kepada suri rumah yang mengalami kemalangan atau apa-apa insiden semasa melakukan urusan berkaitan rumah tangga. SKSSR juga melindungi suri rumah yang ditimpa penyakit atau keuzuran yang diakses pada https://www.perkeso.gov.my/perkhidmatan-kami/perlindungan/suri-rumah.html .

Secara keseluruhannya, PERKESO memainkan peranan penting dalam memastikan perlindungan sosial pekerja di Malaysia. Ia menyediakan sokongan kewangan, perkhidmatan penjagaan kesihatan, dan program rehabilitasi kepada individu yang terkesan oleh kemalangan berkaitan dengan pekerjaan atau penyakit pekerjaan. Dengan mentadbir program-program keselamatan sosial ini, PERKESO menyumbang kepada kesejahteraan, keselamatan pendapatan, dan kebajikan sosial pekerja dan tanggungan mereka.

1.3 PERMASALAHAN KAJIAN

PERKESO mendapatkan dana untuk memberi pelindungan kepada pekerja adalah melalui caruman pekerja dan majikan yang dikuatkuasakan melalui akta-akta yang ditadbir urus. Untuk memastikan PERKESO terus dapat melaksanakan kewajipan untuk melindungi pekerja, dana caruman yang terkumpul itu perlu cukup untuk menampung keperluan pembiayaan pada masa hadapan. Buat masa sekarang PERKESO cuba mengimbangi pemprosesan tuntutan yang cekap dengan pelarasan dasar atau akta yang proaktif. Antara langkah yang diambil dengan memastikan tuntutan yang dibuat oleh pekerja melalui proses semakan rapi bagi memastikan ianya memenuhi kelayakan yang ditakrifkan oleh akta-akta yang ditadbir urus oleh PERKESO. Proses semakan ini adalah termasuk menyemak caruman pekerja tersebut, jenis kemalangan atau penyakit dan mengikut proses pelaporan yang betul.

Di samping itu, laporan perubatan juga dapat mengesahkan tahap kemalangan dan ketidakupayaan untuk menentukan pampasan yang bersesuaian dengan seseorang pekerja yang tercedera. PERKESO menggunakan khidmat doktor dan hospital yang bertauliah bagi mengesahkan laporan perubatan ini. Manakala bagi pengiraan faedah yang diberikan kepada pekerja yang terbencana, ianya bergantung kepada jenis tuntutan iaitu bencana kerja, ilat, cuti sakit dan banyak lagi. Untuk tujuan ini terdapat formula yang khusus untuk mengira jumlah pampasan berdasarkan faktor seperti gaji, tahap hilang upaya dan bilangan tanggungan. Faedah yang diluluskan akan diagihkan atau dibayar tepat pada masanya mengikut saluran yang ditetapkan. PERKESO akan memantau kes yang berterusan untuk menilai kemajuan pemulihan seseorang pekerja yang ditimpa bencana di samping menyediakan rawatan perubatan, kemudahan pemulihan fizikal dan vokasional serta kemudahan-kemudahan yang lain. Rawatan-rawatan sebegini akan menelan kos yang besar jika tidak dipantau dengan lebih dekat dan efektif.

Sehubungan dengan itu PERKESO melaburkan caruman terkumpul dalam portfolio yang pelbagai untuk memaksimumkan pulangan dan mengurangkan risiko kewangan. Strategi pelaburan akan mempertimbangkan kemampanan jangka panjang dan keselamatan. PERKESO juaga sentiasa menyemak dan mengemas kini dasar peraturannya untuk menyesuaikan dengan perubahan keperluan dan cabaran yang

wujud. Antaranya PERKESO pernah mencadangkan pelarasan kadar caruman pekerja dan majikan. Pelarasan ini adalah untuk memastikan dana yang mencukupi bagi memenuhi liabiliti semasa. Pelarasan ini tidak boleh dilaksanakan secara berkala di mana banyak faktor perlu diambil seperti faktor ekonomi dan proses pelarasan ini melibatkan banyak proses dan kelulusan.

PERKESO perlu mengekalkan kemampanan dananya untuk jangka masa panjang bagi membiayai perbelanjaan mengurus dan tanggungan yang semakin meningkat selari dengan pertambahan bayaran faedah. Strategi memantapkan pengurusan kewangan selain mengoptimumkan pulangan pelaburan yang mampu mengukuhkan dana PERKESO. Bagi merangka strategi pengurusan kewangan PERKESO perlu mempunyai model atau platform yang mampu membuat ramalan atau perangkaan yang efisien. Menurut kajian yang dijalankan oleh Shamshimah Samsuddin, Noriszura Ismail (2018) peningkatan kos perbelanjaan PERKESO adalah disebabkan oleh peningkatan kemalangan di tempat kerja, terutamanya peningkatan jumlah kemalangan dalam perjalanan pergi dan balik bekerja. Pada masa sekarang, PERKESO membuat kajian aktuari menggunakan analisa deskriptif berdasarkan data caruman pekerja, demografi dan trend ekonomi untuk mengunjurkan keperluan kewangan pada masa hadapan. Justeru itu, PERKESO sentiasa proaktif dalam mencari atau mengkaji kaedah yang lebih efisien dalam membantu pengurusan membuat keputusan kewangan dalam merangka hala tuju PERKESO.

1.4 PERSOALAN KAJIAN

Berdasarkan permasalahan kajian, persoalan kajian dirangka seperti mana berikut :

1. Adakah terdapat cara lain untuk menganggar kos tuntutan bencana kerja(antara faedah yang ditawarkan oleh PERKESO) selain daripada cara yang sedia iaitu menggunakan cara statistik.
2. Apakah faktor-faktor yang menyumbang kepada kes tuntutan faedah Bencana Kerja PERKESO.
3. Bagaimana membuat perancangan strategik dan bersasar dalam mencegah kemalangan tempat kerja. Adakah kemungkinan perancangan pencegahan kemalangan yang sedia ada dapat diperhalusi mengikut sesetengah kawasan

atau jenis perindustrian atau faktor-faktor lain yang banyak mempengaruhi kes tuntutan bencana kerja.

1.5 OBJEKTIF KAJIAN

Konsep Perlindungan Keselamatan Sosial PERKESO adalah berteraskan konsep tanggungjawab bersama menerusi *"pooling resources, sharing of risk and replacement of income"* Perlindungan keselamatan sosial adalah asas yang perlu dipenuhi sebagaimana yang disepakati di bawah Konvensyen Geneva Pertubuhan Buruh Antarabangsa (ILO) 1952 iaitu Konvensyen 102; *Minimum Standard for Social Security.* Dalam memenuhi matlamat tersebut, fungsi utama PERKESO adalah memberi perlindungan keselamatan sosial kepada pekerja dan tanggungannya menerusi Skim-skim yang ditawarkan oleh PERKESO.

Terdapat beberapa objektif utama yang telah digariskan untuk dicapai melalui kajian ini iaitu :

1. Mengkaji kebolehgunaan Pembelajaran Mesin yang merupakan suatau pendekatan kecerdasan pengiraan untuk meramal faktor kos tuntutan insurans dalam bidang keselamatan sosial.
2. Membandingkan prestasi algoritma atau model Pembelajaran Mesin dalam membuat ramalan faktor kos tuntutan yang dikemukakan kepada PERKESO menggunakan data PERKESO sendiri.
3. Membangunkan model ramalan untuk mengenalpasti faktor yang menyumbang kepada kos tuntutan bencana kerja di PERKESO dan mengkaji bahagian anggota badan yang kerap terlibat dalam kemalangan supaya dapat merancang program *Return To Work* atau Program Pencegahan yang lebih efektif .

1.6 METODOLOGI KAJIAN

Tujuan utama kajian adalah untuk mengkaji faktor yang menyumbang ke arah tuntutan faedah PERKESO sekaligus dapat meramal keseluruhan kos tuntutan faedah bagi merancang kedudukan kewangan organisasi. Bagi mencapai tujuan tersebut, beberapa langkah atau metodologi kajian akan dilaksanakan berdasarkan proses berikut :-

1. Pengumpulan Data

Pengumpulan data dari sumber yang boleh dipercayai adalah amat penting bagi kajian ini berikutan model pembelajaran mesin banyak bergantung kepada data yang disediakan. Kuantiti dan kualiti data yang diperlukan juga perlu mencukupi dan tepat bagi mendapatkan hasil yang betul. Untuk proses ini, beberapa data daripada legasi sistem telah dikenalpasti. Data ini perlu diekstraks dan diproses dengan teliti.

2. Pra-pemprosesan data

Selepas kesemua data diperolehi dan dikumpulkan dari beberapa sumber legasi di PERKESO, data tersebut perlu perlu dibersihkan terlebih dahulu. Data yang tidak diperlukan perlu dibuang. Data yang *redundant* dan *missing value* juga perlu dikenalpasti dan dibuang. Data-data tersebut akan distruktur semula dan diletak dalam bentuk *visulization* seperti carta untuk memahami hubungan di antara pelbagai data yang diperolehi.

Data yang telah dibersihkan sepenuhnya itu akan dibahagikan kepada dua set iaitu set latihan dan set ujian. Data yang terkandung dalam set latihan merupakan data yang akan digunakan oleh pembelajaran mesin untuk membina model ramalan. Manakala set data latihan pula akan digunakan untuk menguji ketepatan model pembelajaran mesin yang telah dibangunkan.

3. Memilih Model Pembelajaran Mesin

Pemilihan model pembelajaran mesin adalah penting kerana ianya menentukan output yang akan diperolehi nanti berdasarkan informasi pada https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/. Bagi kajian ini output yang tepat adalah penting kerana ianya akan digunakan untuk merancang pelan jangka panjang kedudukan kewangan PERKESO. Di antara beberapa pendekatan model pembelajaran mesin yang biasa digunakan untuk membuat ramalan kos insurans adalah menggunakanan model *Linear Regression, Multiple Linear Regression, Decision Tree, K-Nearest Neighbour* dan banyak lagi.

4. Melatih Model Pembelajaran Mesin

Proses ini merupakan salah satu elemen penting dalam membina model pembelajaran mesin di mana model akan mencari corak dan membuat ramalan daripada set data latihan yang telah disediakan. Melatih model pembelajaran mesin akan mengambil masa kerana beberapa algoritma atau model akan dicuba pada peringkat ini.

5. Menilai Model Pembelajaran Mesin

Selepas beberapa model Pembelajaran Mesin telah dibina semasa proses latihan sebelum ini menggunakan set data latihan, seterusnya model tersebut perlu dinilai ketepatannya dalam menghasilkan output dengan melalui proses ujian menggunakan set data ujian yang telah disediakan pada peringkat awal tadi.

6. Membuat Ramalan

Ketepatan setiap pendekatan yang digunakan bagi model pembelajaran mesin untuk menghasilkan output yang dikehendaki perlu dibandingkan. Pendekatan Model Pembelajaran Mesin yang berjaya menghasilkan output yang tepat dan paling rendah *error* akan dipilih untuk diimplementasi.

1.7 SKOP KAJIAN

Pampasan Bencana Kerja secara asasnya menjadi tunggak kepada kebanyakan produk PERKESO seperti Akta Keselamatan Sosial Pekerja(Akta 4), Akta Keselamatan Sosial Pekerjaan Sendiri 2017 (Akta 789) dan Akta Skim Keselamatan Sosial Suri Rumah 2022 (Akta 838). Pampasan Bencana Kerja ini melibatkan permohonan pampasan atau faedah bagi kecederaan yang berpunca daripada pekerjaan. Pampasan ini boleh menyebabkan rentetan kepada pelbagai tuntutan yang di lain di PERKESO. Apabila seorang pekerja mengalami kecederaan berkaitan dengan pekerjaan atau penyakit pekerjaan, mereka mungkin layak membuat beberapa jenis tuntutan di bawah PERKESO. Berikut adalah beberapa tuntutan yang berpotensi timbul daripada kecederaan pekerjaan (Datin Azlaily Binti Abd Rahman et.al,2020) *:*

1. **Pampasan Hilang Upaya Kekal (HUK)** : Jika kecederaan pekerjaan menyebabkan kecacatan kekal atau ketidakupayaan bekerja, pekerja tersebut mungkin layak untuk memohon HUK. Ini menyediakan sokongan kewangan berterusan kepada pekerja disebabkan kecacatan mereka.

2. **Elaun Layanan Sentiasa (ELS)** : Jika pekerja yang cedera memerlukan kehadiran atau bantuan berterusan disebabkan keparahan keadaan mereka, mereka berhak untuk memohon elaun layanan sentiasa. Ini merupakan faedah tambahan untuk membantu menampung kos penjagaan atau pengawasan berterusan. Kadar ELS yang disediakan oleh PERKESO adalah sebanyak RM500.00 sebulan.

3. **Faedah Pendidikan dan Latihan**: Dalam kes kecacatan yang teruk atau ketidakupayaan untuk kembali ke pekerjaan sebelumnya, pekerja yang cedera mungkin layak untuk mendapatkan faedah pendidikan dan latihan. Faedah ini menyokong pekerja dalam memperoleh kemahiran atau pengetahuan baru untuk mengejar peluang pekerjaan alternatif.

Tuntutan-tuntutan yang dinyatakan di atas berbeza bergantung kepada keadaan kecederaan pekerjaan, keparahan kecacatan dan kriteria kelayakan peruntukan yang berkenaan dalam Akta Keselamatan Sosial Pekerja 1969.

Oleh itu skop kajian ini hanya mengunapakai data-data bagi kes tuntutan faedah atau pampasan Bencana Kerja sahaja. Skop kajian bertumpu kepada kepada kes tuntutan faedah yang berlaku ke atas pekerja disebabkan oleh kecederaan atau kematian atau ketidakupayaan. Kes tuntutan faedah kehilangan pekerjaan tidak termasuk dalam skop kajian ini berikutan Skim Insurans Pekerjaan(SIP) baru diwujudkan pada tahun 2018 dan data mengenai kes ini tidak mempunyai volume yang besar. Dalam menghasilkan model pembangunan menggunakan pembelajaran mesin data dalam *volume* diperlukan. Data-data yang didapati melalui sistem legasi sistem PERKESO adalah berbentuk kuantitatif sahaja.

Kajian ini juga tertumpu kepada kes tuntutan yang dikemukakan kepada PERKESO sahaja. Kes tuntutan insurans bagi syarikat insurans yang berdaftar di pasaran tidak termasuk.

1.8 KEPENTINGAN KAJIAN

Proses pembelajaran mesin bermula dengan analisa dan pemerhatian melalui data yang disediakan. Ia mengenalpasti corak dan trend serta pengalaman yang boleh dipelajari melalui data tersebut. Setelah pembelajaran mesin ini meneroka analisis data, model akan dibina menggunakan pendekatan dan algoritma lantas menghasilkan output seperti ramalan data yang boleh digunakan untuk membuat sesuatu keputusan. Kelebihan pembelajaran mesin adalah komputer belajar secara automatik tanpa arahan atau bantuan manusia.

Model Pembelajaran Mesin boleh menganalisis sejumlah data yang besar dari pelbagai sumber termasuk rekod tuntutan pekerja yang mengalami di tempat kerja. Analisis komprehensif ini membolehkan pengenalpastian corak dan sebarang kelemahan berkenaan keselamatan pekerja di tempat kerja atau organisasi masing-masing. Melalui output yang diperolehi dari algoritma Pembelajaran Mesin, sesebuah profil risiko mungkin boleh dihasilkan mengikut populasi yang berbeza di tempat kerja dan sekaligus dapat merangka program pencegahan yang boleh disesuaikan untuk menangani sebarang kecederaan di tempat kerja. Program pencegahan secara bersasar dapat dirangka dengan menumpukan usaha pencegahan terhadap kumpulan fokus yang terdedah dengan risiko kemalangan atau kecederaan. Ianya dapat membantu memperuntukkan sumber pencegahan dengan lebih cekap di samping dapat memaksumumkan keselamatan dan kesihatan pekerja dan tempat kerja (Anurag Yedla et.al,2020). Selain dapat mengenalpasti risiko semasa kecederaan di tempat kerja ianya juga dapat meramalkan risiko yang bakal berlaku pada masa hadapan. Dengan menganalisis data sejarah dan menggabungkan trend, model ini menjangkakan potensi kemalangan dan kecederaan di tempat kerja. Keupayaan ramalan ini membolehkan langkah pencegahan yang lebih proaktif dan dapat mengurangkan kemungkinan kemalangan dan kecederaan.

Ramalan menggunakan kaedah pembelajaran mesin ini cuba untuk diadaptasi ke dalam PERKESO menyedari kepentingan merangka dan merancang strategi kewangan yang jitu agar organisasi sentiasa utuh melindungi pekerja dan tanggungannya(Pertubuhan Keselamatan Sosial,2021). Kemampanan dan kecukupan dana adalah penting bagi PERKESO berikutan salah satu faedah yang ditawarkan kepada pekerja dan tanggungan yang berkelayakan dalam bentuk pencen bulanan. Pencen bulanan ini memerlukan dana dalam jangka masa panjang di mana penerima pencen ini bergantung sepenuhnya kepada pencen ini untuk meneruskan kehidupan berikutan mereka tidak berkemampuan untuk mencari nafkah disebabkan ketidakupayaan ataupun masih di alam persekolahan. Pencen dimaksudkan di sini adalah adalah Pencen Ilat, Pencen Penakat, Faedah Orang Tanggungan dan Elaun Layanan Sentiasa.

Peruntukan dan dana yang mencukupi adalah penting kerana PERKESO bukan sahaja memberi pelindungan keselamatan sosial kepada pekerja dan tanggungannya, akan tetapi PERKESO juga menjaga kebajikan pekerja dan tanggungannya supaya dapat meneruskan kelangsungan hidup seperti biasa setelah sesebuah keluarga punca pencari pendapatan keluarga.

BAB II

KAJIAN LITERATUR

2.1 PENGENALAN

Tujuan utama kajian literatur adalah untuk memahami penyelidikan terdahulu yang berkait rapat dengan topik kajian ini. Perbandingan dan perbezaan boleh dihasilkan dengan melakukan kajian literatur samada hasil daripada penyelidikan yang terdahulu boleh menyokong atau menolak hasil kajian ini. Menurut Boell and O'Connor (2020), kajian literatur adalah asas kepada mana-mana projek penyelidikan. Ia menyediakan konteks, justifikasi dan rangka kerja teori untuk kajian.

Setiap kajian literatur akan diteliti dan dibandingkan untuk melihat samada terdapat perbezaan, perubahan atau persamaan hasil kajian dengan hasil kajian-kajian yang lepas untuk menyokong setiap objektif kajian ini.

2.2 KECEDERAAN PEKERJAAN DAN TUNTUTAN INSURANS

Kecederaan Pekerjaan merujuk kepada sebarang kecederaan, kerosakan atau gangguan terhadap kesejahteraan fizikal, mental atau emosi seseorang pekerja yang berlaku akibat tugas atau aktiviti berkaitan dengan pekerjaan yang dilakukan. Kecederaan ini boleh meliputi daripada kemalangan yang minor atau serius sehingga boleh mengakibatkan kesan kesihatan yang signifikan kepada seseorang pekerja.

Kecederaan Pekerjaan ini mendatangkan impak ataupun kesan berdasarkan informasi yang diperolehi pada https://www.msd.govt.nz/about-msd-and-our-work/publications-resources/journals-and-magazines/social-policy-

journal/spj23/23aftermath-research-workplace-injury-and-illness-pages181-194.html yang mana ianya dapat dikategorikan dalam beberapa dimensi seperti berikut :

1. Impak ke atas Manusia

a. **Kesihatan Fizikal** : Kecederaan pekerjaan boleh mengakibatkan kecederaan fizikal seperti luka, terbakar, patah tulang, regangan otot dan kecederaan yang lebih serius seperti amputasi, kecederaan otak, traumatik atau kehilangan mana-mana bahagian anggota badan. Kecederaan ini boleh menyebabkan kesakitan, ketidakselesaaan dan mungkin mendatangkan kecacatan sementara ataupun kekal.

b. **Kesihatan Mental** : Pekerja yang mengalami kecederaan pekerjaan juga terdedah kepada kemungkinan berlakunya tekanan psikologi, kebimbangan, gangguan selepas trauma (PTSD), kemurungan dan masalah kesihatan mental yang lain berikutan dari trauma kemalangan yang dilalui.

2. Impak Ekonomi

a. **Kos secara langsung :** Kecederaan pekerjaan membawa kepada kos kewangan secara langsung untuk kos rawatan perubatan, rehabilatasi, hospitalitasi dan kos ubat-ubatan. Kos ini menelan belanja yang besar terutama sekali buat pekerja yang mengalami kecederaan yang serius.

b. **Kos secara tidak langsung :** Kos tidak langsung adalah seperti produktiviti, ketidakhadiran bekerja, kos penggantian dan latihan untuk pekerja baru, kos perbelanjaan pentadbiran berkenaan pengurusan tuntutan pampasan dan moral yang berkurangan di kalangan pekerja.

3. Impak Tempat Kerja

a. Produktiviti Berkurang : Kecederaan pekerjaan boleh menyebabkan ketidakhadiran dan pengurangan kapasiti kerja dan ini akan mengakibatkan penurunan produktiviti bagi pekerja yang terlibat dengan kemalangan atau mengalami kecederaan dan ianya sekaligus menjejas produtiviti pasukan ataupun jabatan secara keseluruhan.

b. Gangguan : Kecederaan serius boleh menganggu aliran kerja biasa di mana proses penyiasatan akan berlaku seperti audit keselamatan dan tindakan pembetulan seperti pencegahan kemalangan untuk mengelakkan kemalangan yang sama akan berulang pada masa hadapan.

4. Impak Sosial

a. Keluarga dan sosial : Kecederaan pekerjaan boleh memberi kesan kepada keluarga pekerja di mana keupayaan pekerja yang tercedera dalam menyumbang kepada aktiviti mengurus rumah tangga seperti menjaga anak, mengemas rumah dan tanggungjawab lain.

5. Impak Undang-Undang dan Peraturan :

a. Kepatuhan Undang-Undang : Kecederaan pekerjaan boleh membawa kepada tindakan undang-undang jika didapati tempat kerja tidak mematuhi peraturan dan standard keselamatan yang sepatutnya diamalkan.

b. Ligitasi : Dalam sesetengah kes, pekerja yang cedera mungkin mengambil tindakan undang-undang terhadap majikan di atas kecuaian ataupun ruang tempat kerja yang tidak selamat.

6. Impak Jangka Panjang

a. Isu Kesihatan Yang Kronik : Sesetengah kecederaan pekerjaan boleh mengakibatkan komplikasi kesihatan jangka panjang seperti sakit kronik dan kualiti hidup yang terjejas pada seseorang pekerja.

7. Impak Terhadap Kerjaya : Kecederaan serius memberi impak terhadap keupayaan pekerja untuk terus bekerja dalam pekerjaan yang sama lalu menyebabkan pekerja tersebut perlu menukar kerjaya yang lain.

Secara kesimpulan, kecederaan pekerjaan mempunyai impak yang negatif terhadap pekerja, majikan dan organisasi. Menurut Hollnagel et. al(2008) terdapat keperluan untuk memahami dan menentukan keadaan, ciri-ciri pekerja, persekitaran dan keadaan kerja yang boleh mengakibatkan kecederaan pekerjaan. Stemn et.al (2019) berpendapat kajian mengenai jenis-jenis kecederaan pekerjaan yang pernah berlaku pada masa lampau dan faktor yang menyumbang kepadanya boleh dijadikan sebagai ramalan kecederaan yang penting di tempat kerja. Iyaz et.al (2021) menegaskan bahawa pentingnya untuk memilih kombinasi yang betul bagi faktor-faktor yang mempengaruhi kecederaan pekerjaan yang digunakan sebagai input untuk model ramalan.

Apabila seseorang pekerja mengalami kecederaan di tempat kerja, seseorang pekerja ataupun majikan biasanya membuat tuntutan insurans dengan penyedia perkhidmatan insurans berdasarkan informasi yang didapati pada https://www.nidirect.gov.uk/articles/accidents-workplace. Kebiasaanya majikan akan menyediakan pampasan insurans kerja yang merupakan sebahagian daripada pakej faedah yang ditawarkan kepada pekerja. Insurans pampasan adalah bentuk insurans yang memberikan faedah kepada pekerja yang cedera atau jatuh sakit akibat ativiti atau tugasan yang berkaitan dengan pekerjaan. Insurans dan kecederaan pekerjaan adalah satu mekanisma penting dalam menyediakan sokongan dan pampasan kepada pekerja yang mengalami kecederaan semasa menjalankan tugas kerja mereka. Ini bertujuan untuk melindungi pekerja, menggalakkan keselamatan tempat kerja dan memastikan pematuhan kepada regulasi dan undang-undang.

2.3 KAJIAN BERKAITAN DENGAN KOS INSURANS

Maka wujudlah pelbagai kajian untuk mengkaji kos tuntutan insurans mahupun perubatan berikutan kadar kenaikan kos penjagaan kesihatan yang semakin meningkat di seluruh dunia. Model Pembelajaran Mesin mempunyai keupayaan untuk menggunakan sejumlah data yang lampau yang bersaiz besar dan pelbagai dimensi dan sumber dan mampu meningkatkan ketepatan dalam membuat ramalan atau anggaran bagi membantu dalam menghasilkan sebarang keputusan. Ch. Anwar ul Hassan et al.(2021) membuat ramalan mengenai Kos Insurans Perubatan menggunakan satu set model Pembelajaran Mesin Terselia yang terdiri daripada *Linear Regression, Stochastic Gradient Boosting, XGBoost, Support Vector Regression, K-Nearest Neighbours, Ridge Regressor, Decision Tree, Random Forest Regressor* dan *Multiple Linear Regression.* Keputusan ramalan yang dihasilkan setiap model ini dibandingkan dari segi ketepatan dan *root mean square error* (RMSE) dan model *Stochastic Gradient Boosting* dapat memberi keputusan ramalam Kos Insurans Perubatan dengan ketepatan 86% dengan nilai RMSE adalah 0.340.

Belisario Panay et al.(2019) membentangkan satu kajian mengenai satu kaedah *Regression* yang mempunyai ketelusan dan kebolehtafsiran berdasarkan teori Dempster-Shafer untuk meramal kos penjagaan kesihatan. Model ini telah diuji menggunakan data kesihatan dari Hospital Tsuyama Chou Jepun dan ternyata model *Artifical Neural Network* dan *Gradient Boosting* mempunyai prestasi utk membuat ramalan dengan bacaan R2 bersaman 0.44.

J.Pesantez-Narvaez (2019) et al. menilai kemampuan model Pembelajaran Mesin *Logistic Regression* berbanding XGBoost dalam meramal tuntutan insurans motor dengan menggunakan telematics data. Berdasarkan penemuan melalui kajian ini mendapati bahawa model *logistic regression* lebih sesuai untuk dijadikan model ramalan berikutan kemampuan model membuat kebolehtafsiran dan model tersebut mempunyai kebolehan membuat ramalan yang tepat. Manakala model XGBoost pula memerlukan banyak prosidur *fine-tuning* untuk mendapatkan ramalan yang sesuai.

Kerajaan India menyedari peningkatan perbelanjaan kesihatan berikutan peningkatan jangka hayat penduduk India dan peralihan daripada penyakit pandemik

ke arah penyakit yang tidak berjangkit lantas menjadikan insurans sebagai satu keperluan bagi semua rakyat. Untuk itu kerajaan India membangunkan algoritma komputer dan Pembelajaran Mesin untuk mengkaji dan menganalisa data insurans yang lampau dan meramalkan output yang baru berdasarkan trend tingkah laku pelanggan, polisi insurans dan mendorong kepada suatu proses pembuatan keputusan berasaskan data dalam melahirkan suatu skim insurans yang baharu. Sistem Insurans Kesihatan India juga berharap dengan model ramalan ini ianya dapat meningkatkan operasi dan perkhidmatan insurans. Justeru itu, kajian yang dilaksanakan oleh Sudhir Panda et. Al (2022) membangunkan sistem ramalan harga kos insurans secara *real time* dan sistem ini dinamakan sebagai Sistem Ramalan Insurans Kesihatan Pembelajaran Mesin (MLHIPS) di mana MLHIPS menggunakan algoritma pembelajaran mesin. Algoritma-algoritma yang digunakan sepanjang kajian ini adalah *Ridge Regression, Lasso Regression, Simple Linear Regression, Multiple Linear Regression* dan *Polinomial Regression*. Antara kebanyakan algoritma yang dibangunkan, model *Polinomial Regression* mencapai keputusan yang terbaik dengan nilai RMSE bersamaan 5100.53 dan nila R2 bersamaan 0.80. Oleh itu model Polinomial Regression menjadi asas bagi sistem MLHIPS dan MLHIPS akan membantu syarikat insurans yang berada dalam pasaran untuk menentukan nilai premium insurans dengan cara yang mudah dan pantas dan dapat mengurangkan perbelanjaan kesihatan daripada pihak kerajaan.

Satu kajian empirikal berkenaan Pembelajaran Mesin Regresi untuk meramalkan kos insurans kesihatan telah dicetuskan oleh Y.Angeline Christobel et. al (2022). Kajian ini menggunakan teknik *regression* yang terdiri daripada *Linear Regression, Ridge Regression, Lasso Regression* dan *Polynominal Regression* untuk menganggar kos insurans menggunakan data kesihatan yang berbentuk peribadi. Hasil daripada eksperimen yang dijalankan, *Polinominal Regression* merupakan kaedah terbaik untuk membuat ramalan dengan bacaan ketepatan adalah 88% dan skor R2 yang tertinggi manakala bacaan RMSE yang terendah. Faktor seperti umur, jantina dan sejarah kesihatan ditemui sebagai faktor yang mempengaruhi kos insuran kesihatan melalui kajian ini.

Angela D.Kafuris (2022) membangunkan model ramalan untuk mengira harga premium insurans kesihatan menggunakan algoritma Pembelajaran Mesin. Kajian ini

dibangunkan berikutan terdapat keperluan yang tinggi bagi syarikat insurans untuk membangunkan model yang dapat mengira perbelanjaan perubatan yang tepat untuk keseluruhan populasi dan sekaligus dapat mengurangkan beban orang berinsurans untuk mengeluarkan duit poket sendiri. Algoritma yang digunakan sepanjang kajian ini adalah terdiri daripada *Multiple Linear Regression(MLR), K-Nearest Neighbors(KNN), Least Absolute Shrinkage and Selection Operator (LASSO), Extreme Gradient Boosting (XGBoosting)* dan *Random Forest Regression (RFR).* Prestasi algoritma tersebut dinilai dan didapati algoritma *XGBoosting* adalah model yang terbaik untuk membuat ramalan dengan bacaan R2 bersamaan 85.5%, MAE bersamaan 2688.2, RMSE bersamaan 4748.7 dan algoritma RFR mencatatkan bacaan R2 bersamaan 85.3%, MAE bersamaan 27266.4 dan RMSE bersamaan 4783.8. Kajian ini juga mendapati pembolehubah yang signifikan dalam menentukan kos insurans adalah umur, BMI, status merokok dan *region*. Justeru itu, mana-mana syarikat insurans yang memerlukan model ramalan untuk mengira harga premium insurans yang tepat disarankan untuk menggunakan XGBoosting dan RFR.

Menyedari akan hakikat kos penjagaan kesihatan semakin meningkat hari ke hari disebabkan oleh kemunculan pelbagai virus dalam kehidupan manusia menyebabkan Laksmanarao et. al(2020) melakukan analisis untuk menganggar kos perubatan tersebut. Laksmanarao et. al(2020) memanfaatkan algoritma pembelajaran mesin seperti *Multiple Linear Regression, Support Vector Regression, Decision Tree Regression* dan *Random Forest Regression* ke atas set data yang didapati dari Kaggle yang terdiri dari umur, jantina, status merokok, BMI, bilangan anak dan kawasan. Hasil kajian mendapati *Random Forest Regression* dapat memberikan hasil ramalan yang terbaik berdasarkan data jantina dalam membantu kerajaan membuat keputusan berkaitan dengan isu kesihatan.

Mohammad Amin Morid et. al (2017) mengutarakan kepentingan untuk mempunyai *tool* untuk mengawal kos perubatan dengan keupayaan untuk menganggar kos perubatan secara individu dengan tepat. Kajian membuat tiga perbandingan kajian literatur sistematik yang terdiri daripada kos ramalan menggunakan *non-cost predictors*, kos ramalan secara *bucket* dan kos ramalan menggunaka*n cost predictors.* Menerusi kajian literatur yang dibuat, pendekatan yang paling kerap digunakan dalam

bidang kesihatan adalah kos ramalan menggunakan *cost predictors* seperti kaedah Pembelajaran Mesin Terselia. Kaedah Pembelajaran Mesin Terselia yang digunakan sepanjang kajian ini adalah *Gradient Boosting, Artificial Neural Network, Ridge, Support Vector Machine, Elastic Net*. Kaedah *Gradient Boosting* didapati sebagai model ramalan yang terbaik untuk keseluruhan ramalan samada dari kos insurans yang rendah mahupun tinggi, manakala kaedah *Artificial Neural Network (ANN)* berprestasi untuk mengira kos insurans yang tinggi sahaja.

Keshav Kaushik et. al (2022) memanfaatkan penggunaan *Artificial Neural Network(ANN)* yang berasaskan rangka kerja model *regression* untuk meramal insurans premium kesihatan. Bisnes insurans menggunakan pendekatan pembelajaran mesin supaya dapat menawarkan pelanggan sesuatu pakej perlindungan insurans yang cepat, tepat dan efisien. Untuk kajian ini model ANN dibandingkan dengan model *Linear Regression(LR)* dan diuji pada 1 juta set data. Keputusan yang didapati adalah model ANN dapat memberi ramalan yang lebih tepat dengan ketepatan yang diperolehi adalah 92.72%.

C.Yang et al (2018) membincangkan satu kajian yang menggunakan pendekatan Pembelajaran Mesin untuk menjangkakan perbelanjaan perubatan bagi pesakit yang berada di Unit *High Cost High Need(HCHN).* Kajian ini dijalankan di Negara China menggunakan data Program Texas Medicaid. Menurut C.Yang et al(2018) memilih pendekatan *Linear Regression*(LR) kerana ianya banyak digunakan dalam model ramalan *Regularized Regression*(LASSO) pula dipilih kerana ianya digunakan secara meluas sebagai pendekatan yang *default* dalam kebanyakan tugas-tugasan pembelajaran mesin. Pendekatan *Gradient Boosting Machine*(GBM) juga dipilih dalam kajian ini berikutan model tersebut mampu mengendalikan pembolehubah input berbentuk dimensi tinggi. Model terakhir yang dipilih adalah *Reccurent Neural Net*(RNN) yang merupakan pendekatan yang berbentuk *deep learning* yang dapat menangani pelbagai tugasan yang berjujukan seperti pengecaman bahasa pertuturan. Antara keempat-empat model yang dipilih, RNN mempamerkan prestasi terbaik dalam memberi ketepatan meramal kos perubatan pesakit HCHN berbanding model LR, LASSO dan GBM. Akan tetapi jika diukur dari sudut kebolehtafsiran, model LASSO dan GBM secara konsisten memilih pembolehubah yang sama dan menjana sumbangan

yang stabil tanpa perlu bergantung kepada pensampelan semula. Justeru itu, secara kesimpulannya kajian ini berjaya menunjukkan kolerasi temporal yang signifikan dalam perbelanjaan perubatan pesakit. Model-model pembelajaran mesin banyak membantu untuk meramalkan perbelanjaan perubatan dengan lebih tepat. Melalui hasil kajian ini, ianya dapat membawa bidang perubatan lebih maju ke depan dalam usaha menurunkan kadar kos penjagaan kesihatan secara keseluruhan dan dapat menyampaikan perkhidmatan penjagaan kesihatan dengan lebih efektif.

Sam Goundar et. al (2020) pula memfokuskan kajian menggunakan *Artificial Neural Network* yang terdapat dalam Pembelajaran Mesin untuk meramal insurans kesihatan. Kajian ini membangunkan dua kaedah yang berasaskan *Artificial Neural Network* yang mana Model 1 adalah *Feed Forward Neural Network* dan Model 2 adalah *Reccurent Neural Network.* Model-model ini dikatakan sesuai untuk meramal kos tuntutan perubatan yang dijangkakan. Kesimpulan, model *Artificial Neural Network* yang telah diimplemen terbukti sebagai alat yang berkesan untuk meramal jangkaan tuntutan kos perubatan tahun bagi BSP Life. Model *Reccurent Neural Network* berjaya mengatasi Model *Feed Forward Neural Network* apabila diukur dari segi ketepatan pengiraan yang diperlukan untuk menjalankan ramalan yang dikehendaki.

Satu kajian dijalankan oleh Anurag Yedla et. al (2020) untuk menguji kemampuan Teknik Pembelajaran Mesin dalam mewujudkan persekitaran kerja yang selamat dalam industri Perlombongan. Industri Perlombongan sememangnya diketahui umum yang ianya merupakan salah satu bidang pekerjaan paling berbahaya di dunia. Banyak kemalangan serius masih berlaku walaupun terdapat pelbagai usaha untuk mewujudkan persekitaran kerja yang selamat dalam industri tersebut. Namun usaha tersebut tidak berhenti di situ di mana melalui kajian ini teknik pembelajaran mesin seperti *Decision Teee, Random Forest* dan *Artificial Neural Network* digunakan untuk menganalisis meramal kemalangan berlaku di lombong dan juga untuk meramal hari yang tidak dapat hadir bekerja. Data yang digunakan untuk kajian ini diperoleh daripada pihak Pentadbiran Keselamatan dan Kesihatan Lombong yang mana ianya mengandungi data dalam bentuk *tabular*(berstruktur) dan *narratives.* Teknik sintentik augmentasi data menggunakan *word embedding* bagi data dalam *narratives* untuk menangani masalah ketidakseimbangan data dan teknik ini dapat menambahbaik skor

F1. Secara keseluruhan kajian ini mendapati model yang dilatih menggunakan data *narratives* menunjukkan prestasi yang lebih baik daripada model yang dilatih menggunakan data *tabular*(berstruktur) dalam meramalkan kemalangan. Untuk ramalan hari yang tidak dapat hadir bekerja pula model yang dilatih menggunakan model pada data *tabular*(berstruktur) mempunyai min ralat kuasa dua yang lebih rendah berbanding model yang dilatih pada data *narratives.*

Pekerja di Taman Negara dan Hutan Simpan Afrika Selatan mengalami kecederaan pada bahagian bawah anggota badan (*lower extremity),* bahagian *torso* dan tangan ataupun jari menyebabkan M.Chadyiwa et. al(2022) membuat satu kajian untuk menyiasat kecederaan tersebut menggunakan Pembelajaran Mesin. Algoritma Pembelajaran Mesin yang terdiri daripada *Support Vector Machine, K-Nearest Neighbours , XG Classifier* dan *Deep Neural Networks* dan data daripada Jabatan Dana Pampasan Pekerjaan dan Buruh digunapakai untuk membuat kajian ini. Model SVM menampilkan prestasi yang terbaik dalam meramal kecederaan pada bahagian bawah anggota badan (*lower extremity),* bahagian *torso,* tangan ataupun jari dan *features* jantina menunjukan *features* yang terpenting dalam menentukan ramalan ini. Akan tetapi kajian ini menyarankan agar lebih banyak *feature*s dapat digunakan dalam ramalan akan datang berikutan hanya empat *feature*s yang digunakan iaitu syarikat, jantina, tahun kemalangan dan umur pada kemalangan berlaku.

Membuat ramalan kemalangan di tempat kerja menggunakan teknik automatik telah membuka lebih banyak ruang dan potensi dalam penyelidikan dan kajian yang berasaskan bukti. Oleh itu, Divya Sukumar et. al(2020) membentangkan projek kajian PhD dalam meramal kemalangan tempat kerja menggunakan beberapa algoritma pembelajaran mesin seperti *Random Forest, K-Nearest Neighbour dan Decision Tree.* Kajian menggunakan data kemalangan dan kecederaan daripada Kaggle dan hasil kajian mendapati model *Decision Tree* mempunyai prestasi tertinggi di antara ketiga-tiga model tersebut. Ia juga membuktikan bahawa pekerja dan tempat kerja menyumbang kepada kemalangan yang berlaku di tempat kerja. Terdapat beberapa limitasi dalam kajian ini berikutan tiada data mengenai tahap keseriusan kecederaan, majikan dan industri mana yang terlibat dalam kemalangan tersebut.

Peter H F Ng et. al (2023) memperkenalkan *Smart Work Injury Management System* (SWIM) untuk menganggarkan cuti sakit dan pelan pemulihan atau rehabilitasi apabila seseorang pekerja terlibat dalam kemalangan tempat kerja. SWIM ini dibangunkan berasaskan pendekatan Pembelajaran Mesin iaitu *K-Nearest Neighbours* yang dilatih menggunakan data berstruktur dan data tidak berstruktur. Model ini didapati dapat meramal cuti sakit dan pelan rehabilitasi dengan baik di mana ketepatan yang diperolehi mencapai 90%. Kemudian kajian ini membangunkan *dashboard* sistem yang interaktif untuk memaparkan keputusan model Pembelajaran Mesin yang diperolehi kepada pengguna.

Kajian kos tuntutan insurans tidak terhad terhadap insurans kesihatan sahaja malah ianya telah dilakukan secara meluas sehingga tuntutan insurans kereta atau automotif di mana objektif utama kajian yang diajalankan oleh Shady Abdelhadi et. al (2020) adalah untuk membangunkan satu model yang dapat mengira kos insurans kereta yang tepat menggunakan teknik pembelajaran mesin yang fokus kepada statistik dan algoritma pembelajaran mesin yang dapat mengurus *missing values* dalam data. Terdapat banyak *missing values* dalam data kos insurans kereta dalam dunia yang sebenar dan ini akan menyebabkan berat sebelah dalam membuat keputusan berdasarkan tersebut. Algoritma yang digunakan untuk kajian ini adalah *Artificial Neural Network (ANN), Decision Tree (DT), Naive Bayes* dan *XGBoost* untuk membangunkan model ramalan dan ternyata *XGBoost* dapat mencapai ketepatan yang tertinggi iaitu 92.53% berbanding dengan model lain.

Manakala trend peningkatan kekerapan tuntutan auto insurans yang teruk memerlukan satu kaedah untuk memfailkan kes tuntutan dengan cepat dan tepat. Kaedah yang dikenalpasti adalah menggunakan Kaedah Pembelajaran Mesin kerana kaedah ini menganggap masalah lambakan data tuntutan yang banyak adalah sebagai suatu Pembelajaran Terselia. Tambahan pula, kebiasaanya terdapat data yang tidak sempurna atau hilang di dalam sesuatu tuntutan. Justeru itu Muhammad Arief Fauzan (2018) menyatakan keperluan utk menggunakan Pembelajaran Mesin yang dapat mengurus keadaan data-data yang tidak sempurna ini untuk meramal kes tuntutan insurans. Algoritma XGBoost dipilih sebagai algoritma Pembelajaran Mesin yang paling utama untuk kajian ini. XGBoost merupakan merupakan satu teknik algoritma

Pembelajaran Mesin Terselia yang bersifat *ensemble* baharu. Untuk menguji ketepatan algoritma XGBoost, kajian ini membuat perbandingan dengan algoritma ensemble yang lain seperti *AdaBoost, Stochastic Gradient Boosting, Random Forest* dan *Neural Network* dan ternyata *XGBoost* dapat memberikan ketepatan dalam membuat ramalan berbanding teknik lain.

Satu kajian khas yang dilaksanakan oleh Shamshimah Shamsuddin et. al (2018) menggunakan data tuntutan kecederaan yang diperolehi dari PERKESO. Kajian ini bertujuan untuk mengenalpasti trend ketidakupayaan seseorang pekerja yang dilindungi di bawah Skim Kecederaan Pekerjaan PERKESO bagi tahun 2009 sehingga 2013. Untuk menganalisis trend tersebut kajian ini menggunakan kaedah analisis deskriptif dan mendapati 80% daripada jumlah keseluruhan pekerja yang mengalami kecederaan atau kematian di tempat kerja adalah pekerja lelaki. Selain itu, kajian ini mendapati pekerja yang berumur 25 tahun ke atas lebih terdedah kepada kecederaan dan kematian yang berkaitan dengan pekerjaan. Secara keseluruhan, kajian menunjukkan bahawa faktor sosio-demografi seperti jantina, umur dan tahap ketidakupayaan merupakan faktor beberapa faktor penting yang perlu diambil kira dalam mengenalpasti trend tersebut.

Pada tahun yang sama iaitu pada tahun 2018 Shamshimah Shamsuddin et. al (2018) menjalankan satu lagi kajian yang juga menggunakan data tuntutan dari PERKESO bagi tahun 2009 sehingga 2013. Kajian ini berkenaan pampasan yang perlu dibayar oleh PERKESO terhadap kes hilang upaya yang dialami oleh seseorang pekerja. Kaedah analisis deskriptif juga digunakan untuk menganalisa trend dan didapati pekerja aktif dan jumlah pampasan yang perlu di bayar oleh PERKESO saban tahun makin meningkat. Berdasarkan data Hilang Upaya Sementara (HUS), jumlah hari bekerja yang hilang akibat kemalangan yang dialami adalah berbeza mengikut jantina dan umur di mana pekerja lelaki lebih tinggi perempuan dan pekerja muda(25 -29 tahun) lebih ramai daripada pekerja berumur. Penemuan ini memberi petunjuk kepada pihak yang terlibat seperti pekerja, majikan dan kerajaan bahawa kesihatan pekerja mendatangkan kesan ketara kepada perbelanjaan kesihatan dan boleh memberi kesan besar kepada ekonomi negara.

Jadual 2.1 Kajian Lepas Mengenai Pembelajaran Mesin Dalam Domain Insurans

NO	PENGARANG	TAHUN	METODOLOGI	SET DATA	MODEL DIPILIH
1	Ch. Anwar ul Hassan et al.	2021	*Linear Regression, Stochastic Gradient Boosting, XGBoost, Support Vector Regression, K-Nearest Neighbours, Ridge Regressor, Decision Tree, Random Forest Regressor* dan *Multiple Linear Regression.*	Data Insurans dari Kaggle https://www. kaggle.com/mirichoi 0218/insurance.	*Stochastic Gradient Boosting*
2	Belisario Panay et al	2019	*Regression, Artifical Neural Network* dan *Gradient Boosting*	Rekod Kesihatan Penduduk japan dari Hospital Tsuyama Chou	*Artifical Neural Network, Gradient Boosting*
3	J.Pesantez-Narvaez	2019	*Logistic Regression,* XGBoost	Data Telematic	*Logistic Regression*
4	Sudhir Panda et. Al (2022)	2022	*Linear Regression, Ridge Regression, Lasso Regression* dan *Polynominal Regression*	Data insurans India	*Polynominal Regression*
5	Y.Angeline Christobel et. al	2022	*Linear Regression, Ridge Regression, Lasso Regression* dan *Polynominal Regression*	Data dari Kaggle https://www.kaggle. com/mirichoi0218/i nsurance	*Polinominal Regression*
6	Angela D.Kafuris	2022	*Multiple Linear Regression(MLR), K-Nearest Neighbors(KNN), Least Absolute Shrinkage and Selection Operator (LASSO), Extreme Gradient Boosting (XGBoosting)* dan *Random Forest Regression (RFR)*	Data dari Kaggle https://www.kaggle. com/mirichoi0218/i nsurance	*Extreme Gradient Boosting (XGBoosting)* dan *Random Forest Regression (RFR)*
7	Laksmanarao et al(2020)	2020	*Multiple Linear Regression, Support Vector Regression, Decision tree Regression* dan *Random Forest Regression*	Data dari Kaggle https://www.kaggle. com/mirichoi0218/i nsurance	*Random Forest Regression*

bersambung…

…sambungan

8	Mohammad Amin Morid et. al	2017	*Gradient Boosting, Artificial Neural Network, Ridge, Support Vector Machine, Elastic Net*	Universiti Utah – Pelan Kesihatan 1 juta set data	*Gradient Boosting*
9	Keshav Kaushik 1 et al (2022)	2022	*Artificial Neural Network, Linear Regression(LR)*	Data dari Kaggle https://www. kaggle.com/datasets/ noordeen/insurance-premium-prediction	*Artificial Neural Network*
10	C.Yang et al	2018	*Linear Regression*(LR), *Regularized Regression, Gradient Boosting Machine, Reccurent Neural Net*	Unit *High Cost High Need(HCHN)* yang berdaftar Program Texas MediAid, China	*Regularized Regression, Gradient Boosting Machine,*
11	Sam Goundar et. al	2020	*Artificial Neural Network - Feed Forward Neural Network, Artificial Neural Network - Reccurent Neural Network*	BSP Life (Fiji)	*Artificial Neural Network - Reccurent Neural Network*
12	Anurag Yedla et.al	2020	*Decision Teee, Random Forest* dan *Artificial Neural Network*	Data daripada pihak Pentadbiran Keselamatan dan Kesihatan Lombong	Data *narratives ramal kemalangan.* Data *tabular* ramal hari tidak dapat hadir bekerja
13	M.Chadyiwa et. al	2022	*Support Vector Machine, K-Nearest Neighbours , XG Classifier* dan *Deep Neural Networks*	Jabatan Dana Pampasan Pekerjaan dan Buruh, Afrika	*Support Vector Machine*
14	Divya Sukumar et. al	2020	*Random Forest, K-Nearest Neighbour dan Decision Tree.*	Kaggle	*Decision Tree*
15	Peter H F Ng et. al	2023	*K-Nearest Neighbours*	Data Berstruktur, Data Tidak Berstruktur.	*K-Nearest Neighbours*

bersambung…

…sambungan

16	Shady Abdelhadi et. al	2020	*Artificial Neural Network (ANN), Decision Tree (DT), Naive Bayes* dan *XGBoost*	Data dari Kaggle https://www.kaggle.com/c/porto-seguro-safe- driver-prediction	*XGBoost*
17	Muhammad Arief Fauzan	2018	*AdaBoost, Stochastic Gradient Boosting, Random Forest* dan *Neural Network*	Data dari Kaggle https://www.kaggle.com/c/porto-seguro-safe- driver-prediction	*XGBoost*
18	Shamshimah Samsuddin et. al	2018	Analisis Deskriptif	Data dari Pertubuhan Keselamatan Sosial, Malaysia	Analisis Deskriptif
19	Shamshimah Samsuddin et. al	2018	Analisis Deskriptif	Data dari Pertubuhan Keselamatan Sosial, Malaysia	Analisis Deskriptif

2.4 PEMBELAJARAN MESIN

Majumdar et.al (2022) menyatakan bahawa model pembelajaran mesin mempunyai tahap fleksibiliti yang tinggi berbanding kaedah tradisional model statistik kerana keupayaan meramalkan hubungan kompleks dengan pelbagai dimensi yang bukan linear atau *additive.*

Pembelajaran Mesin ini adalah merupakan subset kepada Kecerdasan Buatan (AI) yang melibatkan pembangunan dan model yang membolehkan komputer belajar daripada data dan membuat ramalan atau keputusan berdasar pembelajaran tersebut. Algoritma pembelajaran mesin boleh mengenalpasti corak, hubungan dan trend dalam data tersebut. Pembelajaran Mesin juga mampu meningkatkan prestasi dari masa ke semasa kerana ia telah terdedah kepada banyak jenis data.

Teras idea di sebalik pembelajaran mesin adalah untuk membolehkan komputer mengenali corak, memahami perhubungan yang kompleks dan membuat keputusan

atau ramalan tanpa campur tangan dari manusia. Algoritma pembelajaran mesin menggunakan teknik statistik untuk belajar daripada data yang bersaiz besar dan tidak perlu mengikut peraturan spesifik yang diprogramkan oleh manusia.

Pembelajaran mesin memainkan peranan yang signifikan dalam bidang Kecerdasan Buatan. Pembelajaran Mesin berjaya menarik minat para penyelidik untuk mengkaji mengenainya dan ianya berjaya diimplementasi dan diadaptasi ke dalam bidang perubatan, pendidikan, membuat perangkaan atau ramalan dan sebagainya.

2.5 KESIMPULAN

Secara kesimpulannya, kajian literatur yang komprehensif mendedahkan bahawa terdapat trend yang ketara di pelbagai negara di mana kajian atau penyelidikan telah menerokai penggunaan Pembelajaran Mesin dalam sektor insurans. Berdasarkan pada kajian literatur juga dapat membuktikan bahawa penggunaan teknologi Pembelajaran Mesin seperti *Linear Regression, Stochastic Gradient Boosting, Support Vector Regression, Artificial Neural Netweok, Decision Tree*, *Random Forest, Naive Bayes* dan sebagainya boleh diaplikasikan dan dilaksanakan dan dapat memberi maanfaat kepada sektor insurans.

Selain itu, penggunaan teknik Pembelajaran Mesin seperti algoritma-algoritma ensemble dan model berdasarkan algoritma *Neural Network* menunjukkan potensi yang besar dalam meramal dan mengurus kos tuntutan insurans perubatan dan kecederaan tempat kerja. Kajian-kajian literatur menunjukkan bahawa Pembelajaran Mesin menjadi alat yang kuat untuk meramal dan mengurus kos tuntutan insurans dalam pelbagai konteks termasuk insurans kesihatan dan kereta. Walaubagaimanapun keputusan yang paling baik yang dihasilkan oleh Pembelajaran Mesin bergantung kepada kesesuaian model atau algoritma Pembelajaran Mesin yang digunakan. Dalam sesetengah kajian, model seperti XGBoost telah tebukti sebagai yang terbaik manakala dalam kajian lain pula mendapati model Neural Network adalah yang terbaik. Oleh itu adalah penting untuk memilih model yang sesuai dengan data matlamat kajian.

Melalui kajian literatur ini juga dapat diperhatikan antara faktor penting dalam membuat ramalan menggunakan Pembelajaran Mesin ini adalah data yang digunakan

sebagai input. Prinsip asas Pembelajaran Mesin adalah mempelajari dari set data yang besar untuk mengenalpasti sebarang bentuk trend, hubungan atau corak yang wujud.

Kajian sektor insurans atau kecederaan pekerjaan yang sebelum ini khususnya di Malaysia tidak menggunapakai teknik pembelajaran mesin dan hanya menggunakan kaedah analisa deskriptif sebagaimana kajian yang dijalankan oleh Shamimah Samsuddin et.al (2018). Ini menunjukkan penggunaan metodologi canggih seperti Pembelajaran Mesin belum diterokai. Ketiadaan pengunaan pembelajaran mesin dalam industri insurans Malaysia khususnya di PERKESO akan membuka ruang dan peluang baharu untuk diterokai. Ia jugak dapat membantu organisasi dan pengurusan dengan mempertingkatkan ketepatan ramalan, meningkatkan keberkesanan dalam membuat keputusan yang lebih baik dalam pengurusan risiko dan perbelanjaan.

Dalam era pembuatan keputusan yang berasaskan data ini, integrasi antara sektor insurans dan pembelajaran mesin dapat membawa banyak kebaikan dan kelebihan yang ketara. Oleh itu adalah penting bagi PERKESO yang mentadbir urus Keselamatan Sosial bagi negara Malaysia untuk mengorak langkah dan mengikuti perkembangan teknologi ini.

BAB III

METODOLOGI KAJIAN

3.0 PENGENALAN

Metodologi Kajian yang digunakan dalam kajian ini adalah merupakan asas permulaan di mana keseluruhan kajian terletak. Ia memainkan peranan penting dalam membolehkan analisis dan tafsiran data yang tepat dan seterusnya membawa kepada pencapaian matlamat kajian.

Dalam bahagian berikut, metodologi kajian akan didedahkan dengan detail termasuk pengumpulan data, praprompresesan, pemilihan dan penilaian model. Setiap langkah telah direka bentuk dengna teliti untuk menyelarasakan dengan objektif kajian dan untuk memastikan kesahihan dan kobolehpercayaan penemuan kajian ini.

3.1 PENGUMPULAN DATA

Data berfungsi sebagai bahan mentah kepada Pembelajaran Mesin di mana algoritma Pembelajaran Mesin akan belajar, menyesuaikan diri dan akhirnya menjana hasil yang bermakna dari data tersebut. Justeru itu, proses mengenalpasti data yang berpotensi menjawab persoalan dan memenuhi objektif kajian yang telah ditetapkan adalah penting. Data yang dikumpul untuk kajian ini hendaklah data yang boleh dipercayai supaya pembelajaran mesin boleh mencari corak data yang betul. Kualiti data yang diberikan kepada pembelajaran mesin untuk diproses juga akan menentukan ketepatan model ramalan.

Selepas diteliti dan dianalisa data yang sesuai untuk kajian ini adalah berkenaan data tuntutan insurans PERKESO untuk mengkaji sebarang corak atau trend. Permohonan secara formal telah dibuat kepada Bahagian Strategi dan Transformasi, PERKESO untuk mendapatkan data ini. Melalu permohonan tersebut, objektif kajian dan saiz data yang diperlukan telah dinyatakan dengan jelas. Di samping itu, manfaat yang akan diperolehi dari hasil kajian ini disertakan di dalam borang permohonan untuk mendapatkan data tuntutan insurans bagi tahun 2017 sehingga tahun 2020.

Jadual 3.1 Senarai data yang dipohon untuk dijadikan sumber kajian

No	Atribut	Jenis	Penerangan
1	Jantina	String	Jantina
2	Umur (TLAPOR – TLAHIR)	Numerik	Umur dari Tarikh Lapor Kemalangan
3	Umur(TKHAKRU – TKHLAHIR)	Numerik	Umur
4	Jenis Kes	String	Jenis Kes Tuntutan Insurans
5	PNPPP	String	Pejabat
6	Kod Industri	Numerik	Kod Industri OB
7	Deskripsi Industri	String	Penerangan Kod Industri
8	Kod Lokasi Kecederaan	Numerik	Kod Lokasi Kecederaan OB
9	Deskripsi Lokasi Kecederaan	String	Penerangan Lokasi Kecederaan
10	Kod Jenis Kemalangan	Numerik	Kod Jenis Kemalangan OB
11	Deskripsi Jenis Kemalangan	String	Penerangan Jenis Kemalangan OB
12	Kod Sebab Kemalangan	Numerik	Kod Sebab Kemalangan OB
13	Deskripsi Sebab Kemalangan	String	Penerangan Sebab Kemalangan OB
14	Tempoh MC	Numerik	Tempoh Cuti Sakit OB
15	Amaun Bayaran	Numerik	Amaun tuntutan insurans

Berdasarkan jadual 3.1 terdapat 15 atribut yang diterima daripada data yang diberikan oleh PERKESO kesemua data tersebut berada dalam fail berbeza mengikut tahun 2017 sehingga 2020. Fail-fail tersebut digabungkan menjadi satu fail dan total bilangan data tuntutan yang diterima adalah 224,548 bagi tahun 2017 sehingga 2020.

Data industri yang diperolehi berada dalam bentuk kod sub industri dan data ini perlu dipetakan kepada kod industri utama untuk memudahkan pemahaman dan pengkelasan. Pemetaan ini dilakukan sebagaimana pemetaan kod industri yang selalu dipraktiskan di PERKESO.

Atribut Kod Industri berserta Deskripsi mengandungi data seperti jadual berikut :

Jadual 3.2 Senarai data bagi atribut Industri

No	Kod Industri	Deskripsi Industri
1	01	Pertanian, Perhutanan dan Perikanan
2	02	Perlombongan dan Pengkuarian
3	03	Pembuatan
4	04	Perkhidmatan Elektrik, Gas, Air & Kebersihan
5	05	Pembinaan
6	06	Perdagangan
7	07	Penginapan dan Aktiviti Perkhidmatan Makanan Minuman
8	08	Pengangkutan dan Penyimpanan
9	09	Aktiviti Kewangan dan Insurans/Takaful
10	10	Aktiviti Hartanah, Penyewaan dan Perniagaan
11	11	Pentadbiran Awam dan Pertahanan/Aktiviti Keselamatan Wajib
		Pendidikan
		Kesihatan dan Kerja Sosial
		Aktivit Perkhidmatan Komuniti, Sosial dan Persendirian Lain
		Isi Rumah Persendirian dengan Pekerja Bergaji
		Aktiviti Badan dan Pertubuhan Luar Wilayah
12	12	Aktiviti yang tidak dapat kenalpasti

Atribut Kod Lokasi Kemalangan berserta Deskripsi Lokasi Kemalangan mengandungi data seperti jadual berikut :

Jadual 3.3 Senarai data bagi atribut Lokasi Kecederaan berserta deskripsi

No	Lokasi Kecederaan	Deskripsi Lokasi Kecederaan
1	11	Sekitar tempurung kepada/otak
2	12	Mata
3	13	Telinga
4	14	Mulut
5	15	Hidung
6	16	Muka,lokasi tidak dinyatakan
7	18	Kepala, lokasi berganda (*multiple locations*)
8	19	Kepala.lokasi tidak dinyatakan
9	20	Leher / *Neck*
10	31	Belakang/*Back*
11	32	Dada/*Chest*
12	33	Abdomen/*Abdoment*
13	34	Tulang punggung/*pelvis*
14	38	Tubuh lokasi berganda (*Trunk, multiple locations*)
15	39	Tubuh lokasi tidak dinyatakan (*Trunk, unspecified location*)
16	41	Bahu/*Shoulder*
17	42	Lengan Atas / *Upper Arm*
18	43	Siku / *Elbow*

bersambung…

…sambungan

19	44	Siku ke pergelangan tangan / *Forearm*
20	45	Pergelangan tangan / *Wrist*
21	46	Tangan (kecuali jari)
22	47	Jari / *Fingers*
23	48	Bahagian Atas Anggota Badan, Lokasi Anggota Badan Berganda
24	49	Bahagian Bawah Anggota Badan, Lokasi Anggota Badan Tidak dinyatakan
25	51	Pinggul / Pangal Peha / *Hip*
26	52	Paha / *Thigh*
27	53	Lutut / *Knee*
28	54	Kaki / *Leg*
29	55	Pergelangan Kaki / *Ankle*
30	56	Kaki / Feet (*except toes alone*)
31	57	Jari Kaki / *Toes*
32	58	Bahagian Bawah Anggota Badan, lokasi anggota badan berganda
33	59	Bahagian Bawah Anggota Badan, lokasi anggota badan tidak berganda
34	61	Kepala dan tubuh, kepala dan satu atau lebih anggota badan / *Head and trunk, head and one or more limbs*
35	62	Tubuh dan satu atau lebih anggota badan / *Trunk and one or more limbs*
36	63	Satu bahagian atas anggota badan, satu bahagian bawah anggota badan atau lebih / *One upper limb, one lower limb or more than two*
37	68	Lokasi Anggota Badan Berganda Yang Lain / *Other multiple locations*
38	69	Anggota Badan Berganda,Lokasi Anggota Badan Tidak Dinyatakan /*Multiple locations, unspecified locations*
39	71	Sistem perjalanan darah /*Circulatory system in general*
40	72	Sistem pernafasan /*Respiratory system in general*
41	73	Sistem penghadaman /*Digestive system in general*
42	74	Sistem saraf / *Nervous system in general*
43	78	Kecederaan am lain / *Other general injuries*
		Kecederaan am, lokasi tidak dinyatakan /*General injuries, unspecified locations*

Atribut Kod Jenis Kemalangan berserta Jenis Kemalangan mengandungi data seperti jadual berikut :

Jadual 3.4 Senarai data bagi atribut Jenis Kemalangan berserta deskripsi

No	Kod Jenis Kemalangan	Deskripsi Jenis Kemalangan
1	10	Keretakan / *Fractures*
2	20	Dislokasi / *Dislocations*
3	25	Tergeliat dan terseliuh / *Sprains and strains*
4	30	Hentaman kuat dan cedera dalaman / *Concussions and other internal injuries*
5	40	Amputasi dan enukelasi / *Amputations and enucleations*
6	41	Kecederaan lain / *Other wounds*
7	50	Luka luaran / *Superficial injuries*
8	55	Kontusi dan kehancuran / *Contusions and crushings*
9	60	Terbakar/ *Burns*

bersambung…

…sambungan		
10	70	Terdedah kepada racun / *Acute poisonings*
11	80	Kesan cuaca / *Effects of weather, exposure and related conditions*
12	81	
13	82	Mati lemas / *Asphyxia*
14	83	Kesan elektrik / *Effects of electric currents*
15	90	Kesan radiasi / *Effects of radiation*
16	99	Kecederaan berganda / *Multiple injuries of different nature*
		Kecederaan lain dan tidak dinyatakan / *Other and unspecified injuries*

Atribut Kod Sebab Kemalangan berserta Jenis Sebab Kemalangan mengandungi data seperti jadual berikut :

Jadual 3.5 Senarai data bagi atribut Sebab Kemalangan berserta deskripsi

No	Kod Sebab Kemalangan	Deskripsi Sebab Kemalangan
1	11	Terjatuh dari Tempat Tinggi / Terjunam ke dalam lubang atau Jurang */Person Failing from height into pits/holes*
2	12	Terjatuh pada aras yang sama*/Person Failing from same level*
3	21	Tertimbus oleh tanah/batu/pasir/S*lides and cave-in under earth/sand*
4	22	Ditimpa bangunan runtuh,dinding atau tangga/*Collapsing of building wall or staircase*
5	23	Dihempap oleh benda yang jatuh semasa penyelenggaraan/*Struck by falling objects during handling*
6	24	Dihempap oleh benda yang jatih, tidak dispesifikasikan/*Struck by falling objects, unspecified*
7	31	Terpijak sesuatu objek*/Stepping on objects*
8	32	Terkena Objek Yang Pegun/*Striking Against Stationary Objects*
9	33	Terkena Objek Yang Bergerak/*Striking against moving object*
10	34	Terhempap oleh benda yang melayang/*Struck by flying object*
11	41	Tersepit di dalam objek/*Caught in an object*
12	42	Tersepit di antara objek pegun dan bergerak/*Caught between a stationary and moving objects*
13	43	Tersepit antara objek yang bergerak/*Caught between moving object*
14	51	Terseliuh apabila mengangkat Objek/*Over-exertion in lifting object*
15	52	Terseliuh apabila menolak/menarik objek/*Over-exertion in pushing or pulling objects*
16	53	Terseliuh semasa mengurus/melontar objek/*Over-exertion in handling or throwing object*
		Pergerakan yang berat/*Strenous Movement*
17	54	Terdedah kepada haba panas/*Exposure to hea*t
18	61	Terdedah kepada hawa dingin beku/*Exposure to cold*
19	62	Tersentuh objek/bahan panas/*Contact with objects/hot substance*s
20	63	Tersentuh objek/bahan sejuk/Contact with objects/cold substances

bersambung…

.... sambungan		
21	64	Terdedah/Tersentuh Elektrik/*Exposed to/Contact with Electric Current*
22	70	Terhidu/Terserap bahan merbahaya/*Contact by inhalation or absorption by harmful substance*
23	81	Terdedah kepada sinaran radiasi ion/*Expose to iniosing*
24	82	Terdedah kepada radiasi selain daripada radiasi ion/*Exposure to radiations othe than ionising radiations*
25	83	
26	91	Lain-lain kemalangan/*Other Type Of Accidents*

3.2 PRA-PEMPROSESAN DATA

Pra-pemprosesan data ialah kaedah untuk menganalisis, menapis, mengubah atau menukar kepada kod supaya algoritma Pembelajaran Mesin boleh memahami dan memproses untuk mengeluarkan output. Proses ini merupakan proses penyediaan dan pembersihan data mentah untuk menjadikannya sesuai untuk melatih model Pembelajaran Mesin. Kualiti data yang dimasukkan ke dalam algoritma Pembelajaran Mesin mempunyai impak yang signifikan ke atas prestasi dan ketepatan model seperti mana kata pepatah *"if garbage goes in, garbage goes out*".

Data yang diekstrak daripada scenario dunia sebenar akan wujud data yang kotor dan data yang tidak lengkap. Ini terjadi disebabkan oleh ralat manual, isu teknikal, peristiwa tidak dijangka berlaku atau disebabkan oleh pelbagai halangan yang lain. Data yang tidak lengkap atau kotor ini tidak boleh diproses oleh algoritma kerana biasanya algoritma tidak dicipta untuk mengendalikan data yang nilainya hilang atau data kotor yang menyebabkan gangguan kepada corak sebenar data tersebut. Pra-pemprosesan data ini bertujuan untuk menyelesaikan ini dengan rawatan menyeluruh terhadap data yang ada.

Data yang telah diperolehi akan diterokai atau diakses bagi mengesan sebarang bentuk corak atau sebarang ketidakkonsisten dalam data tersebut. Data tersebut akan melalui penilaian kualiti data antaranya :

1. **Mendapatkan gambaran keseluruhan data :** memahami format dan keseluruhan struktur di mana data tersebut disimpan. Cari sifat data tersebut

dengan mengeluarkan, min, median, standard kuantil dan sisihan piawai. Butiran ini membantu mengenalpasti sesuatu yang tidak betul.

2. **Kenalpasti data yang hilang :** data yang hilang dalam set data dunia sebenar adalah benda yang biasa berlaku. Ia boleh menggangu corak data sebenar malah boleh menyebabkan lebih banyak data lain yang hilang apabila keseluruhan baris dan lajur dikeluarkan disebabkan data yang hilang tersebut.

3. **Kenalpasti data yang terpencil atau anomali :** beberapa titik data berada jauh dari corak data utama. Titik-titik data ini adalah data terpencil dan mungkin perlu dibuang untk mendapatkan ramalan dengan ketepatan yang lebih tinggi melainkan algoritma tertentu dapat mengesan anomali.

Selepas Penilaian Kualiti Data diperolehi, data perlu dirawat melalui teknik berikut jika terdapat data yang tidak konsisten atau nilai data yang hilang :

1. **Gugurkan sampel data yang mengandungi nilai data yang hilang** : teknik ini syorkan untuk bilangan sampel data yang tinggi dan kiraan nilai data yang hilang dalam sampel juga tinggi. Teknik ini tidak disyorkan untuk sampel data yang rendah.

2. **Gantikan nilai yang hilang dengan sifar** : 0 boleh digunakan sebagai penggantian ke dalam set data apabila penggantian tersebut tidak mendatangkan kesan. Sebagai contoh dalam data bil telefon, nilai data yang hilang dalam amaun bilaangan boleh digantikan dengan 0 di mana ia menandakan pelanggan tidak melanggan pelan komunikasi pada bulan tersebut. Akan tetapi teknik ini tidak sesuai jika nilai 0 memberi makna kepada set data, sebagai contoh bacaan suhu bagi sesebuah sensor. Jika nilai bacaan suhu yang hilang ini digantikan dengan nilai 0, ia akan mengelirukan model.

3. **Gantikan nilai yang hilang dengan min, median atau mod** : menggunakan fungsi statistik seperti min,median atau mode sebagai penggantian kepada nilai

data yang hilang. Walaupun teknik ini adalah merupakan data andaian, akan tetapi nilai yang diandaikan lebih logik.

4. **Interpolasi nilai yang hilang** : interpolasi membantu menjana nilai dalam julat berdasarkan saiz langkah tertentu. Sebagai contoh jika terdapat 9 nilai yang hilang dalam lajur antara sel nilai 0 hingga 10, interpolasi akan mengisi sel yang hilang dengan nombor dari 1 hingga 9. Untuk teknik ini set data perlu disusun sebetulnya.

5. **Ekstrapolasi nilai yang hilang** : esktrapolasi membantu mengisi nilai yang berada di luar julat tertentu, seperti nilai ekstrem sesuatu atribut. Ekstrapolasi menggunakan bantuan dari pembolehubah lain(biasanya pembolehubah sasaran) untuk dijadikan bandingan dan rujukan.

3.3 MODEL PEMBELAJARAN MESIN

Pembelajaran Mesin yang mempunyai keupayaan untuk mendedahkan corak, membuat ramalan dan memberi informasi berharga daripada set data yang kompleks. Pembelajaran Mesin merupakan pendekatan yang ideal untuk menangani pelbagai cabaran yang ditemui semasa kajian ini. Dalam Pembelajaran Mesin terdapat pelbagai algoritma jadi pemilihan algoritma yang terbaik untuk set data yang diperolehi dan cuba menyelesaikan masalah kajian ini adalah merupakan perkara utama semasa membangunkan model ramalan. Empat model telah dipilih dalam kajian ini iaitu LR, DT, RF dan XGBoost.

3.3.1 Linear Regression(LR)

LR merupakan algoritma pembelajaran mesin terselia asas yang selalu digunakan untuk membangun model hubungan antara pembolehubah bergantung dengan satu atau lebih pembolehubah bebas untuk menyesuaikan persamaan linear terhadap data yang diperhatikan. Ianya adalah algoritma yang mudah dan senang untuk ditafsirkan yang mana kebiasaanya LR digunakan untuk membuat ramalan bagi nilai numerik.

Berikut adalah penguraian konsep utama yang berkaitan dengan LR merujuk kepada maklumat pada https://www.javatpoint.com/regression-analysis-in-machine-learning :

1. **Pembolehubah Bersandar (Sasaran) :** adalah pembolehubah yang ingin diramal. Ia dinamakan sebagai Y dan mewakili pembolehubah hasil ataupun respons.

2. **Pembolehubah Bebas :** adalah pembolehubah-pembolehubah yang digunakan untuk meramal pembolehubah bersandar. Ia dinamakan sebagai X dan bilangannya dalam jumlah satu ataupun lebih. Apabila terdapat hanya satu pembolehubah bebas, maka ianya dikenali sebagai *Single Linear Regression.* Jika terdapat banyak pembolehubah bebas, ianya dikenali sebagai *Multiple Linear Regression.*

3. **Persamaan Linear** : LR menganggap bahawa hubungan antara pembolehubah bebas X dan pembolehubah bersandar Y boleh diwakili oleh persamaan linear berikut :
   $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$
   β_0: Intercept (nilai Y apabila semua X adalah 0).
   β_1, β_2, ..., β_k: Pemalar atau kecerunan (menunjukkan perubahan dalam Y untuk setiap perubahan unit dalam setiap X yang berkaitan).
   ε: Pembolehubah ralat (mewakili perbezaan antara Y sebenar dan Y yang diramal).

4. **Objektif** : Matlamat LR adalah untuk mencari nilai-nilai pemalar (β_0, β_1, dan lain-lain) yang mengurangkan jumlah perbezaan kuasa dua antara Y sebenar dan Y yang diramal. Ini dikenali sebagai kaedah *least squared.*

5. **Jenis-Jenis Regresi Linear**
   Single Linear Regression. : Apabila hanya ada satu pembolehubah bebas.
   Multiple Linear Regression : Apabila terdapat banyak pembolehubah bebas.

Polinomial Regression : Sejenis regresi di mana hubungan antara X dan Y dimodelkan sebagai polinomial darjah n.

Ridge Regression dan *Lasso Regression :* Variasi regresi linear yang memperkenalkan pengekalan untuk mencegah penyesuaian berlebihan.

Kebolehtafsiran : LR menghasilkan keputusan yang boleh ditafsir oleh pengguna sebagaimana tafsiran *coefficients.* Sebagai contoh dalam *Single Linear Regression*, *coefficients* mewakili perubahan pada pembolehubah bersandar bagi sesuatu unit perubahan pada pembolehubah bebas.

Dalam set data kajian ini, pembolehubah bersandar adalah amaun kos tuntutan manakala pembolehuabh bebas adalah jantina, lokasi kecederaan, jenis kemalangan, sebab kemalangan dan tempoh cuti sakit. LR boleh digunakan untuk memodelkan hubungan linear antara ciri-ciri lain dalam set data yang menyumbang kepada perubahan dalam pembolehubah sasaran.

3.3.2 Ridge Regression

Ridge Regression ialah regresi linear yang menyebarluaskan regresi *Ordinary Least Square*(OLS). Ianya juga dikenali sebagai Tikhonov regularisasi atau L2 regularisasi yang digunakan dalam pembelajaran mesin untuk mencegah *overfitting* dan meningkatkan *generalization* sesuatu model.

Ridge regression adalah sesuatu kaedah model *tuning* untuk menganalisi data yang terseksa daripada multikolinearity. Kaedah model *tuning* ini melakukan regularisasi L2. Apabila masalah multikolinearity berlaku, *least-squared* tidak berat sebelah dan nilai varians adalah besar, ini akan menghasilkan nilai ramalan yang jauh dari nilai sebenar. *Ridge regression* sangat berguna untuk mengurus masalah multikolinearity di mana pembolehubah sesuatu ramalan mempunyai korelasi. Regularisasi membantu menstabilkan model dengan menghalangnya daripada terlalu bergantung kepada mana-mana pembolehubah yang tunggal.

Dalam regresi linear yang tradisional di mana matlamatnya adalah untuk meminimumkan jumlah perbezaan kuasa dua antara nilai yang diperhatikan dan nilai yang diramalkan. *Ridge regression* memperkenalkan istilah regularisasi kepada regresi linear dengan fungsi objektif yang akan menghukum *coefficients* yang besar.

Fungsi objektif adalah seperti berikut = $\sum_{i=1n}(y_i - y\hat{}_i)_2 + \alpha\sum_{j=1p}\beta_{j2}$
di mana :
n bilangan pemerhatian
p bilangan ciri (pembolehubah bebas)
yi nilai yang diperhatikan untuk pemerhatian ke-i
y^i nilai ramalan untuk pemerhatian ke-i
βj nilai pekali bagi ciri ke-j
α parameter regularisasi yang mengawal kekuatan tempoh penalti

3.3.3 Support Vector Machine(SVM)

SVM ialah algoritma Pembelajaran Mesin Terselia yang digunakan untuk tugas klasifikasi dan regresi berdasarkan informasi yang diperolehi melalui https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm . Untuk kajian ini, SVM digunakan untuk tugas regresi. SVM adalah algoritma yang berkuasa dan terkenal dengan keberkesanannya dalam ruang yang berdimensi tinggi dan keupayaannya dalam mengendalikan perhubungan dalam data linear dan bukan linear.

Matlamat algoritma SVM adalah untuk mencari garis yang terbaik samada dalam bentuk sempadan keputusan yang memisahkan titk data daripada kelas data yang berbeza. Sempadan ini dikenali sebagai *hyperplane* apabila ia bekerja dalam ruang yang berdimensi tinggi. Idea SVM secara lazimnya adalah untuk memaksimumkan margin, iaitu jarak antara hyperplane dan titik terdekat bagi setiap kategori sekaligus memudahkan unttuk membezakan kelas-kelas data.

SVM juga berguna untuk menganalisis data yang kompleks yang tidak boleh dipisahkan dengan garis lurus yang mudah. Untuk kes sebegini ianya dipanggil SVM tidak linear di mana helah matematik digunakan untuk mengubah data menjadi ruang berdimesi lebih tinggi supaya mudah untuk mencari sempadan.

SVM menggunakan fungsi kernel dimana fungsi ini mengira titik produk antara ciri vektor yang diubah secara tersirat. Fungsi pengiraan kernel ini berbeza dengan pengiraan koordinat atau titik ruang yang diubah secara tersurat atau eksplisit. Ini boleh mengelakkan pengendalian pengiraan yang mahal bagi kes yang melampau. Fungsi kernel boleh mengendalikan pemisahan data yang linear dan bukan linear menggunakan fungsi kernel yang berbeza samada fungsi Kernel Linear, Kernel Polinomial, Kernel *Radial Basis Function* (RBF) atau Kernel Sigmoid . Berikut adalah penerang ringkas mengenai fungsi kernel :

1. **Kernel Linear :** fungsi kernel yang paling mudah dan ianya memetakan data ke ruang dimensi yang lebih tinggi di mana data boleh dipisahkan secara linear.

2. **Kernel Polinomial :** Kernel ini lebih berkuasa daripada kernel linear kerana ia boleh digunakan untuk memetakan data ke ruang dimensi yang lebih tinggi dan data tersebut tidak mampu dipisahkan secara linear.

3. **Kernel RBF :** fungsi RBF ini adalah fungsi kernel yang paling popular bagi SVM kerna ianya berkesan untuk pelbagai masalah klasifikasi.

4. **Kernel Sigmoid :** fungsi kernel ini menyamai fungsi kernel RBF, tetapi ianya mempunyai bentuk yang berbeza yang boleh berguna untuk beberapa masalah klasifikasi.

Kesemua kernel ini mampu untuk menerokai perhubungan kompleks dan corak atau trend yang wujud di dalam data. Akan tetapi, pemilihan fungsi kernel untuk algoritma adalah berdasarkan ketepatan dan kompleksiti. Semakin

berkuasa sesuatu fungsi kernel seperti RBF yang mampu mencapai ketepatan yang tinggi, semakin banyak data dan masa pengiraan diperlukan untuk SVM. Setelah dilatih, SVM boleh mengklasifikasikan titik data baharu yang tidak boleh dilihat sebelum ini dengan menentukan sebelah mana sempadan keputusan dibuat. Output SVM adalah label kelas yangdikatkan dengan sisi sempadan keputusan.

3.3.4 Decision Tree(DT)

DT merupakan algoritma pembelajaran mesin yang digunakan untuk *classification* dan regresi menurut maklumat pada https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/. DT adalah model yang mudah difahami, senang untuk ditafsirkan dan menjadikanya sebagai satu pilihan yang ideal dalam pembangunan model peembelajaran mesin. Ia mempunyai struktur pokok hirarki yang terdiri daripada nod akar, dahan, nod dalaman dan nod daun. Berikut adalah detail mengenai DT :

1. **Struktur DT** : DT disusun seperti pokok yang terbalik, dengan akar di bahagian atas dan dahan-dahan yang memegang daun berada di bahagian bawah. Setiap nod dalaman (termasuk akar) mewakili satu keputusan yang berdasarkan atribut atau ciri dan setiap nod daun pula mewakili label kelas (dalam *classfication*) atau nilai numerik(dalam regresi).

2. **Pembahagian Nod** : Proses membahagi nod kepada dua atau lebih sub nod menggunakan kriteria pembahagian dan ciri yang dipilih.

3. **Kedalaman Pokok** : ditentukan oleh bilangan peringkat nod daripada akar sehingga daun yang paling hujung. DT yang mempunyai kedalaman pokok yang tinggi boleh menangkap corak yang kompleks akan tetapi ianya akan cenderung kepada *overfitting.*

4. **Peraturan Keputusan :** Setiap laluan dari akar ke daun mewakili satu peraturan keputusan. Bagi *classification* peraturan ini adalah kelas label manakala untuk regresi peraturan ini adalah nombor yang diramal.

5. ***Impurity* :** adalah untuk mengukur kehomogenan pembolehubah sasaran dalam subset data. Ia merujuk kepada tahap rawak atau ketidapastian dan set data. Indeks Gini dan Entropi adalah dua jenis ukuran *Impurity* yang biasa digunakan dalam DT untuk *classification.*

6. **Varian :** adalah untuk mengukur seberapa banyak pembolehubah yang diramalkan dan sasaran berbeza dalam sampel yang berbeza untuk suatu set data. Ianya digunakan untuk masalah regresi dalam DT. MSE, MAE atau *Half Poisson Deviance* digunakan untuk mengukur varian.

7. ***Information Gain :*** merupakan ukurang bagi pengurangan *impurity* yang dicapai dengan membahagikan set data mengikut citi-ciri tertentu dalam DT. Kriteria pembahagian ditentukan oleh ciri yang mempunyai *information gain* yang tertinggi. *Information Gain* digunakan untuk mengenalpasti ciri yang paling berinformasi bagi pembahagian setiap nod pokok supaya dapat menwujudkan set data yang tulen. *Information Gain* digunakan mengikut formula berikut :

   Information Gain = Entropi(S)- [(Purata Wajaran) *Entropi(setiap ciri)

   Entropi : Entropi ialah metrik untuk mengukur kekotoran dalam atribut tertentu. Ia menentukan rawak dalam data. Entropi boleh dikira sebagai:

   Entropi= -P(ya)log2 P(ya)- P(tidak) log2 P(tidak)

   di mana,

   S= Jumlah bilangan sampel

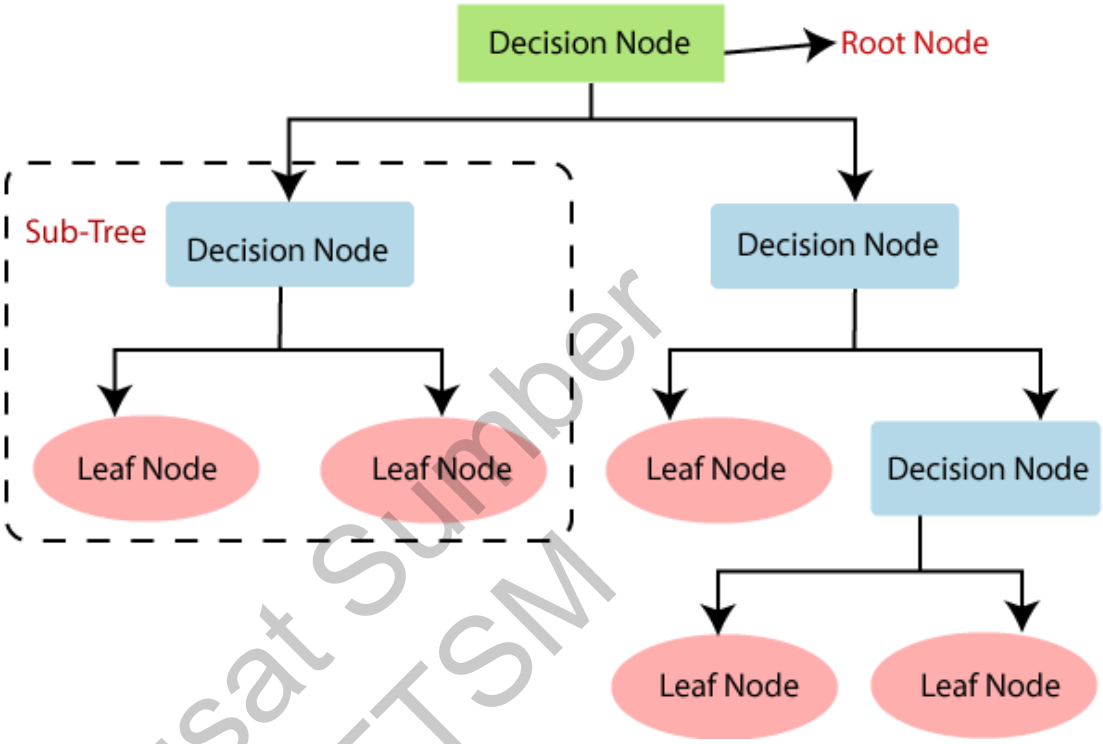
   P(ya)= kebarangkalian ya

   P(tidak)= kebarangkalian tidak

8. ***Pruning :*** Untuk mengelakkan overfitting dari berlaku, DT boleh dipangkas dengan mengeluarkan cabang-cabang yang memberi sedikit kuasa ramalan. Pruning meringkaskan DT sambil mempertahankan ketepatan ramalan yang dibuat.

DT adalah serbaguna dan boleh menangkap hubungan linear dan bukan linear. Ianya sesuai untuk meneroka interaksi kompleks antara atribut-atribut

atau ciri-ciri dalam set data tuntutan. Selain itu, DT juga boleh divisualisasikan untuk memahami kepentingan ciri. Ciri-ciri yang terletak lebih tinggi dalam hierarki pokok atau menghasilkan maklumat yang siginifikan baisanya lebih penting.



Rajah 3.1 Rajah yang menggambarkan DT

Sumber [https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm]

3.3.5 Random Forest(RF)

RF adalah merupakan algoritma Pembelajaran Mesin Terselia yang digunakan secara meluas dalam *Classification* dan Regresi berdasarkan maklumat yang diperolehi pada https://www.ibm.com/topics/random-forest. RF diperkenalkan oleh Leo Breiman dan Adele Cutler yang menggabungkan output daripada berbilang DT untuk mengeluarkan satu keputusan atau output. Menurut Hanafy & Mahmoud (2021), model bagi RF digambarkan seperti berikut :

$g(x)= f_0(x)+f_1(x) +f_2(x) +……+ f_n(x)$…….....

Rajah 3.2 Rajah yang menunjukkan persamaan RF

Di mana 'g' ialah model akhir yang merupakan jumlah semua model dan setiap model f(x) ialah DT. Model ini dipilih untuk set data yang diberikan kerana ia menggabungkan banyak DT untuk meramalkan hasil yang lebih tepat. Setiap pokok digunakan untuk menghasilkan ramalan bagi sampel rawak baharu, kemudian ramalan tersebut dipuratakan untuk membentuk RF menurut N.Yego et. al (2020).

RF popular kerana ianya bersifat mesra pengguna dan sifat kebolehsuaiannya yang memboleh ia menangani masalah *classification* dan regresi dengan berkesan. Kekuatan RF terletak pada kemampuannya untuk mengendalikan set data yang kompleks dan dapat mengurangkan *overfitting*, lalu menjadikan ia suatu alat yang bernilai untuk melakukan pelbagai tugas ramalan dalam Pembelajaran Mesin.

3.3.6 Xgboost

XGBoost merupakang singkatan kepada *Extreme Gradient Boosting* di mana ianya terdiri daripada perpustakaan Pembelajaran Mesin *Gradient Boosting Decision Tree*(GBDT) dan *Ensemble Learning* yang teragih dan mempunyai kobolehskalaan tinggi menurut informasi pada https://www.shiksha.com/online-courses/articles/xgboost-algorithm-in-machine-learning/. GBDT adalah algoritma *Decision Tree Ensemble Learning* yang menyerupai RF untuk melakukan tugas *classification* dan regresi. Algoritma ini menggabungkan beberapa algoritma Pembelajaran Mesin lain dalam menghasilkan model yang lebih baik.

Kedua-dua GBDT dan RF terdiri daripada beberapa DT, dan perbezaan di antara kedua-dua model ini adalah melalui perbezaan bagaimana pokok-pokok dibangunkan dan digabungkan. RF menggunakan teknik *bagging* untuk membuat suatu sesuatu DT yang selari dengan sampel bootstrap rawak bagi set data dan ramalan akhir RF ialah purata semua ramalan DT yang diproses. Manakala GBDT menggunakan teknik *boosting* atau meningkatkan model yang lemah dengan menggabungkan model lemah dengan beberapa model lemah yang lain untuk suatu model yang kukuh secara kolektif. *Gradient Boosting*

adalah lanjutan daripada *boosting* di mana proses tambahan bagi menjana model-meodel yang lemah untuk diformalkan sebagai algoritma *gradient descent. Gradient Boosting* menetapkan hasil yang disasarkan untuk model seterusnya dalam usaha memininumkan ralat. Hasil yang disasarkan untuk setiap kes adalah berdasarkan *gradient error*(dinamakan sebagai *gradient boosting*) yang berkaitan dengan sesuatu ramalan.

GBDT berulang kali melatih himpunan DT yang cetek di mana setiap ulangan menggunakan sisa ralat model sebelumnya agar ianya sesuai dan padan untuk model seterusnya. Ramalan akhir GBDT ialah jumlah wajaran semua ramalan pokok. GBDT ygn menggunakan *boosting* dapat meminimumkan *bias* dan *underfitting* berbeza dengan RF yang menggunakan teknik *bagging* dapat meminimumkan varian dan *overfitting*,.

XGBoost mempraktikkan strategi yang berbeza berbanding dengan GBDT tradisional yang membina pokok secara berurutan. XGBoost membina pokok secara selari dan ia menggunakan strategi kedalaman (juga dikenali sebagai *level-wise).* Strategi kedalaman bermaksud algoritma akan meneroka belahan dengan cara yang lebih seimbang merentasi kedalaman pokok yang berbeza dan bukannnya bergerak berkembang sepenuhnya bagi satu tahap sebelum bergerak ke tahap yang seterusnya.

Dalam konteks kajian ini di mana analisis kos tuntutan insurans dilakukan, pemilihan algoritma yang dipilih adalah berdasarkan faktor seperti kompleksiti hubungan antara ciri-ciri dan pembolehubah sasaran, saiz set data dan kobolehtafsiran output. Secara praktis, kajian akan akan bermula dengan meneroka dengan model yang lebih mudah sepert LR untuk mendapatkan pandangan awal, dan kemudian diterokai dengan model yang lebih kompleks seperti DT, RF atau XGBoost.

3.4 METRIK PRESTASI

Metrik prestasi adalah instrumen penting untuk menilai keberkesanan model pembelajaran mesin. Metrik prestasi menyediakan ukuran yang boleh diukur tentang prestasi model pada tugasan tertentu. Pilihan metrik pretasi bergantung pada jenis

masalah pembelajaran yang cuba ditangani samada *classification*, regresi, *clustering* atau sesuatu yang lain. Berikut ialah beberapa metrik prestasi yang biasa digunakan berdasarkan maklumat diperolehi pada https://medium.com/@brandon93.w/regression-model-evaluation-metrics-r-squared-adjusted-r-squared-mse-rmse-and-mae-24dcc0e4cbd3 :-

1. Means Squared Error (MSE)

Metrik yang digunakan secara meluas untuk menilai prestasi model regresi dalam pembelajaran mesin dan statistik. Ia mengukur purata perbezaan kuasa dua antara nilai yang diramalkan dan sebenar, sekaligus menekankan ralat yang lebih besar. MSE amat berguna dalam aplikasi di mana matlamatnya adalah untuk meminimumkan kesan *outlier* atau ralat pengedaran akan diandaikan sebagai Gaussian.

Taburan Gaussian adalah taburan kebarangkalian yang simetri tentang min yang menunjukkan bahawa data berhampiran min adalah lebih kerap berlaku daripada data yang berada jauh dari nilai min.

Memandangkan set data dengan pemerhatian di mana y_i idialah nilai sebenar dan ŷ_i adalah nilai ramalan untuk pemerhatian ke-i, MSA boleh dikira menggunakan formula berikut :

$$\mathrm{MSE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{Y_i} - \widehat{\mathrm{Y}}_\mathrm{i} \right)^2$$

Rajah 3.3 Rajah yang menunjukkan persamaan MSE

Di sini, perbezaan kuasa dua antara setiap nilai sebenar (y_i) dan nilai ramalan yang sepadan (ŷ_i) dikira, dan jumlah perbezaan kuasa dua ini dibahagikan dengan jumlah bilangan cerapan (n) untuk mendapatkan purata ralat kuasa dua.

MSE menyediakan ukuran prestasi model yang menghukum ralat yang lebih besar dengan lebih teruk daripada yang lebih kecil. MSE yang lebih rendah menunjukkan kesesuaian model yang lebih baik, menunjukkan bahawa ramalan model, secara purata, lebih hampir kepada nilai sebenar. Ia biasanya digunakan apabila membandingkan model yang berbeza pada set data yang sama, kerana ia boleh membantu mengenal pasti model dengan ramalan yang paling tepat.

2. Root Means Squared Error (RMSE)

MSE mengukur purata perbezaan kuasa dua antara nilai yang diramalkan dengan nilai sebenar. RMSE mempunyai unit yang sama dengan pembolehubah sasaran, menjadikannya lebih mudah ditafsir dan lebih mudah untuk dikaitkan konteks masalah berbanding MSE.

Diberi set data dengan n pemerhatian, dengan y_i ialah nilai sebenar dan ŷ_i ialah nilai ramalan untuk pemerhatian ke-i, RMSE boleh dikira menggunakan formula berikut:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} ||y(i) - \hat{y}(i)||^2}{n}}$$

Rajah 3.4 Rajah yang menunjukkan persamaan RMSE

Di sini, perbezaan kuasa dua antara setiap nilai sebenar (y_i) dan nilai ramalan yang sepadan (ŷ_i) dikira, dan jumlah perbezaan kuasa dua ini dibahagikan dengan jumlah bilangan cerapan (n) untuk mendapatkan purata ralat kuasa dua. Punca kuasa dua nilai ini kemudiannya diambil untuk mengira RMSE.

RMSE boleh menyediakan ukuran prestasi model yang mengimbangi penekanan pada ralat yang lebih besar (seperti dalam MSE) dengan kebolehtafsiran (kerana ia mempunyai unit yang sama dengan pembolehubah sasaran). RMSE yang lebih rendah menunjukkan kesesuaian model yang lebih baik, menunjukkan bahawa ramalan model, secara purata, lebih hampir kepada nilai sebenar. Ia biasanya digunakan apabila membandingkan model yang berbeza pada set data

yang sama, kerana ia boleh membantu mengenal pasti model dengan ramalan yang paling tepat.

3. R-Squared (R2)

Metriks R-Kuasa Dua(R2) yang juga dikenali sebagai *coefficient of determination* adalah merupakan ukuran statisktik yang mewakili perkadaran varian dalam pembolehubah bersandar yang boleh diramal daripada pembolehubah bebas dalam model regresi. R2 biasanya digunakan sebagai metrik penilaian untuk model regresi dalam pembelajaran mesin.

Skor R2 adalah nilai di antara 0 dan 1 di mana :

0 - menandakan model tidak mejelaskan sebarang kebolehubahan dalam pembolehubah sasaran.
1 - menandakan model menerangkan dengan sempurna berkenaan kebolehubahan dalam pembolehubah sasaran.
Jika skor R2 lebih hampir kepada 1 menunjukkan bahawa model berjaya membuat penyesuaian dengan baik pada data dan ramalan hampir sepadan dengan nilai sebenar. Manakala jika skor R2 lebih hampir kepada 0 menunjukkan model tidak menunjukkan prestasi yang baik dalam menerangkan varian.

Skor R2 dikira berdasarkan formula berikut :

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals (SSR)}}{\text{Total Sum of Squares (SST)}}$$

Rajah 3.5 Rajah yang menunjukkan persamaan R2

Di mana
SSR : jumlah perbezaan kuasa dua antara nilai yang diramalkan dan nilai sebenar.
SST : jumlah perbezaan kuasa dua antara nilai sebenar dan min nilai sebenar.

Walaupun R2 ialah metriks yang digunakan secara meluas namun ianya mempunyai beberapa had seperti kepekaanya terhadap bilangan model dan

ketidakupayaannya mengesan *overfitting.* Oleh itu, untuk kajian ini ianya akan digunakan bersama dengan metrik penilaian yang lain.

3.5 KESIMPULAN

Bab ini menerangkan tentang metodologi kajian secara keseluruhan. Setiap peringkat dalam kajian ini dibincangkan secara terperinci bermula dari pengumpulan data, pemprosesan data, pembangunan model dan penilaian model. Keputusan bagi satiap peringkat yang diterangkan dalam bab ini akan diulas dengan lebih lanjut dalam bab berikutnya.

BAB IV

DAPATAN KAJIAN

4.1 PENGENALAN

Bab ini memberikan maklumat yang terperinci dan komprehensif mengenai faktor-faktor yang menyumbang kepada kos tuntutan di PERKESO dengan menjalankan analisis terhadap data yang diperolehi. Melalui analisis data yang teliti, kajian ini cuba untuk mengkaji dan mendapatkan jawapan bagi soalan-soalan kajian yang telah digariskan.

Dapatan kajian yang dikemukakan dalam bahagian ini merupakan hasil penemuan daripada penggunaan teknik pembelajaran mesin dalam usaha membawa satu pendekatan baru dalam dunia keselamatan sosial atau insurans. Bahagian ini merangkumi hasil penggunaan kaedah komputasi canggih untuk menganalisis dan menginterpretasikan data dan bagaimana data tersebut dapat menjawab persoalan-persoalan kajian yang telah digariskan.

Pembelajaran Mesin telah muncul sebagai suatu alat yang yang berkesan untuk mengekstrak sebarang corak atau trend, membuat ramalan dan dapat memberi pemahaman yang mendalam daripada data yang kompleks. Dalam konteks ini, perbincangan berikut merangkumi hasil yang diperolehi melalui penggunaan algoritma pembelajaran mesin.

Sepanjang bahagian ini, hasil kajian secara berstruktur akan dipaparkkan untuk mengambarkan objektif utama kajian. Model ramalan akan diterokai, prestasi model

ramalan akan dinilai dan atribut-atribut yang penting akan dikenalpasti dalam membuat keputusan untuk organisasi. Pencapaian ini tidak hanya akan menerangkan masalah yang sedang dihadapi tetapi juga dapat memberi suatu persepsi mengenai keberkesanan dan adaptabiliti pembelajaran mesin dalam sektor insurans.

Asas pencapaian ini berpaksikan dalam penggunaan algoritma pembelajaran mesin yang teliti dalam set data kajian ini. Set data bagi kajian ini merupakan data kes tuntutan PERKESO bagi tahun 2017 sehingga 2020 yang digunakan sebagai asas untuk model pembelajaran mesin dilatih dan dinilai. Metodologi yang digunakan untuk pra-pemprosesan data, pemilihan model telah dijelaskan dengan detail dalam bab yang terdahulu.

4.2 PENGUMPULAN DATA

Proses pengumpulan data adalah proses mengenalpasti sumber-sumber untuk mengumpulkan data yang relevan bagi kajian ini. Sumber yang dikenalpasti adalah data daripada pangkalan data daripada beberapa legasi sistem. Data-data ini diekstrak keluar dan disimpan dalam bentuk fail csv. Data bagi tahun 2017 sehingga 2020 yang diperolehi daripada legasi sistem tersebut berada di dalam fail yang berasingan. Oleh itu kesemua fail tuntutan (dalam bentuk .csv) tersebut digabungkan menjadi satu fail. Gabungan tersebut menggunakan Microsoft Excel.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Tahun	Jantina	Umur (TLA	Umur (TKH	Jenis Kes	PNPP	Kod Industri	Deskripsi I	Kod Sub Lo	Deskripsi	Deskripsi M	Kod Jenis K	Deskripsi J	Kod Sub Se	Deskripsi S	Deskripsi M	Bil MC	Amaun Ba
2	2017	Perempuai	31	0	HUS	A31 KUALA	8200	Aktiviti Ke	68	Other mult	Lokasi Ber	41	Other wou	33	Striking ag	Terpijak di	12	405.33
3	2017	Lelaki	33	0	HUS	A31 KUALA	6141	Perdagang	46	Hand (exc	Anggota A	10	Fractures	33	Striking ag	Terpijak di	95	4095.55
4	2017	Lelaki	0	0	HUS	A31 KUALA	9129	Pentadbira	45	Wrist	Anggota A	30	Concussio	34	Struck by r	Terpijak di	52	4090.66
5	2017	Lelaki	57	0	HUS	A31 KUALA	8200	Aktiviti Ke	53	Knee	Anggota B	30	Concussio	42	Caught bet	Tersepit Di	5	393.33
6	2017	Lelaki	52	0	HUS	A31 KUALA	9200	Pentadbira	47	Fingers	Anggota A	10	Fractures	33	Striking ag	Terpijak di	292	12458.64
7	2017	Lelaki	35	0	HUS	A31 KUALA	8102	Aktiviti Ke	46	Hand (exc	Anggota A	41	Other wou	31	Stepping o	Terpijak di	6	357.33
8	2017	Lelaki	43	0	HUS	A31 KUALA	5001	Constructi	57	Toes	Anggota B	10	Fractures	23	Struck by f	Terhempa	43	3382.66
9	2017	Lelaki	23	0	HUS	A31 KUALA	9599	Pentadbira	54	Leg (lower	Anggota B	41	Other wou	31	Stepping o	Terpijak di	219	6570
10	2017	Lelaki	38	0	HUS	A31 KUALA	5002	Constructi	14	Mouth	Kepala	10	Fractures	31	Stepping o	Terpijak di	11	865.33
11	2017	Perempuai	51	0	HUS	A31 KUALA	2901	Perlombor	41	Shoulder	Anggota A	10	Fractures	31	Stepping o	Terpijak di	38	1469.33
12	2017	Lelaki	43	0	HUS	A31 KUALA	9350	Pentadbira	18	Head, mult	Kepala	20	Dislocatio	33	Striking ag	Terpijak di	34	1020
13	2017	Lelaki	58	0	HUS	A31 KUALA	9599	Pentadbira	41	Shoulder	Anggota A	10	Fractures	31	Stepping o	Terpijak di	32	1123.55
14	2017	Lelaki	34	0	HUS	A31 KUALA	3111	Manufactu	46	Hand (exc	Anggota A	10	Fractures	24	Struck by f	Terhempa	53	3274.22
15	2017	Lelaki	19	0	HUS	A31 KUALA	6219	Perdagang	53	Knee	Anggota B	10	Fractures	34	Struck by r	Terpijak di	27	810
16	2017	Lelaki	49	0	HUS	A31 KUALA	9331	Pentadbira	53	Knee	Anggota B	10	Fractures	31	Stepping o	Terpijak di	84	5301.33
17	2017	Lelaki	27	0	HUS	A31 KUALA	3420	Manufactu	54	Leg (lower	Anggota B	30	Concussio	11	Falls of pe	Terjatuh d	76	5573.34
18	2017	Lelaki	72	0	HUS	A31 KUALA	9599	Pentadbira	56	Feet (exce	Anggota B	10	Fractures	34	Struck by r	Terpijak di	588	23520
19	2017	Perempuai	46	0	HUS	A31 KUALA	9513	Pentadbira	18	Head, mult	Kepala	20	Dislocatio	33	Striking ag	Terpijak di	40	2890.66
20	2017	Lelaki	44	0	HUS	A31 KUALA	2200	Perlombor	41	Shoulder	Anggota A	41	Other wou	31	Stepping o	Terpijak di	63	4956
21	2017	Lelaki	41	0	HUS	A31 KUALA	9600	Pentadbira	56	Feet (exce	Anggota B	25	Sprains an	42	Caught bet	Tersepit Di	23	1809.33
22	2017	Lelaki	49	0	HUS	A31 KUALA	4101	Perkhidma	45	Wrist	Anggota A	20	Dislocatio	33	Striking ag	Terpijak di	38	2989.33
23	2017	Perempuai	44	0	HUS	A31 KUALA	6320	Penginapa	46	Hand (exc	Anggota A	20	Dislocatio	33	Striking ag	Terpijak di	60	4213.33

Rajah 4.1 Rajah yang menunjukkan sekali imbas data yang diperolehi daripada gabungan legasi sistem
[Sumber : PERKESO]

4.3 PRA PEMPROSESAN DATA

Menyediakan data untuk pembelajaran mesin merupakan langkah penting dalam membangunkan model pembelajaran mesin yang efektif dan tepat. Penyediaan data dalam kajian ini melibatkan pembersihan, transformasi dan penyusunan semula supaya data tersebut sesuai untuk digunakan proses latihan dan pengujian algoritma pembelajaran mesin.

4.4 SEMAKAN KELENGKAPAN DATA

Memeriksa kekompleksan data melibatkan pengesahan bahawa set data bebas daripada nilai yang hilang atau tidak lengkap. Ia adalah penting untuk menilai integriti data dengan sebarang kekosongan atau maklumat yang hilang.

Berikut adalah semakan data secara keseluruhan :

Jadual 4.1 Senarai atribut yang terkandung dalam data

Atribut	Bilangan
Tahun	0
Jantina	0
Umur (TLAPOR-TLAHIR)	0
Umur (TKHAKRU-TKHLAHIR)	0
Jenis Kes	0
PNPP	0
Zon	0
Kod Industri	0
Deskripsi Industri	0
Kod Sub Lokasi Kecederaan	0
Deskripsi Sub Lokasi Kecederaan	0
Deskripsi Lokasi Kecederaan Utama	0
Kod Jenis Kemalangan	0
Deskripsi Jenis Kemalangan	0
Kod Sub Sebab Kemalangan	0
Deskripsi Sub Sebab Kemalangan	0
Deskripsi Sebab Utama Kemalangan	0
Tempoh MC	0
Amaun Bayaran	0

Berikut adalah semakan setiap atribut dalam set data :

1. Atribut Jantina

Jadual 4.2 Data bagi atribut Jantina

Atribut	Bilangan
Lelaki	180048
Perempuan	44500
Name: Jantina, dtype: int64	

2. Atribut Pejabat PERKESO

Jadual 4.3 Data bagi atribut Pejabat PERKESO

Atribut	Bilangan
A31 Kuala Lumpur	22193
E11 Johor Bahru	18070
B34 Klang	17471
C52 Butterworth	16261
D41 Ipoh	10147
E23 Melaka	9461
B39 Putrajaya	8691
E21 Seremban	8523
C51 Georgetown	6758
D62 Sg Petani	5408
E13 Muar	5311
E15 Batu Pahat	5061
D61 Alor Setar	5059
B33 Rawang	4594
D42 Taiping	4375
F74 Kuantan	4332
E12 Kluang	3841
F86 Kuching	3372
D63 Kulim	3338
F87 Sibu	2832
E14 Segamat	2566
F91 Kota Bharu	2517
C51 George Town	2489
F73 Temerloh	2470
D44 Teluk Intan	2265

bersambung…

…sambungan

F82 Kuala Terengganu	2159
F96 Kota Kinabalu	1904
F71 Bentong	1893
D46 Seri Manjung	1735
F84 Bintulu	1542
F83 Miri	1298
D45 Kuala Kangsar	1238
D65 Langkawi	1235
F80 Kemaman	1175
D43 Tapah	1114
E22 Kuala Pilah	1105
D64 Kangar	970
F93 Tawau	817
F92 Kuala Krai	783
F81 Dungun	711
F97 Sandakan	697
T38 Sps	429
B29 Tanjung Karang	427
F89 Sarikei	404
F95 Lahad Datu	400
F76 Labuan	300
F88 Kapit	275
F94 Keningau	241
F85 Sri Aman	149
F77 Limbang	109
F78 Serian	106
F90 Mukah	88
F70 Kota Marudu	45
F99 Beaufort	41

Name: PNPP, dtype: int64

3. Atribut Zon Pejabat PERKESO

Jadual 4.4 Data bagi atribut Zon Pejabat PERKESO

Atribut	**Bilangan**
Tengah	77558
Utara	62392
Selatan	53938
Pantai Timur	16040
Sarawak	10175
Sabah	4445

4. Atribut Industri

Jadual 4.5 Data bagi atribut Industri

Atribut	Bilangan
Pentadbiran Awam Dan Pertahanan/Aktiviti Keselamatan Wajib	68211
Manufacturing	53064
Construction	29838
Perdagangan,Trading	27441
Aktiviti Hartanah, Penyewaan & Perniagaan	16048
Pengangkutan dan Penyimpanan	10716
Pertanian, Perhutanan Dan Perikanan	6508
Penginapan dan Aktiviti Perkhidmatan Makanan Minuman	5321
Aktiviti Kewangan dan Dan Insurans/Takaful	3955
Perkhidmatan Elektrik, Gas, Air & Kebersihan	2226
Perlombongan & Pengkuarian	1042
Activities not adequately defined	107
Null	71
Name:Industri, dtype: int64	

Terdapat 71 *null value* bagi atribut Industri

5. Atribut Lokasi Kecederaan Utama

Jadual 4.6 Data bagi atribut Lokasi Kecederaan Utama

Atribut	Bilangan
Bahagian atas anggota badan / *Upper Limb*	66092
Bahagian bawah anggota badan / *Lower Limb*	49044
Lokasi Anggota Badan Berganda / *Multiple Location*	36079
Tubuh	30731
Unspecified location of injury	16816
Kepala	10771
Kecederaan Am / *General Injuries*	9996
Leher (termausk kerongkong dan tulang belakang)	5019
Name: Deskripsi Lokasi Kecederaan Utama, dtype: int64	

6. Atribut Jenis Kemalangan

Jadual 4.7 Data bagi atribut Jenis Kemalangan

Atribut	Bilangan
Other wounds	69546
Fractures	59065
Other and unspecified injuries	30913
Sprains and strains	28294
Multiple injuries of different nature	13185
Concussions and other internal injuries	10183
Superficial injuries	6640
Dislocations	2610
Burns	1748
Contusions and crushings	959
Amputations and enucleations	804
Effects of weather, exposure	437
Acute poisonings	49
Effects of electric currents	44
Asphyxia	36
Effects of radiation	35
Name: Deskripsi Jenis Kemalangan, dtype: int64	

7. Atribut Sebab Utama Kemalangan

Jadual 4.8 Data bagi atribut Sebab Kemalangan Utama

Atribut	Bilangan
Terpijak di atas/terkena/terhempap oleh benda(tidak termasuk benda jatuh)	73781
Terjatuh dari Tempat Tinggi / Terjunam ke dalam lubang atau Jurang	39779
Terjatuh pada aras yang sama	37993
Lain-lain kemalangan	25263
Pergerakan Yang Berat	21794
Terhempap oleh Benda-benda Yang Jatuh	13615
Tersepit Di dalam/Di antara Benda	11099
Terdedah/Tersentuh suhu yang panas	1017
Terdedah/Tersentuh Bahan Merbahaya	151
Terdedah/Tersentuh Elektrik	56
Name: Deskripsi Sebab Kemalangan Utama, dtype: int64	

8. Atribut Tempoh Cuti Sakit

Jadual 4.9 Data bagi atribut Tempoh Cuti Sakit

Cuti Sakit	Bilangan
4	12499
3	9327
6	8377
7	8238
5	8219
...	....
369	1
427	1
553	1
508	1
354	1
Name: Tempoh MC, Length: 430, dtype: int64	

4.5 EKSPLORASI DATA

Periksa data untuk mendapatkan gambaran struktur dan kandungannya.

Jadual 4.10 Statitistik Deskripsi bagi keseluruhan data

Tahun	Umur (TLAPOR-TLAHIR)	Umur (TKHAKRU-TKHLAHIR)	Kod Industri	Kod Sub Lokasi Kecederaan	Kod Jenis Kemalangan	Kod Sub Sebab Kemalangan	Tempoh MC	Amaun Bayaran
224548.00	224548.00	224548.0	224548.00	224548.00	224548.00	224548.00	224548.00	224548.00
2018.49	36.48	0.0	6427.056	52.14	41.52	33.87	35.77	2124.619598
1.09370	13.46	0.0	2430.68	15.83	30.15	24.37	36.15	2513.628823
2017.00	0.00	0.0	1.00	11.00	10.00	11.00	0.00	0.00
2018.00	26.00	0.0	3909.00	41.00	10.00	12.00	9.00	450.00
2019.00	35.00	0.0	6219.00	48.00	41.00	32.00	25.00	1267.56
2019.00	46.00	0.0	9126.00	68.00	41.00	41.00	52.00	2874.67
2020.00	120.00	0.0	9600.00	80.00	99.00	92.00	687.00	56774.71

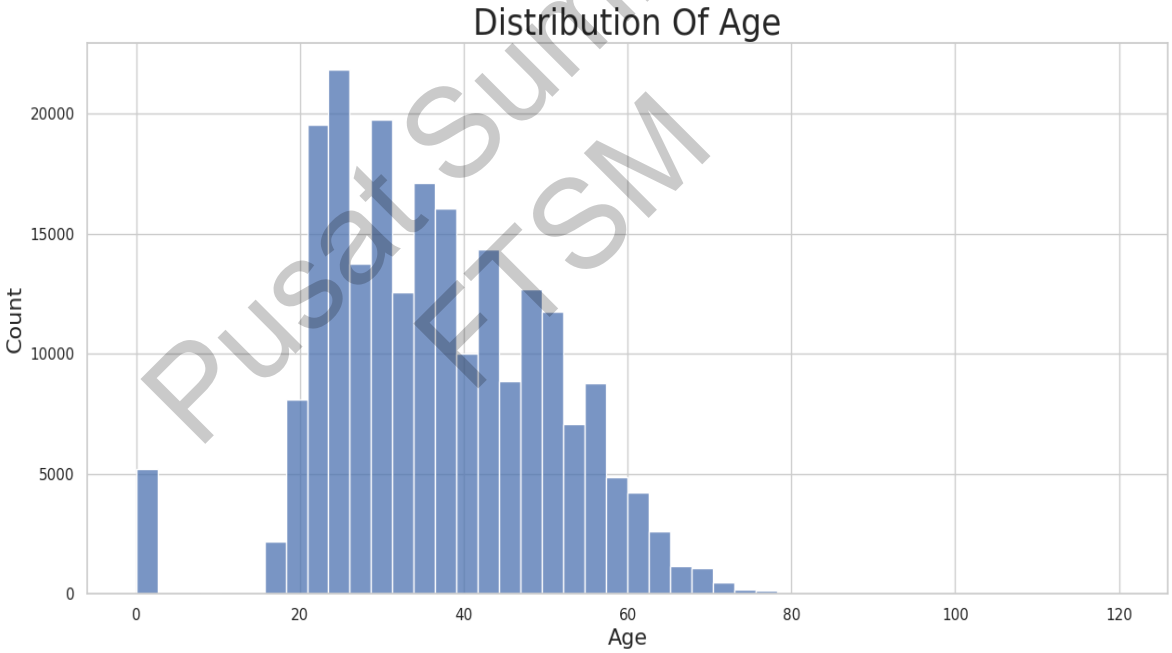
Terdapat tiga aspek yang diperhatikan untuk menganalisis statistik deskripsi ini iaitu :

1. **Nilai minimum dan maksimum** : Ianya boleh memberi idea mengenai julat nilai dan membantu untuk mengesan sebaran outlier. Dalam kajian ini, semua nilai minimum dan maksimum kelihatan munasabah kecuali nilai Umur(TLAPOR – TLAHIR). Nilai minimum dan maksimum bagi atribut Umur(TLAPOR – TLAHIR) adalah tidak munasabah di mana nilai minimum adalah O dan nilai maksimum adalah 120. Manakala nilai Umur (TKHAKRU – TKHLAHIR) boleh diabaikan berikutan atribut tidak akan digunapakai

2. **Nilai *Mean* dan *Standard Deviation*** : Nilai *Mean* menunjukkan kecederungan pusat taburan, manakala nilai *standard deviation* pula megukur jumlah variasinya. Berdasarkan pemerhatian pada statistik deskripsi ini nilai Tempoh MC yang mempunyai nilai *standard deviation* yang rendah di mana nilainya berada hampir dengan nilai *mean.*

3. **Bilangan data** : Ini adalah penting untuk memberi gambaran mengenai *volume* bagi data yang hilang. Di sini, dapat perhatikan tiada data yang hilang bagi kesemua atribut.

Seterusnya adalah langkah untuk merawat nilai Umur(TLAPOR – TLAHIR) yang tidak munasabah.



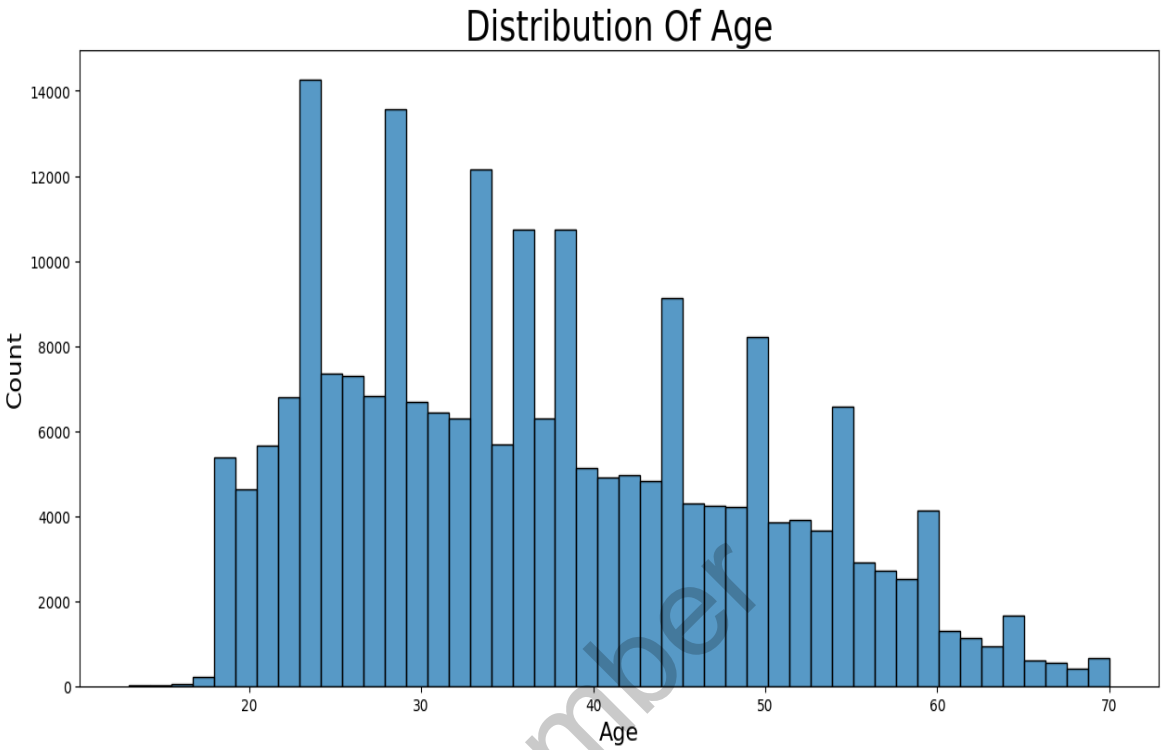
Rajah 4.2 Graf yang menunjukkan taburan data bagi umur bagi senarai OB yang membuat tuntutan di PERKESO

Berdasarkan graf di atas, terdapat taburan bagi nilai 0 dan sehingga umur 120 tahun. Data kajian ini merupakan data tuntutan PERKESO daripada golongan masyarakat yang bekerja di Malaysia, oleh itu adalah mustahil golongan bekerja terdiri daripada umur 0 . Bagi menentukan kadar minimum bagi kadar umum golongan bekerja di Malaysia Akta Kanak-kanak dan Orang Muda (Perkerjaan)(Pindaan) 2019 iaitu Akta A1586) Seksyen 2(2a) menyatakan bahawa umur minimum yang membolehkan

seseorang bekerja ialah 13 tahun di mana hanya kerja ringan sahaja dibenarkan. Justeru itu, kadar minimum tahap umur bekerja dalam kajian adalah 13 tahun merujuk kepada akta yang dinyatakan. Manakala kadar maksimum umur golongan bekerja di Malaysia pada nilai 120 adalah mustahil berikutan pada tahap ini kebanyakan manusia mengalami masalah kemerosotan kesihatan . Tambahan pula, tahap umur bersara atau pencen di Malaysia adalah 60 tahun. Kadar maksimum umur golongan bekerja dalam kajian ini ditetapkan pada 70 dengan mengambil kira golongan yang masih mampu bekerja sehingga umur 70 tahun secara purata. Kaedah yang digunakan untuk menggantikan nilai tidak munasabah bagi Umur(TLAPOR – TLAHIR) adalah menggunakan fungsi statistik *mean*.

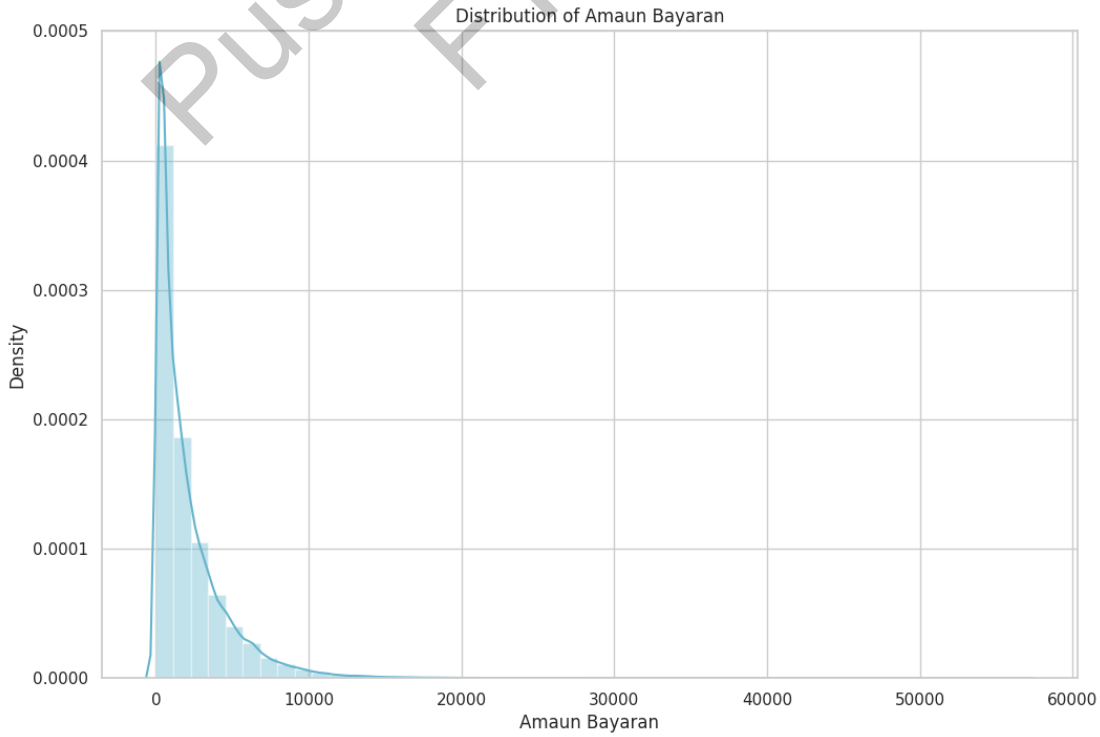
Jadual 4.11 Statitistik Deskripsi bagi umur

	Umur (TLAPOR-TLAHIR)
count	224548.00
mean	37.18
std	11.98
min	13.00
25%	27.00
50%	36.00
75%	46.00
max	70.00



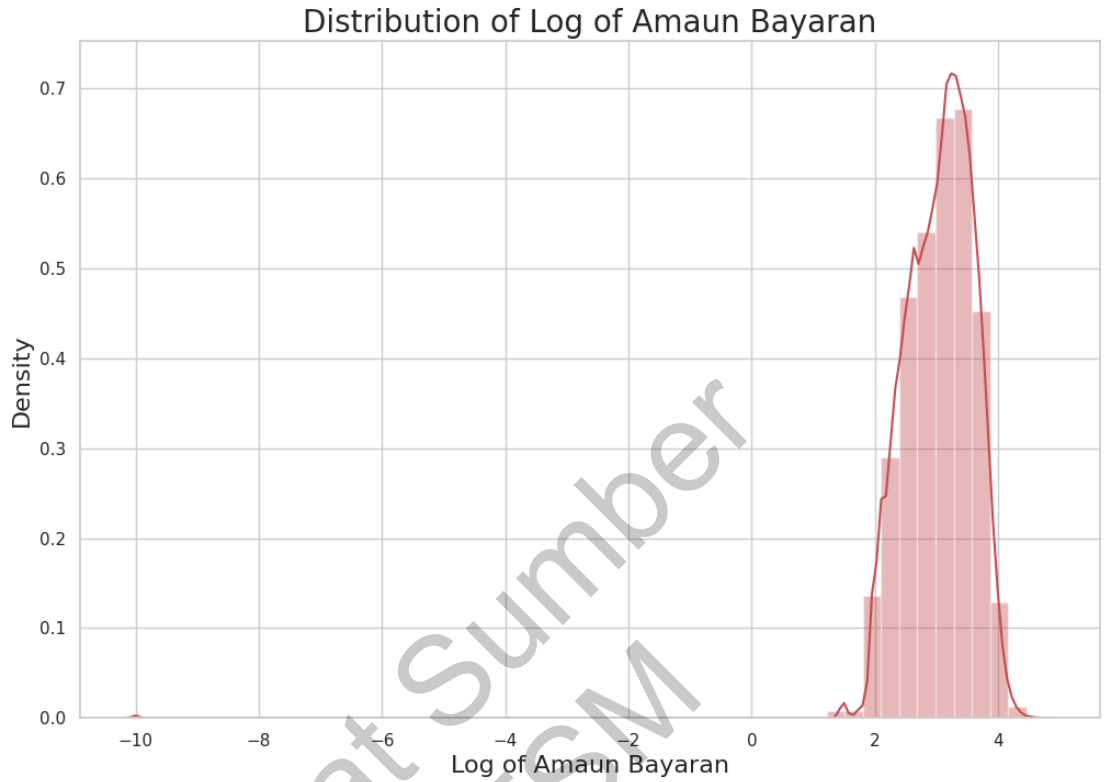
Rajah 4.3 Graf yang menunjukkan taburan data bagi umur bagi senarai OB yang membuat tuntutan di PERKESO (selepas nilai penggantian dibuat)

Berikut adalah analisis yang dilakukan untuk memahami korelasi di antara input atau pembolehubah tidak bergantung.



Rajah 4.4 Graf yang menunjukkan taburan data bagi amaun tuntutan yang dibuat oleh OB

Perhatikan taburan data bagi amaun tuntutan bayaran yang dibuat oleh OB condong ke kanan. Untuk menjadikan lebih normal, log semula jadi digunakan.



Rajah 4.5 Graf yang menunjukkan taburan data bagi amaun tuntutan yang dibuat oleh OB menggunakan log semula jadi.

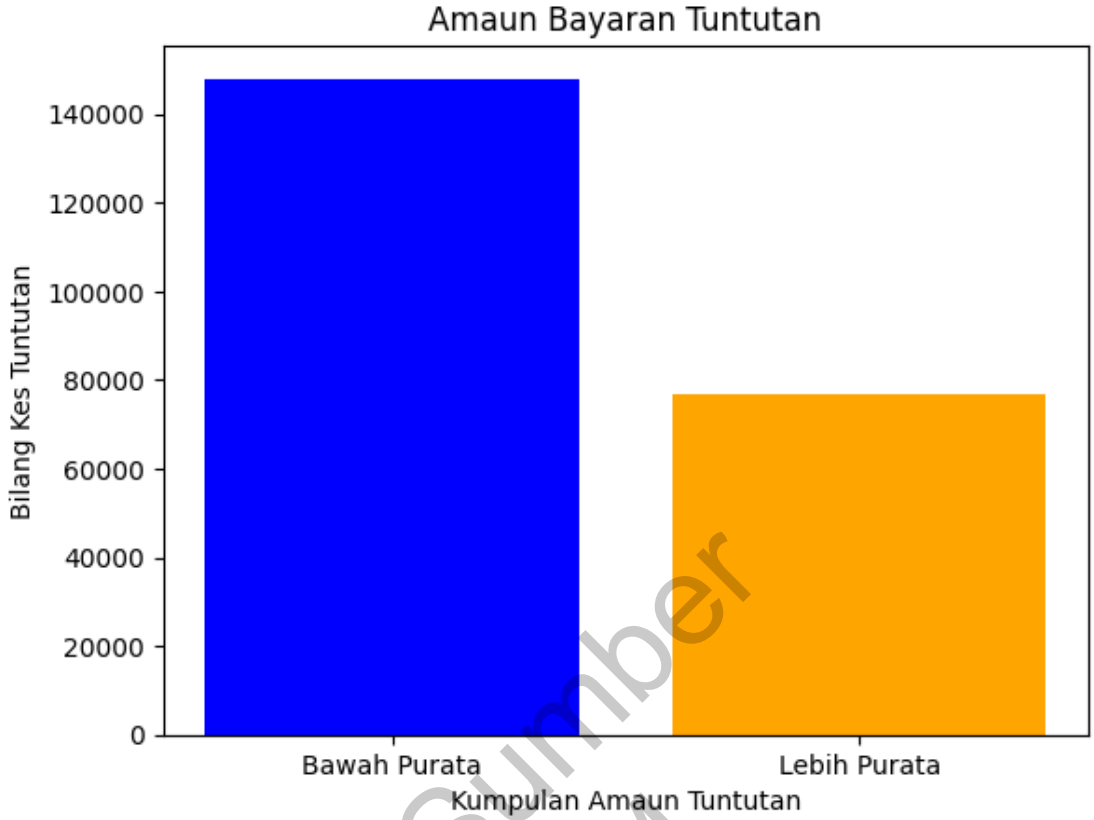
Taburan data bagi amaun tuntutan bayaran yang dibuat oleh OB adalah simetri selepas menggunakan log semula jadi.

4.5.1 Purata Amaun Bayaran

Purata amaun bayaran adalah RM2124.00. Bagi mengenalpasti trend atau corak amaun bayaran yang dituntut oleh OB secara keseluruhan satu graf analisis dilakukan.

Jadual 4.12 Statitistik purata bagi kes tuntutan bayaran

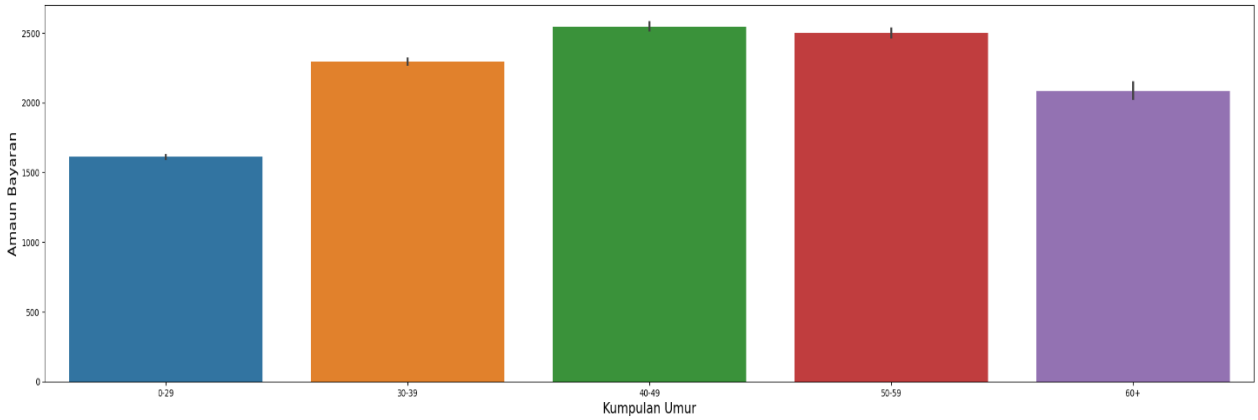
Purata Bayaran	**Bilangan Kes**
Bilangan Kes Tuntutan yang kurang dari keseluruhan purata	147954
Bilangan Kes Tuntutan yang lebih dari keseluruhan purata	76594



Rajah 4.6 Graf yang menunjukkan bilangan kes tuntutan berdasarkan purata bayaran

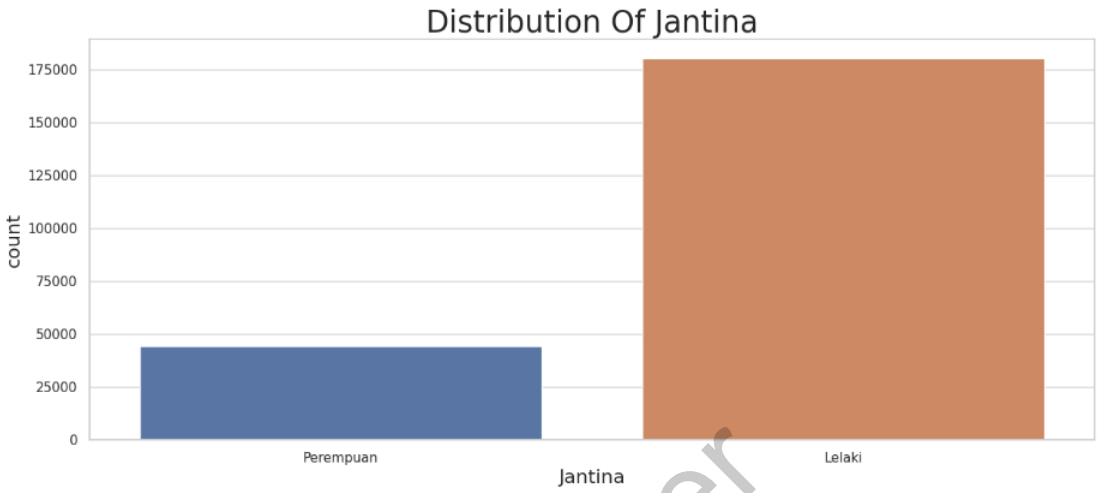
Berdasarkan graf yang dihasilkan kebanyakan amaun tuntutan yang dikemukakan oleh OB berada di bawah purata bayaran.

4.5.2 Hubungan Atribut Umur Dan Amaun Bayaran



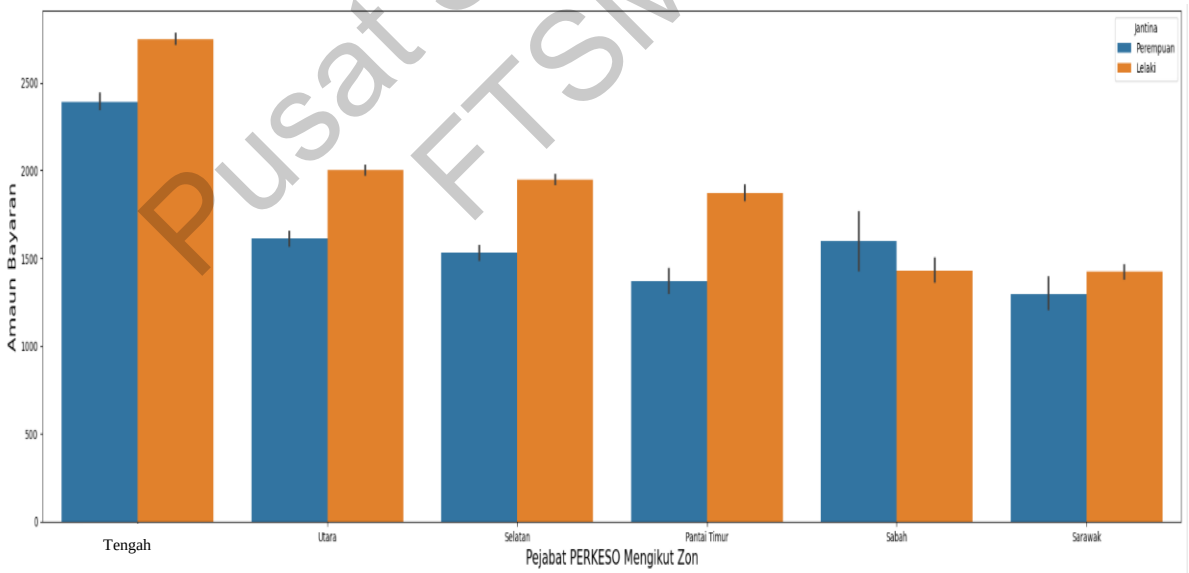
Rajah 4.7 Graf yang menunjukkan taburan amaun bayaran mengikut kumpulan umur

4.5.3 Hubungan Atribut Jantina Dan Amaun Bayaran



Rajah 4.8 Graf yang menunjukkan taburan amaun bayaran mengikut jantina

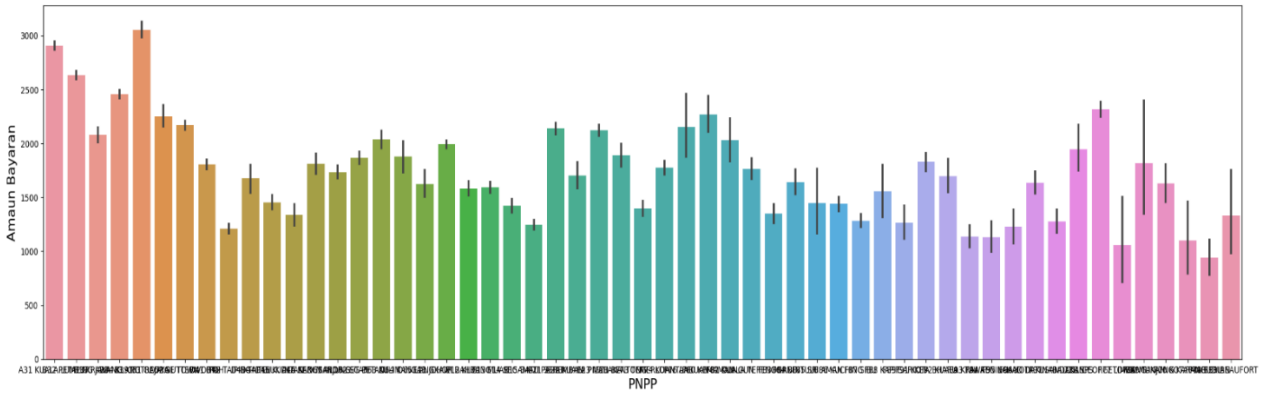
Menurut taburan data kes tuntutan di PERKESO, kes tuntutan banyak berlaku di kalangan pekerja lelaki. Ini berikutan pekerja lelaki merupakan tunggak utama pencari nafkah di dalam keluarga.



Rajah 4.9 Graf yang menunjukkan amaun tuntutan berdasarkan Pejabat PERKESO yang dikelaskan mengikut jantina

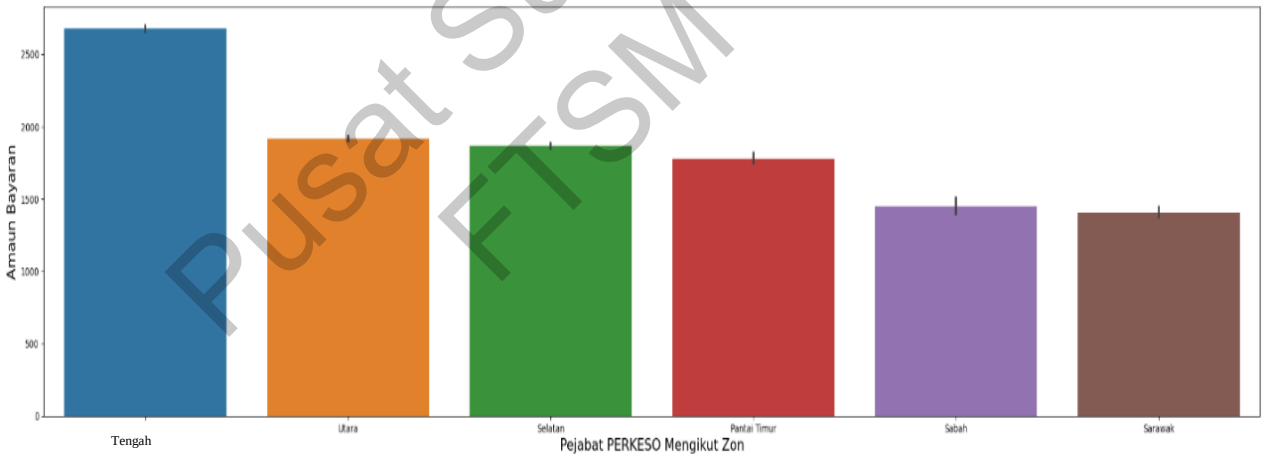
Trend tuntutan kes kemalangan tinggi di kalangan pekerja lelaki bagi semua zon kecuali Sabah. Tuntutan tertinggi berlaku di Zon Tengah berikutan pusat industri dan ekonomi berada di sekitar Zon Tengah.

4.5.4 Hubungan Atribut Pejabat Perkeso Dan Amaun Bayaran



Rajah 4.10 Graf yang menunjukkan amaun tuntutan berdasarkan Pejabat PERKESO

Graf yang dipaparkan pada Rajah 4.10 sukar untuk dibaca berikutan PERKESO mempunyai 53 cawangan pejabat secara keseluruhan. Bagi membolehkan graf lebih mudah untuk difaham atribut Pejabat PERKESO dikelaskan mengikut 6 zon utama iaitu Zon Tengah, Utara, Selatan, Pantai Timur, Sabah dan Sarawak.



Rajah 4.11 Graf yang menunjukkan amaun tuntutan berdasarkan Pejabat PERKESO yang dikelaskan mengikut Zon

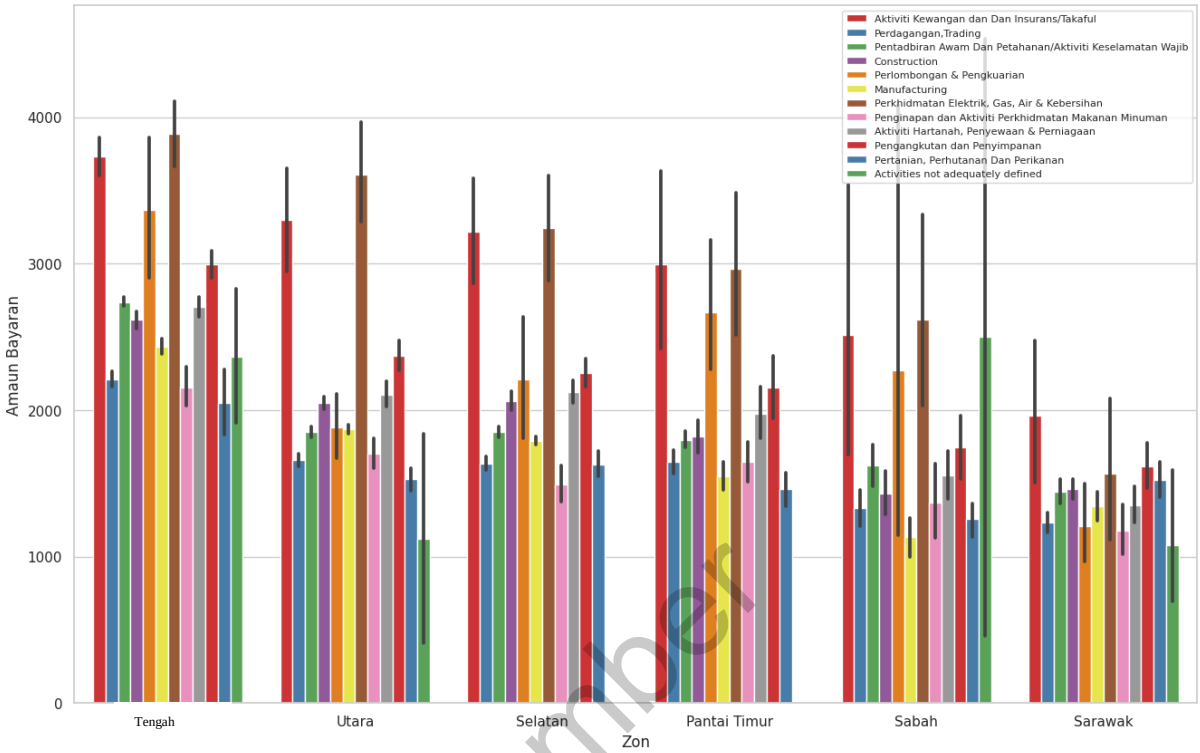
Melalui Rajah 4.11 data ini selari dengan graf-graf yang lain di mana kes tuntutan yang dikemukakan ke Pejabat PERKESO Zon Tengah mencatatkan kes tertinggi.

4.5.5 Hubungan Atribut Industri Dan Amaun Bayaran

Jadual 4.13 Total Amaun Bayaran Tuntutan berdasarkan Jenis Industri

Deskripsi Industri	**Total Amaun Bayaran (RM)**
Pentadbiran Awam Dan Pertahanan/Aktiviti Keselamatan Wajib	153799700
Manufacturing	102600400
Construction	64000250
Perdagangan,Trading	48486240
Aktiviti Hartanah, Penyewaan & Perniagaan	37302120
Pengangkutan dan Penyimpanan	26875270
Pertanian, Perhutanan Dan Perikanan	10111380
Penginapan dan Aktiviti Perkhidmatan Makanan Minuman	9268736
Aktiviti Kewangan dan Dan Insurans/Takaful	14111250
Perkhidmatan Elektrik, Gas, Air & Kebersihan	7885299
Perlombongan & Pengkuarian	2345492
Activities not adequately defined	292979.70
Name:Industri, dtype: int64	

Berdasarkan jadual 4.13 Industri Penginapan dan Aktiviti Perkhidmatan Makanan dan Minuman menjadi pemyumbang terbesar dalam tuntutan di PERKESO. Kemudian diikuti dengan industri Perkhidmatan Elektrik, Gas, Air dan Kebersihan dan Pembinaan.



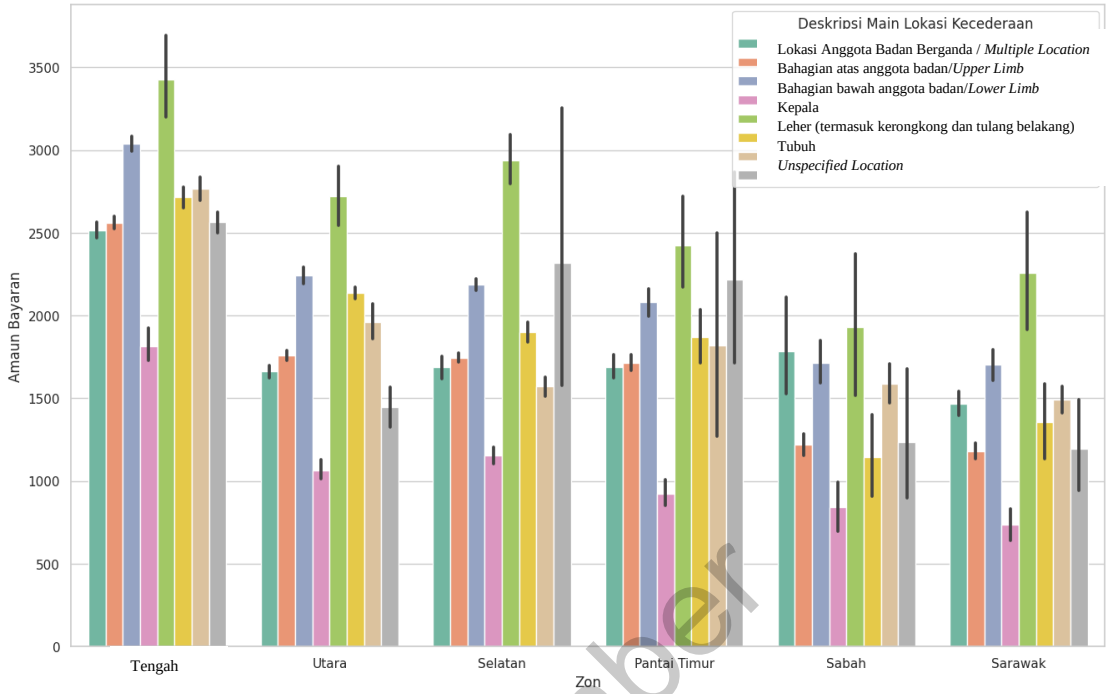
Rajah 4.12 Graf yang menunjukkan amaun tuntutan berdasarkan Pejabat PERKESO yang dikelaskan mengikut Industri

4.5.6 Hubungan Atribut Lokasi Kecederaan Utama Dan Amaun Bayaran

Jadual 4.14 Total Amaun Bayaran Tuntutan berdasarkan Lokasi Kecederaan Utama

Deskripsi Lokasi Kecederaan Utama	**Total Amaun Bayaran (RM)**
Bahagian atas anggota badan / *Upper Limb*	128375000
Bahagian bawah anggota badan / *Lower Limb*	120277200
Lokasi Anggota Badan Berganda / *Multiple Location*	71283020
Tubuh	69001920
Unspecified location of injury	36024180
Kepala	13459340
Kecederaan Am / *General Injuries*	24216810
Leher (termasuk kerongkong dan tulang belakang)	14441650
Name: Deskripsi Lokasi Kecederaan Utama, dtype: int64	

Jadual 4.14 menunjukkan lokasi kecederaan yang berganda menyumbang kepada tuntutan amaun bayaran yang tertinggi.



Rajah 4.13 Graf yang menunjukkan amaun tuntutan berdasarkan Pejabat PERKESO yang dikelaskan mengikut Lokasi Kecederaan Utama

4.5.7 Hubungan Atribut Jenis Kemalangan Dan Amaun Bayaran

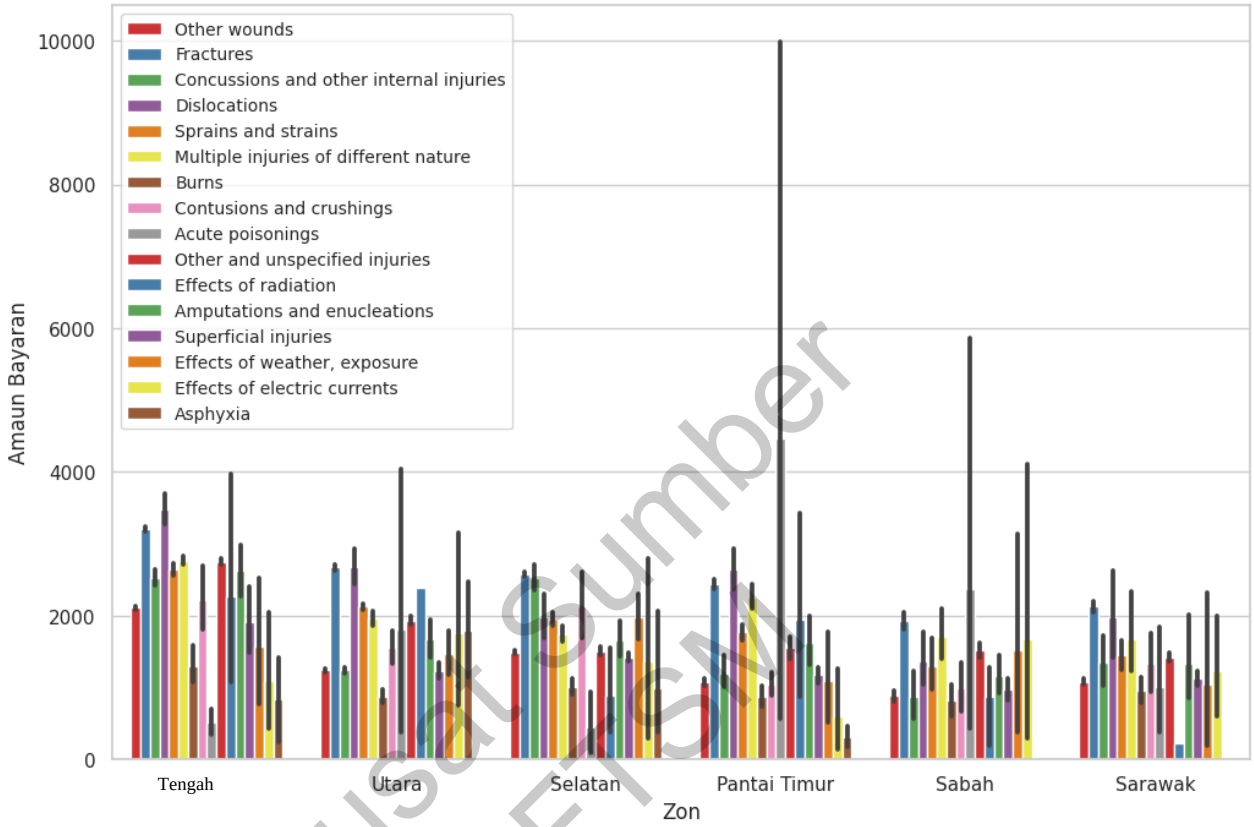
Jadual 4.15 Total Amaun Bayaran Tuntutan berdasarkan Jenis Kemalangan

Deskripsi Jenis Kemalangan	**Total Amaun Bayaran (RM)**
Other wounds	110181900
Fractures	164445400
Other and unspecified injuries	68572700
Sprains and strains	60834070
Multiple injuries of different nature	32795900
Concussions and other internal injuries	18007510
Superficial injuries	8867055
Dislocations	7679070
Burns	1760968
Contusions and crushings	1479099
Amputations and enucleations	1452162
Effects of weather, exposure	728105.30
Acute poisonings	101380.40
Effects of electric currents	58511.31
Asphyxia	54291.84
Effects of radiation	60983.26

Name: Deskripsi Jenis Kemalangan, dtype: int64

Jadual menunjukkan Jenis Kemalangan yang bersifat kecederaan pada permukaan (*Superficial Injuries*) menyumbang kepada tuntutan amaun bayaran yang

tertinggi. Tuntutan kemalangan yang kedua tertinggi dari sudut jenis kemalangan adalah terkehel (*dislocation*) dan diikuti dengan kesan cuaca dan pendedahan (*Effects of weather, exposure).*



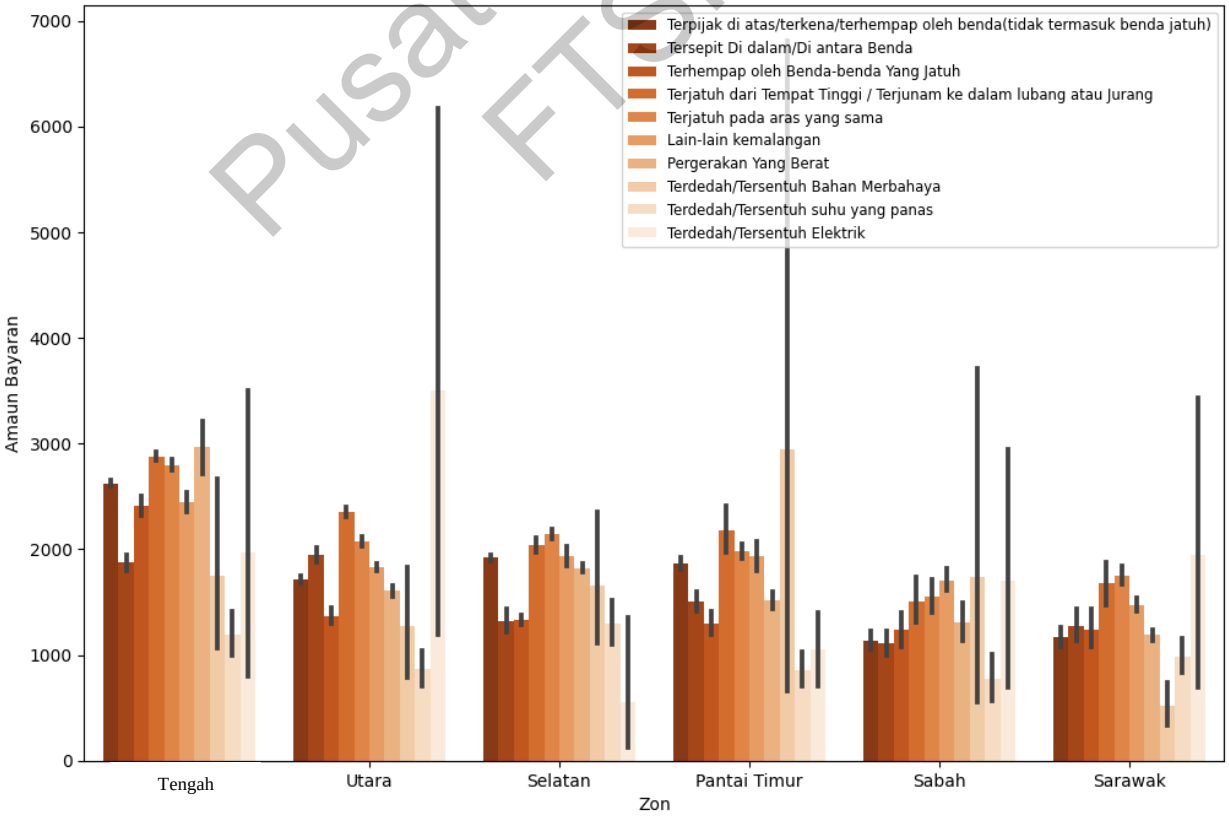
Rajah 4.14 Graf yang menunjukkan amaun tuntutan berdasarkan Pejabat PERKESO yang dikelaskan mengikut Jenis Kemalangan

4.5.8 Hubungan Atribut Sebab Kemalangan Utama Dan Amaun Bayaran

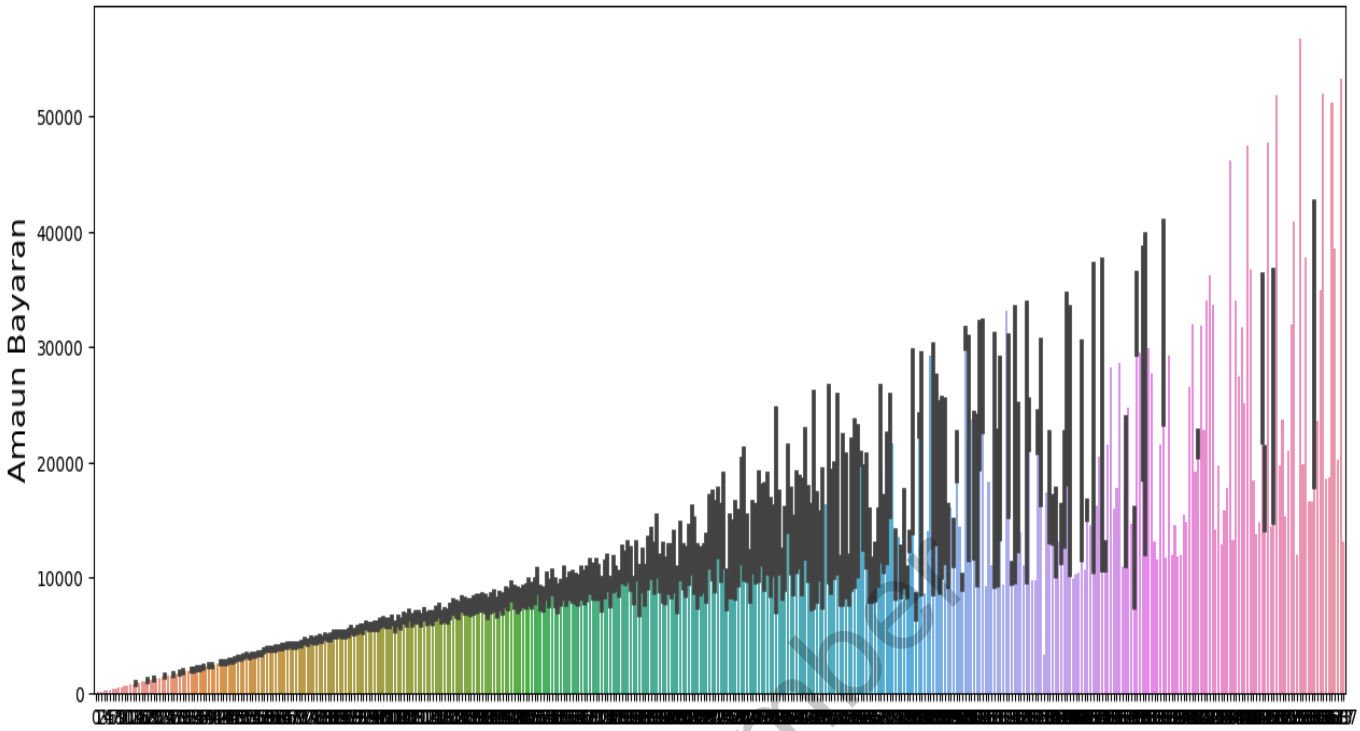
Jadual 4.16 Total Amaun Bayaran Tuntutan berdasarkan Sebab Kemalangan Utama

Deskripsi Sebab Kemalangan Utama	Total Amaun Bayaran (RM)
Terpijak di atas/terkena/terhempap oleh benda(tidak termasuk benda jatuh)	159026700
Terjatuh dari Tempat Tinggi / Terjunam ke dalam lubang atau Jurang	104462800
Terjatuh pada aras yang sama	86023340
Lain-lain kemalangan	47933340
Pergerakan Yang Berat	36431910
Terhempap oleh Benda-benda Yang Jatuh	22053830
Tersepit Di dalam/Di antara Benda	19739480
Terdedah/Tersentuh suhu yang panas	1054191
Terdedah/Tersentuh Bahan Merbahaya	255186.50
Terdedah/Tersentuh Elektrik	98310.22
Name: Deskripsi Sebab Kemalangan Utama, dtype: int64	

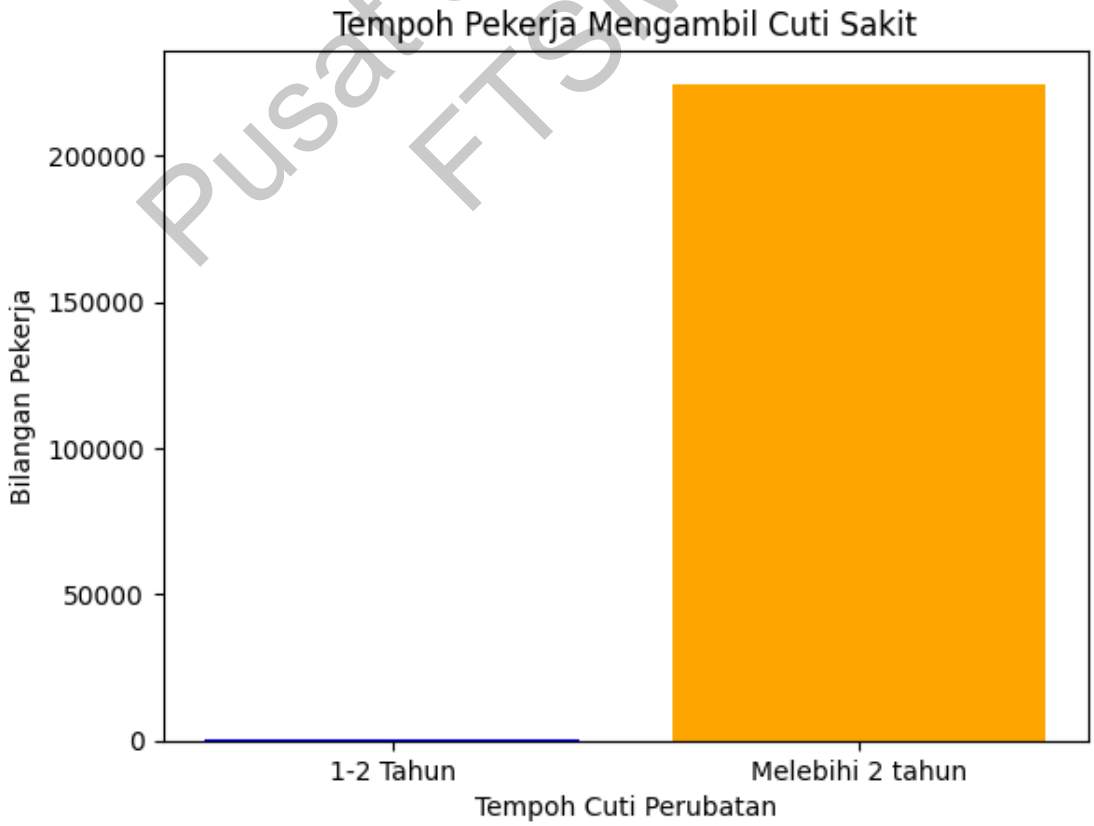
Jadual menunjukkan sebab utama kemalangan ialah terpijak di atas/tekena/terhempap oleh benda. Manakala sebab kemalangan yang tertinggi kedua adalah terjatuh dari tempat/terjunam ke dalam lubang atau jurang.



Rajah 4.15 Graf yang menunjukkan amaun tuntutan berdasarkan Pejabat PERKESO yang dikelaskan mengikut Sebab Kemalangan Utama



Rajah 4.16 Graf yang menunjukkan amaun tuntutan berdasarkan tempoh MC



Rajah 4.17 Graf yang menunjukkan amaun tuntutan berdasarkan Purata Tempoh MC

4.6 PENYEDIAAN DATA INPUT PEMBELAJARAN MESIN

Set data kajian ini mengandungi banyak pembolehubah dalam bentuk kategori. Data bagi pembolehubah seperti Industri, Zon PERKESO, Lokasi Kecederaan Utama, Jenis Kemalangan dan Sebab Kemalangan Utama berada dalam bentuk kategori. Data yang berbentuk kategori adalah tidak sesuai untuk diproses oleh Pembelajaran Mesin. Justeru itu data-data tersebut akan dipetakan atau dikodkan kepada nilai numerik supaya dapat diproses oleh algoritma pembelajaran mesin.

Berikut adalah contoh pemetaan atau pengkodan yang dijalankan ke atas pembolehubah yang menpunyai data bentuk kategori :

```
[ ]  #LOKASI KECEDERAAN
     KEPALA=(df1['Deskripsi Main Lokasi Kecederaan']=='Kepala')
     LEHER=(df1['Deskripsi Main Lokasi Kecederaan']=='Leher(termasuk kerongkong dan tulang belakang)')
     TUBUH=(df1['Deskripsi Main Lokasi Kecederaan']=='Tubuh')
     UPPERLIMB=(df1['Deskripsi Main Lokasi Kecederaan']=='Anggota Atas /Upper Limb')
     LOWERLIMB=(df1['Deskripsi Main Lokasi Kecederaan']=='Anggota Bawah / Lower Limb')
     MULLOC=(df1['Deskripsi Main Lokasi Kecederaan']=='Lokasi Berganda / Multiple Location')
     GENERAL=(df1['Deskripsi Main Lokasi Kecederaan']=='Kecederaan Am / General Injuries')
     UNLOC=(df1['Deskripsi Main Lokasi Kecederaan']=='Unspecified location of injury')

[ ]  df1.loc[KEPALA,'Deskripsi Main Lokasi Kecederaan']=1
     df1.loc[LEHER,'Deskripsi Main Lokasi Kecederaan']=2
     df1.loc[TUBUH,'Deskripsi Main Lokasi Kecederaan']=3
     df1.loc[UPPERLIMB,'Deskripsi Main Lokasi Kecederaan']=4
     df1.loc[LOWERLIMB,'Deskripsi Main Lokasi Kecederaan']=5
     df1.loc[MULLOC,'Deskripsi Main Lokasi Kecederaan']=6
     df1.loc[GENERAL,'Deskripsi Main Lokasi Kecederaan']=7
     df1.loc[UNLOC,'Deskripsi Main Lokasi Kecederaan']=8
```


Rajah 4.18 Graf yang menunjukkan contoh pemetaan atau pengekodan data berbentuk kategori kepada nilai numerik

Beberapa atribut atau pembolehubah yang bertindan dan tidak diperlukan juga dibuang pada peringkat ini. Pembolehubah dan atribut yang dibuang adalah atribut Tahun, Umur (TKKAKRU -TKHLAHIR), Kes, PNPP, Kod Industri, Kod Sub Lokasi Kecederaan, Deskripsi Sub Lokasi Kecederaan, Kod Jenis Kemalangan, Kod Sub Sebab Kemalangan dan Deskripsi Sub Sebab Kemalangan.

Matriks Korelasi dihasilkan untuk memahami hubungan di antara pembolehubah yang berada di dalam *dataframe* yang akan diproses oleh algoritma Pembelajaran Mesin. Matriks Korelasi membantu memahami pembolehubah berkait antara satu sama lain dan boleh menjadi panduan dalam pemilihan ciri (*features)*serta