BAB II PROSEDUR KERJA

2.1 Deskripsi Penugasan Kerja

Python for Data Science adalah program pembelajaran dari PT. Hacktivate Teknologi Indonesia yang diadakan selama delapan minggu yang terdiri dari enam belas pertemuan dimana terdapat dua pertemuan reguler per minggunya yang membahas mengenai materi *data science* dan lima pertemuan *engineering empathy* di minggu pertama yang membahas mengenai kondisi mental. Program ini berfokus pada *Python* dan *toolkit Python* populer seperti *Pandas, Matplotlib, Scikit-Learning,* dan *Seaborn, Data Cleaning,* Visualisasi Data, Statistik, *Machine learning,* dan *Deployment*. Setelah pembelajaran reguler selesai sesuai jamnya yaitu dari jam 19.00 hingga jam 22.00, maka setiap mahasiswa wajib meng*upload* materi yang telah dipelajari ke github. Selama belajar *python for data science* mahasiswa mendapatkan berbagai macam tugas yang menunjang pembelajaran, diantaranya: tugas *resume* video dari *website* kode.id, tiga tugas *assessment*, dan empat *final project.*

Dalam tugas *resume* video, mahasiswa wajib me*resume* beberapa materi yang terdapat di *website* kode.id diantaranya: Algoritma *Machine learning*, Komputasi Matematika Dengan Pustaka NumPy, Memahami Jupyter Notebook, Mempelajari *Machine learning* Dengan *Python*, dan Menguasai *Python* Tingkat Pemula. Setelah materi di *resume* maka mahasiswa wajib meng*upload*nya ke google classroom.

Selanjutnya dalam tugas *assessment* terdapat tiga tugas yang memiliki tujuan untuk memperkenalkan mahasiswa mengenai visualisasi data dengan *Python*. Dalam tugas pertama mahasiswa menggunakan dataset Himpunan Data Kejahatan London dari Kaggle. Mahasiswa harus menganalisis dataset berdasarkan kriteria yaitu harus terdapat visualisasi *area plot, box plot, histogram, bar chart, pie chart, scatter plot, word clouds,* dan *folium maps,* serta kesimpulan yang didapatkan. Pada tugas kedua mahasiswa menggunakan dataset NYC *Property Sales* dari Kaggle. Pada tugas kedua ini mahasiswa harus menganalisa untuk mendapatkan informasi yang terkandung di dalam dataset terdebut. Pada tugas

ketiga mahasiswa menggunakan dataset bank dari UCI *Machine learning Repository*. Dalam tugas ini mahasiswa harus mencari algoritma manakah yang cocok untuk memprediksi apakah *client* akan *subscribe (yes/no)* sebuah deposito dari data bank tersebut.

Pada *final project* terdapat 4 kasus yang berbeda - beda dan harus diselesaikan dengan menggunakan algoritma pemodelan yang berbeda – beda pula. Dalam *final project* 1, mahasiswa harus memprediksi harga taksi *online* uber dan lyft menggunakan dataset Uber and Lyft Dataset Boston, MA | Kaggle [1]. Final project 1 inilah yang penulis pakai sebagai topik laporan akhir MBKM untuk Fakultas Teknik Telekomunikasi dan Elektro. Final Project 1 ini dibuat guna mengevaluasi konsep Regression sebagai berikut:

1. Mampu memahami konsep regression dengan Linear Regression.
2. Mampu mempersiapkan data untuk digunakan dalam model Linear Regression.
3. Mampu mengimplementasikan Linear Regression untuk membuat prediksi

Dataset Uber and Lyft Dataset Boston, MA | Kaggle [1] ini memiliki 57 atribut, tetapi yang paling relevan ada 10 atribut dari semuanya, yaitu:

1. id
2. timestamp
3. hour
4. day
5. month
6. datetime
7. timezone
8. source: destinasi awal
9. destination: destinasi akhir
10. cab_type: tipe transportasi (uber / lyft)

Kemudian dalam *final project* 2, mahasiswa harus memprediksi besok hujan atau tidak berdasarkan dataset hujan di Australia. Dalam *final project* 3, mahasiswa harus mengklasifikasikan pasien gagal jantung dengan menggunakan dataset *heart*

failure dari Kaggle. Dalam *final project* 4, mahasiswa harus mengelompokkan pengguna kartu kredit berdasarkan dataset *credit card* yang didapatkan dari Kaggle.

2.2 Teori Dasar Pendukung

2.2.1 *Artificial Intelligence*

Kecerdasan buatan atau disebut dengan (*Artifical Intelligence* atau AI) didefinisikan dengan kecerdasan yang ditujukan oleh suatu objek buatan. Kecerdasan buatan ini umumnya dikenal sebagai komputer. Pada umumnya, kecerdasan buatan di kombinasikan ke dalam komputer agar bisa membantu pekerjaan seperti yang dapat diinginkan manusia.

Menurut John McCarthy, 1956, AI: untuk mengetahui dan memodelkan proses-proses berpikir manusia dan mendesain mesin agar dapat menirukan perilaku manusia. Cerdas, memiliki pengetahuan ditambah pengalaman, penalaran (bagaimana membuat keputusan mengambil tindakan), moral yang baik. Manusia yang cerdas dalam menyelesaikan permasalahan adalah manusia yang mempunyai pengetahuan dan pengalaman. Pengetahuan dapat diperoleh melalui pembelajaran. Semakin banyaknya bekal pengetahuan yang dimiliki tentu akan lebih mampu menyelesaikan permasalahan. Tetapi bekal pengetahuan saja tidak cukup, manusia juga diberikan akal untuk melakukan penalaran, mengambil kesimpulan berdasarkan pengetahuan dan pengalaman yang dimiliki.

Maka dari itu AI dapat membantu manusia untuk menyelesaikan masalah, karena AI mencakup bidang yang cukup besar. Mulai dari yang paling umum sampai yang spesifik. Contoh dari kecerdasan buatan itu sendiri diantaranya yaitu *Virtual Reality* (VR), Aplikasi ojek *online*, *drone, e-commerce, platform* google, mobil pintar, serta beberapa negara sudah mengembangkan rumah pintar [2].

Agar komputer bisa bertindak seperti dan sebaik manusia, maka komputer juga harus diberi bekal pengetahuan dan mempunyai kemampuan untuk menalar. Untuk itu AI akan mencoba untuk memberikan beberapa metoda untuk membekali komputer dengan kedua komponen tersebut agar komputer bisa menjadi mesin pintar. Lingkup utama kecerdasan buatan:

1. *Natural Language Processing* (NLP) cabang ilmu komputer dan linguistik yang mengkaji interaksi antara komputer dengan bahasa (alami) manusia. NLP sering dianggap sebagai cabang dari kecerdasan buatan dan bidang kajiannya bersinggungan dengan linguistik komputasional. Kemajuan dibidang ini membuat komputer dapat melakukan penerjemahan dari satu bahasa manusia ke bahasa manusia yang lain. Inti pengolahan bahasa alami ada dalam *parser*. *Parser* adalah bagian yang membaca kalimat dari bahasa sumber dan menguraikan serta menganalisis kata-kata yang terdapat di dalam kalimat tersebut dan mencocokkan dengan tata bahasa yang benar. Pendukung *parser* adalah kamus yang berisis kosa kata. Keluaran *parser* akan diproses oleh bagian yang disebut representasi pengetahuan, yang berperan dalam mengartikan kalimat masukan. Pada aplikasi penerjemahan,setelah makna kalimat diketahui bagian penerjemah keluaran akan menghasilkan keluaran berupa teks dalam bahasa alami.
2. *Knowledge Representation* (KR) adalah suatu proses untuk menangkap sifat-sifat penting pada sebuah permasalahan dan membuat informasi tersebut dapat diakses oleh prosedur pemecahan permasalahan.
3. *Automated Reasoning* (AR) Penalaran otomatis merupakan bidang ilmu komputer dan logika matematika didedikasikan untuk memahami berbagai aspek penalaran.
4. *Machine Learning* (ML) adalah sebuah cabang dari kecerdasan buatan, Inti dari mesin belajar berkaitan dengan representasi dan generalisasi.
5. *Computer Vision* (CV) adalah bidang yang mencakup metode untuk memperoleh, mengolah, menganalisis, dan pemahaman gambar dan, secara umum, data dimensi tinggi dari dunia nyata untuk menghasilkan informasi numerik atau simbolis, misalnya dalam bentuk keputusan.
6. *Robotic* (R) adalah cabang dari teknologi yang berhubungan dengan desain, konstruksi, operasi, dan aplikasi robot.

Untuk penyelesaian kecerdasan buatan mempunyai 4 teknik sebagai berikut:

1. *Searching* (teknik pencarian) yaitu teknik penyelesaian masalah yang mempresentasikan masalah ke dalam ruang keadaan (*state*) dan secara

sistematis melakukan pembangkitan dan pengujian *state-state* dari *initial state* sampai ditemukan suatu *goal state*.

2. *Reasoning* (teknik penalaran) yaitu teknik penyelesaian masalah yang mempresentasikan masalah masalah ke dalam *logic* (*mathematics tools* yang digunakan untuk mempresentasikan dan memanipulasi fakta dan aturan).
3. *Planning* (Perencanaan) suatu metode penyelesaian masalah dengan cara memecahkan masalah ke dalam *sub - sub* masalah yang lebih kecil, menyelesaikan masalah satu demi satu, kemudian menggabungkan solusi - solusi dari satuan terkecil menjadi konfrehensif.
4. *Learning*. Secara otomatis menerapkan aturan yang diharapkan bisa berlaku umum untuk data-data yang belum pernah kita ketahui [3].

Agar mesin bisa cerdas (bertindak seperti dan sebaik manusia) maka harus diberi bekal pengetahuan, sehingga mempunyai kemampuan untuk menalar. Untuk membuat aplikasi kecerdasan buatan ada 2 bagian utama yang sangat dibutuhkan:

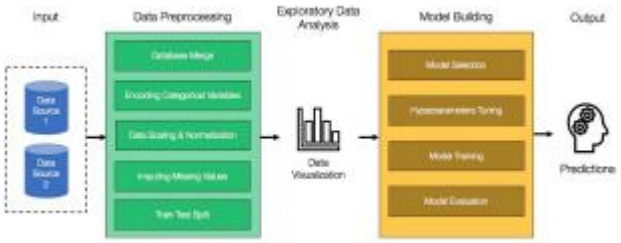
1. Basis Pengetahuan (*knowledge base*), bersifat fakta-fakta, teori , pemikiran dan hubungan antar satu dengan yang lainnya.
2. Motor Inferensi (*inference engine*), kemampuan menarik kesimpulan berdasarkan pengetahuan dan pengalaman [4].

2.2.2 *Machine Learning*

Istilah *Machine learning* pertama kali disebutkan oleh Arthur Samuel pada tahun 1959, pada saat itu ia menjelaskan dalam konteks menyelesaikan permainan catur dengan mesin. Secara istilah *machine learning* merupakan sebuah model komputasi statistik, yang berfokus pada prediksi menggunakan komputer. Algoritma *machine learning* membangun model matematika dari data sampel, yang dikenal sebagai "data pelatihan atau data *training*", untuk membuat prediksi atau keputusan tanpa diprogram secara eksplisit untuk melakukan tugas.

Inti alur kerja dari *machine learning* adalah tentang mengambil data mentah sebagai *input* dan menghasilkan prediksi sebagai *output*. Alur kerja *machine learning* dapat tersebut dapat dilihat pada Gambar 2.1 Kemudian, secara luas

algoritma *machine learning* dapat diklasifikasikan menjadi tiga jenis, yaitu *supervised learning, unsupervised learning,* dan *reinforcement learning* [5].



Gambar 2.1 Alur Kerja *Machine learning* [5].

Machine learning merupakan ilmu dan seni tentang pemrograman komputer yang bisa belajar dari data. Machine learning secara umum dibagi menjadi 4, yaitu *supervised learning, unsupervised learning, semi unsupervised learning,* dan *reinforcement learning.*

1. *Supervised learning*, dataset yang digunakan memiliki label. Label adalah tag pengenal dari data. Klasifikasi email spam adalah contoh *supervised learning.*
2. *Unsupervised learning*, dataset yang digunakan tidak memiliki label. Model ini melakukan belajar sendiri untuk melabeli atau mengelompokkan data.
3. *Semi Supervised* merupakan gabungan dari *supervised learning* dan *unsupervised learning.* Pada model *semi supervised*, dataset untuk pelatihan sebagian memiliki label dan sebagian tidak.
4. *Reinforcement Learning* adalah model belajar menggunakan sistem *reward and penalties*. Model belajar mendapatkan *reward* dan menghindari *penalties*.

Sebuah model *machine learning* tidak mampu untuk langsung mengolah data yang ditemukan dari berbagai sumber. Ada istilah *Garbage In - Garbage Out* yang berarti hasil dari *machine learning* akan buruk jika *input* yang dimasukkan juga buruk. Berikut adalah tahapan yang harus dilakukan dalam analisis teks dengan menggunakan *machine learnin*g:

1. *Data Cleaning/Preparation.* Sebelum membuat model *machine learning*, data *training* yang ada harus dilakukan proses data *cleaning*. Tahapan yang dilakukan dalam proses ini meliputi pengecekan konsistensi format, skala data,

duplikasi data, *missing value*, dan *skewness* (kemencengan yang menyebabkan distribusi data tidak seimbang).

2. *Data Preprocessing*. Dalam tahap ini dilakukan berbagai proses dimulai dari *case folding* (membuat teks berhuruf kecil semua). Kemudian, *tokenizing* yaitu ini akan dilakukan proses penghapusan angka, kata sambung, tanda baca, dan memecah kalimat ke dalam *tokens* (kata-kata penyusunnya). Kemudian dilakukan penghapusan kata-kata *stopword* seperti yang, dengan, dan kata penghubung lainnya.
3. *Data Normalization*. Dalam tahap ini dilakukan normalisasi teks ke dalam bentuk baku sesuai dengan kaidah bahasa (Inggris atau Indonesia). Semua kata-kata dalam bahasa *slank word* (gaul) akan diubah terlebih dahulu menjadi bentuk bakunya. Sehingga pada saat membangun model nanti akan mengurangi bias dan diharapkan bisa meningkatkan akurasi model yang akan dibuat.
4. *Data Classifying.* Klasifikasi adalah teknik untuk menentukan kelas atau kategori berdasarkan atribut yang diberikan. Klasifikasi masuk dalam kategori *supervised learning.* Sebuah model *classification* bertujuan untuk menentukan kelas berdasarkan atribut tertentu [6].

Machine learning merupakan salah satu cabang dari kecerdasan buatan dan untuk implementasi dapat menggunakan berbagai bahasa pemorgraman. Salah satu bahasa pemrograman yang memberikan kemudahan adalah bahasa *Python*. Dukungan bahasa *Python* dalam *machine learning* diantaranya banyaknya *library* yang sangat mendukung dalam mengimplementasikan algoritma di *machine learning* [7].

2.2.3 Prediksi

Prediksi (*prediction*) adalah memperkirakan nilai-nilai data bertipe apa saja dan kapan saja (masa lalu, sekarang, dan masa depan). Terdapat satu istilah yang mirip dengan prediksi, yaitu peramalan (*forecasting*) adalah memperkirakan nilai-nilai data *time series* dimasa depan [8].

Peramalan adalah kegiatan memperkirakan apa yang akan terjadi pada masa yang akan datang. Sedangkan ramalan adalah sesuatu situasi atau kondisi yang

diperkirakan akan terjadi pada masa yang akan datang, ramalan tersebut dapat didasarkan atas bermacam-macam cara yang dikenal dengan metode peramalan. Metode peramalan adalah cara memperkirakan secara kuantitatif apa yang akan terjadi pada masa depan, berdasarkan data yang relevan pada masa lalu. Maka metode peramalan ini digunakan dalam peramalan yang obyektif.

Sebagaimana diketahui bahwa metode peramalan merupakan cara berpikir yang sistematis dan pragmatis atas pemecahan suatu masalah. Dengan dasar ini, maka metode peramalan merupakan cara memperkirakan apa yang akan terjadi dimasa depan secara sistematis dan pragmatis melalui data yang relevan di masa yang lalu, dari hal ini metode peramalan diharapkan dapat memberikan objektivitas yang lebih besar. Berdasarkan uraian tersebut maka didapatkan suatu gambaran bahwa metode peramalan sangat berguna, karena akan membantu dalam mengadakan pendekatan analisa terhadap tingkah laku atau pola dari data yang lalu, sehingga dapat memberikan cara pemikiran, pengerjaan dan pemecahan yang sistematis dan pragmatis, serta memberikan tingkat keyakinan yang lebih besar atas ketetapan hasil ramalan yang dibuat atau disusun.

Pada umumnya peramalan dapat dibedakan dari beberapa segi tergantung dari cara melihatnya. Jika diklasifikasikan maka jenis-jenis peramalan dikelompokkan menjadi tiga macam sifat yang mendasarinya, yaitu:

1. Peramalan menurut sifat penyusunannya.
   Apabila dilihat dari sifat penyusunannya, maka peramalan dapat dibedakan atas dua macam, yaitu:
   a. Peramalan subyektif, yaitu peramalan yang didasarkan atas perasaan atau intuisi dari orang yang menyusunnya. Dalam hal ini pandangan atau "*judgement*" dari orang yang menyusunnya sangat menentukan baik tidaknya hasil ramalan tersebut.
   b. Peramalan obyektif, adalah peramalan yang didasarkan atas data yang relevan pada masa lalu, dengan menggunakan teknik-teknik dan metode-metode dalam penganalisaan data tersebut.
2. Peramalan menurut jangka waktu ramalan yang disusunnya.

Jika dilihat dari jangka waktu ramalan yang disusun, maka peramalan dapat dibedakan atas dua macam, yaitu:

a. Peramalan jangka panjang, yaitu peramalan yang dilakukan untuk penyusunan hasil ramalan yang jangka waktunya lebih dari satu setengah tahun atau tiga semester. Peramalan seperti ini misalnya diperlukan dalam penyusunan rencana pembangunan suatu negara atau daerah, *corporate planning*, rencana investasi atau rencana ekspansi perusahaan.
b. Peramalan jangka pendek, yaitu peramalan yang dilakukan untuk penyusunan hasil ramalan dengan jangka waktu yang kurang dari satu setengah tahun, peramalan seperti ini diperlukan dalam penyusunan rencana tahunan, rencana kerja operasional dan anggaran.

3. Peramalan menurut kategori jenis data yang digunakan.

   Metode peramalan jika dilihat dari jenis data yang digunakan dapat diklasifikasikan dalam dua kategori, yaitu:

   a. Metode Kualitatif. Metode ini digunakan tanpa ada model matematik, biasanya disebabkan oleh data yang ada tidak cukup representatif untuk meramalkan masa yang akan datang (*long term forecasting*). Peramalan kualitatif menggunakan pertimbangan pendapat-pendapat para pakar yang ahli atau *expert* di bidangnya. Adapun kelebihan dari metode ini adalah biaya yang dikeluarkan sangat murah (tanpa data) dan cepat diperoleh. Sementara kekurangannya yaitu bersifat subyektif sehingga seringkali dikatakan kurang ilmiah.
   b. Metode Kuantitatif Penggunaan metode ini didasari ketersediaan data mentah disertai serangkaian kaidah matematis untuk meramalkan hasil di masa depan. Terdapat beberapa macam model peramalan yang tergolong metode kuantitatif, yaitu:
      - Model-model Regresi. Perluasan dari metode *linear regression*, yaitu meramalkan suatu variabel yang memiliki hubungan secara linier dengan variabel bebas yang diketahui atau diandalkan.
      - Model Ekonometrik. Menggunakan serangkaian persamaan-persamaan *regresi*, yaitu terdapat variabel-variabel tidak bebas yang menstimulasi segmen-segmen ekonomi seperti harga dan lainnya.

- Model *Time Series Analysis* (Deret Waktu). Memasang suatu garis trend yang representatif dengan data-data masa lalu (*historis*) berdasarkan kecenderungan datanya dan memproyeksikan data tersebut ke masa yang akan dating [9].

Adapun tahapan peramalan secara ringkas terdapat tiga tahapan yang harus dilalui dalam perancangan suatu metode peramalan, yaitu:

1. Melakukan analisa pada data masa lampau. Langkah ini bertujuan untuk mendapatkan gambaran pola dari data bersangkutan.
2. Memilih metode yang akan digunakan. Terdapat bermacam-macam metode yang tersedia dengan keperluannya. Pemilihan metode dapat mempengaruhi hasil ramalan. Hasil ramalan diukur dengan menghitung *error* atau kesalahan terkecil. Oleh karena itu, tidak ada metode peramalan yang pasti baik untuk semua jenis data.
3. Proses transformasi dari data masa lampau dengan menggunakan metode yang dipilih. Apabila diperlukan maka diadakan perubahan sesuai kebutuhannya [10].

2.2.4 *Linear Regression*

Metode *Linear Regression* adalah regresi yang melibatkan hubungan antara satu variabel dependen dengan satu variabel independen atau variabel dependen (Y) dan variabel independen (X). Hubungan variabel dependen dan varibel independen tergantung dalam beberapa bentuk persamaan, sebagai berikut hubungan linear, eksponensial dan yang terakhir berganda. Tujuan penggunaan analisis regresi adalah untuk mengestimasi nilai variabel dependen yang didasarkan pada nilai variabel independen. Metode *linear regression* didasarkan pada pola hubungan data terkait masa lalu. Secara umum variabel yang dapat diprediksi yang diwakili oleh variabel yang direpresentasikan oleh variabel (seperti persediaan) dipengaruhi oleh besar kecilnya variabel bebas. Hubungan yang terjadi antara variabel independen dengan variabel yang akan ditemukan adalah sebuah fungsi [11].

Rumus untuk Regresi Linear dengan metode kuadrat terkecil atau sederhana adalah:

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n\,(\sum x^2) - (\sum x)^2} \quad (1)$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n\,(\sum x^2) - (\sum x)^2} \quad (2)$$

$$y = a + bx \quad (3)$$

Dengan y adalah kuantiti penjualan, x adalah periode penjualan atau bulan penjualan, a adalah konstanta yang menunjukan besarnya nilai y apabila x = 0, dan b adalah besaran perubahan nilai y[12].

2.2.4 *Random Forest*

Metode *Random Forest* (RF) adalah pengembangann dari metode *Classification and Regression Tree* (CART), yaitu dengan menerapkan metode *bootstrap aggregating* (*bagging*) dan *random feature selection*. *Random Forest* merupakan salah satu metode yang digunakan untuk klasifikasi dengan membangun banyak pohon klasifikasi. Metode ini dapat meningkatkan hasil akurasi, dengan cara membangkitkan simpul anak untuk setiap *node* (simpul diatasnya) dan dilakukan pemilihan secara acak,

Kemudian hasil klasifikasi dari setiap pohon diakumulasikan dan dipilh hasil klasifikasi yang paling banyak muncul. Metode ini terdiri dari *root node*, *internal node*, dan *leaf node*. *Root node* merupakan simpul yang terletak paling atas, atau biasa disebut sebagai akar dari pohon keputusan. *Internal node* adalah simpul percabangan, dimana *node* ini mempunyai *output* minimal dua dan hanya ada satu *input*. Sedangkan *leaf node* atau *terminal node* merupakan simpul terakhir yang hanya memiliki satu *input* dan tidak mempunyai *output*. Pohon keputusan dimulai dengan cara menghitung nilai *entropy* sebagai penentu tingkat ketidakmurnian atribut dan nilai *information gain* [13].

$$Entropy(Y) = -\sum_i p\,(c|Y) log_2 p(c|Y) \quad (2)$$

dengan Y adalah himpunan kasus dan $p(c|Y)$ adalah proporsi nilai Y terhadap kelas c.

$$Information\ Gain\ (Y,a) = Entropy(Y) - \sum_{v \epsilon Values(a)} \frac{|Yv|}{|Ya|} Entropy(Y_v) \quad (3)$$

dengan $Values(a)$ adalah semua nilai yang mungkin dalam himpunan kasus a, Y_v adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a, dan Y_a adalah semua nilai yang sesuai dengan a [14].

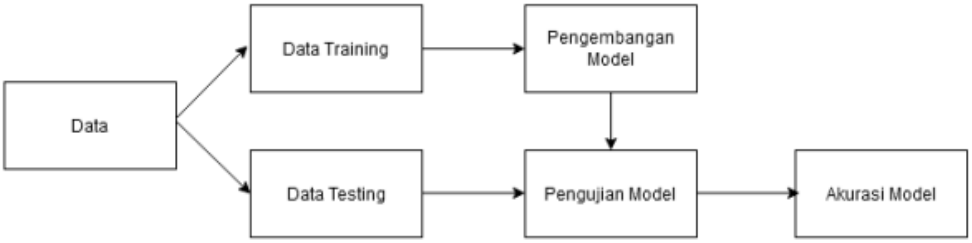
2.2.5 *Hold Out Validation*

Pada kondisi terbatasnya data yang digunakan untuk *training* dan *testing*, diperlukan metode untuk mendapatkan hasil tingkat akurasi dari sebuah metode pada *machine learning*. Salah satu cara untuk validasi adalah dengan menggunakan metode *holdout*. Metode *holdout* adalah metode yang akan menyediakan sejumlah data untuk digunakan sebagai data *testing*, dan sisanya sebagai data *training*.

Saat proses pengacakan data untuk dibagi sebagai data *training* dan *testing*, sangat mungkin terjadi *overrepresented* pada salah satu atau lebih klasifikasi. Dalam artian bahwa klasifikasi tersebut dominan dibandingkan klasifikasi lainnya, sehingga data *training* dan *testing* yang tercipta menjadi tidak representatif. Maka dari itu diperlukan prosedur *stratification holdout*, dimana dengan prosedur ini dapat dijamin bahwa setiap klasifikasi dapat terwakili pada data *training* dan *testing* yang tercipta secara proporsional. Menurut Freitas, 2002, kelas yang terbagi dari hasil proses *holdout* proporsinya harus sedekat mungkin dengan proporsi aslinya. Dilakukan perulangan terhadap seluruh proses *training* dan *testing* beberapa kali dengan data *training* dan *testing* yang teracak. Kemudian diambil nilai rata-ratanya. Prosedur ini dikatakan sebagai *repeated holdout* [15].

Data yang sudah diketahui klasifikasinya dapat dibagi menjadi dua bagian yaitu data *training* yang digunakan untuk membuat model dan data *testing* yang digunakan untuk menguji model. Perbandingannya biasanya adalah 80% sebagai data *training* dan sisanya 20% sebagai data *testing*. Akurasi model dihitung dengan cara membandingkan antara hasil klasifikasi model yang diujikan ke data *testing*

dengan klasifikasi sebenarnya dari data *testing*. Gambar 2.2 memperlihatkan pendekatan holdout.



Gambar 2.2 Pendekatan *Hold Out* [16].

Pendekatan *holdout* merupakan pendekatan yang berguna karena sangat mudah digunakan, cepat dan fleksibel. Meskipun demikian, pendekatan *holdout* mempunyai kelemahan mendasar yaitu pemilihan data *training* dan data *testing.* Dengan variasi data *training* dan data *testing* yang berbeda, akan dihasilkan nilai akurasi yang berbeda pula [16].

2.2.6 *Explained Variance Score*

Explained variance score menjelaskan dispersi kesalahan dari himpunan data yang diberikan, dan rumusnya ditulis sebagai berikut:

$$\textit{Explained variance}\ (y,\hat{y}) = 1 - \frac{Var\ (y-\hat{y})}{Var\ (y)} \quad (4)$$

Di sini, $Var\ (y - \hat{y})$ dan $Var\ (y)$ adalah varians dari kesalahan prediksi dan nilai aktual masing-masing. Skor mendekati 1,0 sangat diinginkan, menunjukkan kuadrat yang lebih baik dari simpangan baku kesalahan [17].

2.2.7 *Mean Square Error*

Mean Square Error (MSE) adalah metode lain untuk mengevaluasi metode peramalan. Masing-masing kesalahan atau sisa dikuadratkan. Pendekatan ini mengatur kesalahan peramalan yang besar karena kesalahan-kesalahan itu dikuadratkan. Metode itu menghasilkan kesalahan-kesalahan sedang yang kemungkinan lebih baik untuk kesalahan kecil, tetapi kadang menghasilkan perbedaan yang besar. MSE merupakan cara kedua untuk mengukur kesalahan

peramalan keseluruhan. MSE merupakan rata-rata selisih kuadrat antara nilai yang diramalkan dan yang diamati. Kekurangan penggunaan MSE adalah bahwa MSE cenderung menonjolkan deviasi yang besar karena adanya pengkuadratan. Rumus untuk menghitung MSE adalah sebagai berikut:

$$MSE = \frac{\sum_{t=1}^{n}(Xt-Ft)^2}{n} \quad (5)$$

Dimana:

X_t = data aktual pada periode t.

F_t = nilai peramalan pada periode t.

n = jumlah data [18].

Salah satu ciri estimator yang baik yaitu memiliki MSE terkecil dan sifat estimator yang efisien. Untuk mengetahui estimator mana yang merupakan lebih efisien dan model terbaik. Untuk menentukan jenis uji mana yang paling mendekati kebenaran dilakukan dengan mengukur *error* (kesalahan). Untuk mengukur *error* biasanya digunakan *Mean Square Error*. Pengujian yang menghasilkan *error* terkecil adalah uji yang dipilih [19].

2.2.8 *Mean Absolute Error*

Mean absolute error (MAE) merupakan rata-rata nilai kesalahan yang bernilai mutlak positif dari jumlah data, sesuai persamaan berikut:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (6)$$

Hal tersebut bertujuan untuk mengantisipasi kesalahan atau *error* yang bernilai negatif, sehingga dapat menentukan nilai rata-rata kesalahan secara tepat [18].

Jika Xi merupakan data aktual untuk periode i dan Fi merupakan ramalan (atau nilai kecocokan/fitted value) untuk periode yang sama, kesalahan didefinisikan sebagai ei = Xi-Fi [20].