BAB III

METODOLOGI PENELITIAN

3.1. Kedudukan dan Koordinasi

Pelaksanaan program kerja magang ini berupa penelitian. Penulis berkedudukan sebagai peneliti dari Universitas Multimedia Nusantara yang memiliki sejumlah peran sebagai berikut:

1. Melakukan tahapan-tahapan *data mining* yang dimulai dari memilih topik permasalahan yang akan diteliti, mendefinisikan latar belakang masalah, mengumpulkan dan memahami data, memilih metode penyelesaian masalah berupa model *machine learning*, membangun model klasifikasi kanker payudara berdasarkan data analisis darah *Breast Cancer Coimbra Dataset* salah satu rumah sakit di Portugal, yaitu *Coimbra Hospital and University Centre* (CHUC) (Patrício et al., 2018).
2. Mengevaluasi model yang dibangun, dan memilih serta mengajukan model klasifikasi terbaik dari hasil penelitian.
3. Menyusun jurnal penelitian dengan judul "Penerapan Algoritma *Data Mining Decision Tree* untuk Prediksi Hasil Diagnosa Kanker Payudara".

Selama pelaksanaan program kerja magang berlangsung, penulis dibimbing dan diarahkan oleh bapak Iwan Prasetiawan selaku dosen pembimbing lapangan yang mengawasi pelaksanaan program kerja magang ini.

3.2. Tugas yang Dilakukan

Selama pelaksanaan program kerja magang berupa proyek penelitian independent, penulis memiliki beberapa tugas dan tanggung jawab yang harus dijalani. Tugas-tugas tersebut dapat dibagi menjadi enam bagian berdasarkan tahapan-tahapan pada metode data mining CRISP-DM antara lain:

1. *Business Understanding*

   Pada tahapan ini, penulis berusaha memahami dan mendalami latar belakang permasalahan sehingga perlu dilakukannya penelitian. Penulis juga menetapkan tujuan dari penelitian ini guna menyelesaikan permasalahan yang ada.

2. *Data Understanding*

   Pada tahapan ini, penulis mengumpulkan dan mempelajari data yang digunakan, mengidentifikasi kualitas data, dan melakukan sejumlah analisis deskriptif dan korelasi fitur dari data tersebut.

3. *Data Preparation*

   Pada tahapan ini, penulis menerapkan sejumlah teknik pra-pemrosesan data, pemilihan fitur, dan ekstraksi fitur yang akan digunakan pada pembuatan model klasifikasi.

4. *Modeling*

   Pada tahapan ini, penulis menentukan teknik pemodelan, algoritma *machine learning*, dan parameter terbaik untuk membangun model.

5. *Evaluation*

   Pada tahapan ini, penulis mengevaluasi model yang telah dibangun pada tahap sebelumnya berdasarkan performa akurasi model.

6. *Deployment*

   Pada tahapan ini, penulis mengimplementasikan model klasifikasi yang telah dibangun untuk memprediksi seseorang terserang kanker payudara atau tidak.

3.3. Uraian Pelaksanaan Kerja Magang

Pelaksanaan program kerja magang ini berlangsung selama delapan minggu terhitung mulai dari tanggal 21 September 2020 hingga 13 November 2020. Berikut disajikan pada Tabel 3.1 Uraian Pelaksanaan Kerja Magang yang menjabarkan kegiatan-kegiatan setiap minggunya selama melaksanakan program kerja magang.

Tabel 3.1 Uraian Pelaksanaan Kerja Magang

Minggu Ke-	Kegiatan	Mulai	Selesai
1 dan 2	*Business Understanding*	21 September 2020	2 Oktober 2020
2	*Data Understanding*	29 September 2020	2 Oktober 2020
2	*Data Preprocessing*	1 Oktober 2020	2 Oktober 2020
3, 4, dan 5	*Data Modeling*	5 Oktober 2020	23 Oktober 2020
6 dan 7	*Evaluation*	26 Oktober 2020	7 November 2020
8	*Deployment*	9 November 2020	13 November 2020

3.3.1. *Business Understanding*

Kanker payudara merupakan salah satu jenis penyakit kanker yang menempati urutan jumlah kasus kedua tertinggi dan telah menjadi salah satu penyebab utama kematian yang semakin meningkat setiap tahunnya (Chaurasia et al., 2018). Oleh karena itu, kemampuan untuk memprediksi keberadaan kanker payudara sejak dini dengan cara menganalisis tanda-tanda seseorang terkena kanker payudara perlu ditingkatkan (Ganggayah et al., 2019). Dengan adanya kemampuan yang dapat memprediksi kanker payudara sejak dini, tindak lanjut perawatan baik operasi maupun terapi yang lebih efektif dapat dilakukan serta meningkatkan kelangsungan hidup para penderita kanker payudara (Kaya Keleş, 2019).

Salah satu cara untuk memeriksa keberadaan kanker payudara yang umumnya dilakukan adalah melalui analisis darah, yaitu dengan cara mengukur kadar zat-zat tertentu yang terkandung dalam darah, seperti *Glucose*, *Insulin*, HOMA, *Leptin*, *Adiponectin*, *Resistin*, dan MCP-1 (Patrício et al., 2018). Dengan menggunakan data tersebut, salah satu hal yang dapat dilakukan untuk meningkatkan kemampuan prediksi seseorang terserang kanker payudara dengan lebih optimal adalah menerapkan teknologi *Data Mining* mengidentifikasi pola atau karakteristik seseorang terserang kanker payudara atau tidak sehingga dapat menjadi bahan pertimbangan untuk melakukan tindak lanjut pencegahan maupun perawatan yang tepat sedini mungkin agar dapat meningkatkan kelangsungan hidup para penderita kanker payudara.

Berdasarkan permasalahan tersebut, penelitian ini membangun model klasifikasi menggunakan beberapa algoritma klasifikasi *machine learning* yang dapat memprediksi seseorang merupakan penderita kanker payudara atau tidak melalui hasil analisis darah yang dilakukan sebelumnya. Proses pelaksanaan penelitian ini akan mengikuti tahapan-tahapan kerangka kerja *data mining*, yaitu *Cross-Industry Standard Process for Data Mining* (CRISP-DM).

3.3.2. *Data Understanding*

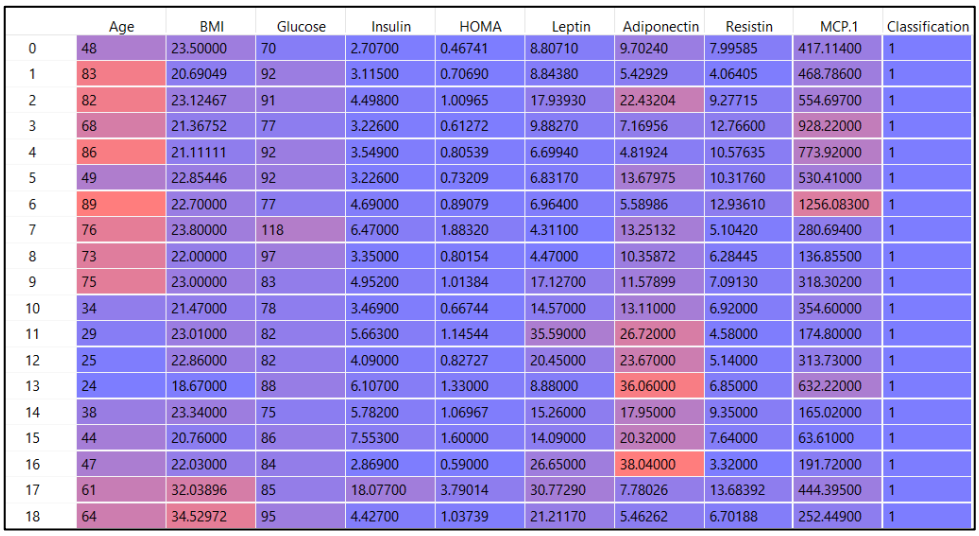
Dalam penelitian proyek independen ini, data yang digunakan adalah *Breast Cancer Coimbra Dataset* (Patrício et al., 2018). Data tersebut merupakan data dari hasil analisis darah dari pasien-pasien kanker payudara dan orang-orang yang dinyatakan sehat yang dikumpulkan oleh salah satu rumah sakit di Portugal, yaitu *Coimbra Hospital and University Centre* (CHUC) untuk penelitian sebelumnya yang dapat diakses melalui internet pada https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra. Data ini memiliki 10 atribut yang terdiri dari 9 atribut berupa fitur dan 1 atribut berupa variabel target. Pada Tabel 3.2 Atribut pada *Breast Cancer Coimbra Dataset* dapat dilihat atribut yang terdapat pada *Breast Cancer Coimbra Dataset*.

Tabel 3.2 Atribut pada *Breast Cancer Coimbra Dataset*

No	Nama Atribut	Keterangan
1	*Age*	Usia dari partisipan
2	*BMI*	*Body Mass Index* (Berat Badan) dari partisipan
3	*Glucose*	Jumlah kadar zat glukosa (gula darah) yang terkandung dalam tubuh partisipan.
4	*Insulin*	Jumlah kadar insulin yang terkandung dalam tubuh partisipan.
5	*HOMA*	Jumlah kadar *Homeostatic Model Assessment* (HOMA) yang terkandung dalam tubuh partisipan.
6	*Leptin*	Jumlah kadar leptin yang terkandung dalam tubuh partisipan.
7	*Adiponectin*	Jumlah kadar adiponektin yang terkandung dalam tubuh partisipan.
8	*Resistin*	Jumlah kadar resistin yang terkandung dalam tubuh partisipan.
9	*MCP-1*	Jumlah kadar *Monocyte Chemoattractant Protein-1* (*MCP-1*) yang terkandung dalam tubuh partisipan.
10	*Classification*	Kelas *Healthy* (Orang Sehat) atau *Patients* (Penderita Kanker Payudara)

Dapat dilihat pada Tabel 3.2 Atribut pada *Breast Cancer Coimbra Dataset*, atribut Nomor 1-9 merupakan fitur dan atribut Nomor 10 merupakan variabel target yang terdiri dari dua kelas, yaitu *healthy* (H) yang berarti kelas orang yang dinyatakan sehat dan *patients* (P) yang berarti pasien penderita kanker payudara. *Dataset* ini terdiri 116 sampel data hasil analisis darah yang terbagi menjadi 64 sampel data dari pasien-pasien penderita kanker payudara dan 52 sampel data dari orang-orang yang dinyatakan sehat. Sampel data para pasien penderita kanker payudara diambil sebelum dilakukan perawatan apapun dan sampel data orang-orang sehat diambil dari orang-orang yang dinyatakan tidak mengidap penyakit

apapun. Berikut contoh sampel data dapat dilihat pada Gambar 3.1 Contoh Sampel Data.



	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
0	48	23.50000	70	2.70700	0.46741	8.80710	9.70240	7.99585	417.11400	1
1	83	20.69049	92	3.11500	0.70690	8.84380	5.42929	4.06405	468.78600	1
2	82	23.12467	91	4.49800	1.00965	17.93930	22.43204	9.27715	554.69700	1
3	68	21.36752	77	3.22600	0.61272	9.88270	7.16956	12.76600	928.22000	1
4	86	21.11111	92	3.54900	0.80539	6.69940	4.81924	10.57635	773.92000	1
5	49	22.85446	92	3.22600	0.73209	6.83170	13.67975	10.31760	530.41000	1
6	89	22.70000	77	4.69000	0.89079	6.96400	5.58986	12.93610	1256.08300	1
7	76	23.80000	118	6.47000	1.88320	4.31100	13.25132	5.10420	280.69400	1
8	73	22.00000	97	3.35000	0.80154	4.47000	10.35872	6.28445	136.85500	1
9	75	23.00000	83	4.95200	1.01384	17.12700	11.57899	7.09130	318.30200	1
10	34	21.47000	78	3.46900	0.66744	14.57000	13.11000	6.92000	354.60000	1
11	29	23.01000	82	5.66300	1.14544	35.59000	26.72000	4.58000	174.80000	1
12	25	22.86000	82	4.09000	0.82727	20.45000	23.67000	5.14000	313.73000	1
13	24	18.67000	88	6.10700	1.33000	8.88000	36.06000	6.85000	632.22000	1
14	38	23.34000	75	5.78200	1.06967	15.26000	17.95000	9.35000	165.02000	1
15	44	20.76000	86	7.55300	1.60000	14.09000	20.32000	7.64000	63.61000	1
16	47	22.03000	84	2.86900	0.59000	26.65000	38.04000	3.32000	191.72000	1
17	61	32.03896	85	18.07700	3.79014	30.77290	7.78026	13.68392	444.39500	1
18	64	34.52972	95	4.42700	1.03739	21.21170	5.46262	6.70188	252.44900	1

Gambar 3.1 Contoh Sampel Data

Penelitian ini selanjutnya melakukan analisis deskriptif terhadap fitur-fitur yang ada pada *dataset* untuk melihat perbedaan karakteristik antara pasien kanker payudara dan orang yang dinyatakan sehat. Hasil analisis deskriptif yang dilakukan pada penelitian ini dapat dilihat pada Tabel 3.3 Hasil Analisis Deskriptif Fitur.

Tabel 3.3 Hasil Analisis Deskriptif Fitur

Atribut Fitur	***Mean***		***Median***	
	Healthy	***Patients***	***Healthy***	***Patients***
Age	58	56,6	65	53
BMI	28,3	26,9	27,7	27,4
Glucose	88,2	105,6	87	98,5
Insulin	6,9	12,5	5,4	7,6
Homa	1,5	3,6	1,1	2
Leptin	26,6	26,6	21,5	18,9
Adiponectin	10,3	10	8,1	8,4
Resistin	11,6	17,2	8,9	14,4
MCP-1	499,7	563	471,3	465,4

Dapat dilihat dari Tabel 3.3 Hasil Analisis Deskriptif Fitur, beberapa atribut fitur memiliki perbedaan yang signifikan antara pasien penderita kanker payudara dengan orang yang sehat. Dari hasil perhitungan rata-rata masing-masing fitur, setidaknya terdapat lima fitur yang memiliki perbedaan jumlah kadar zat dalam tubuh yang signifikan dengan urutan dari yang paling signifikan, yaitu kadar HOMA dengan perbedaan sebesar 140%, *Insulin* sebesar 81,2%, *Resistin* sebesar 49,1%, *Glucose* sebesar 19,7%, dan MCP-1 sebesar 12.67%.

Selanjutnya, penelitian ini juga melakukan analisis korelasi setiap atribut fitur dengan variabel target. Berikut hasil analisis korelasi tersebut disajikan pada Tabel 3.4 Hasil Analisis Korelasi Fitur.

Tabel 3.4 Hasil Analisis Korelasi Fitur

Atribut Fitur	Nilai Korelasi
Age	-0.044
BMI	-0.133
Glucose	0.384
Insulin	0.277
HOMA	0.284
Leptin	-0.001
Adiponectin	-0.0195
Resistin	0.227
MCP-1	0.0914

Dapat dilihat pada Tabel 3.4 Hasil Analisis Korelasi Fitur bahwa hanya terdapat beberapa fitur yang menunjukkan korelasi dengan variabel target, yaitu *Glucose* sebesar 0.384, HOMA sebesar 0.284, *Insulin* sebesar 0.277, dan *Resistin* sebesar 0.227, sedangkan fitur-fitur lainnya hanya menunjukkan korelasi yang sangat rendah. Beberapa fitur juga

menunjukkan angka negatif yang berarti hubungan antara fitur-fitur tersebut dengan variabel target memiliki sifat negatif, semakin tinggi nilai fitur-fitur tersebut maka semakin rendah kemungkinan subjek atau orang tersebut terserang kanker payudara, dan begitu juga sebaliknya untuk angka korelasi yang bersifat positif.

3.3.3. *Data Preparation*

Selanjutnya, tahapan data preparation dilakukan untuk mempersiapkan data sebelumnya menjadi input bagi model yang dibangun. Pada *dataset* yang digunakan tidak terdapat *missing values* ataupun *outlier* sehingga dapat dilanjutkan ke proses pembagian data. Pada penelitian ini, sampel data yang ada dibagi menjadi dua bagian, yaitu *training* dan *testing*. Karena jumlah data yang terbatas, penelitian ini melakukan pembagian data dengan menerapkan teknik *K-Fold Cross Validation* dengan nilai K = 5. Data dibagi menjadi 5 bagian dengan kurang lebih sama banyak, 4 bagian akan berperan sebagai data untuk *training* dan 1 bagian akan berperan sebagai data untuk *testing* secara bergantian. Selain itu, pada penelitian ini semua fitur yang ada digunakan untuk membangun model klasifikasi. Berikut kode Python yang digunakan untuk melakukan *5-Fold Cross Validation* dapat dilihat pada Gambar 3.2 Kode Python *K-Fold Cross Validation*.

```
from sklearn.model_selection import KFold

cv5 = KFold(n_splits=5, shuffle=True, random_state=1000)
```

Gambar 3.2 Kode Python *K-Fold Cross Validation*

3.3.4. *Modeling*

Tahapan selanjutnya yang dilakukan pada penelitian ini adalah membangun model klasifikasi menggunakan beberapa algoritma *machine learning*, yaitu *Decision Tree* (DT), *Random Forest* (RF), *K-Nearest Neighbors* (k-NN), *Support Vector Machine* (SVM), dan *Naïve Bayes* (NB) untuk memprediksi seseorang terserang penyakit kanker payudara atau orang sehat. Perbandingan antar performa akurasi dilakukan untuk menentukan algoritma dan model terbaik dalam memprediksi seseorang terserang penyakit kanker payudara atau tidak. Setiap algoritma juga dilakukan optimalisasi parameter untuk mendapatkan performa terbaik menggunakan teknik *Grid Search*. Berikut kode Python yang digunakan untuk melakukan *Grid Search* dapat dilihat pada Gambar 3.3 Kode Python Optimalisasi Parameter.

```python
# Decision Tree
def dtree(x, y):
    criterion = ['gini']
    splitter = ['best', 'random']
    max_depth = [int(x) for x in np.linspace(start=2, stop=30, num=29)]
    max_depth.append(None)
    min_samples_split = [int(x) for x in np.linspace(start=2, stop=40, num=39)]
    min_samples_leaf = [int(x) for x in np.linspace(start=1, stop=20, num=20)]
    max_leaf_nodes = [int(x) for x in np.linspace(start=2, stop=50, num=25)]
    max_leaf_nodes.append(None)
    max_features = ['auto', 'sqrt', 'log2']
    parameters = {'criterion': criterion,
                  'splitter': splitter,
                  'max_depth': max_depth,
                  'min_samples_split': min_samples_split,
                  'min_samples_leaf': min_samples_leaf,
                  'max_leaf_nodes': max_leaf_nodes,
                  'max_features': max_features}
    model = DecisionTreeClassifier()
    gs = GridSearchCV(estimator=model, param_grid=parameters, cv=cv5, scoring='accuracy', refit=False, verbose=10)
    gs.fit(x, y)
    print(gs.best_score_)
    print(gs.best_params_)
    return gs

# Random Forest
def rforest(x, y):
    n_estimators = [int(x) for x in np.linspace(start=500, stop=2000, num=10)],
    max_features = ['auto', 'sqrt']
    max_depth = [int(x) for x in np.linspace(start=10, stop=110, num=10)]
    min_samples_split = [2, 5, 10]
    max_leaf_nodes = [1, 2, 4, 5, 10]
    bootstrap = [True, False]
    parameters = {'n_estimators': n_estimators,
                  'max_features': max_features,
                  'max_depth': max_depth,
                  'min_samples_split': min_samples_split,
                  'max_leaf_nodes': max_leaf_nodes,
                  'bootstrap': bootstrap}
    model = RandomForestClassifier()
    gs = GridSearchCV(estimator=model, param_grid=parameters, cv=cv5, scoring='accuracy', refit=False, verbose=10)
    gs.fit(x, y)
    return gs
# Support Vector Machine
def svm(x, y):
    c = [0.1, 1, 10, 100, 1000]
    gamma = [1, 0.1, 0.01, 0.001, 0.0001]
    kernel = ['linear', 'rbf']
    parameters = {'C': c,
                  'gamma': gamma,
                  'kernel': kernel}
    model = SVC()
    gs = GridSearchCV(estimator=model, param_grid=parameters, cv=cv5, scoring='accuracy', refit=False, verbose=10)
    gs.fit(x, y)
    return gs

# Naive Bayes
def nbayes(x, y):
    var_smoothing = np.logspace(start=0, stop=-9, num=100)
    parameters = {'var_smoothing': var_smoothing}
    model = GaussianNB()
    gs = GridSearchCV(estimator=model, param_grid=parameters, cv=cv5, scoring='accuracy', refit=False, verbose=10)
    gs.fit(x, y)
    return gs

# K-Nearest Neighbor
def knn(x, y):
    parameters = {'n_neighbors': [3, 5, 11, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37],
                  'weights': ['uniform', 'distance'],
                  'metric': ['euclidean', 'manhattan']}
    model = KNeighborsClassifier()
    gs = GridSearchCV(estimator=model, param_grid=parameters, cv=cv5, scoring='accuracy', refit=False, verbose=10)
    gs.fit(x, y)
    return gs
```

Gambar 3.3 Kode Python Optimalisasi Parameter

Hasil dari dilakukannya teknik *Grid Search* adalah parameter terbaik untuk menghasilkan performa model terbaik. Berikut dapat dilihat pada Tabel 3.5 Hasil Optimalisasi Parameter merupakan parameter terbaik hasil

dari optimalisasi parameter yang dilakukan menggunakan teknik *Grid Search*.

Tabel 3.5 Hasil Optimalisasi Parameter

Algoritma	**Parameter**	
	Nama Parameter	**Nilai Terbaik**
Decision Tree	*Criterion*	*Gini*
	Max Depth	4
	Min Samples Split	2
	Min Samples Leaf	1
	Max Leaf Nodes	10
Random Forest	*N-Estimators*	*500*
	Max Features	*Auto*
	Max Depth	*87*
	Min Samples Split	*5*
	Min Samples Leaf	*1*
	Bootstrap	*False*
K-Nearest Neighbors	*N-Neighbors*	1
	Weights	*Uniform*
	Metric	*Euclidean*
Support Vector Machine	*C* (*Penalty Parameter*)	1
	Gamma	1
	Kernel	*Linear*
Naïve Bayes	*Variable Smoothing*	5.3366992312063

Setelah optimalisasi parameter, parameter terbaik digunakan untuk membangun model menggunakan masing-masing algoritma. Selanjutnya, evaluasi performa akurasi masing-masing model dilakukan untuk menentukan model terbaik pada tahapan *Evaluation*.

3.3.5. *Evaluation*

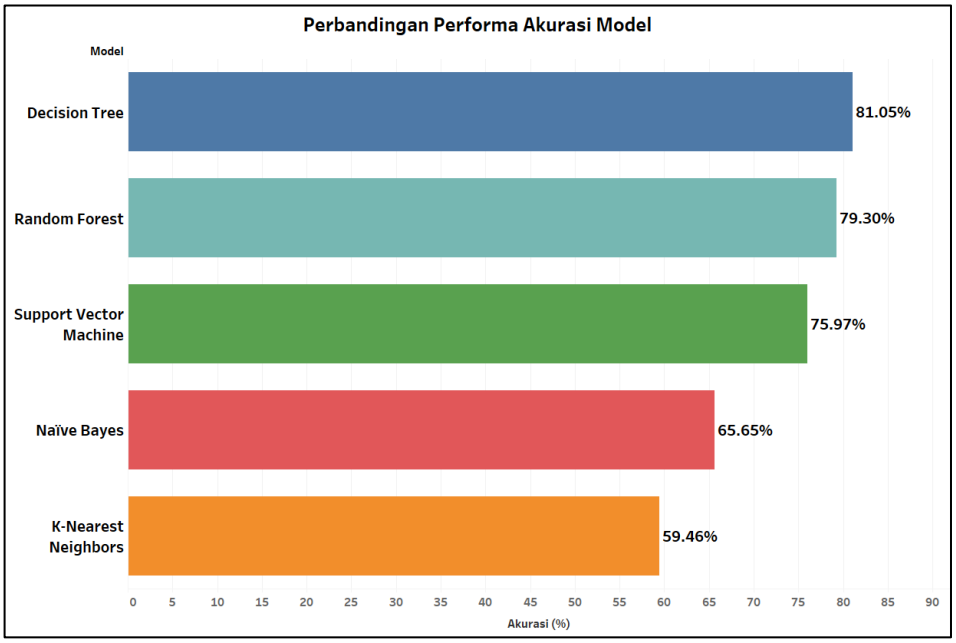
Pada tahapan ini, model-model yang sebelumnya telah dibangun menggunakan masing-masing algoritma dan parameter terbaik dievaluasi menggunakan data *testing* dengan menerapkan *5-Fold Cross Validation*.

Berikut hasil performa akurasi terbaik dari masing-masing model disajikan pada Tabel 3.6 Nilai Akurasi Model.

Tabel 3.6 Nilai Akurasi Model

Algoritma	Tingkat Akurasi
Decision Tree	81,05%
Random Forest	79,30%
Support Vector Machine	75,97%
Naïve Bayes	65,65%
K-Nearest Neighbor	59,46%

Dapat dilihat pada Tabel 3.6 Nilai Akurasi Model bahwa model dengan algoritma *Decision Tree* merupakan model dengan performa terbaik dalam memprediksi sampel data termasuk orang yang terserang penyakit kanker payudara atau orang sehat. Model yang dibangun dengan algoritma *Random Forest* juga memiliki performa yang hampir sama baiknya dengan model *Decision Tree*, sedangkan untuk model dengan algoritma SVM, KNN, dan Naïve Bayes cenderung memiliki performa yang rendah jika dibandingkan dengan model *Decision Tree*. Pada Gambar 3.4 *Bar Chart* Perbandingan Hasil Nilai Akurasi disajikan perbandingan performa akurasi dari masing-masing model dalam bentuk *Bar Chart*.



Gambar 3.4 *Bar Chart* Perbandingan Hasil Nilai Akurasi

Selanjutnya, hasil dari pemodelan yang dibangun pada penelitian ini dibandingkan dengan model yang dibangun pada penelitian-penelitian sebelumnya dari segi performa akurasi. Penelitian-penelitian pembanding dilakukan dalam kondisi penggunaan parameter dan rasio pembagian data yang berbeda, tetapi tetap menyertakan semua fitur atau atribut yang ada sebagai variabel penentu.

Berikut perbandingan performa akurasi model pada penelitian ini dengan penelitian terdahulu dapat dilihat pada Tabel 3.7 Perbandingan Nilai Akurasi dengan Penelitian Lain.

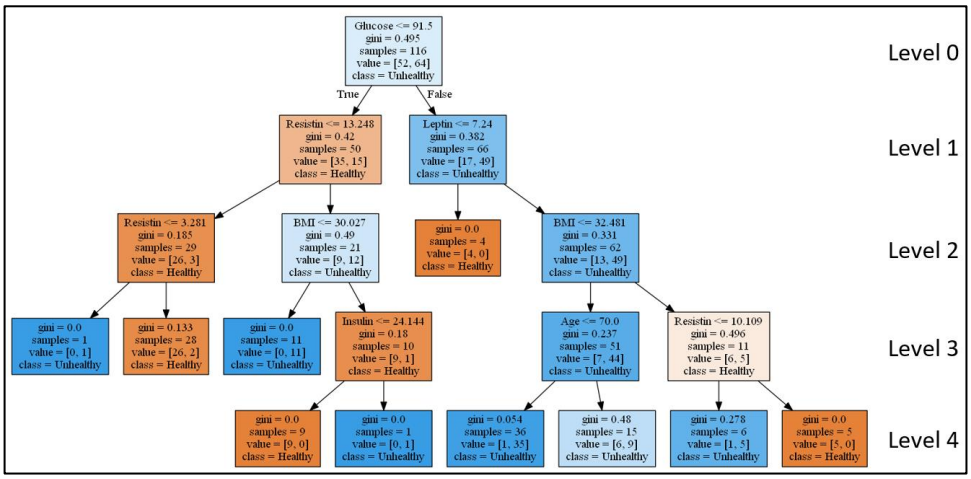
Tabel 3.7 Perbandingan Nilai Akurasi dengan Penelitian Lain

Algoritma	Tingkat Akurasi (%)		Referensi
	Penelitian Ini	Penelitian Terdahulu	
Decision Tree	81,05	69	(Austria et al., 2019)
		70	(Ray et al., 2019)
		72	(Cruz & Bernardino, 2019)
Random Forest	79,30	66	(Cruz & Bernardino, 2019)
		70	(Ray et al., 2019)
		74	(Austria et al., 2019)
Support Vector Machine	75,97	71	(Sardouk et al., 2019)
		72	(Austria et al., 2019)
		73	(Aslan et al., 2018)
Naïve Bayes	65,65	62	(Austria et al., 2019)
		66	(Ray et al., 2019)
		68	(Cruz & Bernardino, 2019)
K-Nearest Neighbour	59,46	48	(Ray et al., 2019)
		58	(Austria et al., 2019)

Dapat dilihat dari hasil perbandingan performa akurasi model yang telah dibangun pada penelitian ini dengan penelitian terdahulu untuk algoritma yang sama menunjukkan bahwa penelitian ini telah membangun model yang memiliki performa yang lebih baik. Model *Decision Tree* yang dibangun pada penelitian ini juga telah menunjukkan performa yang lebih baik dibandingkan model dengan algoritma yang lain dan model yang dibangun pada penelitian terdahulu. Untuk itu, model *Decision Tree* ini dapat digunakan pada tahap *deployment* untuk memprediksi data sampel baru berdasarkan hasil analisis darah.

3.3.6. *Deployment*

Pada tahapan ini, model terbaik hasil dari evaluasi pada tahapan sebelumnya adalah model *Decision Tree* dengan performa tingkat akurasi 81,05%. Model *Decision Tree* yang dibangun dapat dilihat pada Gambar 4.1 Model *Decision Tree*.



Gambar 4.1 Model *Decision Tree*

Pada Gambar 4.1 Model *Decision Tree* dapat dilihat terdapat dapat dilihat bahwa fitur utama (*root node*) yang menjadi faktor penentu seseorang terserang penyakit kanker payudara atau orang sehat adalah *Glucose* yang memiliki nilai *gini index*, yaitu sebesar 0.495. Beberapa fitur penentu seseorang terserang penyakit kanker payudara atau tidak yang disajikan melalui Tabel 4.1 Fitur Penentu Berdasarkan Nilai Indeks Gini.

Tabel 4.1 Fitur Penentu Berdasarkan Nilai Indeks Gini

Node	Fitur	Indeks Gini
Root Node (Level 0)	*Glucose*	0.495
Node 1 (Level 1)	*Resistin*	0.42
Node 2 (Level 1)	*Leptin*	0.382
Node 1 (Level 2)	*Resistin*	0.185
Node 2 (Level 2)	*BMI*	0.49
Node 3 (Level 2)	*BMI*	0.331
Node 1 (Level 3)	*Insulin*	0.18
Node 2 (Level 3)	*Age*	0.237
Node 3 (Level 3)	*Resistin*	0.496

3.4. Kendala yang Dihadapi

Berikut terdapat beberapa kendala yang dihadapi penulis selama pelaksanaan program kerja magang antara lain:

1. Penulis mengalami kesulitan untuk mendapatkan sumber data atau mengumpulkan data mengenai kanker payudara yang baru (5 tahun terakhir).
2. Prosedur pelaksanaan program kerja magang proyek independen. Penulis mengikut program *Batch* 1 dan masih sedikit informasi yang tersedia mengenai program ini serta cenderung berubah-ubah dari segi *timeline* karena adanya penyesuaian dengan peserta proyek independen lainnya.

3.5. Solusi atas Kendala

Berikut terdapat beberapa kendala yang dihadapi penulis selama pelaksanaan program kerja magang antara lain:

1. Solusi atas kendala pertama adalah menggunakan salah satu data yang ada dalam kurun waktu 5 tahun terakhir tentang kanker

payudara dan yang juga sekaligus menjadi data yang digunakan pada penelitian ini adalah *Breast Cancer Coimbra Dataset* (Patrício et al., 2018).

2. Solusi atas kendala kedua adalah adalah bertanya pada pihak *Student Development* dan *Human Resources Development* Universitas Multimedia Nusantara untuk arahan mengenai pelaksanaannya.