Diabetes Dataset

Link

امیر مهدی اعرابی | ۹۹۵۲۲۲۷۵

معرفی مجموعه داده

1. Pregnancies (بارداری‌ها)

- تعداد دفعات بارداری فرد.
- این ویژگی برای زنان در نظر گرفته شده و نشان‌دهنده تأثیر تعداد بارداری‌ها بر دیابت است.

2. Glucose (گلوکز)

- غلظت گلوکز پلاسما طی یک آزمون تحمل گلوکز خوراکی (OGTT).
- این ویژگی یکی از مهم‌ترین عوامل تعیین‌کننده دیابت است.

3. BloodPressure (فشار خون)

- فشار خون دیاستولیک (mm Hg).
- سطح فشار خون می‌تواند با دیابت و سایر مشکلات قلبی مرتبط باشد.

4. SkinThickness (ضخامت پوست)

- ضخامت چین پوستی (mm).
- معیاری برای تخمین سطح چربی بدن که ممکن است با دیابت مرتبط باشد.

5. Insulin (انسولین)

- سطح انسولین سرم (mu U/ml).
- انسولین نشان‌دهنده‌ی نحوه‌ی پاسخ بدن به گلوکز است و در تشخیص مقاومت به انسولین نقش دارد.

6. BMI (شاخص توده بدنی)

- شاخص وزن به قد (kg/m^2)، محاسبه شده به صورت وزن تقسیم بر قد به توان 2.
- BMI نشان‌دهنده وضعیت وزنی فرد (کم‌وزن، نرمال، اضافه‌وزن یا چاق) است و یکی از عوامل خطر دیابت است.

7. DiabetesPedigreeFunction (تابع شجره‌نامه دیابت)

- نشان‌دهنده احتمال ابتلا به دیابت بر اساس سابقه خانوادگی.

- این ویژگی تأثیر ژنتیکی و ارثی را بر احتمال ابتلا به دیابت ارزیابی می‌کند.

8. Age (سن)

- سن فرد به سال.
- دیابت نوع 2 معمولاً در افراد مسن‌تر شایع‌تر است، و این ویژگی به عنوان یکی از متغیرهای کلیدی در نظر گرفته شده است.

9. Outcome (نتیجه)

- متغیر هدف که نشان‌دهنده ابتلا به دیابت است:
  - مقدار **0**: فرد به دیابت مبتلا نیست.
  - مقدار **1**: فرد به دیابت مبتلا است.

مراحل تحلیل داده‌های اکتشافی و تمیز کردن داده‌ها

1. بارگذاری و پیش‌نمایش داده‌ها

ابتدا، داده‌ها از فایل CSV بارگذاری شده و ساختار کلی آن مشاهده شد:

- تعداد ویژگی‌ها و نوع داده‌ها بررسی شد.
- چند ردیف ابتدایی داده‌ها نمایش داده شد.

2. بررسی مقادیر گم‌شده

برای اطمینان از کیفیت داده‌ها:

- ستون‌هایی که دارای مقادیر گم‌شده بودند شناسایی شدند.
- هیچ مقدار گم‌شده صریح (NaN) در داده‌ها وجود نداشت، اما مقادیر صفر در ستون‌هایی نظیر **گلوکز**، **فشار خون**، **ضخامت پوست**، **انسولین**، و **BMI** به عنوان داده‌های غیرواقعی شناسایی شدند که در مراحل بعدی اصلاح شدند.

3. آمار توصیفی

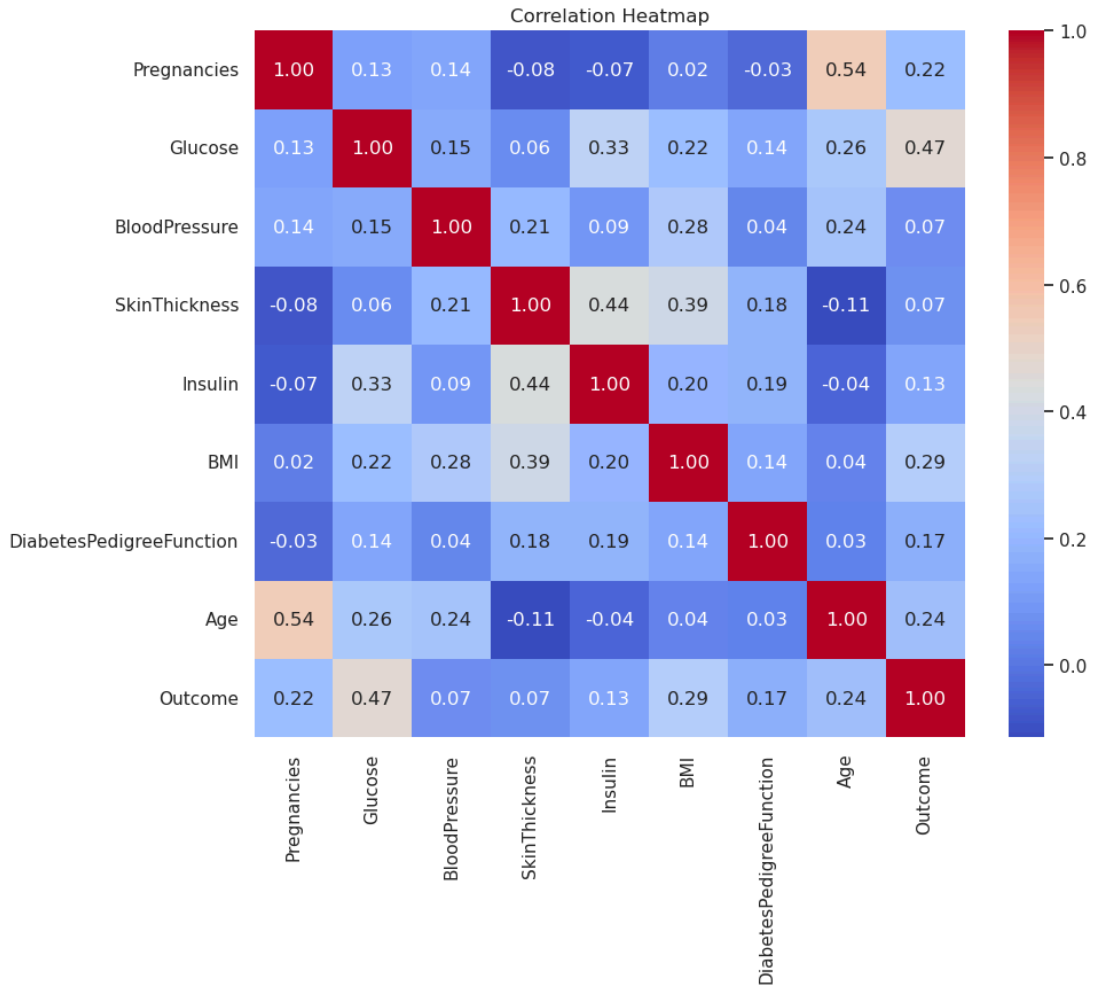
خلاصه‌ای از مقادیر آماری ویژگی‌های عددی:

- میانگین، انحراف معیار، کمینه و بیشینه مشاهده شدند.
- این آمار به شناسایی دامنه مقادیر و مشکلات احتمالی مانند وجود داده‌های پرت کمک کردند.

تجسم داده‌ها

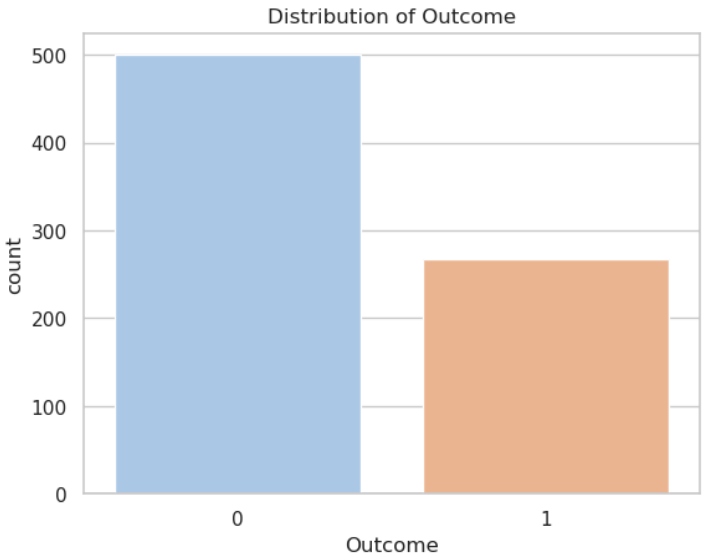
1. نمودار همبستگی (Correlation Heatmap)

- ماتریس همبستگی بین ویژگی‌ها ترسیم شد.
- همبستگی بالای مثبت و منفی بین ویژگی‌های مختلف شناسایی شد. این اطلاعات برای انتخاب ویژگی‌ها در مدلسازی مفید است.



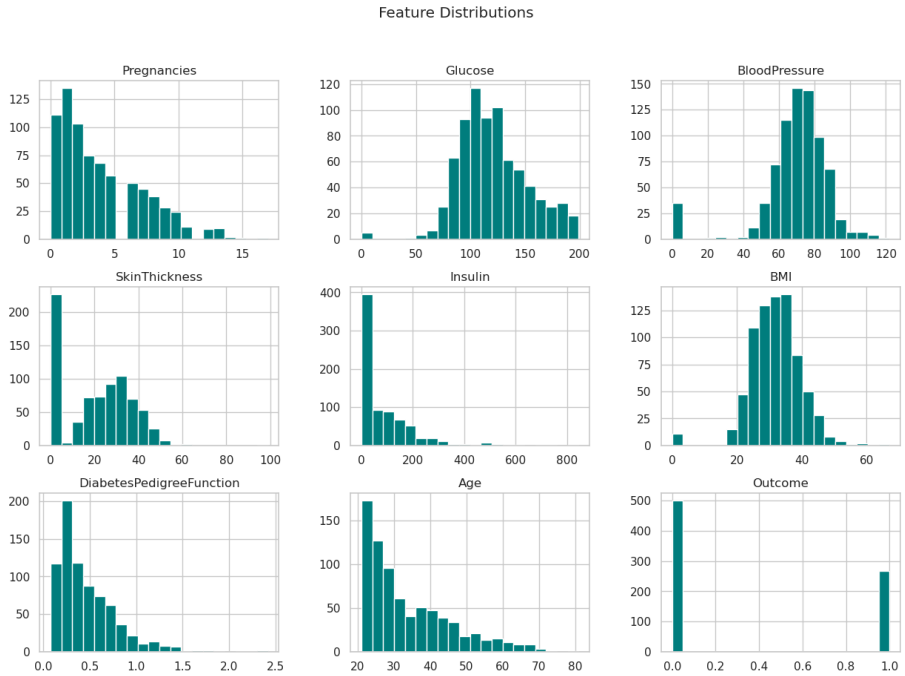
2. توزیع متغیر هدف (Outcome)

- نمودار میله‌ای برای توزیع متغیر هدف نشان داد که کلاس‌ها نسبتاً نامتعادل هستند، با تعداد بیشتری از نمونه‌های کلاس "0" نسبت به کلاس "1".



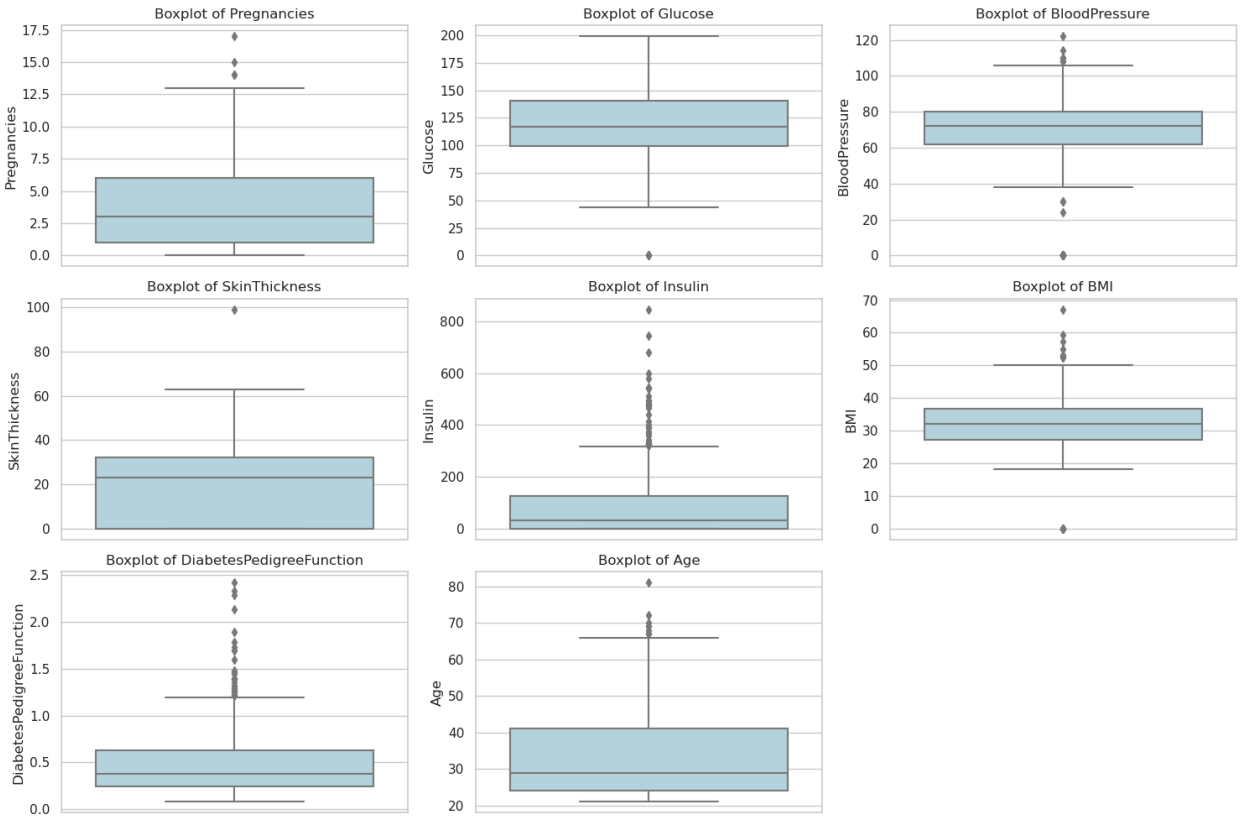
3. توزیع ویژگی‌ها

- هیستوگرام‌هایی برای تمام ویژگی‌های عددی رسم شد تا توزیع داده‌ها مشاهده شود.
- برخی ویژگی‌ها توزیع نامتقارن داشتند که به بررسی و اصلاح آن‌ها کمک کرد.



4. نمودار جعبه‌ای (Boxplot)

- نمودار جعبه‌ای برای شناسایی داده‌های پرت رسم شد.
- داده‌های پرت در ستون‌هایی نظیر **انسولین**، **ضخامت پوست** و **BMI** مشاهده شد.



تمیز کردن داده‌ها

1. جایگزینی مقادیر صفر

- مقادیر صفر در ستون‌های کلیدی با مقدار **NaN** جایگزین شدند، زیرا این مقادیر غیرواقعی بودند.

2. درمان داده‌های پرت

- با استفاده از روش فاصله بین چارکی (IQR)، داده‌های پرت شناسایی و به نزدیک‌ترین مقدار منطقی محدود (Capping) شدند.

مهندسی ویژگی‌ها

1. ایجاد دسته‌بندی BMI

- یک ستون جدید به نام **BMI_Category** بر اساس شاخص توده بدنی ایجاد شد:
  - کمتر از 18.5: **کم‌وزن**
  - بین 18.5 و 25: **وزن نرمال**
  - بین 25 و 30: **اضافه وزن**
  - بالای 30: **چاق**

2. ایجاد گروه‌های سنی

- ستون دیگری به نام **Age_Group** اضافه شد که افراد را به سه گروه تقسیم کرد:
  - زیر 30 سال: **جوان**
  - بین 30 و 50 سال: **میانسال**
  - بالای 50 سال: **سالمند**

3. ایجاد نسبت گلوکز به انسولین

- یک ویژگی جدید به نام **Glucose_Insulin_Ratio** ایجاد شد که نسبت گلوکز به انسولین را نشان می‌دهد.

4. ایجاد ویژگی تعامل BMI و سن

- یک ستون دیگر به نام **BMI_Age_Interaction** اضافه شد که تعامل بین BMI و سن را نشان می‌دهد.

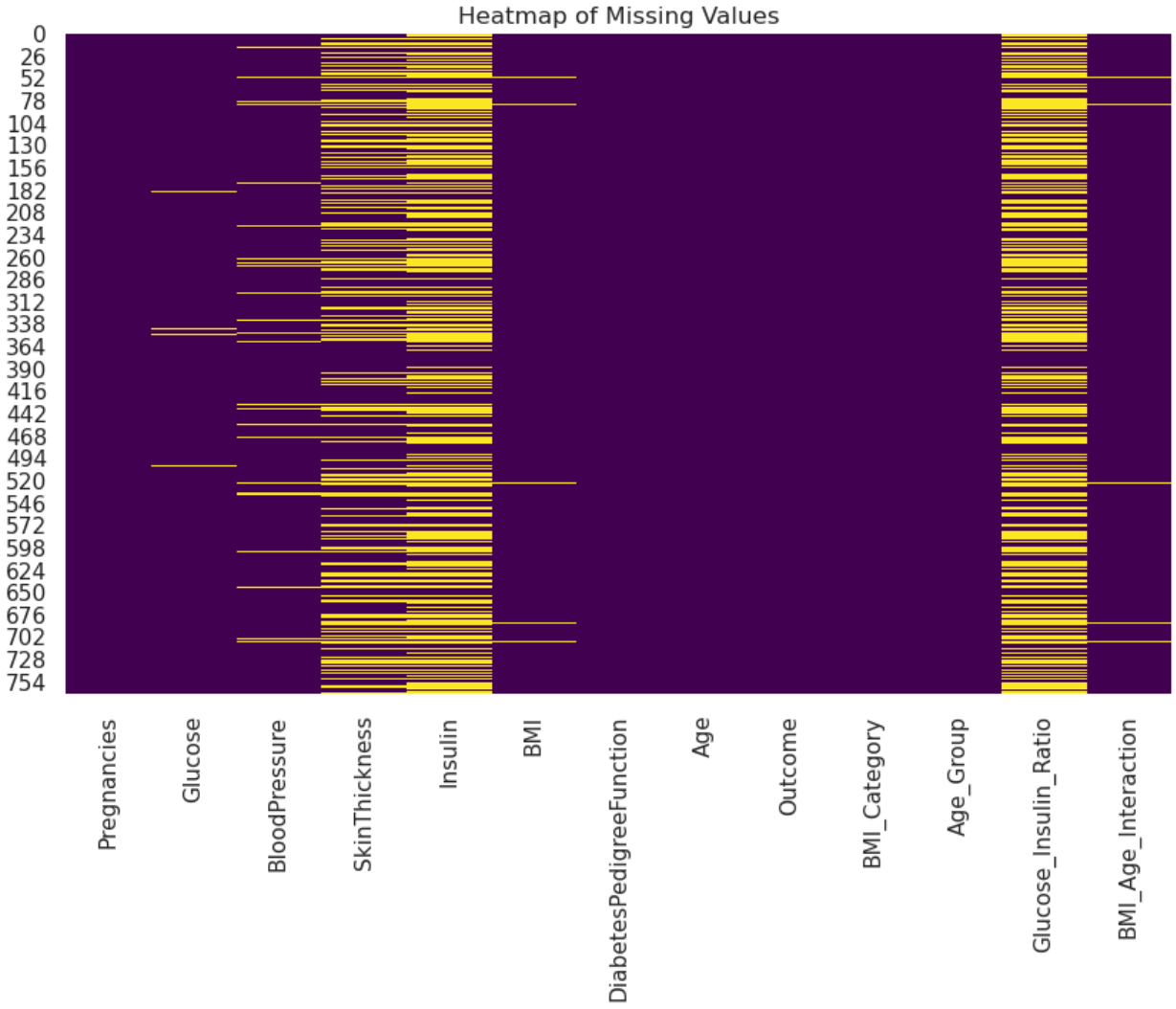
نرمال‌سازی داده‌ها

- برای مقیاس‌بندی ویژگی‌های عددی از **MinMaxScaler** استفاده شد. این کار باعث شد که مقادیر ویژگی‌ها در بازه‌ی [0, 1] قرار گیرند و مدلسازی با الگوریتم‌هایی که به مقیاس حساس هستند، بهبود یابد.

تجسم نهایی داده‌ها

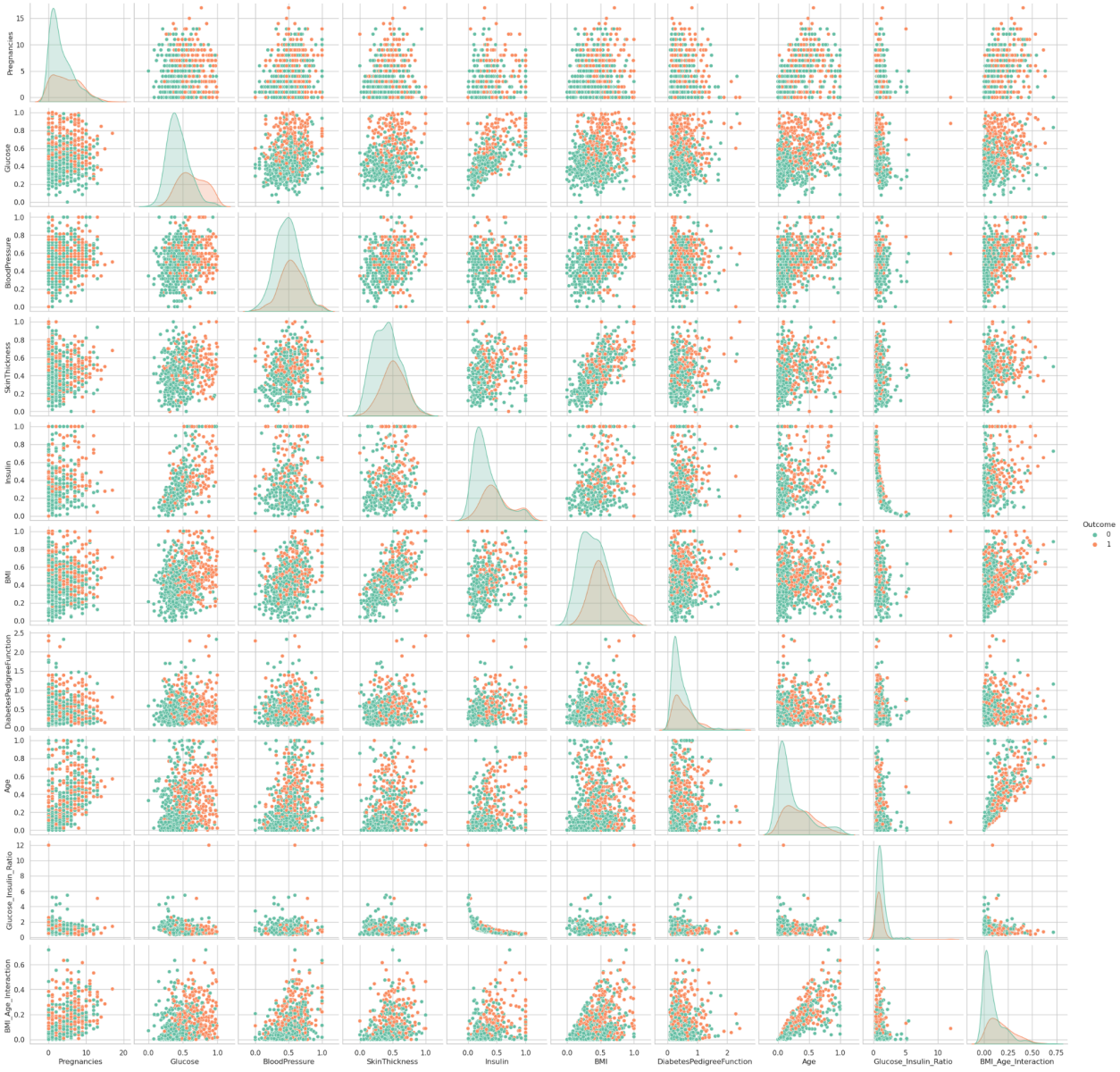
1. نمایش مقادیر گم‌شده

- نمودار Heatmap مقادیر گم‌شده برای بررسی کامل بودن داده‌ها رسم شد. این مرحله تأیید کرد که مقادیر صفر با مقادیر مناسب جایگزین شده‌اند.



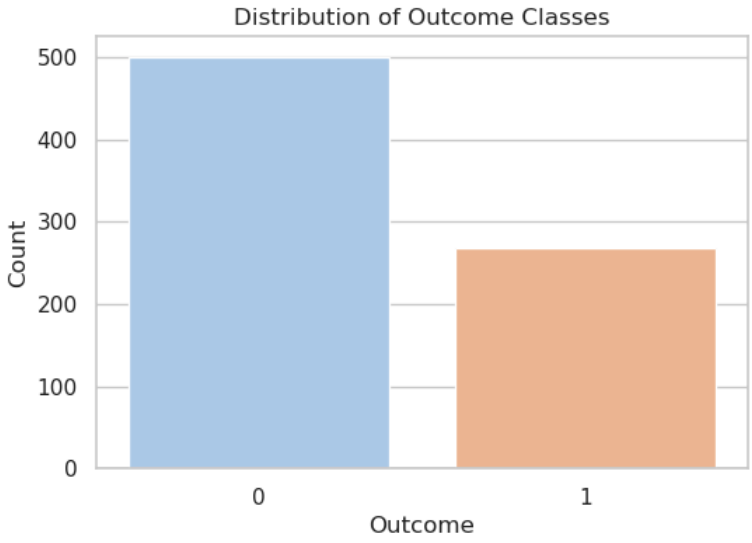
2. نمودار جفتی (Pairplot)

- روابط بین ویژگی‌ها با نمودار جفتی مشاهده شد. این نمودار نشان داد که برخی ویژگی‌ها مانند **گلوکز** و **BMI** تأثیر قابل‌توجهی بر متغیر هدف دارند.



3. ماتریس همبستگی نهایی

- ماتریس همبستگی نهایی ترسیم شد تا تأثیر تمیز کردن داده‌ها و ایجاد ویژگی‌های جدید بر روابط بین ویژگی‌ها مشاهده شود.



ذخیره داده‌های تمیزشده

در نهایت، مجموعه داده‌های تمیزشده به یک فایل CSV ذخیره شد تا در مراحل بعدی مدلسازی مورد استفاده قرار گیرد.
(cleaned_diabetes_data.csv)

نتیجه‌نهایی

- مقادیر گم‌شده و غیرواقعی برطرف شدند.
- داده‌های پرت مدیریت شدند.
- ویژگی‌های جدید با هدف بهبود دقت مدل اضافه شدند.
- داده‌ها برای مدلسازی آماده و ذخیره شدند.(cleaned_diabetes_data.csv)