



CHRONIC KIDNEY DISEASE PREDICTION

1. 2702248591 - Antony Wijaya
2. 2702234592 - Laurentius Brandon Vikario
3. 2702243520 - Annabelle Fevriane
4. 2702223802 - Chingtya Tjoeng
5. 2702229913 - Nadine Beatricia
6. 2702259600 - Samantha Niandra
7. 2702338284 - Jessica Debora



LATAR BELAKANG

- **Penyakit Ginjal Kronis atau Chronic Kidney Disease (CKD)** adalah kondisi progresif yang berdampak besar terhadap kesejahteraan dan aktivitas harian pasien.
- CKD umumnya **disebabkan oleh diabetes, hipertensi, serta kondisi medis lainnya.**
- **Deteksi dini CKD sangat penting untuk mencegah komplikasi** serius seperti gagal ginjal dan penyakit kardiovaskular. Namun, karena CKD sering tidak menunjukkan gejala pada tahap awal, **diagnosis melalui metode konvensional menjadi tantangan.**
- Menerapkan dua model machine learning — **Random Forest dan XGBoost** — dalam memprediksi keberadaan CKD berdasarkan data klinis dari **dataset publik "Chronic Kidney Disease" di Kaggle.**
- **Tujuan** dari penelitian ini, **membandingkan kinerja kedua model dalam mengklasifikasikan pasien dengan CKD** secara akurat.



DATASETS

Dataset berasal dari Kaggle yang berisi catatan pasien dengan berbagai fitur seperti tekanan darah, kadar glukosa darah, hemoglobin, dan kreatinin serum, dan lain-lain

Dataset:

<https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis>



DATASETS ATTRIBUTES

Atribut yang memiliki korelasi dengan Diagnosis lebih dari 0.02

'SocioeconomicStatus'
'EducationLevel'
'BMI'
'Smoking'
'PhysicalActivity'
'FamilyHistoryKidneyDisease'
'FamilyHistoryHypertension'
UrinaryTractInfections'
'SystolicBP'
'DiastolicBP'
'FastingBloodSugar'
'HbA1c'
"SerumCreatinine"

"BUNLevels"
'GFR'
'ProteinInUrine',
'SerumElectrolytesSodium',
'SerumElectrolytesPotassium',
'HemoglobinLevels',
'CholesterolTotal'
'CholesterolHDL'
'Diuretics'
'Edema'
'NauseaVomiting'
'MuscleCramps'
'Itching'

DATA CLEANING

MENGHAPUS DUPLICATE DATA

Duplikasi baris pada dataset menyebabkan model “belajar” dari data yang sama persis berulang kali, sehingga membuat bias pada pattern tertentu. Dengan menghapus menggunakan `'df.drop_duplicate()'`, dataset menjadi lebih bersih dan representatif

MENGHAPUS KOLOM YANG TIDAK RELEVAN

Beberapa kolom dalam dataset mungkin tidak relevan ataupun bersifat bias dengan proses prediksi, misalnya PatientID atau nama, atau variabel yang tidak bersifat medis



DATA PREPROCESSING

PEMISAHAN FITUR DAN LABEL

```
x = df_numeric.drop(columns='Diagnosis')  
y = df_numeric['Diagnosis']
```

x untuk bagian data fitur pasien (input)
y untuk bagian diagnosis CKD atau tidak
(target/ label)

MENANGANI CLASS IMBALANCE

```
smote = SMOTE(random_state=42)  
X_resampled, y_resampled =  
smote.fit_resample(X_scaled, y)
```

jika jumlah pasien CKD jauh lebih sedikit daripada non-CKD, model bisa kesulitan mengenali pattern minoritas. SMOTE membuat data sintetis untuk minoritas CKD agar distribusinya seimbang

PEMBAGIAN DATASET

Data dibagi menjadi 3 bagian yaitu:

- 70% untuk training
- 20% untuk testing
- 10% untuk validasi model



DATA PREPROCESSING

HANDLING MISSING VALUES

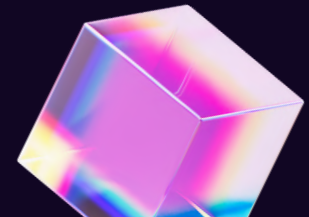
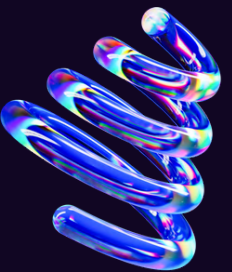
Dalam dataset CKD, beberapa atribut pasien mungkin tidak terisi. Data yang kosong ini perlu diolah, karena jika dibiarkan bisa mempengaruhi hasil pelatihan model. Penanganan missing values ini membantu menjaga kualitas dan konsistensi data.

NORMALIZING NUMERICAL ATTRIBUTES

Selanjutnya dilakukan normalisasi pada data numerik. Beberapa atribut seperti kadar kolesterol, tekanan darah, atau gula darah memiliki skala nilai yang berbeda-beda. Normalisasi dilakukan agar semua data numerik memiliki skala yang seragam, sehingga model dapat memahami pola data dengan lebih baik.

ENCODING CATEGORICAL VARIABLES WHERE NECESSARY

Atribut yang bersifat kategorikal perlu diubah menjadi angka. Model Machine Learning tidak bisa memproses data dalam bentuk teks seperti “Ya” atau “Tidak”, sehingga perlu diubah menjadi nilai numerik misalnya “Ya” menjadi 1 dan “Tidak” menjadi 0. proses ini disebut dengan encoding



MODEL PREDICTION

01

RANDOM FOREST

- Metode pembelajaran supervised yang membangun beberapa decision trees untuk model prediksi, di mana setiap pohon dilatih pada subset data pelatihan yang dipilih secara acak.
- Random Forest merupakan algoritma yang serbaguna karena kesederhanaan dan keragamannya, sehingga dapat diterapkan pada masalah klasifikasi maupun regresi.
- Random Forest juga mampu mengolah dataset dengan variabel kontinu dan kategorikal.

02

XGBOOST

- Kerangka kerja gradient boosting yang dioptimalkan untuk efisiensi dan skalabilitas. Algoritma ini meningkatkan metode boosting tradisional dengan memperkenalkan algoritma yang peka terhadap sparsity untuk menangani missing data dan teknik weighted quantile sketching untuk pembangunan decision tree yang efisien
- Lebih baik dalam menangani high-dimensional data dan menerapkan shrinkage serta column subsampling untuk mengurangi overfitting.

HYPERPARAMETER TUNING (RANDOMIZEDSEARCHCV)

Random Forest

N estimators	antara 100 sampai 299
Max depth	antara 5 sampai 29
Min sample split	antara 2 sampai 9
Min sample leaf	antara 1 sampai 4

XGBoost

N estimators	antara 100 sampai 299
Max depth	antara 3 sampai 9
Learning rate	antara 0.01 sampai 0.21
Subsample	antara 0.5 sampai 1
Col sample by tree	antara 0.5 sampai 1

SELECTED HYPERPARAMETERS

Random Forest

N estimators	157
Max depth	25
Min sample split	5
Min sample leaf	1

XGBoost

N estimators	163
Max depth	9
Learning rate	0.04993475643167195
Subsample	0.73338144662399
Col sample by tree	0.8925879806965068

EVALUATION ACCURATION

accuracy_score() from sklearn.metrics

```
accuracy_rf = accuracy_score(y_test, y_pred_rf)
accuracy_xgb = accuracy_score(y_test, y_pred_xgb)

print(f"Random Forest Accuracy (Test Set): {accuracy_rf * 100:.4f}%")
print(f"XGBoost Accuracy (Test Set): {accuracy_xgb * 100:.4f}%")
```

[257] ✓ 0.0s

... Random Forest Accuracy (Test Set): 98.1785%
XGBoost Accuracy (Test Set): 97.4499%

```
# Prediksi dan evaluasi di validation set
val_pred_rf = best_rf.predict(X_val)
val_pred_xgb = best_xgb.predict(X_val)

print(f"Random Forest Accuracy (Validation Set): {accuracy_score(y_val, val_pred_rf) * 100:.4f}%")
print(f"XGBoost Accuracy (Validation Set): {accuracy_score(y_val, val_pred_xgb) * 100:.4f}%")
```

[258] ✓ 0.0s Python

... Random Forest Accuracy (Validation Set): 97.0492%
XGBoost Accuracy (Validation Set): 96.3934%

classification_report() from sklearn.metrics

```
# Classification report RF
rf_report = classification_report(y_test, y_pred_rf, output_dict=True)
rf_df_report = pd.DataFrame(rf_report).transpose()
print("Random Forest Classification Report:")
print(rf_df_report)
```

[261] ✓ 0.0s

... Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.970480	0.992453	0.981343	265.000000
1	0.992806	0.971831	0.982206	284.000000
accuracy	0.981785	0.981785	0.981785	0.981785
macro avg	0.981643	0.982142	0.981775	549.000000
weighted avg	0.982029	0.981785	0.981790	549.000000

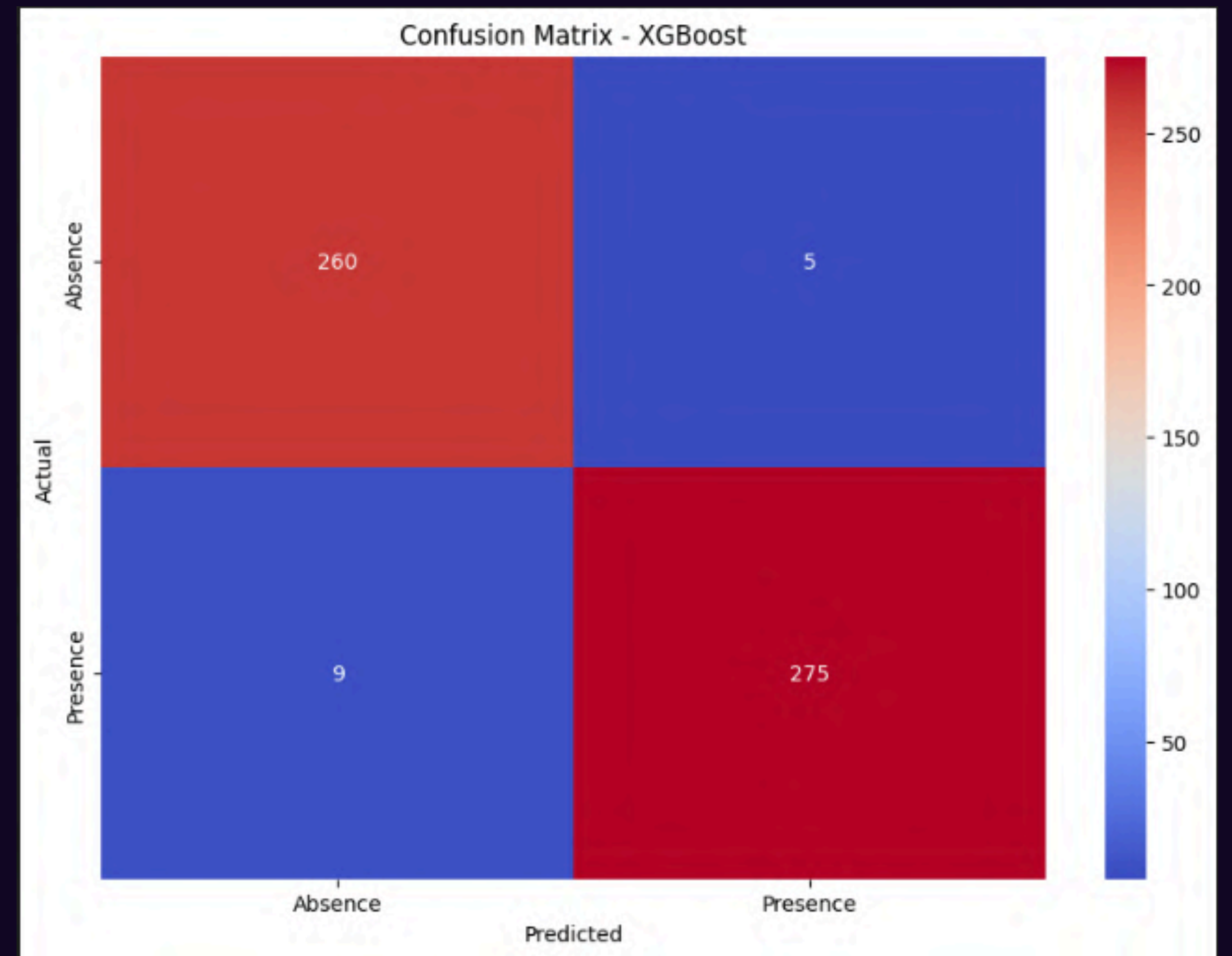
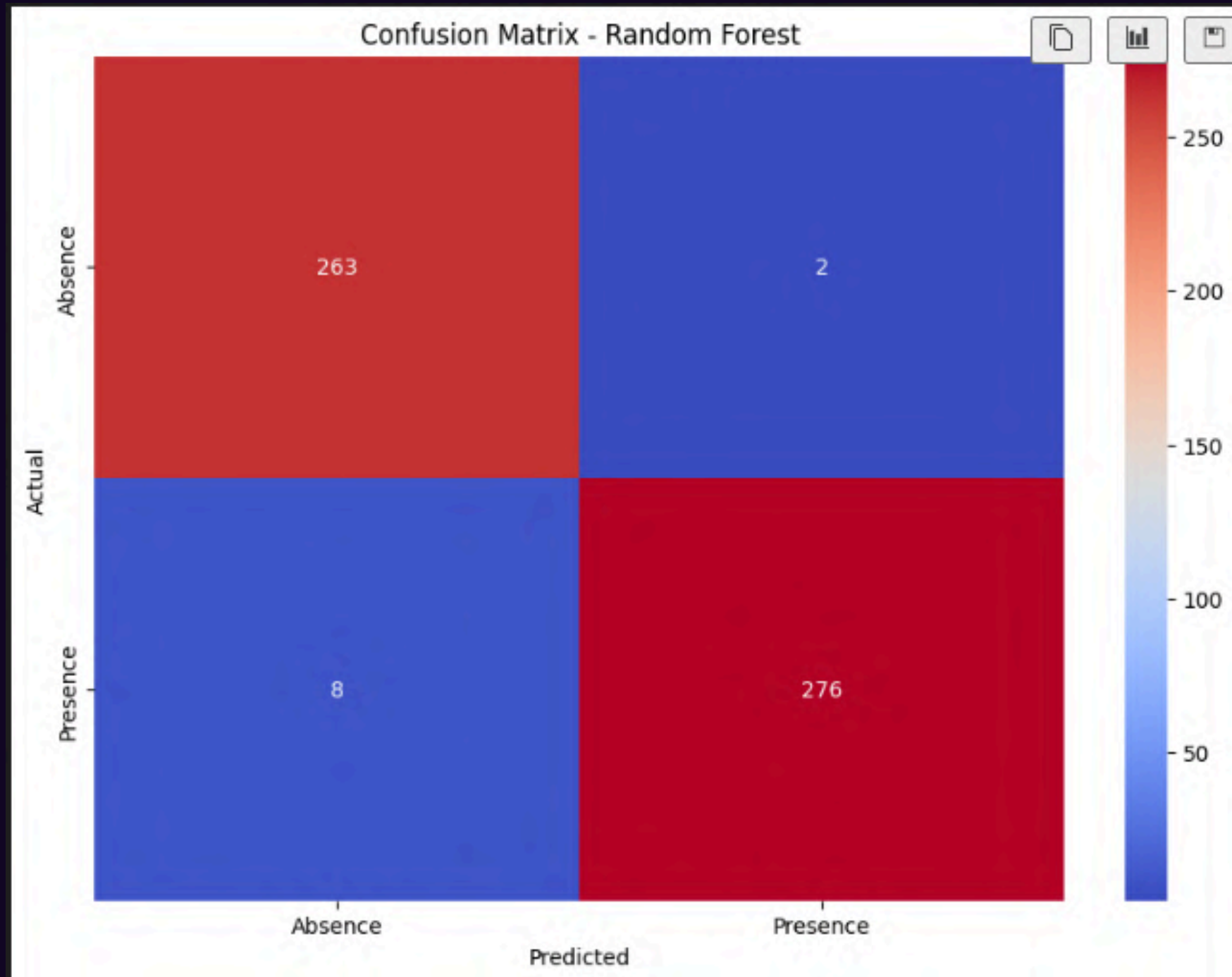
```
# Classification report XGB
xgb_report = classification_report(y_test, y_pred_xgb, output_dict=True)
xgb_df_report = pd.DataFrame(xgb_report).transpose()
print("XGBoost Classification Report:")
print(xgb_df_report)
```

[262] ✓ 0.0s

... XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.966543	0.981132	0.973783	265.000000
1	0.982143	0.968310	0.975177	284.000000
accuracy	0.974499	0.974499	0.974499	0.974499
macro avg	0.974343	0.974721	0.974480	549.000000
weighted avg	0.974613	0.974499	0.974504	549.000000

CONFUSION MATRIX



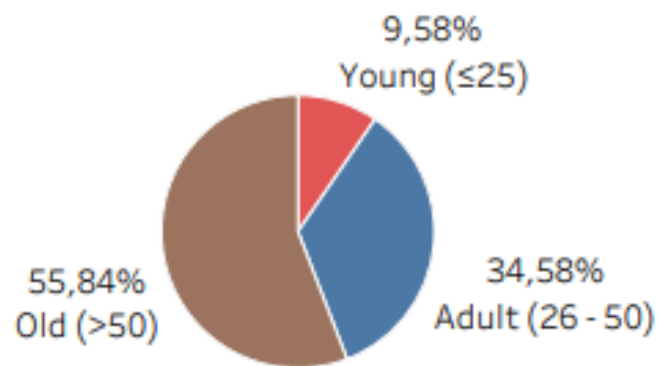
DATA VISUALIZATION

Chronic Kidney Disease Analysis Dashboard

Age Category

- Young (≤25)
- Adult (26 - 50)
- Old (>50)

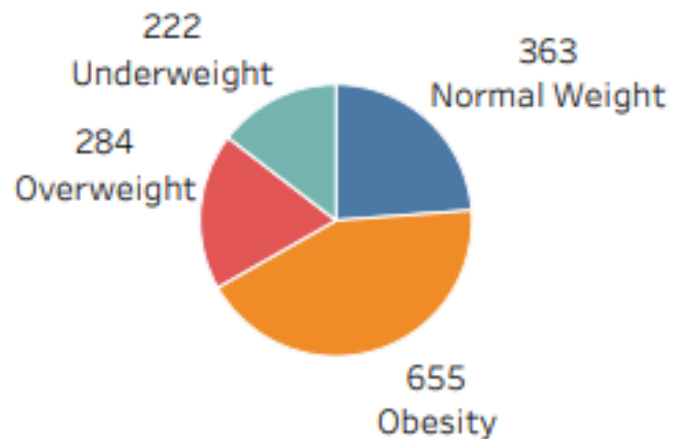
Age Distribution of Positive Chronic Kidney Disease Diagnoses



BMI Category

- Normal Weight
- Obesity
- Overweight
- Underweight

BMI Distribution of Positive Chronic Kidney Disease Diagnoses



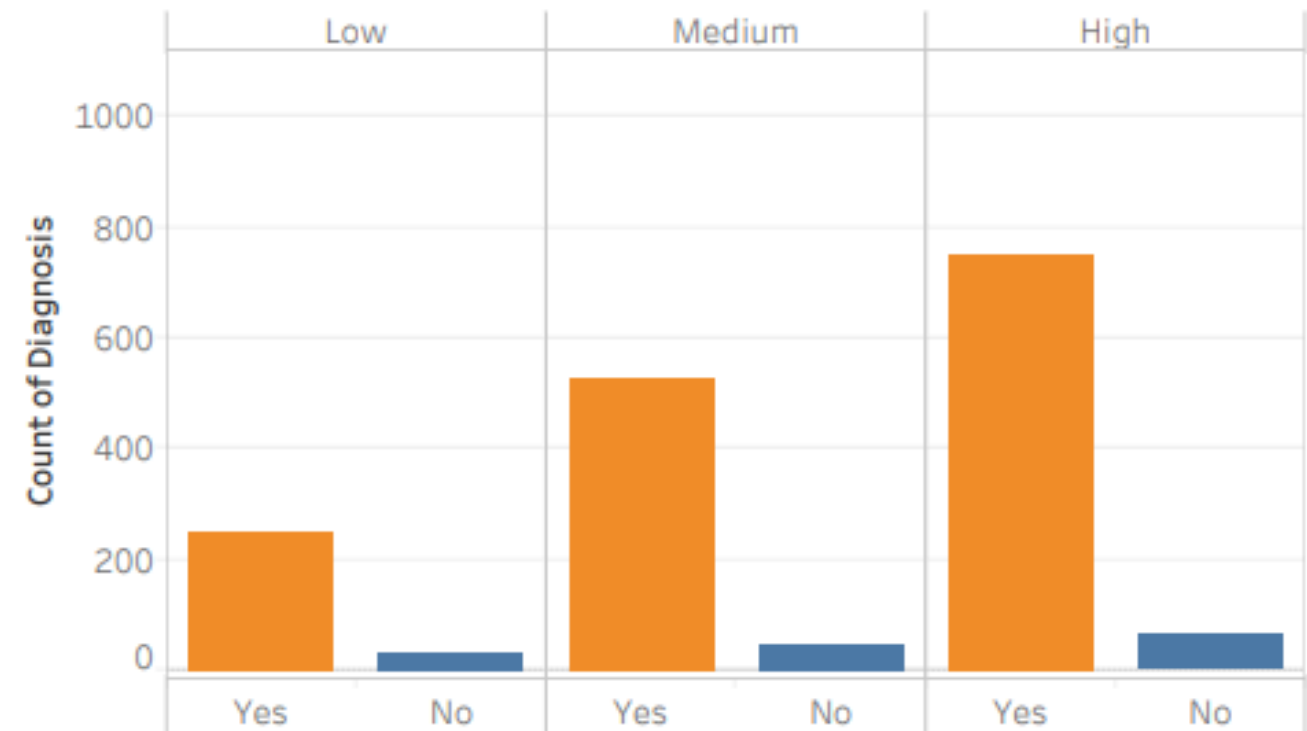
Diagnosis - Dimension

- Yes
- No

Choose Category

Sleep Quality

Chronic Kidney Disease Diagnosis Across Lifestyle Factors



Health Literacy Category

- Low
- Medium
- High

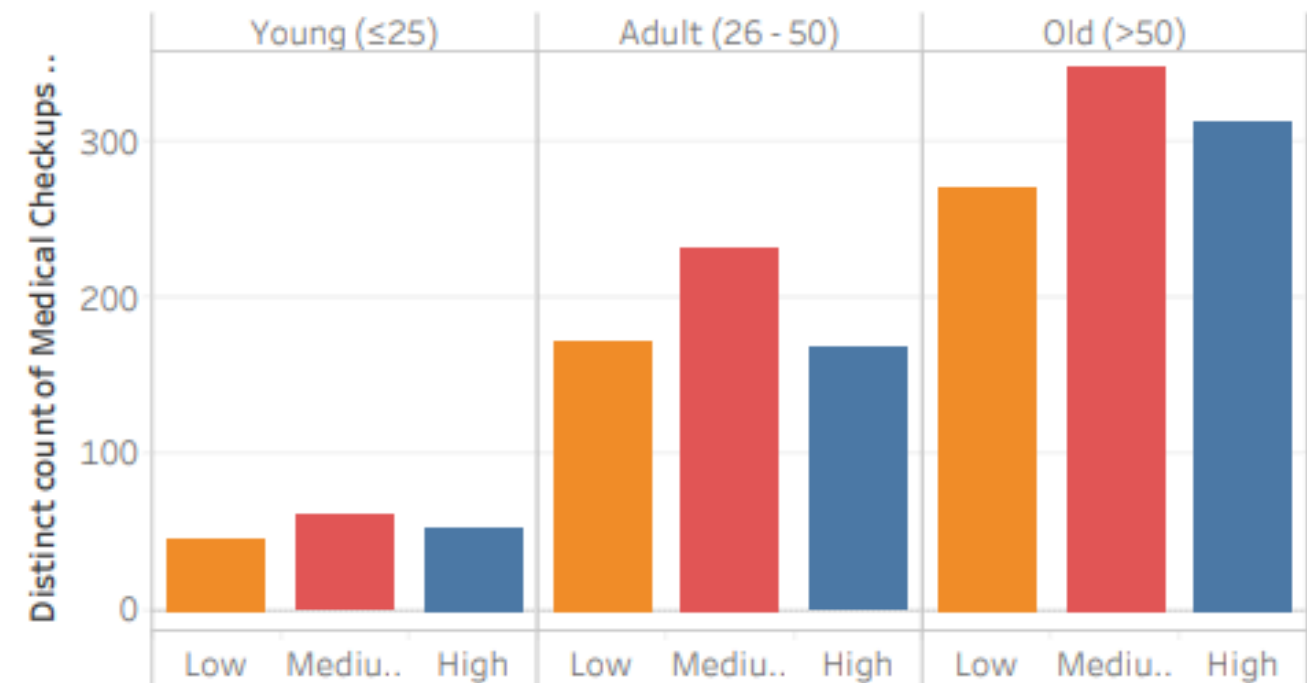
Gender

- Female
- Male

Diagnosis

- No
- Yes

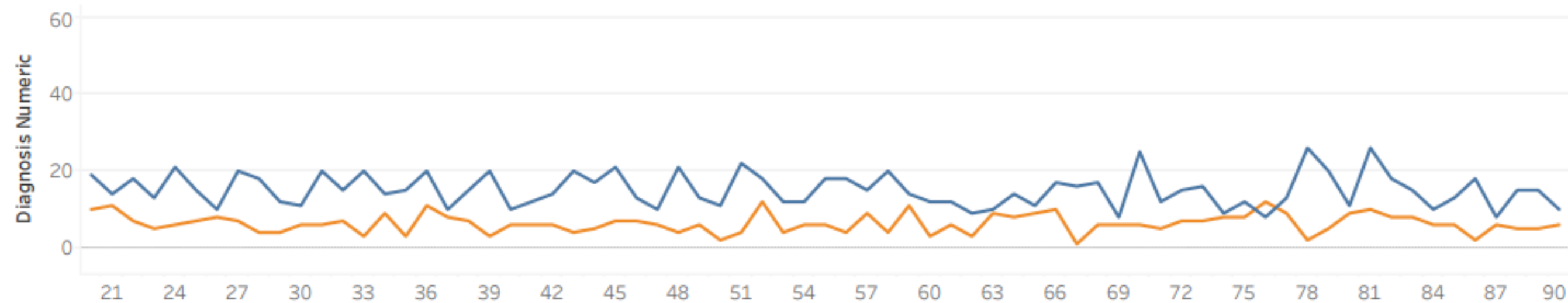
Frequency of Medical Checkups by Age, Gender, and Health Literacy



Smoking

- No
- Yes

Trend of Kidney Disease Diagnoses Across Age Groups: Smokers vs Non-Smokers



THANK YOU!