

DIABETES DATASET WITH CNN

About me


Saya Galih Hanggara. Siswa dari SMK Telkom Purwokerto. saat ini saya menduduki di bangku kelas 11 dengan jurusan

Pengembangan Perangkat Lunak dan Gim (PPLG)



Dataset

Dataset yang digunakan dalam penelitian ini adalah Pima Indians Diabetes Dataset, yang sering digunakan dalam penelitian Machine Learning untuk prediksi diabetes. Dataset ini berisi informasi medis dari pasien perempuan keturunan suku Pima Indian yang berusia 21 tahun ke atas. Tujuan utama dari dataset ini adalah untuk memprediksi kemungkinan seorang pasien menderita diabetes berdasarkan berbagai parameter medis yang telah dikumpulkan.



LOAD & PREPROCESSING DATASET

```
# Handle missing values (jika ada)
df = df.fillna(df.median())

# Split data into features and target
X = df.drop(columns=['Outcome'])
y = df['Outcome']

# Normalize data
scaler = StandardScaler()
X = scaler.fit_transform(X)

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Cek beberapa baris pertama dari dataset setelah preprocessing
print("Sample data after preprocessing:")
print(pd.DataFrame(X_train).head())

# Cek distribusi label target
print("\nLabel distribution in training set:")
print(pd.Series(y_train).value_counts())
```

```
Sample data after preprocessing:
   0      1      2      3      4      5      6  \
0 -0.547919 -1.154694 -3.572597 -1.288212 -0.692891 -4.060474 -0.507006
1  1.530847 -0.278373  0.666618  0.217261 -0.692891 -0.481351  2.446670
2 -0.844885  0.566649 -1.194501 -0.096379  0.027790 -0.417892  0.550035
3 -1.141852  1.255187 -0.987710 -1.288212 -0.692891 -1.280942 -0.658012
4  0.639947  0.410164  0.563223  1.032726  2.519781  1.803195 -0.706334

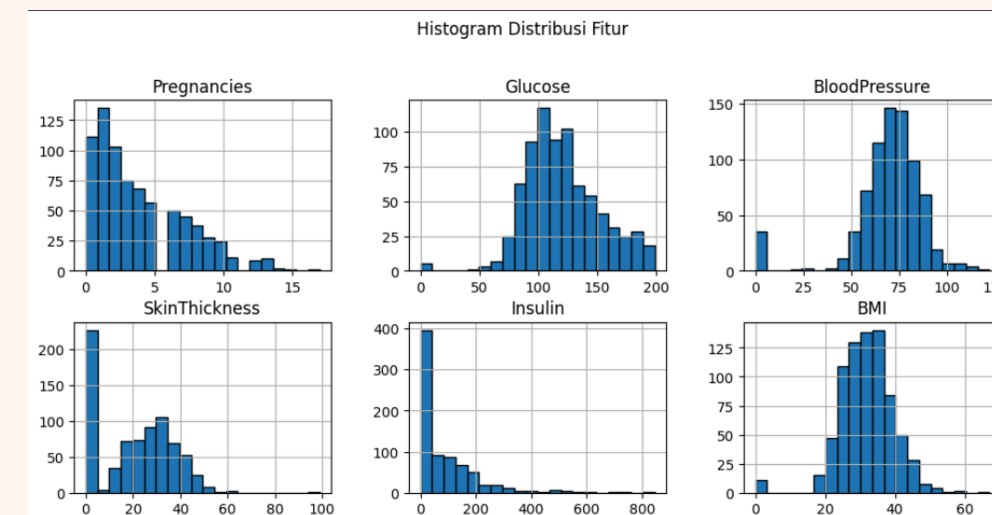
   7
0 -1.041549
1  1.425995
2 -0.956462
3  2.702312
4  1.085644

Label distribution in training set:
Outcome
0    401
1    213
Name: count, dtype: int64
```

Kode ini digunakan untuk membersihkan dan menyiapkan data sebelum digunakan dalam model pembelajaran mesin, khususnya Convolutional Neural Network (CNN).

VISUALISASI DATA 1

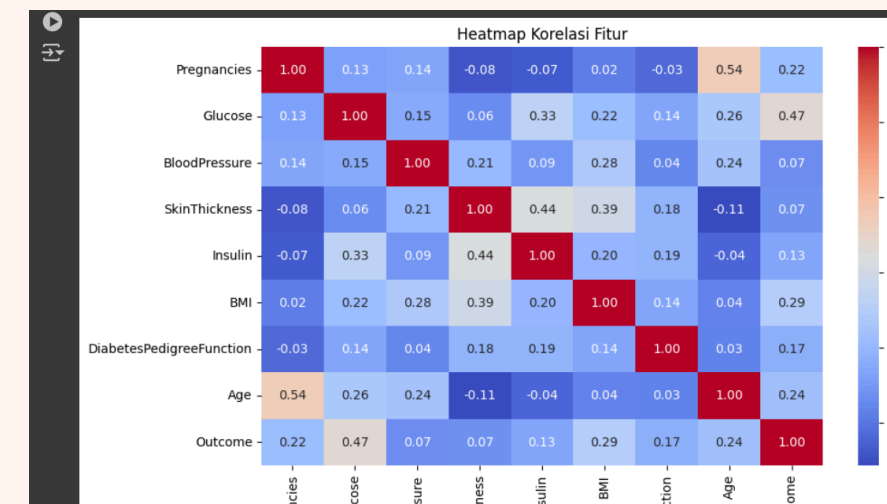
```
[17] Generated code may be subject to a license | 100rab-S/ga-learner-dsb-repo
# Visualisasi 1: Histogram distribusi setiap fitur
plt.figure(figsize=(12, 8))
df.hist(figsize=(12, 8), bins=20, edgecolor='black')
plt.suptitle('Histogram Distribusi Fitur')
plt.show()
```



Histogram ini menunjukkan pola distribusi data

VISUALISASI DATA 2

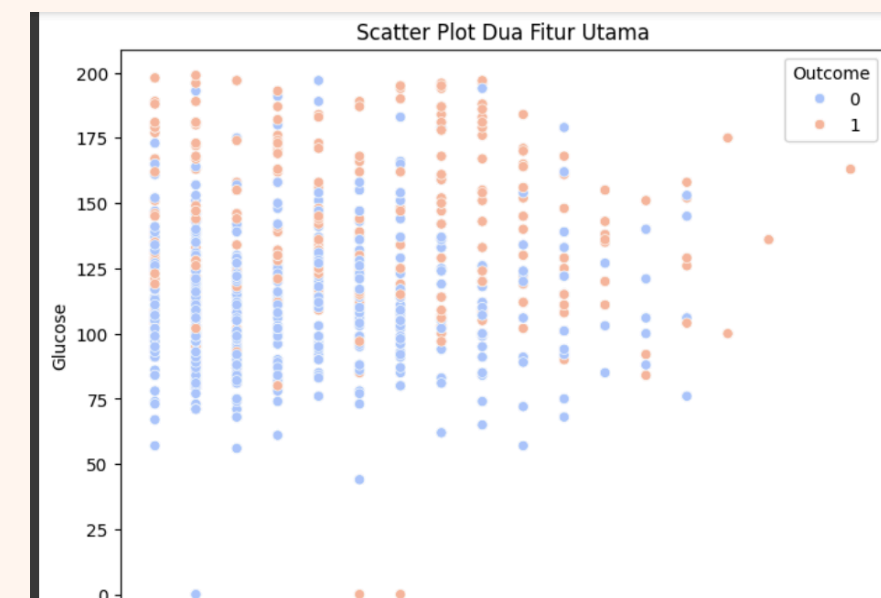
```
# Visualisasi 2: Heatmap korelasi antara fitur
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Heatmap Korelasi Fitur')
plt.show()
```



Menampilkan korelasi antar fitur dalam dataset. Warna menunjukkan hubungan positif atau negatif, membantu mengidentifikasi fitur yang berkaitan erat atau redundan dalam analisis data.

VISUALISASI DATA 3

```
[19] # Visualisasi 3: Scatter plot antara dua fitur utama
plt.figure(figsize=(8, 6))
sns.scatterplot(x=df.iloc[:, 0], y=df.iloc[:, 1], hue=y, palette='coolwarm')
plt.xlabel(df.columns[0])
plt.ylabel(df.columns[1])
plt.title('Scatter Plot Dua Fitur Utama')
plt.show()
```



Untuk melihat hubungan antara dua fitur utama dalam dataset. Warna (hue) menunjukkan perbedaan kategori dalam variabel target, membantu dalam analisis pola dan klasifikasi data.



THANK YOU