BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil

Berdasarkan metodologi yang telah dirancang pada kasus pemodelan prediksi risiko *Alzheimer disease* (AD) menggunakan algoritma *random forest* didapatkan hasil sebagai berikut.

Tabel 4. 1 Dataset Fitur

No	Fitur
1	*Age*
2	*Gender*
3	*Ethnicity*
4	*Education Level*
5	*BMI*
6	*Smoking*
7	*Alcohol Consumption*
8	*Physical Activity*
9	*Diet Quality*
10	*Sleep Quality*
11	*Family History Alzheimers*
12	*Cardiovascular Disease*
13	*Diabetes*
14	*Depression*
15	*HeadInjury*
16	*Hypertension*
17	*Systolic BP*
18	*Diastolic BP*
19	*Cholesterol Total*
20	*Cholesterol LDL*
21	*Cholesterol HDL*
22	*CholesterolTriglycerides*
23	*MMSE*
24	*FunctionalAssessment*
25	*MemoryComplaints*
26	*BehavioralProblems*
27	*ADL*
28	*Confusion*

29	*Disorientation*
30	*PersonalityChanges*
31	*DifficultyCompletingTasks*
32	*Forgetfulness*
33	*Diagnosis*

a. *Exploratory Data Analysis* (EDA)

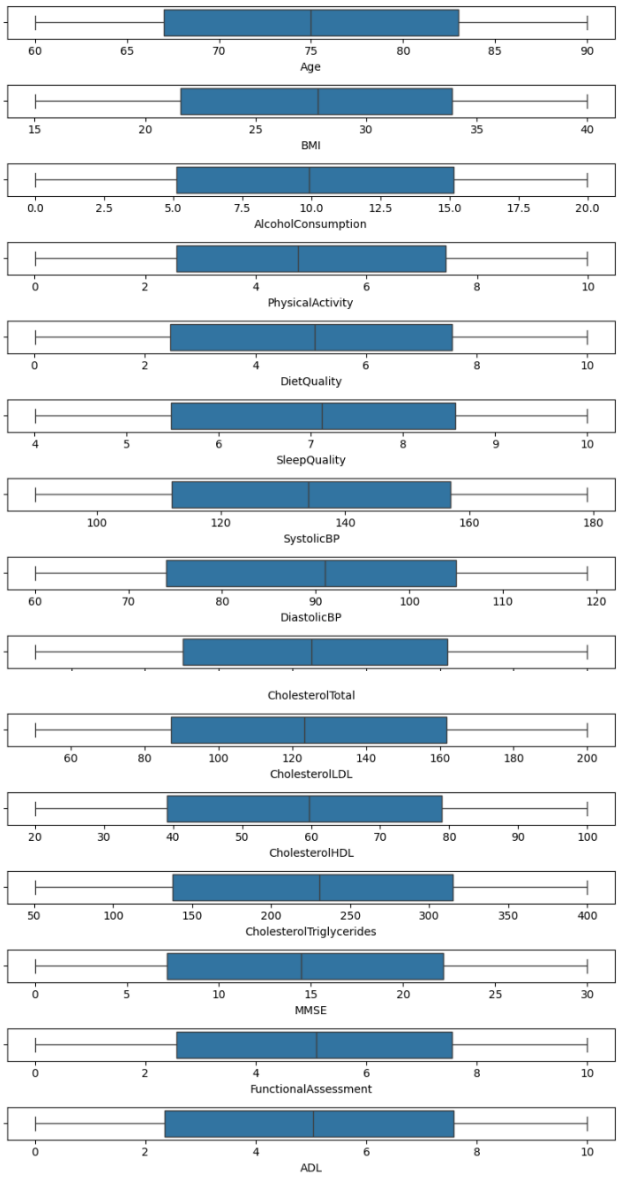
1) *Missing Value*

Age	0
Gender	0
Ethnicity	0
EducationLevel	0
BMI	0
Smoking	0
AlcoholConsumption	0
PhysicalActivity	0
DietQuality	0
SleepQuality	0
FamilyHistoryAlzheimers	0
CardiovascularDisease	0
Diabetes	0
Depression	0
HeadInjury	0
Hypertension	0
SystolicBP	0
DiastolicBP	0
CholesterolTotal	0
CholesterolLDL	0
CholesterolHDL	0
CholesterolTriglycerides	0
MMSE	0
FunctionalAssessment	0
MemoryComplaints	0
BehavioralProblems	0
ADL	0
Confusion	0
Disorientation	0
PersonalityChanges	0
DifficultyCompletingTasks	0
Forgetfulness	0
Diagnosis	0

Gambar 4. 1 *Missing Value*

Berdasarkan hasil analisis dengan dilakukan pengecekan *missing value* bahwa semua kolom dalam dataset semuanya menunjukkan angka 0 artinya tidak terdapat *missing value* dalam variable-variabel tersebut.

2) *Outlier*



Gambar 4. 2 *Outlier*

Berdasarkan analisis *outlier* menggunakan *boxplot* pada variable tidak menunjukkan adanya *outlier*, karena tidak ada titik data yang berada di luar batas garis whisker pada boxplot. Artinya data pada variable-variabel tersebut terdistribusi dengan baik dan tidak ada nilai yang menyimpang jauh dari pola umum.

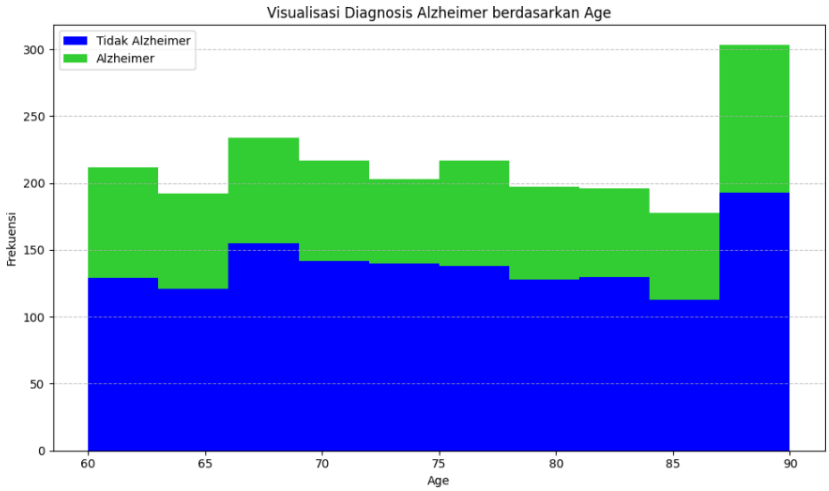
3) Distribusi Data

Gambar 4. 3 Distribusi Data Histogram

Berdasarkan visualisasi distribusi data tersebut, terlihat bahwa data memiliki beragam jenis distribusi. Beberapa menunjukkan distribusi miring dengan sebagian besar data terkonsentrasi di satu sisi, baik di sisi rendah maupun tinggi. Terdapat juga distribusi yang cukup seimbang atau merata di seluruh rentang nilainya. Selain itu, beberapa distribusi terlihat berbentuk bimodal atau memiliki dua puncak, sementara sebagian lainnya menunjukkan pola yang cenderung datar atau seragam. Secara keseluruhan, dataset ini mencerminkan campuran distribusi yang padat di satu kategori, distribusi merata, serta distribusi yang miring, menunjukkan adanya variasi pola data di setiap kelompok.

4) Visualisasi Fitur antar Target

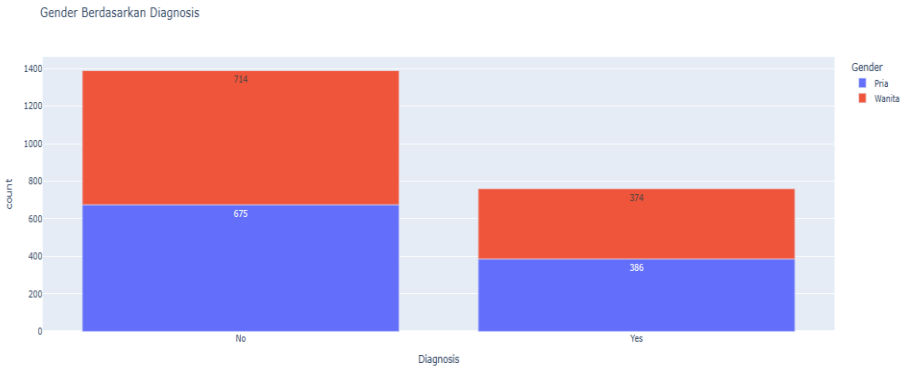
a) *Age* Berdasarkan Diagnosis



Gambar 4. 4 *Visualisasi Histogram Age* Berdasarkan *Diagnosis*

Berdasarkan visualisasi data, jumlah cenderung lebih tinggi pada yang tidak terdiagnosis AD (ditandai warna biru) dibandingkan dengan individu yang didiagnosis AD (ditandai warna hijau). Peningkatan jumlah AD cukup signifikan pada usia mendekati 90 tahun. Hal ini menunjukkan bahwa resiko atau kejadian AD cenderung meningkat pada usia lanjut, meskipun jumlah individu tanpa AD juga tetap tinggi.

b) *Gender* Berdasarkan Diagnosis

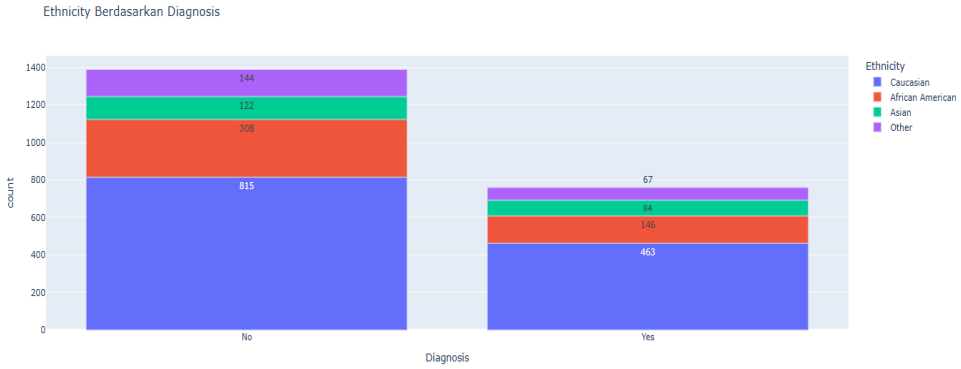


Gambar 4. 5 Visualisasi Bar Chart Gender Berdasarkan *Diagnosis*

Dari data tersebut, jumlah orang yang tidak terdiagnosis AD ("*No*") jauh lebih banyak dibandingkan dengan yang terdiagnosis AD ("*Yes*"). Pada kelompok yang tidak terdiagnosis AD, ada sekitar 675 pria dan 714 wanita. Sementara itu, pada kelompok yang terdiagnosis AD, ada sekitar

365 pria dan 374 wanita. Secara keseluruhan, data ini menunjukkan bahwa jumlah wanita dalam penelitian ini lebih tinggi, dan mayoritas tidak terdiagnosis AD. Meskipun begitu, di antara yang terdiagnosis AD, jumlah wanita juga sedikit lebih banyak. Maka dari itu pada kelompok yang tidak terdiagnosis maupun yang terdiagnosis AD, jumlah wanita sedikit lebih banyak daripada pria.

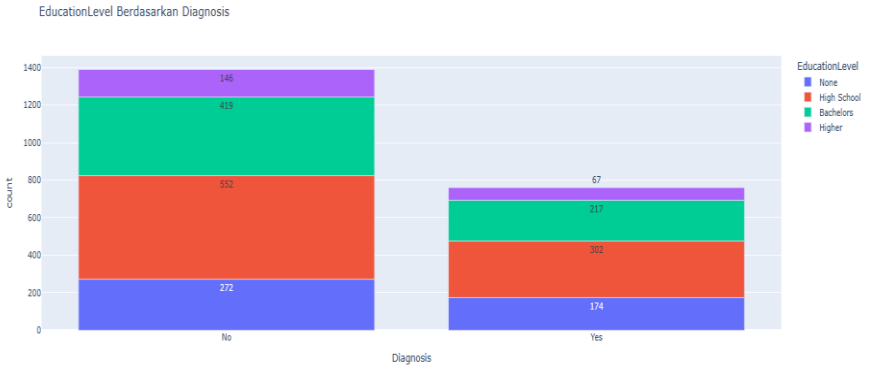
c) *Ethnicity* Berdasarkan Diagnosis



Gambar 4. 6 *Visualisasi Bar Chart Ethnicity* Berdasarkan *Diagnosis*

Dari data tersebut, terlihat bahwa jumlah orang yang tidak terdiagnosis AD ("*No*") jauh lebih banyak dibandingkan dengan yang terdiagnosis AD ("*Yes*"). Pada kelompok yang tidak terdiagnosis AD, kelompok etnis Kaukasia memiliki jumlah terbesar (sekitar 815 orang), diikuti oleh Afrika Amerika (sekitar 308 orang), Asia (sekitar 122 orang), dan kelompok etnis lainnya (sekitar 144 orang). Pola yang serupa terlihat pada kelompok yang terdiagnosis AD, di mana etnis Kaukasia juga memiliki jumlah terbesar (sekitar 463 orang), diikuti oleh Afrika Amerika (sekitar 146 orang), Asia (sekitar 94 orang), dan kelompok etnis lainnya (sekitar 67 orang). Secara keseluruhan, data ini menunjukkan bahwa kelompok etnis Kaukasia merupakan yang paling banyak, baik pada kelompok yang tidak terdiagnosis maupun yang terdiagnosis AD.

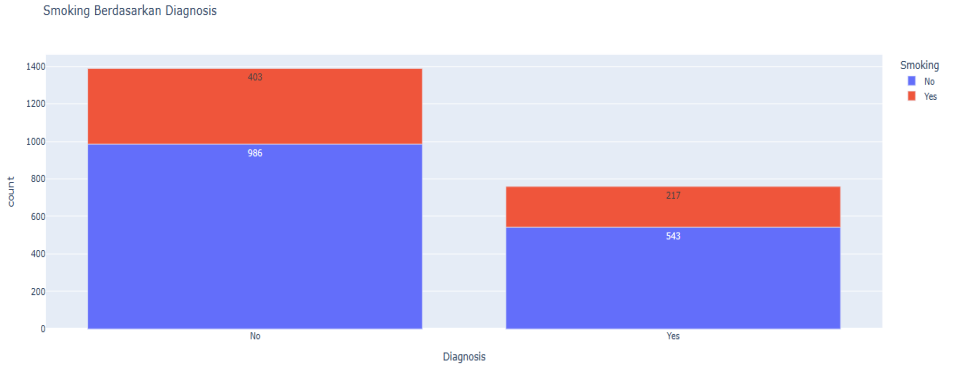
d) *Education Level* Berdasarkan Diagnosis



Gambar 4. 7 *Visualisasi Bar Chart Education Level* Berdasarkan *Diagnosis*

Dari data ini, jumlah orang yang tidak terdiagnosis AD ("*No*") jauh lebih banyak dibandingkan dengan yang terdiagnosis AD ("*Yes*"). Pada kelompok yang tidak terdiagnosis AD, jumlah orang dengan tingkat pendidikan "*None*" adalah yang paling sedikit (sekitar 272 orang), diikuti oleh "*Higher*" dengan sekitar 419 orang, "*Bachelors*" dengan sekitar 552 orang, dan "*High School*" dengan jumlah terbanyak yaitu sekitar 146 orang. Sebaliknya, pada kelompok yang terdiagnosis AD, jumlah orang dengan tingkat pendidikan "*None*" juga merupakan yang paling sedikit (sekitar 174 orang) sedangkan "*Higher*" dengan sekitar 217 orang, "*Bachelors*" dengan sekitar 302 orang, dan "*High School*" dengan jumlah terbanyak yaitu sekitar 67 orang.

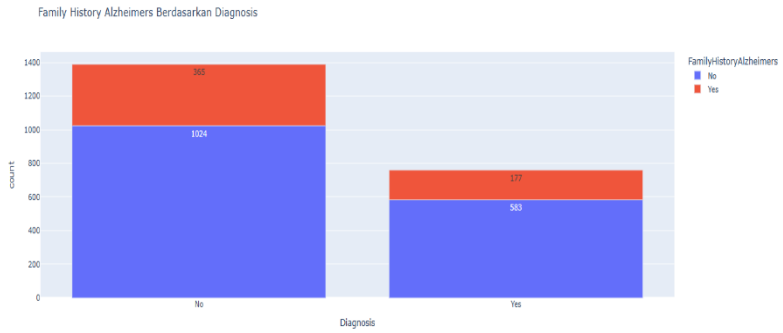
e) *Smoking* Berdasarkan Diagnosis



Gambar 4. 8 *Visualisasi Bar Chart Smoking* Berdasarkan *Diagnosis*

Dari data tersebut, Perbandingan jumlah orang yang merokok berdasarkan diagnosis AD. Pada kelompok orang yang tidak terdiagnosis AD ("*No*"), terdapat 986 orang yang tidak merokok dan 403 orang yang merokok. Sementara itu, pada kelompok orang yang terdiagnosis AD ("*Yes*"), terdapat 543 orang yang tidak merokok dan 217 orang yang merokok. Secara sederhana, terlihat bahwa jumlah orang yang tidak merokok jauh lebih banyak dibandingkan dengan yang merokok, baik pada kelompok yang tidak terdiagnosis AD maupun yang terdiagnosis AD.

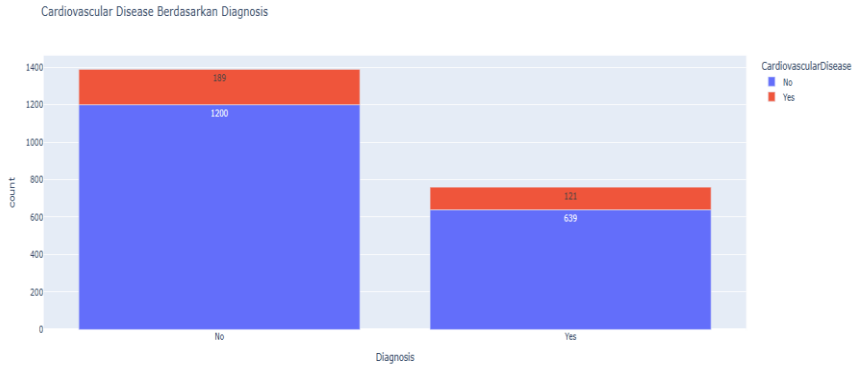
f) *Family History Alzheimer* Berdasarkan Diagnosis



Gambar 4. 9 *Visualisasi Bar Chart Family History AD* Berdasarkan *Diagnosis*

Grafik ini menunjukkan pada kelompok yang tidak terdiagnosis AD ("*No*"), terdapat 1024 orang yang tidak memiliki riwayat keluarga AD dan 365 orang yang memiliki riwayat keluarga AD. Sementara itu, pada kelompok yang terdiagnosis AD ("*Yes*"), terdapat 583 orang yang tidak memiliki riwayat keluarga AD dan 177 orang yang memiliki riwayat keluarga AD. Pada kedua kelompok diagnosis, jumlah orang yang tidak memiliki riwayat keluarga AD lebih banyak daripada yang memiliki.

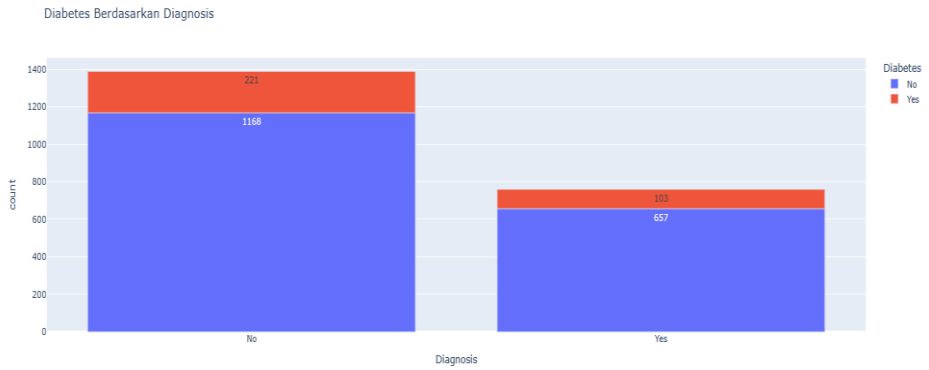
g) *Cardiovascular Disease* Berdasarkan Diagnosis



Gambar 4. 10 *Visualisasi Bar Chart Cardiovascular Disease* Berdasarkan *Diagnosis*

Pada kelompok orang yang tidak terdiagnosis AD ("*No*"), terdapat 1200 orang yang tidak memiliki penyakit kardiovaskular dan 189 orang yang memiliki penyakit kardiovaskular. Sementara itu, pada kelompok orang yang terdiagnosis AD ("*Yes*"), terdapat 639 orang yang tidak memiliki penyakit kardiovaskular dan 121 orang yang memiliki penyakit kardiovaskular. Secara garis besar, terlihat bahwa jumlah orang yang tidak memiliki penyakit kardiovaskular jauh lebih banyak dibandingkan dengan yang memiliki penyakit kardiovaskular, baik pada kelompok yang tidak terdiagnosis AD maupun yang terdiagnosis AD.

h) *Diabetes* Berdasarkan Diagnosis

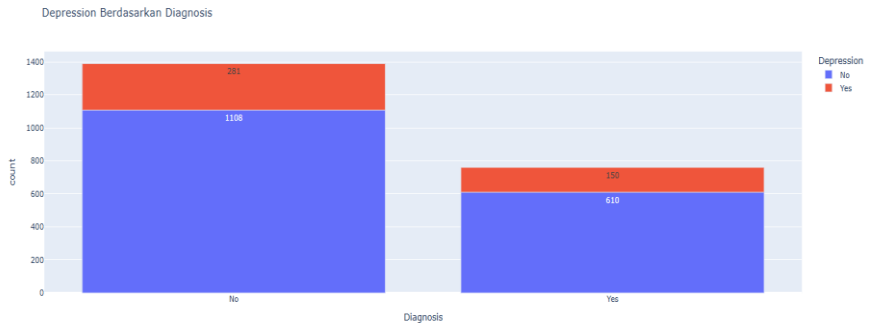


Gambar 4. 11 *Visualisasi Bar Chart Diabetes* Berdasarkan *Diagnosis*

Berdasarkam data tersebut, Kelompok orang yang yang tidak terdiagnosis AD (“*No*”) yang tidak diabetes terdapat 1168 orang dan 221 orang yang diabetes. Sementara itu, pada kelompok terdiagnosis AD (“*Yes*”) terdapat 543 orang yang tidak diabetes dan 217 orang yang

diabetes. Terlihat bahwa jumlah orang yang diabetes jauh lebih sedikit dibandingkan dengan yang tidak diabetes, baik pada kelompok yang tidak terdiagnosis maupun terdiagnosis AD. Perbandingan orang dengan diabetes terlihat sedikit lebih tinggi pada kelompok yang tidak terdiagnosis AD dibandingkan dengan kelompok yang terdiagnosis AD.

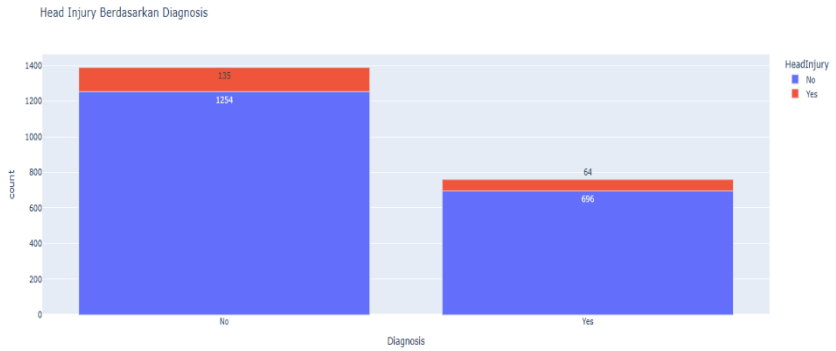
i) *Depression* Berdasarkan Diagnosis



Gambar 4. 12 *Visualisasi Bar Chart Depression* Berdasarkan *Diagnosis*

Dari data tersebut, terdapat 1108 orang yang tidak memiliki depresi yang tidak terdiagnosis AD ("No") dan 281 orang memiliki depresi. Sementara itu pada kelompok terdiagnosis AD ("Yes") tidak memiliki depresi terdapat 610 orang dan 150 orang yang memiliki depresi. Perbandingan orang yang mengalami depresi tampak sedikit lebih tinggi pada kelompok yang terdiagnosis AD dibandingkan dengan kelompok yang tidak terdiagnosis AD.

j) *Head Injury* Berdasarkan Diagnosis

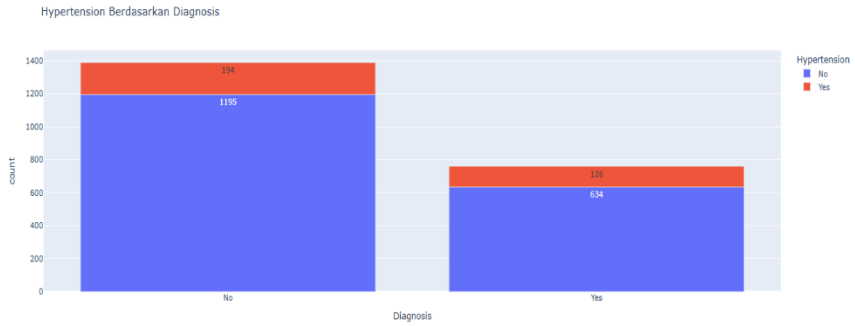


Gambar 4. 13 *Visualisasi Bar Chart Head Injury* Berdasarkan *Diagnosis*

Berdasarkan data tersebut, pada kelompok orang yang tidak terdiagnosis AD ("*No*") yang tidak memiliki riwayat cedera kepala terdapat 1254 orang dan 135 orang yang memiliki riwayat cedera kepala.

Sementara itu, kelompok yang terdiagnosis AD ("*Yes*") terdapat 696 orang tidak memiliki riwayat cidera kepala dan 64 orang memiliki riwayat cedera kepala. Perbandingan orang yang memiliki riwayat cedera kepala tampak sedikit lebih tinggi pada kelompok yang tidak terdiagnosis AD dibandingkan dengan kelompok yang terdiagnosis AD.

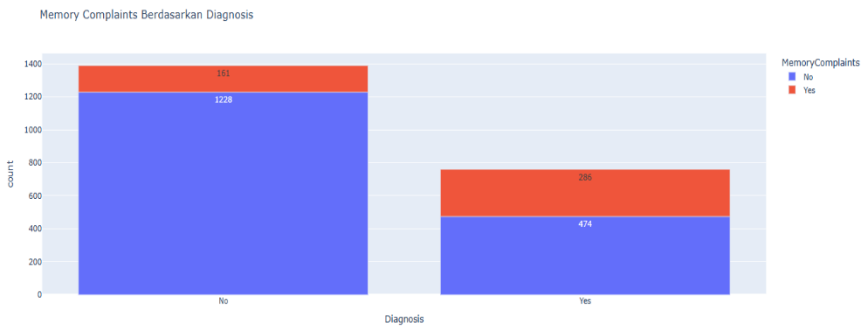
k) *Hypertension* Berdasarkan Diagnosis



Gambar 4. 14 *Visualisasi Bar Chart Hypertension* Berdasarkan *Diagnosis*

Berdasarkan data tersebut, 1195 orang tidak memiliki hipertensi dan 194 orang memilikinya di kelompok yang tidak terdiagnosis AD ("*No*"). Di sisi lain, 634 orang di kelompok yang terdiagnosis AD ("*Yes*") tidak memiliki hipertensi dan 126 orang memilikinya. Data menunjukkan bahwa pada kedua kelompok diagnosis, ada lebih banyak orang yang tidak memiliki hipertensi daripada yang memilikinya. Dalam perbandingan kedua kelompok, terlihat bahwa jumlah orang yang memiliki hipertensi sedikit lebih tinggi pada kelompok yang tidak terdiagnosis AD dibandingkan dengan kelompok yang terdiagnosis AD.

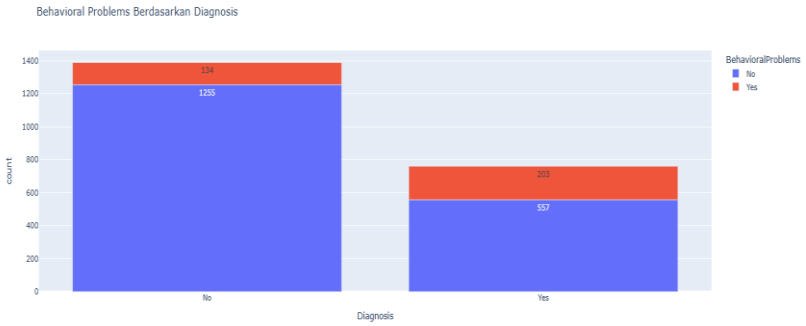
l) *Memory Complaints* Berdasarkan Diagnosis



Gambar 4. 15 Visualisa*si Bar Chart Memory Complaints* Berdasarkan *Diagnosis*

Dari data tersebut, 1228 orang dalam kelompok yang tidak terdiagnosis AD ("*No*") tidak memiliki keluhan memori dan 161 orang memilikinya. Sementara itu, 474 orang dalam kelompok yang terdiagnosis AD ("*Yes*") tidak memiliki keluhan memori dan 286 orang memilikinya. Sebagian besar orang dalam kelompok yang tidak terdiagnosis AD tidak mengalami masalah memori. Sebaliknya, dalam kelompok yang terdiagnosis AD, ada lebih banyak orang yang mengalami masalah memori dibandingkan dengan mereka yang tidak memiliki masalah memori.

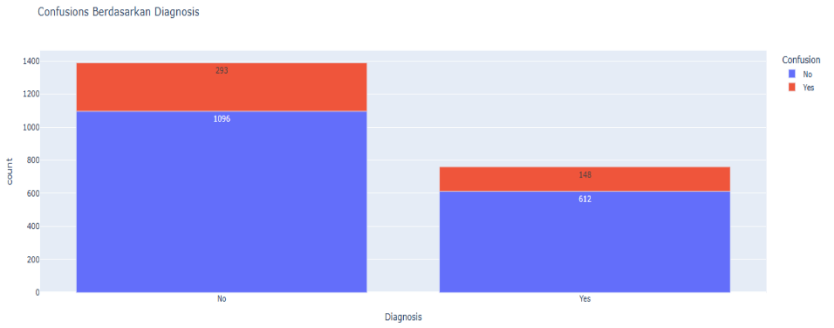
m) *Behavioral Problems* berdasarkan Diagnosis



Gambar 4. 16 Visualisasi *Bar Chart Behavioral Problems* Berdasarkan *Diagnosis*

Dari data yang ada, terdapat 1255 orang dalam kelompok yang tidak terdiagnosis AD ("*No*") tidak memiliki masalah perilaku dan 134 orang memilikinya. Di sisi lain, 557 orang dalam kelompok yang terdiagnosis AD ("*Yes*") tidak memiliki masalah perilaku dan 203 orang memilikinya. Data yang ada menunjukkan bahwa ada lebih banyak individu pada kedua kelompok diagnosis yang tidak memiliki masalah perilaku daripada individu yang memiliki masalah perilaku. Tapi dalam perbandingan kedua kelompok, kelompok yang terdiagnosis AD tampaknya memiliki lebih banyak masalah perilaku daripada kelompok yang tidak terdiagnosis AD.

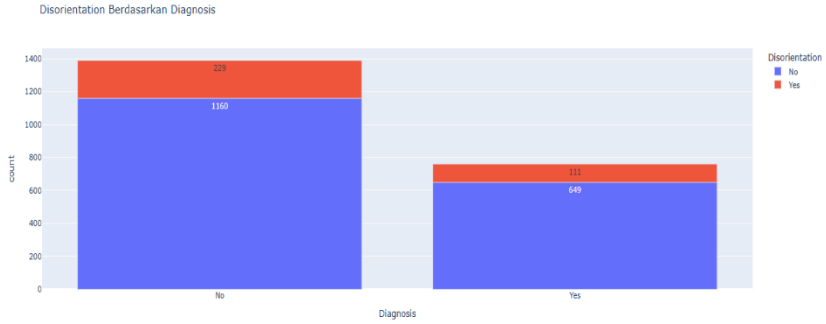
n) *Confusions* Berdasarkan Diagnosis



Gambar 4. 17 Visualisasi *Bar Chart Confusions* Berdasarkan *Diagnosis*

Berdasarkan data yang ada, terdapat 1096 orang dalam kelompok yang tidak terdiagnosis AD ("*No*") tidak mengalami kebingungan, dan 293 orang mengalaminya. Sedangkan, 612 orang dalam kelompok yang terdiagnosis AD ("*Yes*") tidak mengalami kebingungan, dan 148 orang mengalaminya. Data ini menunjukkan bahwa pada kedua kelompok diagnosis, ada lebih banyak orang yang tidak mengalami kebingungan daripada yang mengalaminya. Perbandingan kedua kelompok, orang yang mengalami kebingungan jauh lebih tinggi pada kelompok yang terdiagnosis AD dibandingkan dengan kelompok yang tidak terdiagnosis AD.

o) *Disorientation* Berdasarkan Diagnosis

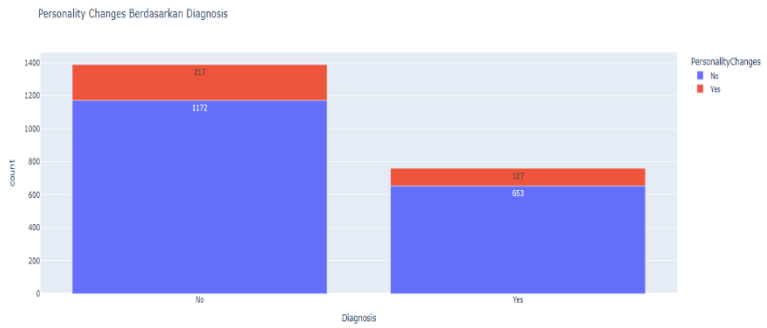


Gambar 4. 18 Visualisasi *Bar Chart Disorientation* Berdasarkan *Diagnosis*

Dari data tersebut, terdapat 1160 orang dalam kelompok yang tidak terdiagnosis AD ("*No*") tidak mengalami disorientasi, dan 229 orang mengalaminya, Sementara itu, terdapat 649 orang dalam kelompok yang terdiagnosis AD ("*Yes*") tidak mengalami disorientasi, dan 111 orang mengalaminya. Data tersebut menunjukkan bahwa pada kedua kelompok diagnosis, ada lebih banyak orang yang tidak mengalami disorientasi daripada yang mengalaminya. Namun, dalam perbandingan kedua

kelompok, jumlah orang yang mengalami disorientasi lebih tinggi pada kelompok yang tidak terdiagnosis AD.

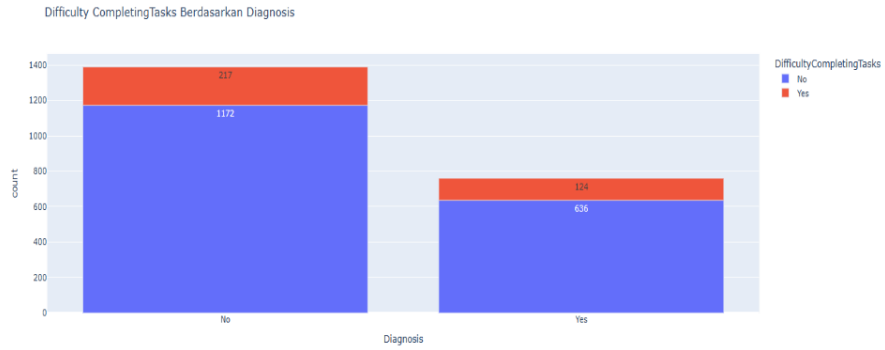
p) *Personality Changes* Berdasarkan Diagnosis



Gambar 4. 19 Visualisasi *Bar Chart Personality Changes* Berdasarkan *Diagnosis*

Berdasarkan data yang ada, 1172 orang dalam kelompok tidak terdiagnosis AD ("*No*") tidak mengalami perubahan kepribadian dan 217 mengalaminya. Sementara itu, dalam kelompok yang terdiagnosis AD ("*Yes*") terdapat 653 orang tidak mengalami perubahan kepribadian dan 107 mengalaminya. Data menunjukkan bahwa ada lebih banyak individu pada kedua kelompok diagnosis yang tidak mengalami perubahan kepribadian daripada individu yang mengalaminya. Dalam perbandingan, individu dalam kelompok yang tidak terdiagnosis AD sedikit lebih tinggi daripada dalam kelompok yang terdiagnosis AD.

q) *Difficulty Completing Tasks* Berdasarkan Diagnosis

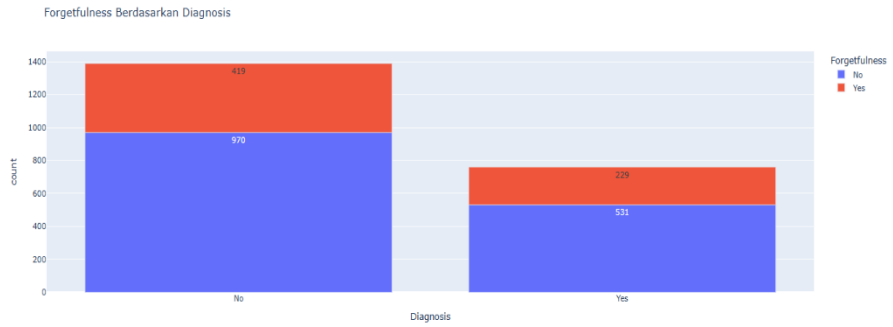


Gambar 4. 20 Visualisasi *Bar Chart Difficulty Completing* Berdasarkan *Diagnosis*

Terdapat 1.172 orang dalam kelompok yang tidak terdiagnosis AD ("*No*") yang tidak mengalami kesulitan menyelesaikan tugas dan 217 orang dalam kelompok yang terdiagnosis AD ("*Yes*"). Dari data ini, terlihat bahwa jumlah orang yang tidak mengalami kesulitan menyelesaikan tugas

lebih banyak daripada orang yang mengalami kesulitan. Namun, jika dibandingkan dengan kelompok yang tidak terdiagnosis AD, proporsi orang yang mengalami kesulitan menyelesaikan tugas tampaknya sedikit lebih tinggi di kelompok yang terdiagnosis AD.

r) *Forgetfulness* Berdasarkan Diagnosis



Gambar 4. 21 Visualisasi *Bar Chart Forgetfulness* Berdasarkan *Diagnosis*

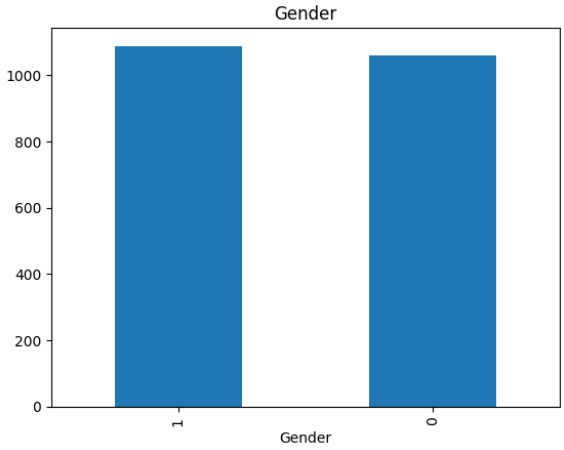
Terdapat 970 orang yang tidak mengalami kelupaan dan 419 orang yang mengalami kelupaan di kelompok yang tidak terdiagnosis AD ("*No*"). Sementara itu, di kelompok yang terdiagnosis AD ("*Yes*") terdapat 531 orang tidak mengalami kelupaan dan 229 orang mengalaminya. Data menunjukkan bahwa jumlah orang yang tidak mengalami kelupaan lebih banyak daripada yang mengalami kelupaan pada kedua kelompok diagnosis. Namun, jika dibandingkan dengan kelompok yang tidak terdiagnosis AD, proporsi orang yang mengalami kelupaan tampaknya lebih tinggi pada kelompok yang terdiagnosis AD.

5) *Univariate Analysis*

Dilakukan untuk melihat hubungan antara masing-masing variabel untuk menggambarkan karakteristik satu variabel tunggal dalam dataset.

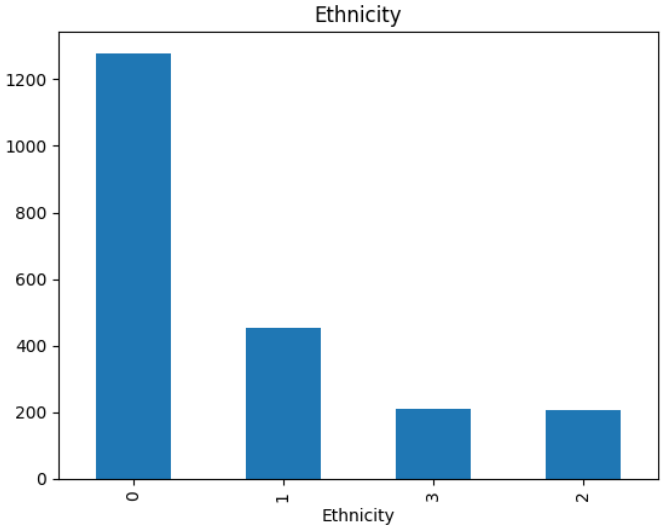
a) Categorical Features

Memahami distribusi kategori dalam dataset dengan menghitung frekuensi dari setiap kategori.



Gambar 4. 22 *Bar Chart Univariate Categorical-Gender*

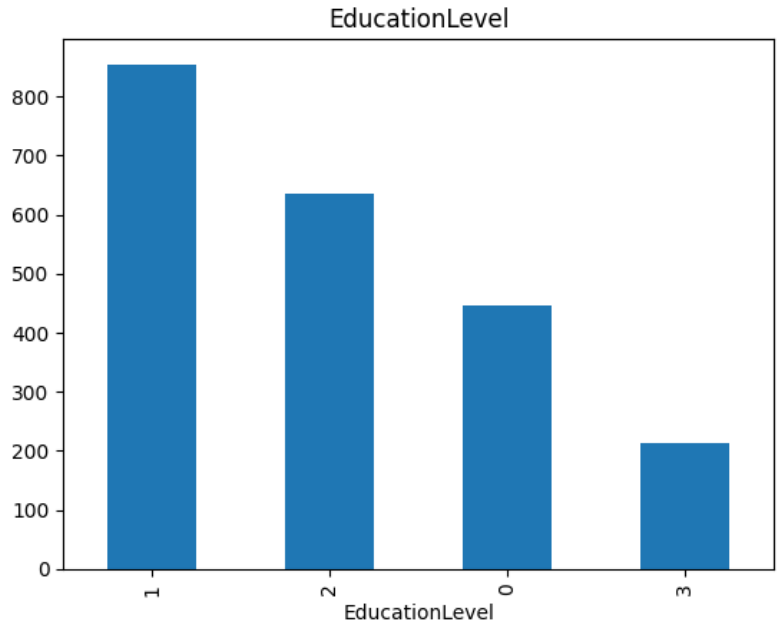
Dalam dataset, ada dua kategori gender dengan label 0 untuk *Male* dan 1 untuk *Female*. Jumlah sampel untuk *Female* adalah 1.088, atau sekitar 50,6% dari total data, sedangkan sampel untuk *Male* adalah 1.061, atau sekitar 49,4%. Ada perbedaan kecil dalam jumlah sampel antara kedua kategori ini, menunjukkan bahwa distribusi *gender* dalam dataset cukup seimbang.



Gambar 4. 23 *Bar Chart Univariate Categorical-Ethnicity*

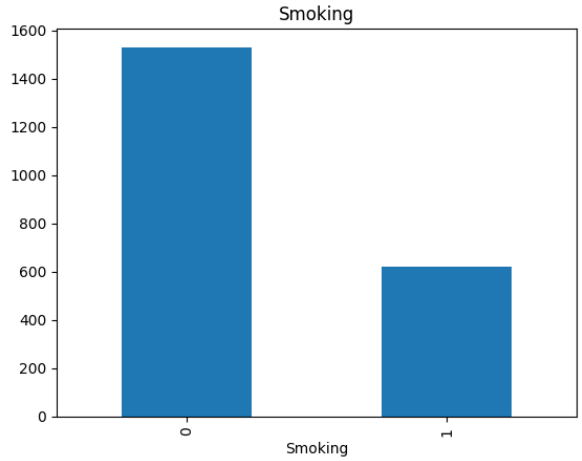
Menurut visualisasi distribusi etnis pada dataset, mayoritas sampel berasal dari kelompok Caucasian (0), yang menghasilkan 1.278 sampel, atau 59,5% dari total sampel; diikuti oleh kelompok Afrika Amerika (1), yang menghasilkan 454 sampel, atau 21,1%; kelompok lain (3), yang menghasilkan 211 sampel, atau 9,8%; dan kelompok Asia (2), yang

menghasilkan 206 sampel, atau 9,6% dari total sampel. Semua ini menunjukkan dominasi etnis Caucasian dalam data, sementara tiga kategori lainnya memiliki proporsi yang lebih kecil.



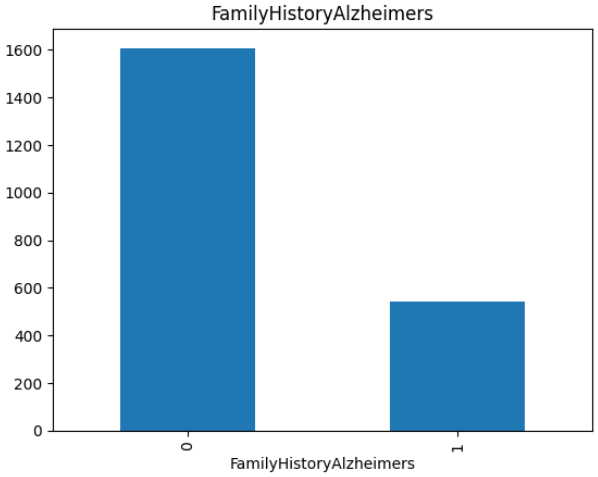
Gambar 4. 24 *Bar Chart Univariate Categorical-Education Level*

Berdasarkan visualisasi distribusi tingkat pendidikan dalam dataset, *High School* (1) adalah kategori yang paling dominan, terdapat 854 sampel, atau 39,7% dari total data, *Bachelors* (2) terdapat 640 sampel, atau 29,7%; dan *None* (0) terdapat 450 sampel, atau 20,9 persen. Kategori yang paling sedikit, *Higher* (3), terdapat 213 sampel, atau 9,9 persen. Distribusi ini menunjukkan bahwa sebagian besar individu dalam dataset memiliki tingkat pendidikan menengah, sementara proporsi individu dengan pendidikan lebih tinggi relatif rendah.



Gambar 4. 25 *Bar Chart Univariate Categorical-Smoking*

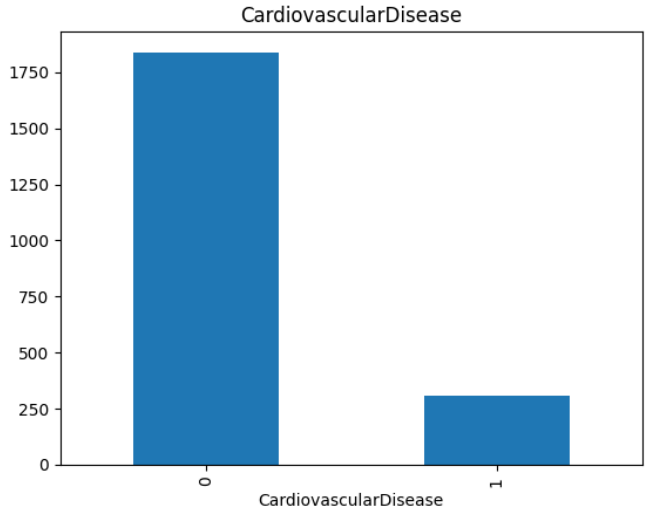
Menurut data kebiasaan merokok, mayoritas berada dalam dataset yang tidak merokok (kategori 0), dengan 1.529 sampel, yang merupakan 71,1% dari total sampel. Sebaliknya, 620 sampel, atau 28,9% dari responden, berada dalam kategori merokok (kategori 1). Distribusi ini menunjukkan bahwa sebagian besar individu dalam dataset memiliki gaya hidup tanpa merokok.



Gambar 4. 26 *Bar Chart Univariate Categorical- Family History ADs*

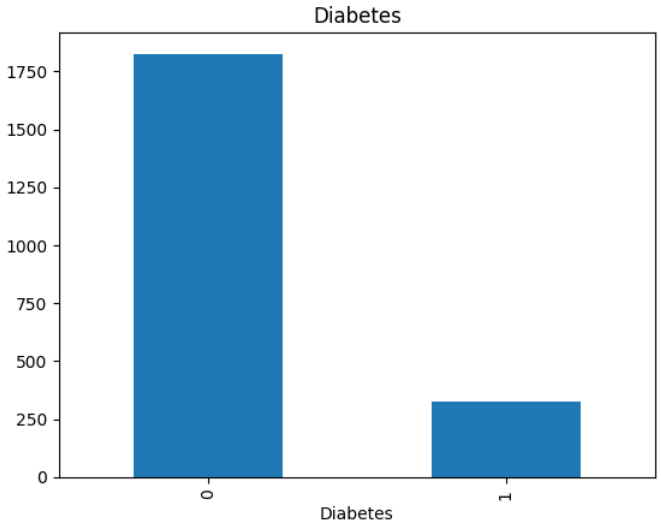
Menurut data riwayat keluarga dengan AD (*Family History Alzheimer*), sebanyak 1.607 sampel, atau 74,8% dari total data, menunjukkan bahwa sebagian besar responden tidak memiliki riwayat keluarga dengan AD (kategori 0). Sementara itu, 542 sampel, atau 25,2% dari responden, menunjukkan bahwa mereka memiliki riwayat keluarga dengan AD (kategori 1). Data menunjukkan bahwa sebagian

besar responden tidak memiliki faktor risiko genetik yang terkait dengan AD.



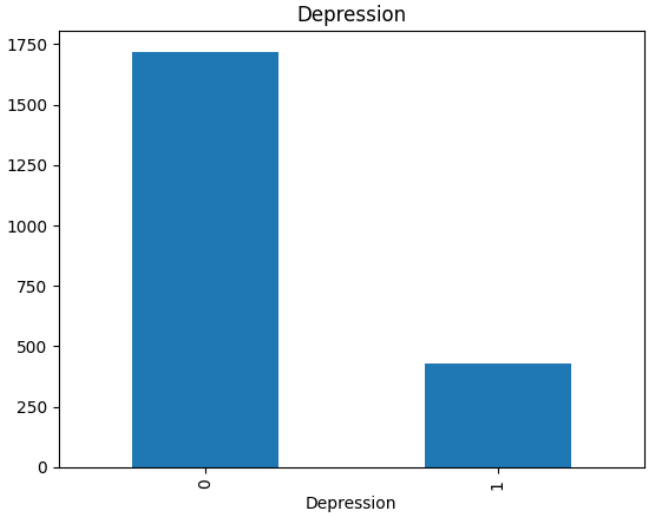
Gambar 4. 27 *Bar Chart Univariate Categorical-Cardivascular Disease*

Terdapat 1.839 sampel dalam kategori 0 (tidak ada penyakit kardiovaskular) dan 310 sampel dalam kategori 1 (ada penyakit kardiovaskular), yang mencakup sekitar 85,6% dari total data. Visualisasi ini juga menunjukkan distribusi data berdasarkan keberadaan penyakit kardiovaskular. Distribusi ini menunjukkan ketidakseimbangan antara kedua kelompok, dengan lebih banyak orang yang tidak memiliki penyakit jantung dibandingkan dengan mereka yang memilikinya.



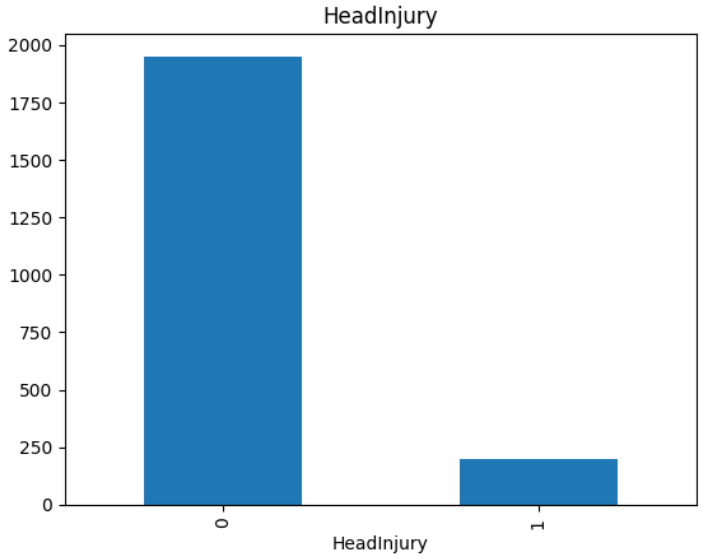
Gambar 4. 28 *Bar chart univariate categorical-diabetes*

Berdasarkan data yang tertera, dapat diamati bahwa terdapat 1825 sampel yang termasuk dalam kategori tidak diabetes (0), yang mencakup 84.9% dari total keseluruhan sampel. Sementara itu, terdapat 324 sampel yang termasuk dalam kategori diabetes (1), yang mencakup 15.1% dari total keseluruhan sampel. Secara keseluruhan, visualisasi ini menunjukkan adanya ketidakseimbangan distribusi data, di mana jumlah individu yang tidak memiliki diabetes jauh lebih besar dibandingkan dengan jumlah individu yang memiliki diabetes dalam kumpulan data ini.



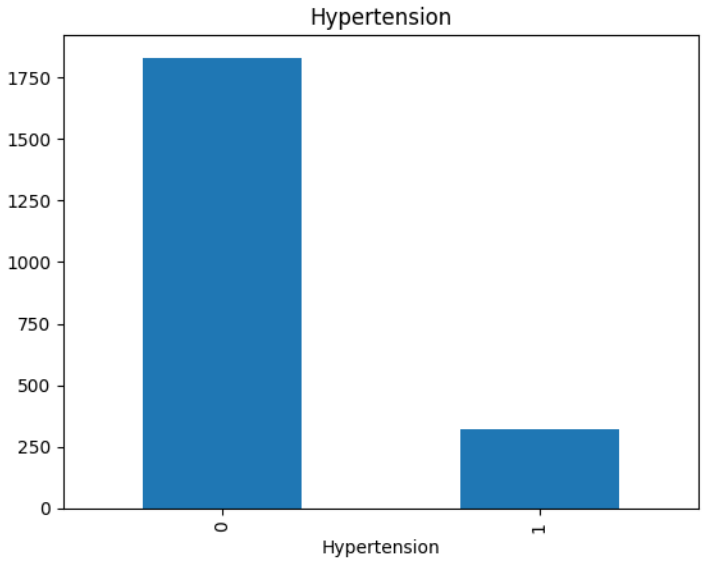
Gambar 4. 29 *Bar chart univariate categorical-depression*

Terlihat bahwa terdapat 1718 sampel yang termasuk dalam kategori tidak depresi (0), yang mewakili 79.9% dari total keseluruhan sampel. Sementara itu, terdapat 431 sampel yang termasuk dalam kategori depresi (1), yang mewakili 20.1% dari total keseluruhan sampel. Secara keseluruhan, visualisasi ini memperlihatkan bahwa mayoritas sampel dalam kumpulan data ini tidak mengalami depresi, dengan proporsi yang jauh lebih besar dibandingkan dengan sampel yang mengalami depresi.



Gambar 4. 30 *Bar Chart Univariate Categorical-Head Injury*

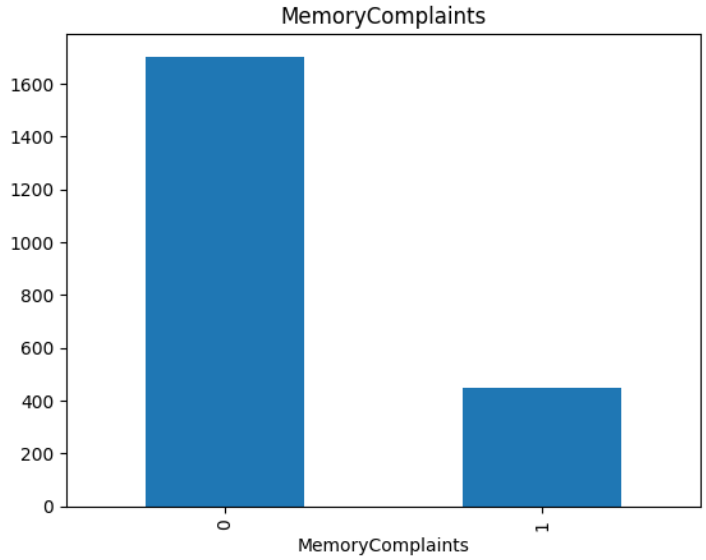
Mayoritas sampel dalam dataset ini tidak memiliki riwayat cedera kepala 199 sampel memiliki riwayat cedera kepala (1), yang merupakan 9.3% dari total sampel, dan 1950 sampel tidak memiliki riwayat cedera kepala (0), yang merupakan 90.7% dari total sampel. Diagram batang dan data numerik yang tertera menunjukkan fakta bahwa proporsi sampel yang tidak memiliki riwayat cedera kepala jauh lebih tinggi daripada proporsi sampel yang memiliki riwayat cedera kepala.



Gambar 4. 31 *Bar Chart Univariate Categorical-Hypertension*

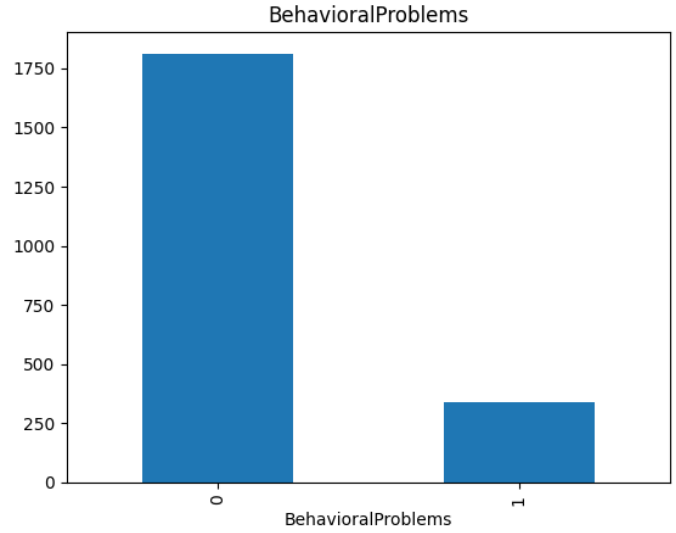
Berdasarkan data yang ada, terdapat 1829 sampel dalam kumpulan data ini yang termasuk dalam kategori tidak hipertensi (0), yang

merupakan 85.1% dari total sampel, dan 320 sampel termasuk dalam kategori hipertensi (1), yang merupakan 14.9% dari total sampel. Secara keseluruhan, visualisasi ini memperlihatkan bahwa mayoritas sampel dalam kumpulan data ini tidak memiliki hipertensi, dengan proporsi yang jauh lebih besar dibandingkan dengan sampel yang memiliki hipertensi.



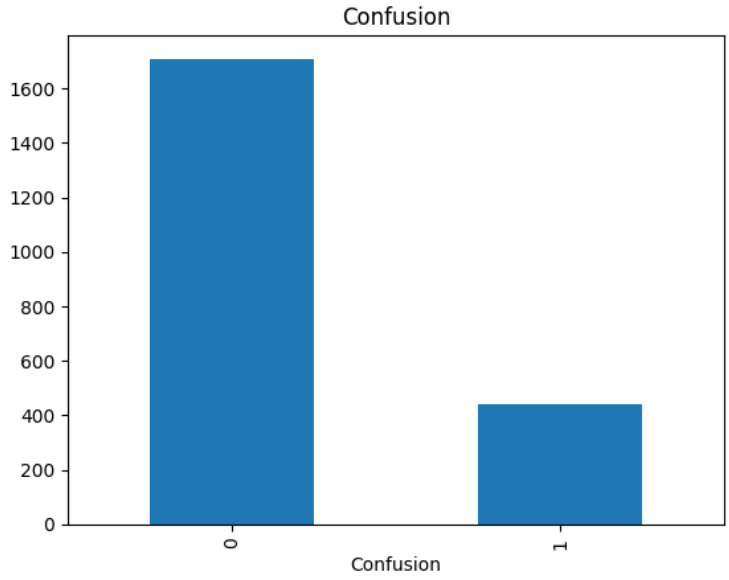
Gambar 4. 32 *Bar Chart Univariate Categorical-Memory Complaints*

Mayoritas sampel dalam dataset ini tidak memiliki keluhan memori, dengan 447 sampel termasuk dalam kategori ada keluhan memori (1), yang mencakup 20,8% dari total sampel, dan 1.702 sampel termasuk dalam kategori tidak ada keluhan memori (0), yang mencakup 79.2% dari total sampel.. Berdasarkan data menunjukkan bahwa mayoritas sampel dalam dataset ini tidak memiliki keluhan memori.



Gambar 4. 33 *Bar Chart Univariate Categorical-Behavioral Problems*

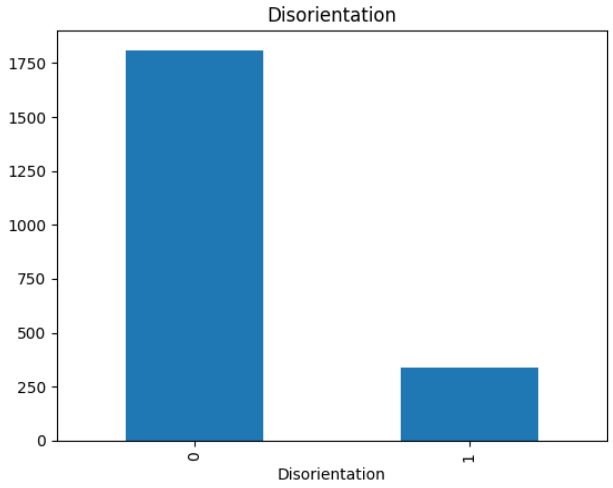
Mayoritas sampel dalam kumpulan data ini tidak menunjukkan masalah perilaku. Ini ditunjukkan oleh diagram batang dan data numerik yang disertakan terdapat 1812 sampel termasuk dalam kategori tidak ada masalah perilaku (0), yang merupakan 84,3% dari total sampel, dan 337 sampel termasuk dalam kategori ada masalah perilaku (1), yang merupakan 15,7% dari total sampel.



Gambar 4. 34 *Bar Chart Univariate Categorical-Confusion*

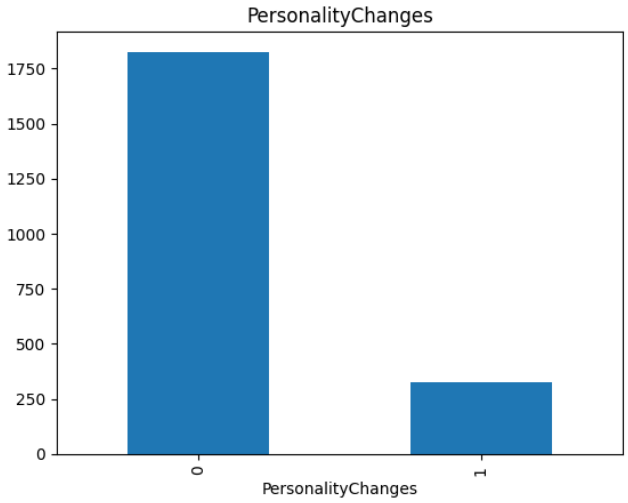
Dari data yang ada, terdapat 1708 sampel termasuk dalam kategori tidak bingung (0), yang merupakan 79.5% dari sampel keseluruhan, dan 441 sampel termasuk dalam kategori bingung (1), yang merupakan

20.5% dari sampel keseluruhan, seperti yang ditunjukkan oleh diagram batang dan data statistik. Secara keseluruhan, visualisasi ini menunjukkan bahwa sebagian besar sampel dalam kumpulan data ini tidak mengalami kebingungan dalam proporsi yang jauh lebih besar daripada sampel yang mengalami kebingungan. Sehingga ada ketidakseimbangan distribusi pada variabel kebingungan



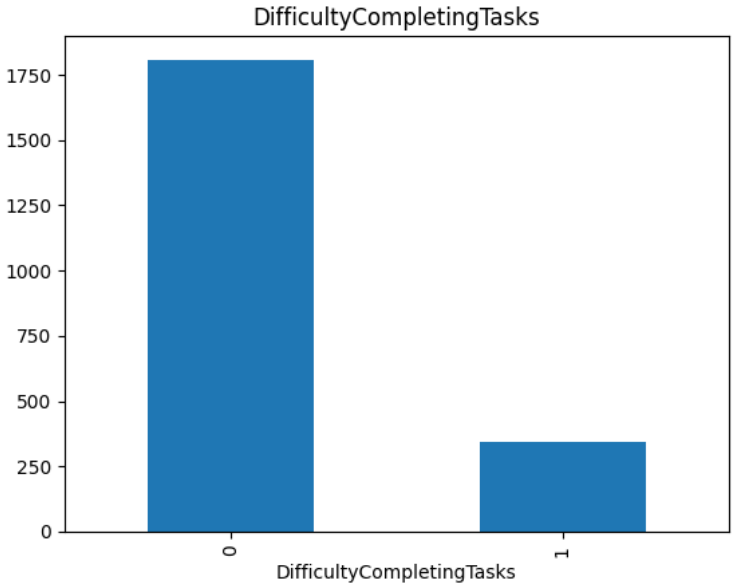
Gambar 4. 35 *Bar Chart Univariate Categorical-Disorientation*

Berdasarkan data tersebut, terdapat 1809 sampel termasuk dalam kategori tidak disorientasi (0), yang merupakan 84,2% dari total sampel, dan 340 sampel termasuk dalam kategori disorientasi (1), yang merupakan 15,8% dari total sampel. Perbedaan ini ditunjukkan oleh diagram batang dan data numerik yang disertakan. Sebagian besar sampel tidak mengalami disorientas dalam proporsi yang jauh lebih besar dibandingkan dengan sampel yang mengalaminya. Ini menunjukkan bahwa ada ketidakseimbangan distribusi.



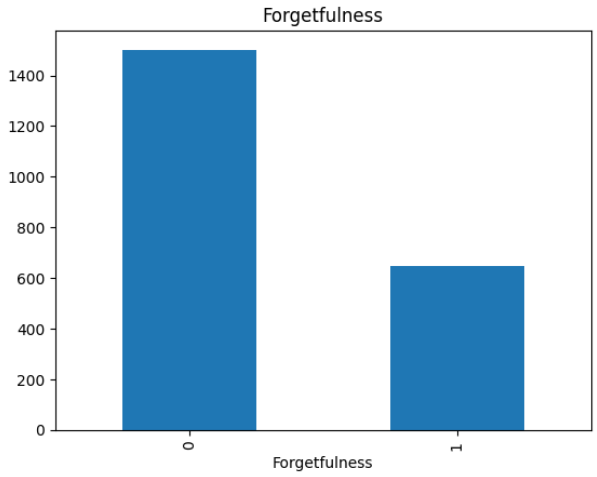
Gambar 4. 36 *Bar Chart Univariate Categorical-Personality Changes*

Terdapat 1825 sampel termasuk dalam kategori tidak ada perubahan kepribadian (0), yang merupakan 84.9% dari total sampel, dan 324 sampel termasuk dalam kategori ada perubahan kepribadian (1), yang merupakan 15.1% dari total sampel. Perbedaan ini ditunjukkan oleh diagram batang dan data statistik yang disertakan. Pada variabel perubahan kepribadian ini, sebagian besar sampel tidak menunjukkan perubahan kepribadian, dengan proporsi yang jauh lebih besar dibandingkan dengan sampel yang menunjukkan perubahan. Ini menunjukkan bahwa ada ketidakseimbangan distribusi.



Gambar 4. 37 *Bar Chart Univariate Categorical-Difficulty Completing Task*

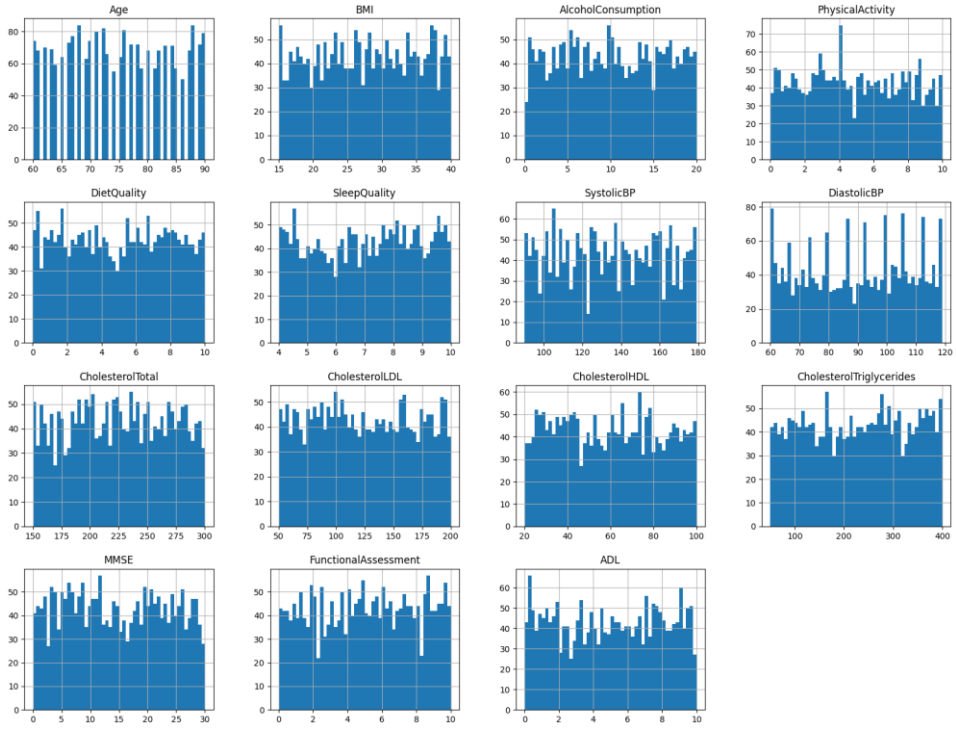
Dari data yang ada, terdapat 1808 sampel termasuk dalam kategori tidak ada kesulitan menyelesaikan tugas (0), yang merupakan 84,1% dari total sampel, dan 341 sampel termasuk dalam kategori ada kesulitan menyelesaikan tugas (1), yang merupakan 15,9% dari total sampel. Sebagian besar sampel tidak mengalami kesulitan menyelesaikan tugas, dengan proporsi yang jauh lebih besar dibandingkan dengan sampel yang mengalami kesulitan menunjukkan bahwa ada ketidakseimbangan distribusi.



Gambar 4. 38 *Bar Chart Univariate Categorical-Forgetfulness*

Berdasarkan data, terdapat 1501 sampel termasuk dalam kategori tidak mengalami kelupaan (0), yang merupakan 69.8% dari total sampel, dan 648 sampel termasuk dalam kategori mengalami kelupaan (1), yang merupakan 30.2% dari total sampel. Hal ini ditunjukkan oleh diagram batang dan data numerik yang disertakan. Meskipun proporsi sampel yang mengalami kelupaan cukup signifikan dibandingkan dengan variabel sebelumnya, Pada variabel kelupaan ini menunjukkan bahwa sebagian besar sampel tidak mengalami kelupaan. Distribusi data untuk variabel kelupaan ini tidak seseimbang, tetapi tidak seekstrem variabel sebelumnya.

b) Numerical Feature

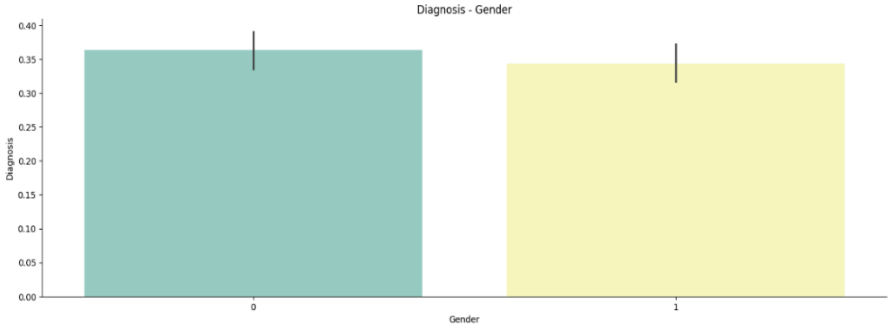


Gambar 4. 39 *Histogram Univariate Numerical Feature*

Berdasarkan visualisasi distribusi fitur numerik, terlihat bahwa sebagian besar fitur seperti BMI, konsumsi alkohol, aktivitas fisik, kualitas diet dan tidur, tekanan darah sistolik dan diastolik, serta berbagai jenis kolesterol cenderung terdistribusi merata atau mendekati distribusi normal. Hal ini mengindikasikan variasi yang cukup besar dalam populasi untuk fitur-fitur tersebut. Sementara itu, fitur-fitur yang berkaitan dengan fungsi kognitif dan aktivitas sehari-hari seperti MMSE, FunctionalAssessment, dan ADL menunjukkan kecenderungan nilai yang lebih tinggi, yang mungkin mengindikasikan sebagian besar partisipan dalam dataset memiliki fungsi kognitif dan kemampuan fungsional yang relatif baik

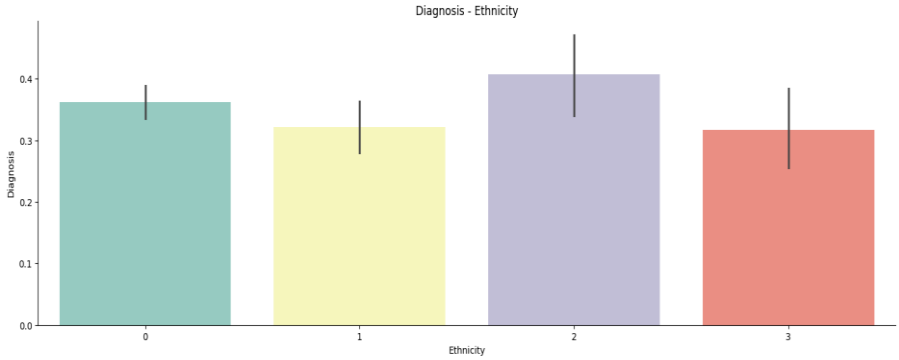
6) *Multivariate Analysis*

a) Categorical features



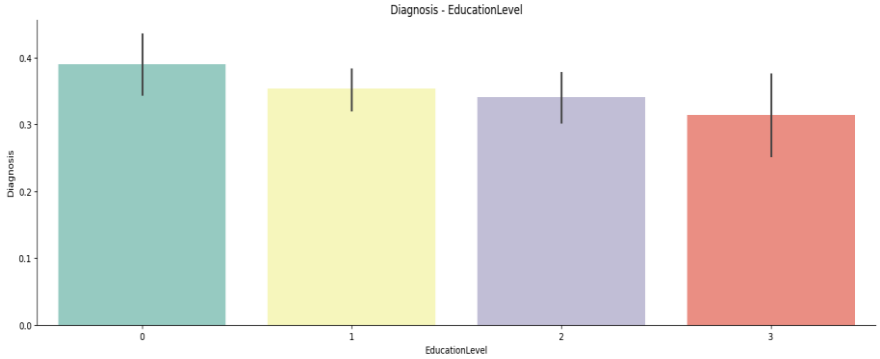
Gambar 4. 40 *Diagram Catplot Diagnosis-Gender*

Berdasarkan grafik *gender* dan diagnosis, bahwa tidak ada perbedaan yang signifikan dalam proporsi diagnosis AD antara pria (0) dan wanita (1). Pria memiliki rata-rata diagnosis yang hampir sama, meskipun wanita memiliki sedikit lebih banyak diagnosis daripada pria. Namun, perbedaan ini masih berada dalam rentang batas kesalahan (*error bar*) yang saling tumpang tindih.



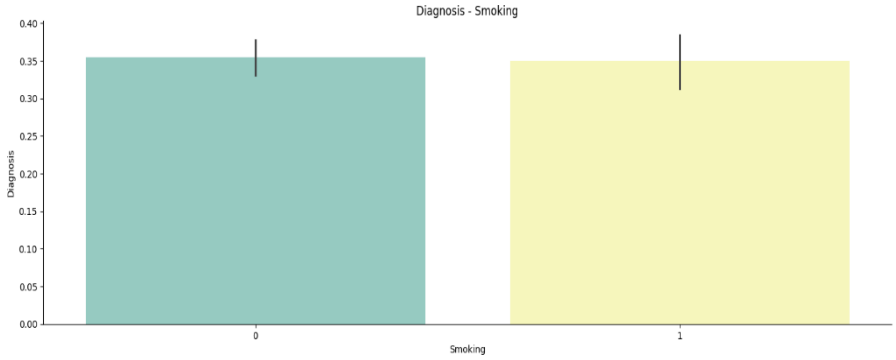
Gambar 4. 41 *Diagram Catplot Diagnosis-Ethnicity*

Berdasarkan data antara etnisitas (*ethnicity*) dan diagnosis, terlihat bahwa etnis Asia (2) memiliki proporsi diagnosis tertinggi, diikuti oleh etnis Kaukasia (0). Sementara itu, proporsi diagnosis AD dari etnis Afrika Amerika (1) dan kategori Lain (3) lebih rendah. Error bar yang saling tumpang tindih menunjukkan bahwa variasi ini belum tentu signifikan secara statistik, meskipun ada perbedaan antar kelompok etnis.



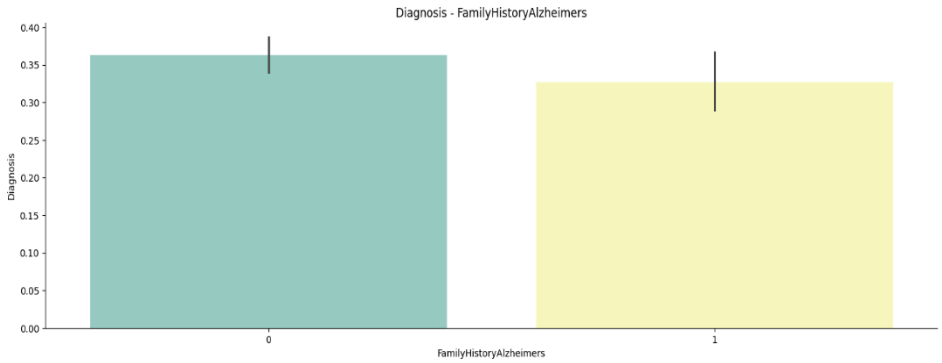
Gambar 4. 42 *Diagram Catplot Diagnosis-Education Level*

Analisis menunjukkan bahwa ada korelasi positif antara tingkat pendidikan dan prevalensi diagnosis. Kelompok dengan tingkat pendidikan "*High School*" (1), "*Bachelor*" (2), dan "*Higher*" (3) menunjukkan proporsi diagnosis AD yang sedikit lebih rendah daripada kelompok dengan tingkat pendidikan "*None*" (0). Namun, garis *error bar* menunjukkan bahwa perbedaan di antara kelompok pendidikan ini tidak terlalu besar.



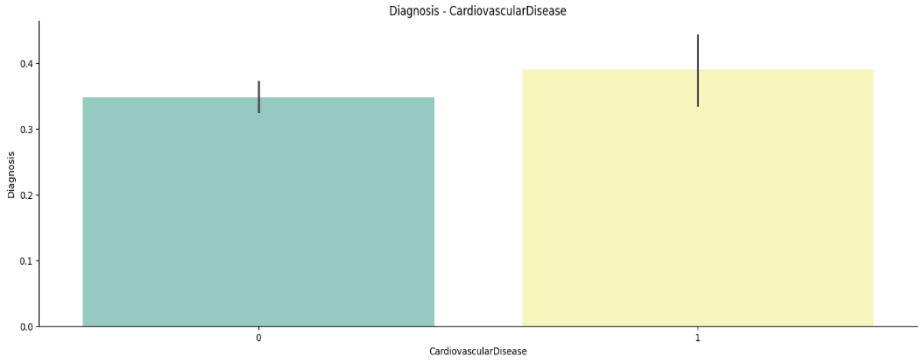
Gambar 4. 43 *Diagram Catplot Diagnosis-Smoking*

Analisis menunjukkan bahwa kelompok yang tidak merokok (0) dan kelompok yang merokok (1) tidak memiliki perbedaan yang signifikan dalam proporsi diagnosis AD. Garis *error bar* menunjukkan bahwa proporsi diagnosis AD di kedua kelompok hampir sama, dengan perbedaan kecil di masing-masing kelompok. Oleh karena itu, informasi ini menunjukkan bahwa kebiasaan merokok secara langsung terkait dengan peningkatan atau penurunan risiko terkena diagnosis AD.



Gambar 4. 44 *Diagram Catplot Diagnosis-Family History AD*

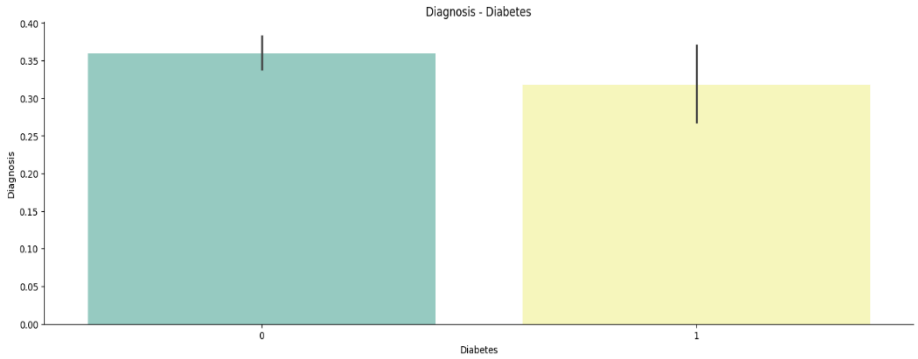
Menunjukkan bahwa ada perbedaan dalam proporsi diagnosis AD antara kelompok yang tidak memiliki riwayat AD (0) dan kelompok yang memiliki riwayat AD (1). Kelompok yang memiliki riwayat AD cenderung memiliki proporsi diagnosis yang sedikit lebih rendah dibandingkan dengan kelompok yang tidak memiliki riwayat AD. Berdasarkan data ini, riwayat keluarga AD tampaknya tidak secara langsung meningkatkan risiko diagnosis penyakit AD, meskipun perbedaan dalam setiap kelompok tidak terlalu besar dan variasi ditunjukkan oleh garis *error bar*.



Gambar 4. 45 *Diagram Catplot Diagnosis-Cardiovascular Disease*

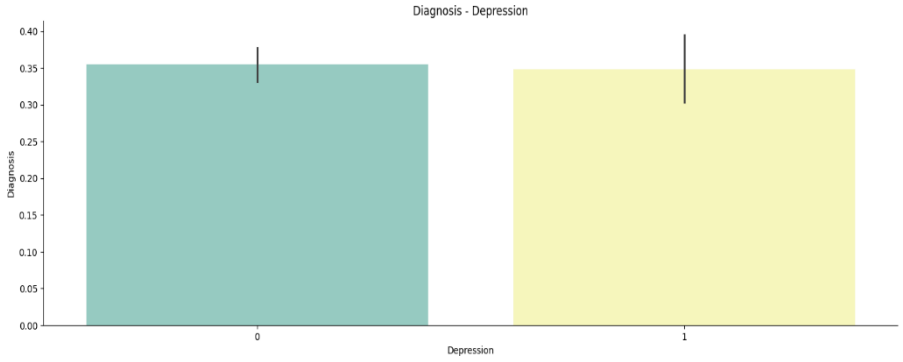
Analisis menunjukkan bahwa ada perbedaan dalam proporsi diagnosis AD antara kelompok yang tidak memiliki riwayat penyakit jantung (0) dan kelompok yang memiliki riwayat penyakit jantung (1). Kelompok dengan riwayat penyakit jantung menunjukkan proporsi diagnosis AD yang sedikit lebih tinggi daripada kelompok yang tidak memiliki riwayat penyakit jantung. Seperti yang ditunjukkan oleh garis *error bar*, data ini menunjukkan adanya korelasi antara riwayat

penyakit kardiovaskular dan risiko diagnosis AD yang lebih tinggi. Hal ini menunjukkan bahwa ada hubungan antara kesehatan otak dan kesehatan pembuluh darah dan jantung.



Gambar 4. 46 *Diagram Catplot Diagnosis-Diabetes*

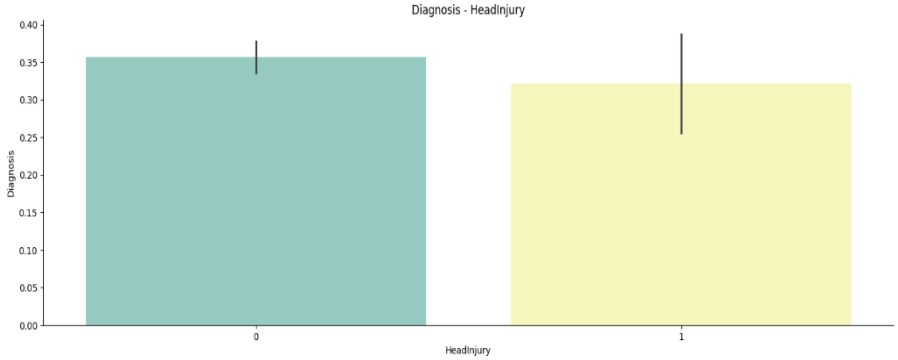
Data menunjukkan bahwa kelompok yang tidak memiliki riwayat diabetes (0) dan kelompok yang memiliki riwayat diabetes (1) memiliki perbedaan dalam proporsi diagnosis AD. Kelompok yang tidak memiliki riwayat diabetes menunjukkan proporsi diagnosis AD yang sedikit lebih tinggi dibandingkan dengan kelompok yang memiliki riwayat diabetes. Data ini menunjukkan bahwa individu dengan riwayat diabetes tidak memiliki risiko AD yang lebih tinggi, meskipun kelompoknya berbeda seperti yang ditunjukkan oleh garis *error bar*.



Gambar 4. 47 *Diagram Catplot Diagnosis-Depression*

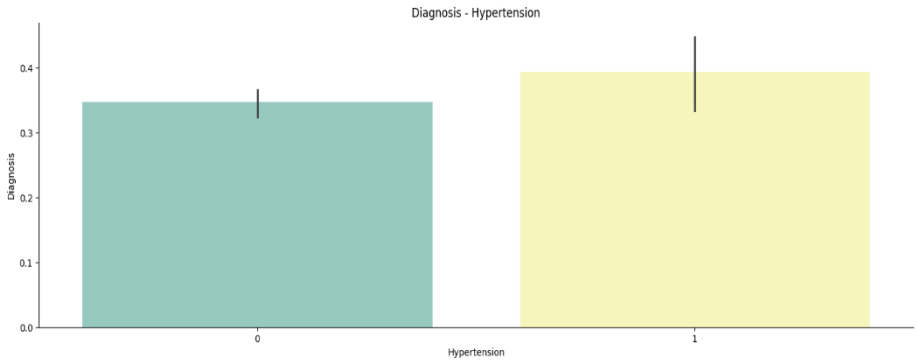
Berdasarkan analisis, terlihat bahwa tidak terdapat perbedaan yang signifikan dalam proporsi diagnosis AD antara kelompok yang tidak memiliki riwayat depresi (0) dan kelompok yang memiliki riwayat depresi (1). Kedua kelompok menunjukkan proporsi diagnosis AD yang hampir serupa, dengan sedikit variasi dalam masing-masing kelompok

seperti yang ditunjukkan oleh garis *error bar*. Oleh karena itu, riwayat depresi tidak menunjukkan secara langsung terkait dengan peningkatan atau penurunan risiko menderita diagnosis AD.



Gambar 4. 48 *Diagram Catplot Diagnosis-Head Injury*

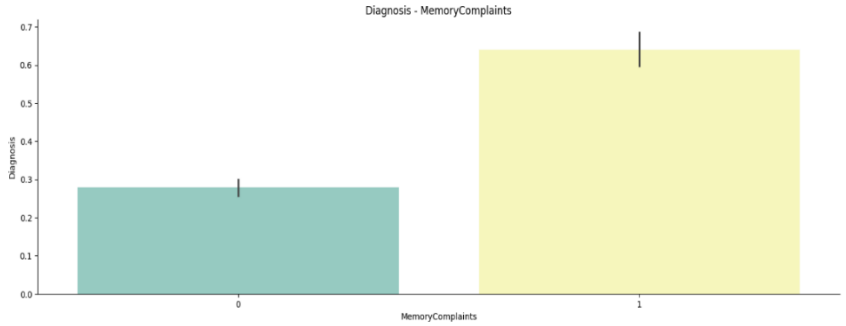
Menunjukkan perbedaan dalam proporsi diagnosis AD antara kelompok yang tidak memiliki riwayat cedera kepala (0) dan kelompok yang memiliki riwayat cedera kepala (1). Kelompok yang memiliki riwayat cedera kepala menunjukkan proporsi diagnosis AD yang sedikit lebih rendah daripada kelompok yang tidak memiliki riwayat cedera kepala. Data ini menunjukkan bahwa individu dengan riwayat cedera kepala tidak memiliki risiko AD lebih tinggi, meskipun kelompoknya berbeda seperti yang ditunjukkan oleh garis *error bar*. Bahkan, data menunjukkan kecenderungan yang berlawanan.



Gambar 4. 49 *Diagram Catplot Diagnosis-Hypertension*

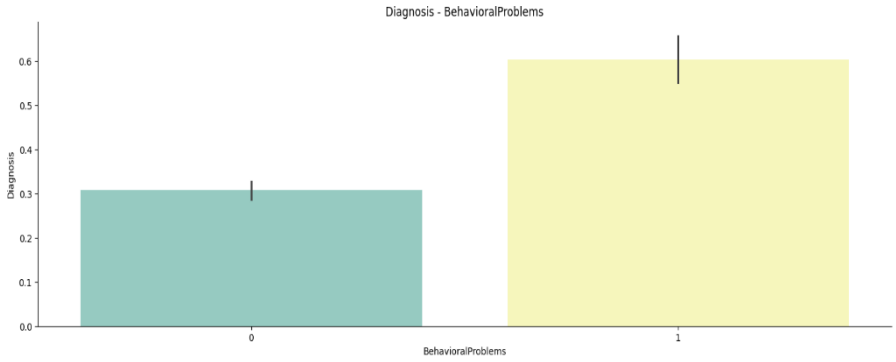
Analisis menunjukkan bahwa ada perbedaan dalam proporsi diagnosis AD antara kelompok yang tidak memiliki riwayat hipertensi (0) dan kelompok yang memiliki riwayat hipertensi (1). Kelompok yang memiliki riwayat hipertensi menunjukkan proporsi diagnosis AD

yang sedikit lebih tinggi dibandingkan dengan kelompok yang tidak memiliki riwayat hipertensi. Setiap kelompok memiliki variasi yang berbeda, seperti yang ditunjukkan oleh garis *error bar*, tetapi informasi ini menunjukkan bahwa riwayat hipertensi dapat berkorelasi dengan peningkatan risiko diagnosis AD.



Gambar 4. 50 Diagram Catplot Diagnosis-Memory Complaints

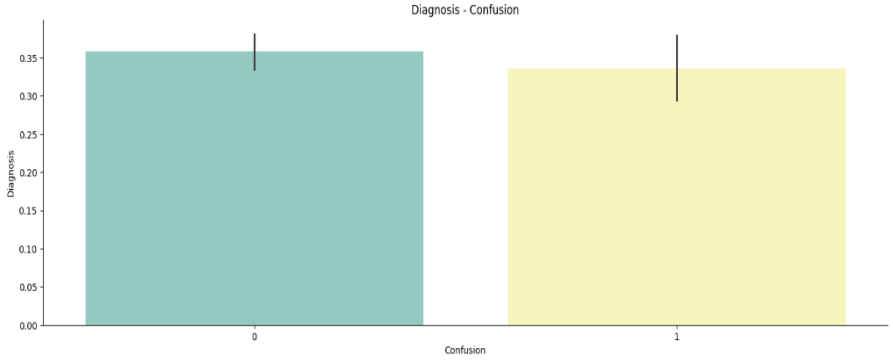
Menunjukkan perbedaan yang cukup signifikan dalam proporsi diagnosis AD antara kelompok yang tidak memiliki keluhan memori (0) dan kelompok yang memiliki keluhan memori (1). Kelompok yang melaporkan keluhan memori menunjukkan proporsi diagnosis AD yang jauh lebih tinggi dibandingkan dengan kelompok yang tidak memiliki keluhan memori. Ada korelasi yang kuat antara keluhan memori dan kemungkinan diagnosis AD, seperti yang ditunjukkan oleh perbedaan yang jelas ini dan *error bar* yang sedikit.



Gambar 4. 51 *Diagram Catplot Diagnosis-Behavioral Problems*

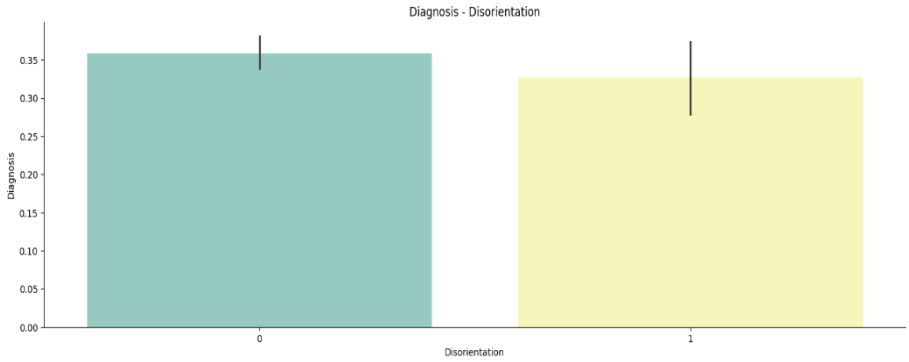
Analisis menunjukkan perbedaan yang cukup signifikan dalam proporsi diagnosis AD antara kelompok yang tidak memiliki masalah perilaku (0) dan kelompok yang memiliki masalah perilaku (1). Kelompok yang dilaporkan memiliki masalah perilaku menunjukkan

proporsi diagnosis AD yang lebih tinggi daripada kelompok yang tidak memiliki masalah perilaku. Ada korelasi yang kuat antara masalah perilaku dan kemungkinan diagnosis AD, seperti yang ditunjukkan oleh perbedaan yang jelas ini dan tidak tumpang tindih yang ditunjukkan oleh *error bar*.



Gambar 4. 52 *Diagram Catplot Diagnosis-Confusion*

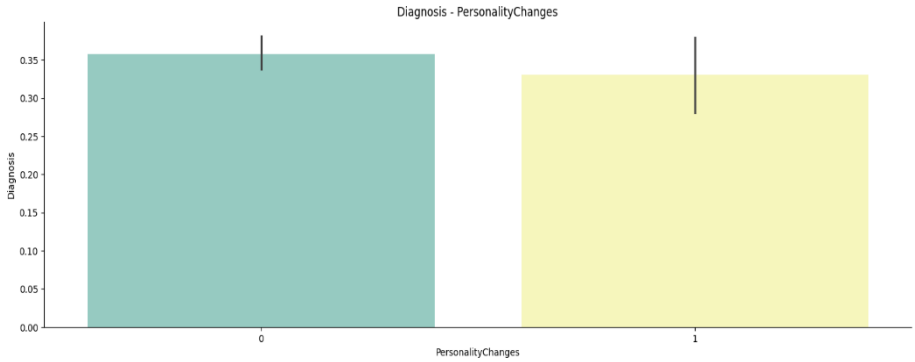
Dengan melihat hubungan antara gejala kebingungan (*confusion*) dan diagnosis AD, terlihat bahwa mereka yang tidak mengalami kebingungan memiliki rata-rata diagnosis AD sedikit lebih tinggi daripada mereka yang mengalami kebingungan (*confusion* = 1). Namun, perbedaan ini tidak signifikan secara statistik karena *error bar* dari kedua kelompok saling tumpang tindih. Ini berarti bahwa gejala kebingungan tidak secara jelas membedakan kemungkinan seseorang didiagnosis AD atau tidak.



Gambar 4. 53 *Diagram Catplot Diagnosis-Disorientation*

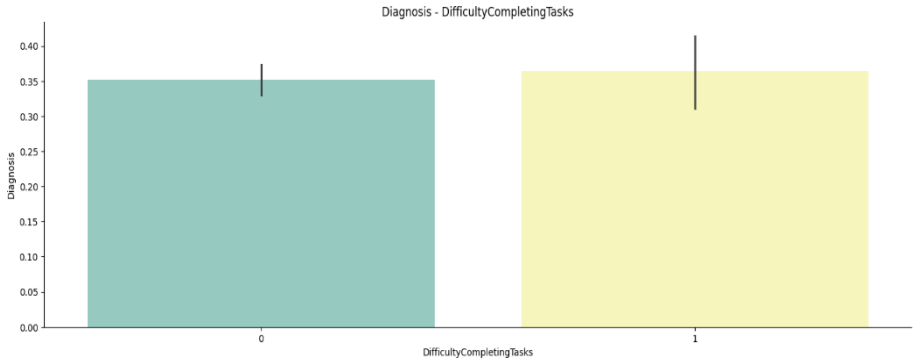
Grafik hubungan antara *Disorientation* (Disorientasi) dan Diagnosis AD menunjukkan bahwa individu yang tidak mengalami disorientasi (Disorientation = 0) memiliki diagnosis AD sedikit lebih tinggi

daripada individu yang mengalami disorientasi (*Disorientation* = 1). Namun, perbedaan ini tidak signifikan secara statistik karena rentang *error bar* kedua kelompok saling tumpang tindih. Data menunjukkan bahwa gejala disorientasi tidak berdampak yang signifikan atau konsisten pada kemungkinan diagnosis AD.



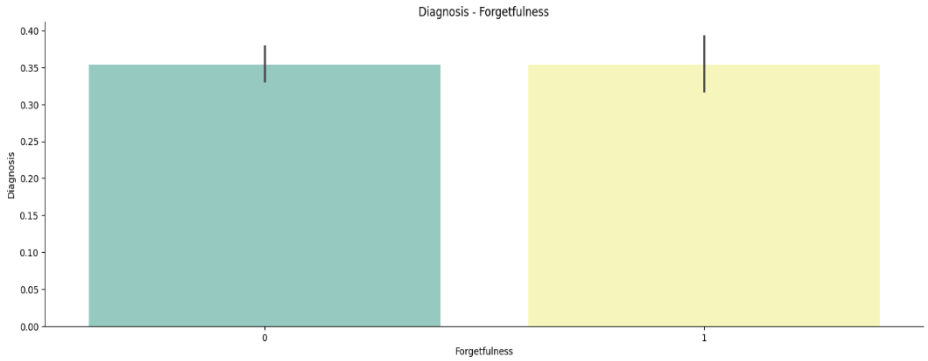
Gambar 4. 54 *Diagram Catplot Diagnosis-Personality Changes*

Berdasarkan grafik hubungan antara perubahan kepribadian (*Personality Changes*) dan diagnosis, terlihat bahwa rata-rata diagnosis AD lebih tinggi pada individu yang tidak mengalami perubahan kepribadian (*Personality Changes* = 0) dibandingkan dengan yang mengalami perubahan (*Personality Changes* = 1). Namun, perbedaan tersebut tidak mencolok, dan *error bar* kedua kelompok tumpang tindih, yang menunjukkan bahwa perbedaan ini tidak signifikan secara statistic dan variabel perubahan kepribadian tidak menunjukkan hubungan kuat atau konsisten dengan kemungkinan seseorang didiagnosis AD.



Gambar 4. 55 *Diagram Catplot Diagnosis-Difficulty Completing Task*

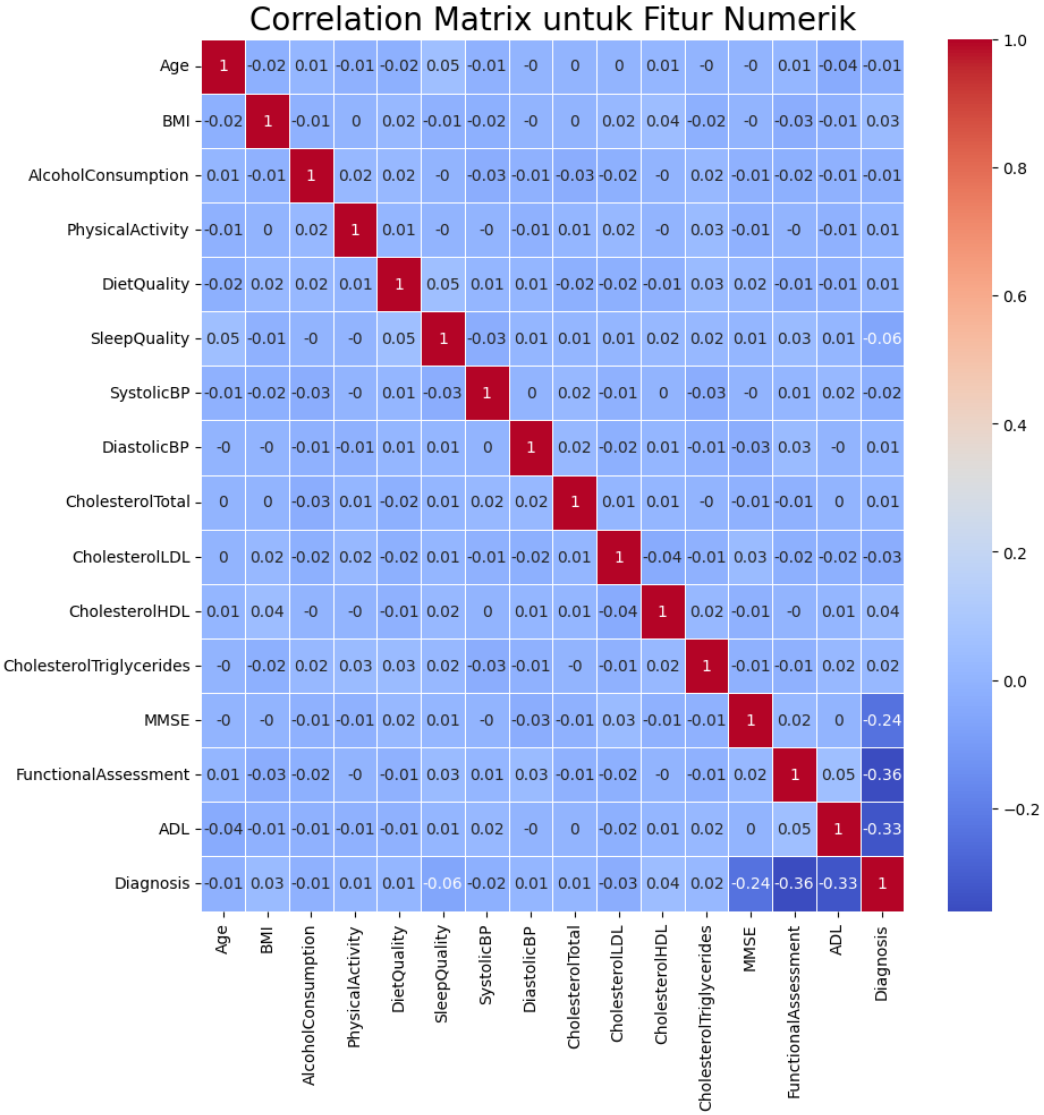
Analisis menunjukkan bahwa, antara kelompok yang tidak mengalami kesulitan menyelesaikan tugas (0) dan kelompok yang mengalami kesulitan menyelesaikan tugas (1), proporsi diagnosis AD sedikit lebih rendah dibandingkan dengan kelompok yang tidak mengalami kesulitan. Namun, perlu diperhatikan bahwa ada hubungan yang cukup besar antara *error bar* kelompok kedua. Hal ini menunjukkan bahwa perbedaan rata-rata proporsi diagnosis AD antara kedua kelompok tersebut mungkin tidak terlalu signifikan secara statistik, dan kesulitan menyelesaikan tugas mungkin merupakan salah satu gejala yang menyertainya. Namun, berdasarkan data ini, gejala ini tidak menjadi perbedaan yang signifikan dalam diagnosis AD.



Gambar 4. 56 *Diagram Catplot Diagnosis-Forgetfullnes*

Analisis menunjukkan bahwa, antara kelompok yang tidak mengalami masalah lupa (0) dan kelompok yang mengalami masalah lupa (1), proporsi diagnosis AD sedikit lebih rendah dibandingkan dengan kelompok yang tidak mengalami masalah lupa. Namun, grafik sebelumnya menunjukkan tumpang tindih yang signifikan antara *error bar* kedua kelompok.

b) Numerical Faetures



Gambar 4. 57 *Matrix korelasi fitur numerik*

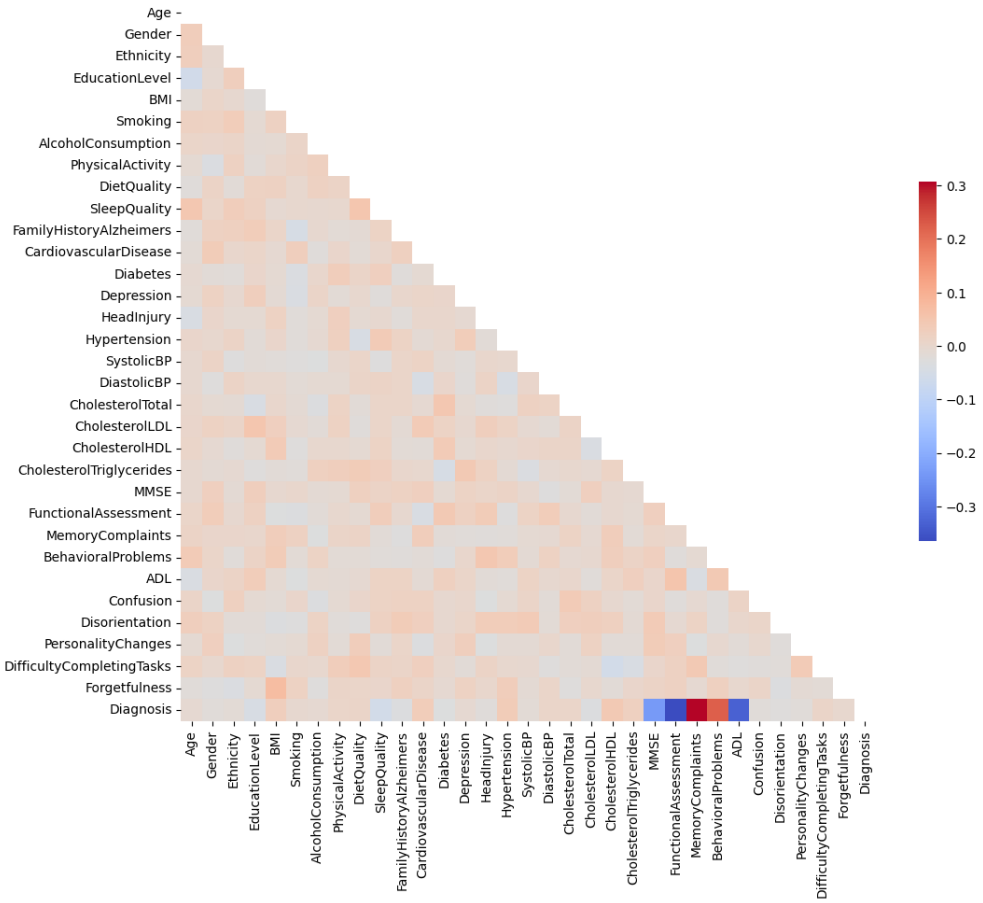
Berdasarkan *matrix korelasi fitur numerik*, menunjukkan bahwa perbedaan antara kedua kelompok dalam proporsi rata-rata diagnosis AD mungkin tidak terlalu signifikan secara statistik.

Terdapat korelasi positif antara *cholesterol total* dan *cholesterol LDL* yang menunjukkan korelasi positif lemah atau tidak memiliki hubungan *linear* yang kuat sehingga peningkatan kadar *cholesterol total* tidak secara signifikan diikuti oleh peningkatan kadar *cholesterol LDL*.

Pada hubungan antara *systolicBP* dan dan *diastolicBP* yang termasuk dalam kategori tidak ada korelasi atau korelasi 0, yang mengindikasikan bahwa perubahan pada tekanan darah *systolic* dan *diastolic* tidak selalu terjadi bersamaan.

Korelasi *negative* yang mendekati -1 terdapat pada *functional assessment* dan diagnosis sebesar -0.36 yang mengindikasikan bahwa semakin tinggi skor pada penilaian fungsional (*functional assessment*) semakin kecil kemungkinan terdiagnosis AD. Pada ADL dan diagnosis juga menunjukkan korelasi yang cukup signifikan sebesar -0.33 yang berarti semakin baik kemampuan seseorang dalam melakukan aktivitas sehari-hari semakin rendah kemungkinan diagnosis AD. Sementara itu pada MMSE dan diagnosis memiliki korelasi negatif lemah sebesar -0.24 ini menunjukkan adanya kecenderungan bahwa skor tes kognitif yang lebih rendah.

7) Matrix Korelasi

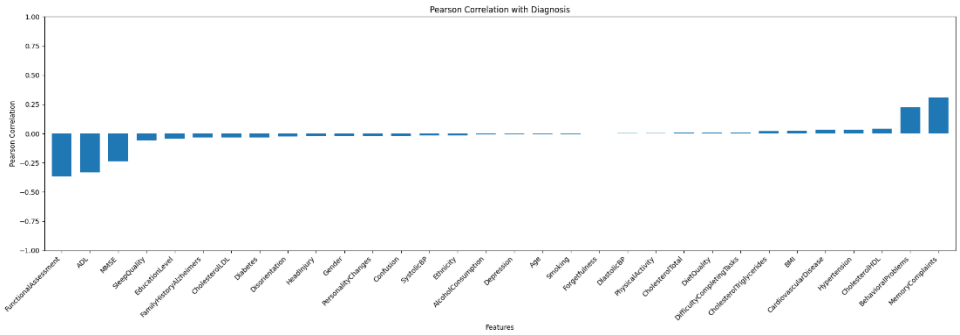


Gambar 4. 58 *Matrix korelasi*

Terdapat tiga fitur numerik *Functional Assessment* (Penilaian Fungsional), ADL (Aktivitas Kehidupan Sehari-hari), dan MMSE (*Mini-Mental State Examination*) berkorelasi negatif dengan diagnosis AD, dengan koefisien korelasi masing-masing -0,36, -0,33, dan -0,24. Hal ini menunjukkan bahwa nilai yang lebih rendah dalam penilaian ini dikaitkan dengan kemungkinan diagnosis AD yang lebih tinggi.

Selain itu, dua variabel kategoris *Behavioral Problems* (Masalah Perilaku) dan *Memory Complaints* (Keluhan Memori) berkorelasi positif dengan diagnosis dengan koefisien korelasi masing-masing sebesar 0,2 dan 0,30. Ini berarti keberadaan masalah-masalah ini dikaitkan dengan kemungkinan diagnosis AD yang lebih tinggi, yang menyoroti pentingnya masalah-masalah ini dalam proses diagnostik.

8) Selection feature



Gambar 4. 59 *Pearson Corelation*

Seleksi fitur yang digunakan adalah metode *pearson correlation* untuk mengukur hubungan linier antara setiap fitur numerik dengan target. Nilai korelasi berkisar dari -1 hingga 1, di mana nilai mendekati 1 atau -1 menunjukkan hubungan linier yang kuat, dan nilai mendekati 0 menunjukkan hubungan yang lemah atau tidak ada hubungan linier. *Functional Assessment*, MMSE dan ADL memiliki korelasi negatif yang cukup kuat dengan diagnosis yang mengindikasikan bahwa semakin tinggi nilai pada fitur-fitur ini, kemungkinan besar diagnosisnya adalah non-positif. Sementara itu, fitur *Behavioral Problems* dan *Memory Complaints* yang berarti semakin tinggi nilainya, semakin besar kemungkinan diagnosis menunjukkan kondisi positif.

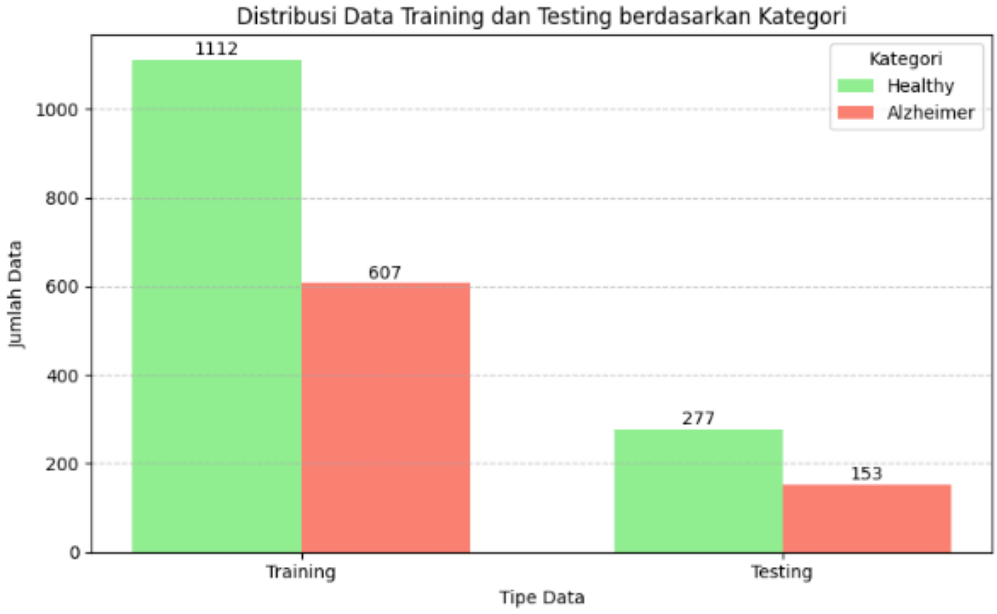
Tabel 4. 2 Hasil *Feature Selection Pearson's Correlation*

No	Fitur
1	*MMSE*
2	*FunctionalAssessment*
3	*MemoryComplaints*
4	*BehavioralProblems*
5	*ADL*
6	*Diagnosis*

b. Data Preprocessing

1) Splitting Data

Pembagian dataset menggunakan rasio sebesar 80:20 untuk *training data* dan *test data* untuk melatih model dan menguji model.



Gambar 4. 60 *Splitting data*

2) *Encoding Fitur*

	FunctionalAssessment	ADL	MMSE	BehavioralProblems	MemoryComplaints
0	0.652102	0.172486	0.715606	0.0	0.0
1	0.712108	0.259154	0.687251	0.0	0.0
2	0.589697	0.711936	0.245145	0.0	0.0
3	0.896823	0.648094	0.466410	1.0	0.0
4	0.604699	0.001341	0.450619	0.0	0.0

Gambar 4. 61 *Encoding Fitur Category*

Encoding () dilakukan untuk mengubah fitur-fitur kategorikal yang ada pada *dataframe* menjadi bentuk numerik menggunakan metode *Label Encoding.*

3) *Normalisasi Data*

Menggunakan metode *Min-Max Scaling* dengan mengubah skala nilai-nilai dalam kolom tertentu agar berada dalam rentang 0 hingga 1 untuk melakukan normalisasi data pada kolom-kolom numerik.

	FunctionalAssessment	ADL	MMSE	BehavioralProblems	MemoryComplaints
count	2149.000000	2149.000000	2149.000000	2149.000000	2149.000000
mean	0.508162	0.498244	0.491889	0.156817	0.208004
std	0.289390	0.295023	0.287238	0.363713	0.405974
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.256685	0.234191	0.238854	0.000000	0.000000
50%	0.509601	0.503846	0.481435	0.000000	0.000000
75%	0.754954	0.758137	0.738867	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000

Gambar 4. 62 Normalisasi Data

c. Modelling

Penelitian ini menggunakan metode *random forest*, Adapun tahapan yang dilakukan dengan melakukan *training data* dan *hyperparameter tuning* sebagai berikut.

a) Training

Training dilakukan untuk melatih model agar dapat mengenali pola atau hubungan antar fitur *input* dan target *output* dengan mencatat hasil evaluasi peforma model *random forest* terhadap *training* seperti akurasi,

presisi, *recall*, dan f1-*score* baik untuk data pelatihan (*train*) maupun data pengujian (*test*).

b) Hyperparameter Tuning

Tuning hyperparameter dilakukan pada model *random forest classifier* menggunakan teknik *Grid Search* dengan 5-*fold cross-validation*. Dengan melakukan model *RandomForestClassifier()* diinisialisasi sebagai model dasar. Kemudian, *param_grid* didefinisikan sebagai kumpulan kombinasi nilai *hyperparameter* yang ingin diuji, seperti jumlah pohon *(n_estimators)* dengan nilai yang diuji 50, 100, dan 200 untuk melihat berapa banyak pohon yang optimal dalam membentuk model yang baik tanpa menyebabkan *overfitting* atau memperlama waktu komputasi., cara pemilihan fitur terbaik (*max_features*) dengan nilai yang diuji *'sqrt'* dan *'log2'* untuk menghindari korelasi antar pohon agar model lebih kuat, kedalaman maksimum pohon (*max_depth*) dengan nilai yang diuji *none,* 10, 20, dan 30 dan untuk mengontrol kompleksitas model semakin dalam pohonnya, semakin kompleks modelnya. kriteria pemisahan (*criterion*) dengan nilai yang diuji 'gini' (*impurity*) untuk mengukur kualitas *split*, dan menggunakan teknik *bootstrap* (*bootstrap*) dengan nilai yang diuji *True*, artinya model menggunakan teknik *bootstrap sampling* yaitu mengambil sampel acak dengan pengembalian dari data pelatihan.

1) Hyperparameter Search

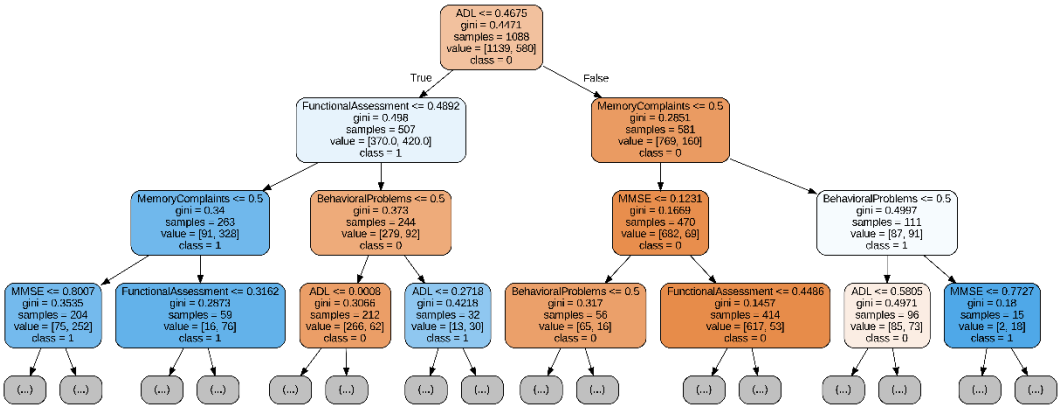
GridSearchCV digunakan untuk mengevaluasi semua kombinasi parameter tersebut melalui validasi silang sebanyak 5 kali (cv=5), sehingga model dapat diuji secara menyeluruh pada data yang berbeda. Dengan *n_jobs*=1, proses berjalan di satu *core* CPU, dan *verbose*=2 digunakan untuk menampilkan proses secara lebih rinci. Selain itu, *return_train_score=True* memungkinkan kita melihat skor performa model pada data pelatihan untuk setiap kombinasi parameter. Tujuan utama dari proses ini adalah menemukan kombinasi *hyperparameter* terbaik yang menghasilkan performa model paling optimal.

Tabel 4. 3 Hasil Tuning Parameter Terbaik

boostrap	*criterion*	*max_depth*	*max_features*	*n_estimators*
True	gini	10	*Sqrt*	100

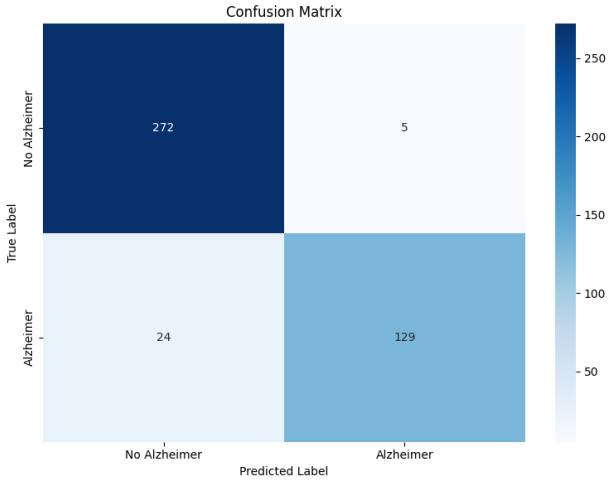
2) Re-Training

Re-training dilakukan untuk memastikan model akhir benar-benar dilatih menggunakan konfigurasi terbaik setelah parameter terbaik ditemukan, model perlu dilatih dan dilakukan visualisasi pohon dalam *random forest.*



Gambar 4. 63 *Random Forest*

d. Evaluation

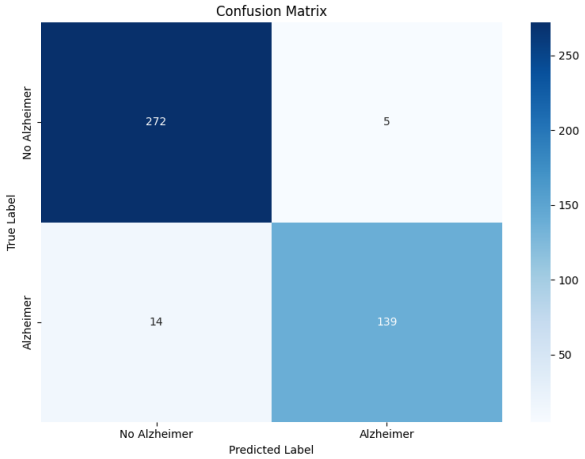


Gambar 4. 64 *Confusion matrix random forest*

Berdasarkan *confusion matrix* model klasifikasi yang digunakan untuk mendeteksi AD menunjukkan performa yang cukup baik. Dari total 430 data yang diuji, model berhasil mengklasifikasikan 272 individu yang tidak mengidap AD dengan benar (*True Negatives*), dan 129 individu yang mengidap AD juga terklasifikasi dengan benar (*True Positives*). Namun, terdapat 24 kasus di mana model gagal mendeteksi individu yang sebenarnya mengidap AD (*False Negatives*), dan 5 kasus di mana individu yang sehat justru diprediksi mengidap AD (*False Positives*).

Secara keseluruhan, model mencapai akurasi sebesar 93,26%, yang menunjukkan bahwa sebagian besar prediksi model sudah tepat. Presisi model untuk mendeteksi AD sebesar 96,26%, artinya ketika model

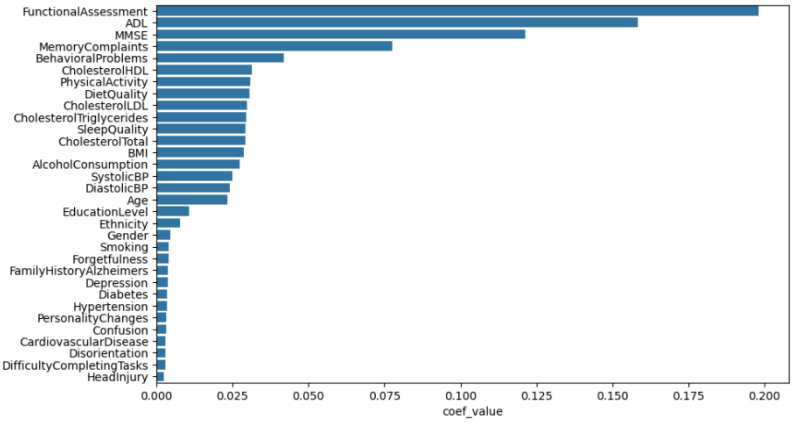
menyatakan seseorang mengidap AD, kemungkinan besar prediksi tersebut benar. Namun, *recall* model sebesar 84,31% menunjukkan bahwa masih ada sejumlah kasus AD yang tidak berhasil terdeteksi oleh model. Dengan F1-*score* sebesar 89,90%, yang merupakan rata-rata harmonis dari presisi dan *recall*, model dapat dikatakan seimbang dalam hal ketepatan dan kemampuan mendeteksi pasien yang benar-benar mengidap AD.



Gambar 4. 65 *Confusion matrix random forest + pearson corelation*

Berdasarkan confusion matrix model klasifikasi menunjukkan dapat membedakan antara penderita AD dan tidak AD. Dari total data uji 430 data uji, model berhasil memprediksi 272 kasus "*No alzheimer*" dan 139 kasus "*Alzheimer*" dengan benar, serta hanya membuat kesalahan yaitu 5 kasus *false positive* dan 14 kasus *false negative*. Model memiliki akurasi sebesar 95,58%, presisi 96,53%, *recall* 90,85%, dan F1-*Score* 93,55%.

9) Menampilkan *Best Feature*

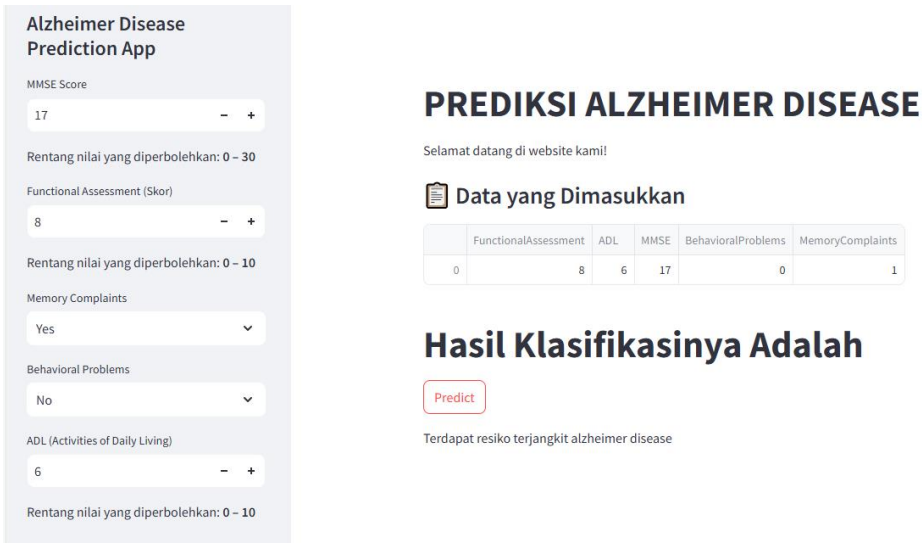


Gambar 4. 66 *best feature*

Menampilkan dan mengurutkan pentingnya fitur-fitur (variabel *input*) dalam sebuah model *random forets*. fitur-fitur yang paling berpengaruh terhadap kinerja model *random forest* adalah *Functional Assessment*, ADL, dan MMSE yang masing-masing memiliki nilai kontribusi (*importance*) tertinggi. Ini menunjukkan bahwa aspek-aspek fungsional dan kognitif individu sangat penting dalam memprediksi *outcome* yang sedang dianalisis kemungkinan besar terkait dengan kondisi kesehatan atau fungsi kognitif lansia.

e. Testing/Deployment

Berdasarkan hasil performa maka akan diuji coba melalui *platform streamlit* untuk memprediksi seseorang terkena *alzheimer disease* (AD) atau tidak.



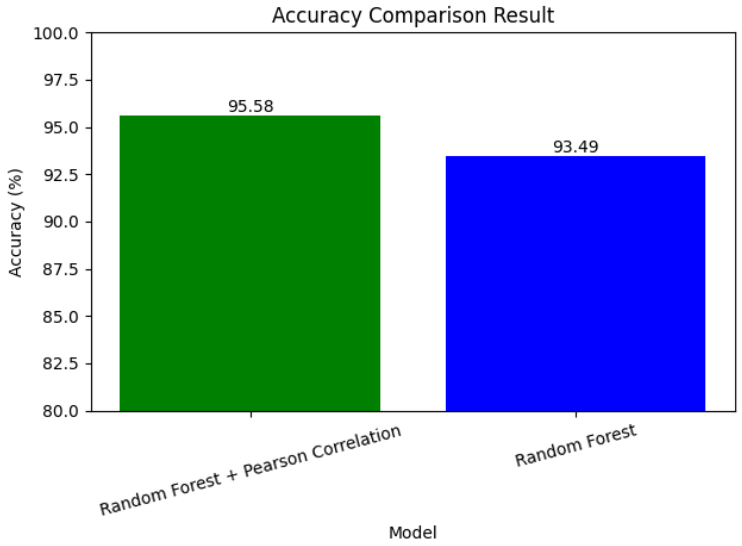
Gambar 4. 67 *Deployment*

Gambar 4.67 merupakan hasil uji coba *deployment* model dengan memasukkan inputan sesuai yang telah disediakan, setelah memasukkan semua indikator lalu klik bagian *predict* sehingga tampil klasifikasi yang terdapat pada gambar 4.69 yang menunjukkan bahwa seseorang tidak terindikasi *alzheimer disease* (AD) berdasarkan indikator yang dimasukkan.

4.2 Pembahasan

Prediksi risiko *alzheimer disease* (AD) dengan metode random forest memberikan performa yang cukup baik dalam menentukan seseorang terkena AD atau tidak. Hal ini didasarkan pada nilai metrik *precision, recall,* f1-*score*. Yang mencapai performa nilai metrik 96,30% *precision*, 84,97% *recall*, dan 90,28% f1-*Score*. Beberapa metrik tersebut di komparasi dengan peforma model prediksi risiko *alzheimer disease* (AD) menggunakan metode random forest dengan kombinasi seleksi fitur *pearson corelation* yang menunjukkan performa model lebih baik dengan performa metrik masing-masing memiliki nilai 96,53% precision, 90,85% recall, dan 93,60% f1-*score.*

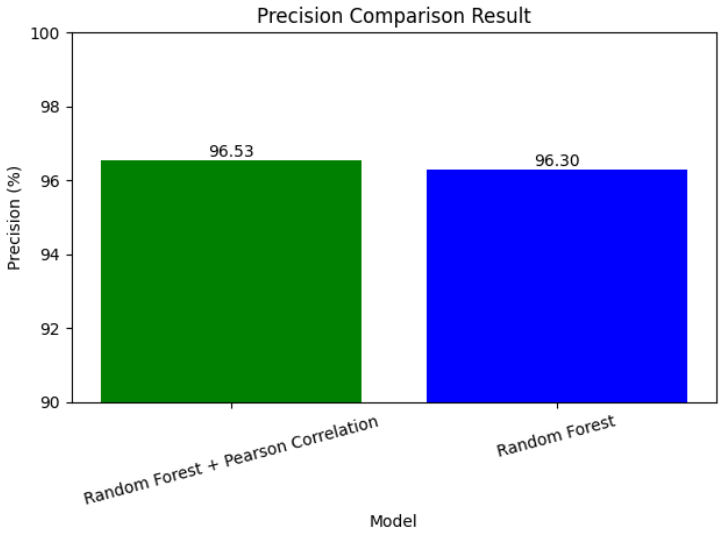
a. Accuracy



Gambar 4. 68 *Bar Chart Comparation Accuracy*

Nilai *accuracy* menunjukkan bahwa penggunaan teknik seleksi fitur dengan *pearson correlation* dapat meningkatkan akurasi model secara signifikan. Model dengan kombinasi *pearson correlation* menghasilkan akurasi sebesar 95,58%, lebih tinggi dibandingkan model *random forest* tanpa seleksi fitur yang hanya mencapai 93,49%. Hal ini mengindikasikan bahwa proses pemilihan fitur yang relevan dapat membantu model mempelajari pola data secara lebih efektif, sehingga memberikan hasil prediksi yang lebih akurat.

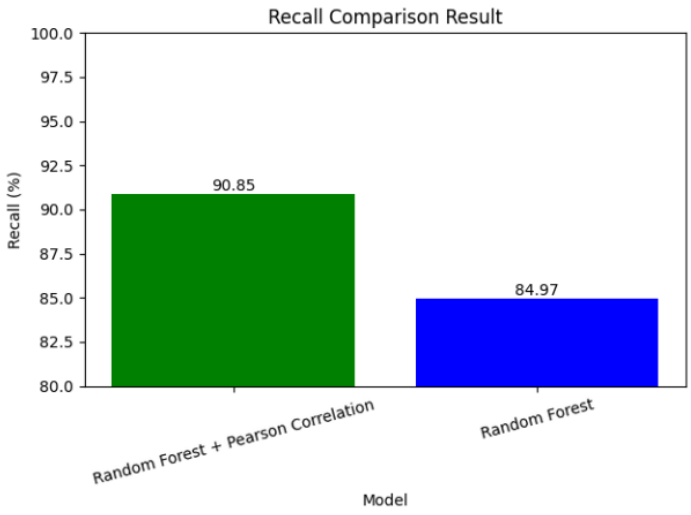
b. *Precision*



Gambar 4. 69 *Bar Chart Comparation Precission*

Perbandingan nilai presisi antara dua model, yaitu "*Random Forest*" dan "*Random Forest + Pearson Correlation*". model yang menggunakan seleksi fitur dengan *pearson correlation* memiliki presisi sedikit lebih tinggi, yaitu 96,53%, dibandingkan dengan *random forest* tanpa seleksi fitur dengan presisi sebesar 96,30%. Meskipun perbedaannya tidak terlalu besar, hasil ini menunjukkan bahwa pemilihan fitur yang tepat dapat memberikan kontribusi positif terhadap kemampuan model dalam mengidentifikasi prediksi yang benar, khususnya dalam mengurangi kesalahan positif palsu.

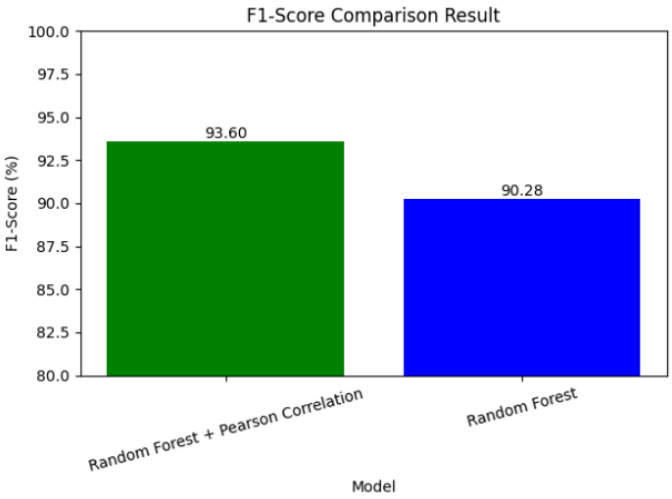
c. *Recall*



Gambar 4. 70 *Bar Chart Comparation Recall*

Model *random forest* yang dikombinasikan dengan seleksi fitur menggunakan *pearson correlation* memiliki performa yang lebih baik dibandingkan dengan model *random forest* tanpa seleksi fitur. Model dengan *pearson correlation* mencapai *recall* sebesar 90,85%, sedangkan model tanpa seleksi fitur hanya memperoleh 84,97%.

d. F1-*Score*



Gambar 4. 71 *Bar Chart Comparation Precission*

Model *random forest* yang menggunakan seleksi fitur *pearson correlation* menunjukkan performa yang lebih unggul dibandingkan dengan model *random forest* tanpa seleksi fitur. Model dengan *pearson correlation* mencapai F1-*score* sebesar 93,60%, sementara model tanpa seleksi fitur hanya memperoleh 90,28%.

Tabel 4. 4 Hasil Perbadingan Model

No	Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
1	*Random Forest*	93,49%	96,30%	84,97%	90,28%
2	*Random Forest Feature Selection Pearson Correlation*	95,58%	96,53%	90,85%	93,60%