

موسسه آموزش عالی خاوران

پروژه پایانی دوره کارشناسی

مهندسی نـرم افزار کامپیوتر

عنوان: پیش‌بینی دیابت به کمک ماشین یادگیری

استاد راهنما: مهندس هوبخت عطاران

تهیه کننده: مریـم سادات رمضان‌زاده

تابستان 1402

فهرست مطالب

[چکیده](#چکیده).......................................................................................................... 4

[فصل اول: تاریخچه](#فصل1).................................................................................... 5

[فصل دوم: ضرورت مسئله](#فصل2)........................................................................ 9

[فصل سوم: حل مسئله](#فصل3)............................................................................ 14

[فصل چهارم: نتیجه گیری](#فصل4)..................................................................... 40

[منابع و مراجع](#منابع).......................................................................................... 41

**چکیده**

در این پروژه، ما به تحلیل یک مجموعه داده می‌پردازیم که ویژگی‌های مختلفی را شامل می‌شود. این ویژگی‌ها عبارتند از تعداد بارداری، سطح گلوکز، فشار خون، ضخامت پوست، مقدار انسولین، شاخص توده بدنی، DPF و سن. نکته مهم این مجموعه داده این است که پیش‌پردازش‌های مورد نیاز از قبل انجام شده است و داده‌ها در حالت تمیز قرار دارند.

در فصل اول این پروژه، ما به تحلیل این مجموعه داده می‌پردازیم. در این مرحله، اطلاعاتی را درباره توزیع و ویژگی‌های مختلف ویژگی‌ها استخراج می‌کنیم تا به درک عمیق‌تری از داده‌ها برسیم.

در فصل دوم، به مدل‌سازی می‌پردازیم. در این بخش، از مدل‌های مختلفی برای پیش‌بینی یک متغیر یا نتایج مورد نظر استفاده می‌کنیم. این مدل‌ها از الگوریتم‌های یادگیری ماشینی متنوعی نظیر درخت یادگیری، رگرسیون لجستیک، KNN، SVM هستند. هدف ما در این مرحله، ایجاد یک مدل دقیق برای پیش‌بینی و تشخیص بیماری دیابت است.

هدف:

هدف مجموعه داده دیابت ارائه اطلاعات و بینش در مورد دیابت و مدیریت آن است. این مجموعه داده به محققان کمک می‌کند تا بیماری را درک کنند، الگوها و عوامل خطر را شناسایی کنند، مدل‌های پیش‌بینی را توسعه دهند و مراقبت از دیابت را بهبود بخشند. از این مجموعه داده برای مطالعه اثربخشی درمان‌ها، پیش‌بینی پیشرفت دیابت، پشتیبانی از تصمیم‌گیری‌های بالینی و برنامه‌ریزی استراتژی‌های بهداشت عمومی استفاده می‌شود.

**فصل 1: تاریخچه**

* + یادگیری ماشین چیست؟

یادگیری ماشین یک حوزه مهم در علوم کامپیوتر و هوش مصنوعی (AI) است که به ماشین‌ها و سیستم‌های کامپیوتری قابلیت یادگیری از داده‌ها و تجربیات بدون برنامه‌ریزی صریح انسانی را می‌دهد. در واقع، در یادگیری ماشین، سیستم‌های کامپیوتری به کمک الگوریتم‌ها و مدل‌های ریاضی توانایی تشخیص الگوها، انجام پیش‌بینی‌ها، استخراج اطلاعات و اتخاذ تصمیم ‌ها را از داده ‌های ورودی خود به دست می‌آورند.

* + یادگیری ماشین از زمانی شروع شد؟

یادگیری ماشین به صورت رسمی از دهه 1950 و 1960 میلادی شروع شد اما پیش از آن نیز تحقیقات مرتبط با آن در حوزه هوش مصنوعی وجود داشت و از آن زمان به سرعت رشد کرده و توسعه یافته است. این زمینه به ویژه در دهه ‌های اخیر با پیشرفت فناوری، موج‌ های جدیدی از یادگیری ماشین مانند یادگیری عمیق (Deep Learning) ظاهر شده است که توانایی پیچیده‌تری در تشخیص الگوها و انجام کار های هوش مصنوعی دارد.

* + یادگیری ماشین چه کاربرد هایی دارد؟

یادگیری ماشین به کسب و کارها و جوامع در افزایش بهره‌وری، تشخیص الگوهای پیچیده و تصمیم‌گیری مبتنی بر داده کمک می‌کند و به عنوان ابزاری قدرتمند در ادامه تکامل علوم کامپیوتر و هوش مصنوعی مورد استفاده قرار می‌گیرد.

کاربردهای یادگیری ماشین بسیار گسترده است و در انواع حوزه‌ها مورد استفاده قرار می‌گیرد. برخی از کاربردهای رایج یادگیری ماشین عبارتند از:

1. پردازش تصویر: در تشخیص و تحلیل تصاویر و ویدیوها، تشخیص چهره، خودروهای خودران و بسیاری از کاربردهای دیگر.
2. زبان‌شناسی محاسباتی: در ترجمه ماشینی، تولید متن خودکار، تحلیل احساسات متن و سیستم‌های گفتاری.
3. پزشکی: در تشخیص بیماری‌ها، پیش‌بینی نتایج بالینی، تحلیل تصاویر پزشکی و توصیه سیستم‌های درمانی.
4. مالی و بازارهای مالی: از یادگیری ماشین برای پیش‌بینی قیمت‌های سهام، تجزیه و تحلیل بازارهای مالی و تشخیص فعالیت‌های مالی ناهنجار استفاده می‌شود.
5. صنعت و تولید: در کنترل کیفیت محصولات، بهینه‌سازی زنجیره تأمین و مانیتورینگ فرآیندهای تولید.
6. خدمات مشتریان و تجارت الکترونیک: تجزیه و تحلیل رفتار مشتریان و سیستم‌های پشتیبانی مشتریان.
7. خودرو های خودران: در تکنولوژی ‌های خودرو های خودران برای تشخیص محیط و کنترل خودرو و همینطور برای بهبود ایمنی استفاده می‌شود.
8. توصیه‌گرها: در سیستم‌ های توصیه محتوا و محصولات به کاربران.
9. متن‌کاوی و پردازش زبان طبیعی: از طریق یادگیری ماشین می‌توان متون را تجزیه و تحلیل کرده و مسائلی مانند ترجمه ماشینی، تشخیص متن و سیستم‌های گفتار به متن را پشتیبانی کرد.
10. بازی‌ها و سرگرمی: در بازی‌ها و سرگرمی‌های رایانه‌ای، یادگیری ماشین برای تجسم رفتار شخصیت‌ها و بهبود تجربیات بازیکنان استفاده می‌شود.
    * الگوریتم ها و مدل های یادگیری ماشین

الگوریتم ‌ها و مدل‌ های یادگیری ماشین انواع مختلفی از روش‌ها و الگوریتم‌ ها هستند که به ماشین‌ها و سیستم‌های کامپیوتری اجازه می‌دهند از داده ‌ها یاد بگیرند و الگوها و اطلاعات مفهومی را استخراج کنند. این الگوریتم ‌ها و مدل‌ ها در مسائل پیش‌بینی، دسته‌بندی، خوشه‌بندی، تصمیم‌گیری و تطبیق داده‌ های بازیابی و یادگیری از تجربه (رینفورسمنت لرنینگ) مورد استفاده قرار می‌گیرند. در ادامه به توضیح مختصری از برخی از مهمترین الگوریتم‌ها و مدل‌های یادگیری ماشین پرداخته می‌شود.

1. درخت تصمیم گیری (Decision Tree):

درخت تصمیم یک مدل یادگیری ماشین است که برای تصمیم‌گیری و دسته‌بندی داده‌ها از ساختار درختی استفاده می‌کند. درخت شامل گره‌ها (نقاط تصمیم) و لیستی از سوالات است که با پاسخ دادن به آن‌ها به تصمیم‌گیری نهایی می‌رسیم.

- کاربردها: تصمیم‌گیری در زمینه‌های مختلف از پزشکی تا تجارت و بهبود سیستم‌ های

تصمیم گیری.

1. ماشین ‌های پشتیبانی (Support Vector Machines):

SVM یک مدل دسته ‌بندی است که داده‌ ها را در فضای چند بعدی می‌نگارد و سعی دارد یک صفحه یا فضای تصمیم را برای جدا کردن داده‌های مختلف ایجاد کند. هدف این مدل افزایش فاصله بین دسته‌ ها است.

- کاربردها: دسته‌بندی تصاویر، تشخیص چهره، پیش‌بینی انواع بیماری‌ها.

1. رگرسیون خطی (Linear Regression):

رگرسیون خطی برای مدل کردن روابط بین یک متغیر وابسته و یک یا چند متغیر مستقل به صورت خطی استفاده می‌شود. مدل، معادله خطی را برای پیش‌بینی مقدار وابسته تولید می‌کند.

- کاربردها: پیش‌بینی قیمت‌ها، مدلسازی اقتصادی، تجزیه و تحلیل داده‌های عددی.

1. شبکه‌ های عصبی مصنوعی (Artificial Neural Networks):

ANN مدل‌های متشکل از لایه‌های نورونی مصنوعی هستند که به صورت مشابه به ساختار مغز انسان عمل می‌کنند. این مدل‌ها برای مسائل پیچیده و تشخیص الگو های پیچیده مناسب هستند.

- کاربردها: تصویربرداری مغزی، ترجمه ماشینی، تشخیص گفتار و تبدیل گفتار به متن.

1. رگرسیون لجستیک (Logistic Regression):

یک مدل یادگیری ماشین است که برای مدلسازی روابط بین یک متغیر وابسته دودویی (با دو کلاس) و یک یا چند متغیر مستقل به صورت خطی استفاده می‌شود. این مدل معادله لجستیک را برای پیش‌بینی احتمال تعلق به کلاس مثبت (مثلاً بودن دیابت) تولید می‌کند.

رگرسیون لجستیک می‌تواند احتمال تعلق به یک کلاس خاص را محاسبه کرده و اطلاعاتی در مورد تأثیر متغیرهای مستقل بر روی این احتمال ارائه دهد. در واقع، این مدل به ما اجازه می‌دهد بفهمیم که هر یک از متغیرهای مستقل چقدر می‌توانند بر تصمیم نهایی در مورد وقوع یا عدم وقوع یک رویداد تأثیر بگذارند. این مدل به دلیل توانایی در پردازش داده‌های دودویی و تعاملی، سادگی مدل‌سازی و تفسیری که ارائه می‌دهد، بسیار مورد توجه قرار گرفته است.

-کاربرد: تصمیم‌گیری در زمینه‌های پزشکی، تحلیل مخاطرات مالی، مدل‌سازی اقتصادی.

1. نزدیک ترین همسایه (K-Nearest Neighbors):

KNN یک الگوریتم یادگیری ماشین برای دسته ‌بندی و پیش ‌بینی است. این مدل بر اساس نزدیک ‌ترین نقاط در مجموعه داده عمل می‌کند و می‌تواند بدون پارامتر اجرا شود. ابتدا تعداد همسایه‌ها (K) انتخاب می‌شود. سپس با محاسبه فاصله بین نقاط، همسایه‌های نزدیک‌تر به نقطه مورد نظر انتخاب می‌شوند. نقطه مورد نظر به کلاسی تعلق می‌گیرد که در اکثریت همسایه‌ها قرار دارد.

-کاربردها: مراقبت‌های پزشکی، تصمیم‌گیری تجاری، تصمیم‌گیری مالی و زمینه‌های دیگر.

1. نوع‌های دیگر: علاوه بر مدل‌های فوق، هنگام مواجهه با مسائل خاص می‌توان از مدل‌هایی مانند Random Forest، Naive Bayes و ... استفاده کرد.

هر یک از این مدل‌ها و الگوریتم‌ها دارای قابلیت‌ها و محدودیت‌های خود هستند و انتخاب مدل مناسب بستگی به نوع داده و مسئله مورد نظر دارد. ترکیب مدل‌های مختلف یادگیری ماشین نیز در بسیاری از مسائل مفید است.

**فصل 2: ضرورت مسئله**

* شرح مختصری درباره بیماری دیابت

دیابت یک بیماری مزمن است که نحوه فرآیند پردازش قند در بدن را تحت تأثیر قرار می‌دهد. قند یک نوع شکر است که به عنوان منبع اصلی انرژی عمده عمل می‌کند. وقتی ما غذا می‌خوریم، سیستم گوارش ما غذا به ویژه کربوهیدرات‌ها را به گلوکز تبدیل می‌کند. در پاسخ به این گلوکز خونی، پانکراس هورمونی به نام انسولین ترشح می‌کند که به انتقال گلوکز از خون به سلول‌های بدن کمک می‌کند تا از آن به عنوان منبع انرژی استفاده کنند.

اما در افراد مبتلا به دیابت، تولید یا استفاده از انسولین به نحوی مختل شده است که باعث افزایش میزان گلوکز خون می‌شود. این اتفاق به یکی از دلایل زیر رخ می‌دهد:

1. دیابت نوع 1: این نوع دیابت یک بیماری خود ایمنی است که سیستم ایمنی به اشتباه به سلول‌های تولید کننده انسولین در پانکراس حمله کرده و آنها را نابود می‌کند. به عبارت دیگر، بدن به میزان کم یا تقریباً هیچ انسولینی تولید نمی‌کند. دیابت نوع 1 معمولاً در دوران کودکی یا جوانی تشکیل می‌شود و نیاز به درمان مادام‌العمر با انسولین دارد.
2. دیابت نوع 2: این نوع دیابت رایج ‌ترین نوع این بیماری است و بیشترین آمار دیابت را تشکیل می‌دهد. در دیابت نوع 2، بدن به اثرات انسولین مقاوم می‌شود و پانکراس ممکن است مقدار کافی انسولین تولید نکند تا جبران این مقاومت شود. این اغلب در افرادی اتفاق می‌افتد که اضافه وزن دارند، فعالیت بدنی کافی انجام نمی‌دهند، یا تاریخچه خانوادگی دیابت دارند. دیابت نوع 2 می‌تواند از طریق تغییرات در سبک زندگی مانند رژیم غذایی سالم، ورزش منظم، داروهای خوراکی و در برخی موارد تزریقات انسولین مدیریت شود.
3. دیابت بارداری: این نوع دیابت زنان باردار را تحت تأثیر قرار می‌دهد که سطح قند خونی بالا دارند. معمولاً پس از زایمان از بین می‌رود، اما زنانی که دیابت بارداری داشته‌اند در آینده با خطر ابتلا به دیابت نوع 2 روبه‌رو هستند.

* مدیریت دیابت

دیابت بدون کنترل می‌تواند به انواع عوارضی منجر شود که بخش‌ های مختلف بدن را تحت تأثیر قرار می‌دهد که شامل قلب، عروق خونی، چشم ‌ها، کلیه‌ ها و اعصاب می‌باشد. عوارض متداول این بیماری، شامل بیماری ‌های قلبی عروقی، دیابتی رتینوپاتی (مشکلات دیدی)، دیابتی نوروپاتی (آسیب اعصاب)، بیماری کلیه و بهبود ناپذیری زخم‌ها هستند.

در نتیجه تشخیص به موقع دیابت اهمیت زیادی دارد زیرا:

1. مدیریت به موقع: تشخیص دیابت به افراد امکان می‌دهد تا به مدیریت مناسب این بیماری بپردازند، از جمله کنترل میزان قند خون و استفاده از انسولین یا داروهای دیگر.
2. پیشگیری از عوارض: اگر دیابت به موقع تشخیص داده نشود و مدیریت نشود، می‌تواند به عوارض جدی از جمله مشکلات قلبی، کلیوی، نوروپاتی (آسیب به عصب‌ها) و اختلالات دید منجر شود.
3. بهبود کیفیت زندگی: مدیریت صحیح دیابت می‌تواند بهبود کیفیت زندگی فرد مبتلا را ایجاد کند و از وقوع مشکلات جدی جلوگیری کند.
4. پیش ‌بینی خطرات: تشخیص دیابت به افراد کمک می‌کند تا خطر ابتلا به عوارض جانبی آن را درک کنند و اقدامات پیشگیرانه را انجام دهند.

* ماشین یادگیری (Machine Learning) و کمک به تشخیص دیابت:

ماشین یادگیری به تشخیص دیابت در افراد از طریق تحلیل و پردازش دقیق داده‌ های مرتبط با آن بیماری کمک می‌کند. ماشین یادگیری بهبود مدیریت و تشخیص دقیق‌تر دیابت را فراهم می‌کند و برای متخصصان سلامت در تشخیص و پیشگیری از این بیماری ارزشمند می‌باشد.

در زیر توضیح داده شده است که چگونه ماشین یادگیری می‌تواند در تشخیص دیابت مفید باشد:

* 1. تجزیه و تحلیل دقیق داده‌ها: ماشین یادگیری قادر به تجزیه و تحلیل دقیق داده‌ های مربوط به دیابت است. این داده‌ها ممکن است شامل اطلاعاتی مانند سطوح گلوکز خون، شاخص توده بدنی، سن، فشار خون، تاریخچه خانوادگی، ویژگی‌های فیزیولوژیکی و غیره باشند. ماشین یادگیری به صورت خودکار و بدون تعامل انسانی می‌تواند این داده ‌ها را تجزیه و تحلیل کرده و الگوهایی را در آنها تشخیص دهد.
  2. پیدا کردن الگوهای پنهان: ماشین یادگیری قادر به شناسایی الگوهای پنهان و روابط پیچیده در داده‌ ها است. این الگو ها ممکن است نشان ‌دهنده عوامل خطرزای دیابت یا نشانگر های زودرس دیابت باشند که توسط انسان به راحتی تشخیص داده نمی‌شوند.
  3. پیش‌بینی خطر دیابت: با استفاده از مدل ‌های یادگیری ماشین، می‌توان دقت خطر دیابت در افراد را پیش ‌بینی کرد. این به افراد اجازه می‌دهد تا اقدامات پیشگیرانه انجام دهند و از وقوع دیابت جلوگیری کنند. به عنوان مثال، مدل ممکن است افرادی را که در خطر دیابت هستند شناسایی کند و به آنها توصیه کند که سبک زندگی سالم ‌تری داشته باشند یا آزمایش‌های دوره ‌ای برای نظارت بر سلامتی خود انجام دهند.
  4. تشخیص زودرس دیابت: ماشین یادگیری می‌تواند به تشخیص زودرس دیابت کمک کند. این به معنای شناسایی افرادی است که علائم دیابت را نشان می‌دهند اما هنوز به عنوان دیابتی تشخیص داده نشده ‌اند. با تشخیص زودرس، درمان به موقع آغاز می‌شود و عوارض جانبی دیابت کاهش می‌یابد.
  5. بهبود دقت تشخیص: الگوریتم‌ های یادگیری ماشین ممکن است دقت تشخیص دیابت را نسبت به روش‌ های سنتی بهبود بخشند. این به معنای داشتن مدل ‌های پیش ‌بینی دقیق ‌تری است که توانایی تشخیص دقیق ‌تری را دارند.

در اینجا توضیح مفصلی درباره مجموعه داده ارائه شده:

* معرفی مجموعه داده:

مجوعه داده‌ای که در این پروژه استفاده کرده‌ایم با نام diabetes.csv یک فایل اکسل است که 768 نمونه را شامل میشود. این دیتا همانطور که در قبل گفته شد ویژگی هایی نظیر تعداد بارداری، سطح گلوکز، فشار خون، ضخامت پوست، مقدار انسولین، شاخص توده بدنی، DPF و سن دارد. لازم به ذکر است که این مجموعه داده هیچ مقادیر گم ‌شده‌ای ندارد، بنابراین تکمیل مجموعه داده با استفاده از الگوریتم ‌ها لازم نخواهد بود. همینطور تمام ویژگی ها از جنس عددی هستند.

این مجموعه داده در اصل از موسسه ملی دیابت و بیماری های گوارشی و کلیوی تهیه شده است. هدف مجموعه داده این است که بر اساس اندازه‌گیری ‌های تشخیصی موجود در مجموعه داده، پیش‌بینی کند که آیا بیمار مبتلا به دیابت است یا خیر. چندین محدودیت برای انتخاب این نمونه ها از یک پایگاه داده بزرگتر قرار داده شده است. به طور خاص، همه بیماران در اینجا زنان حداقل 21 ساله از میراث هندی پیما هستند.

* ویژگی های دیتا:

Pregnancies (تعداد بارداری): تعداد دفعاتی که شخص باردار شده است را توصیف می‌کند.

Glucose (گلوکز): سطح قند خون در زمان آزمایش را توصیف می‌کند.

Blood Pressure (فشار خون): فشار خون دیاستولیک را توصیف می‌کند.

Skin Thickness (ضخامت پوست): ضخامت چربی پوست در منطقه تریسپس را توصیف می‌کند.

Insulin(انسولین): مقدار اینسولین در یک آزمایش سرمی 2 ساعته را توصیف می‌کند.

BMI (شاخص توده بدنی): معیار اندازه گیری تناسب قد و وزن افراد را توصیف می‌کند.

Diabetes Pedigree Function (عملکرد خانوادگی دیابت): تاریخچه خانوادگی شخص در خصوص دیابت را توصیف می‌کند.

Age (سن): سن شخص را توصیف می‌کند.

Outcome (نتیجه): توصیف می‌کند که آیا شخص دیابت دارد یا نه.

**فصل 3: حل مسئله**

این فصل به ما کمک می‌کند تا داده‌ های خود را بهتر بفهمیم و الگوها و ویژگی‌های مهم را در داده‌ها شناسایی کنیم. در واقعیت، تجزیه و تحلیل داده در ماشین لرنینگ بسیار مهم است.

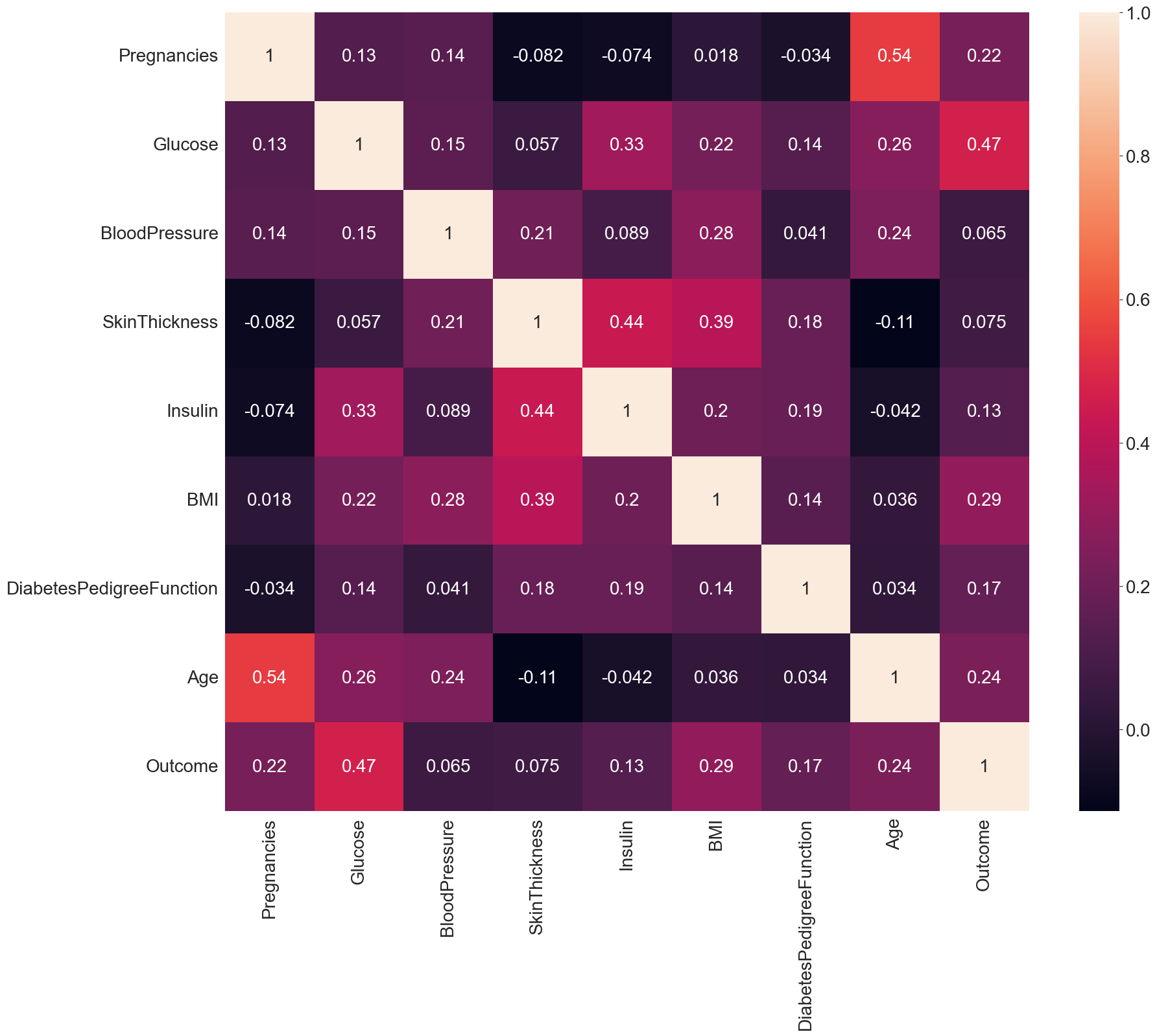
دلایل اصلی تجزیه وتحلیل شامل موارد زیر میشود:

* درک بهتر داده‌ ها
* شناسایی الگو ها و ارتباط ‌ها
* پیش پردازش داده
* انتخاب ویژگی‌ ها
* ارتقاء اطمینان در تصمیم‌ گیری
* کاهش خطرات
* برنامه‌ ریزی بهتر آزمایش‌

مراحل تحلیل دیتا:

1. حذف داده های تکراری به منظور اجرا بهتر و بالا رفتن سرعت محاسبات
2. چک کردن دیتا، در هر ستون برای پیدا کردن تعداد ورودی های خالی
3. بررسی مقادیر حداقل، حداکثر و میانگین به منظور تعیین دامنه‌ی مقادیر ممکن در هر ستون. در صورتی که دامنه‌ی اعداد به طور معقولی تعیین نشده باشد، بررسی می‌کنیم آیا نیاز به بهبود دادن داده‌ ها در این ستون وجود دارد یا خیر.
4. نقشه گرما :(Heat Map) نمایش دوبعدی اطلاعات با کمک رنگ‌هاست. نقشه ‌های گرما به کاربر کمک می‌کنند تا اطلاعات ساده یا پیچیده را تجسم کنند.

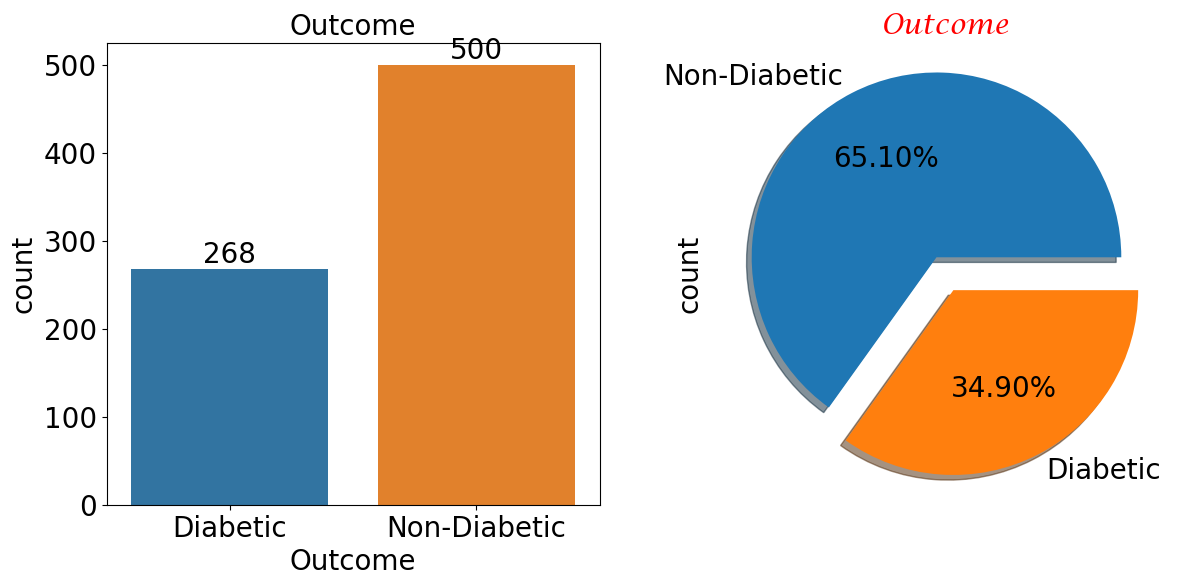
نکته‌ای که مشاهده خواهیم کرد، رابطه ضریب همبستگی 'Outcome' با ویژگی‌ های مختلف است. مشاهده می‌شود که گلوکز، شاخص توده بدنی و سن بیشترین همبستگی را با 'Outcome' دارند. در حالی که فشار خون، انسولین و عملکرد خانوادگی دیابت کمترین همبستگی را دارند.



ضریب همبستگی پیرسون :(Pearson's Correlation Coefficient) به شما کمک می‌کند تا رابطه بین دو متغیر را پیدا کنید. این شاخص اندازه‌گیری قدرت ارتباط بین دو متغیر را به شما می‌دهد. مقدار ضریب همبستگی پیرسون می‌تواند بین -1 تا +1 باشد. مقدار 1 به معنای ارتباط بسیار قوی بین دو متغیر است و مقدار 0 به معنای عدم ارتباط است.

1. اکنون، در هر ستون از مجموعه داده، ویژگی ‌ای که مقدار صفر دارد را با مقدار میانگین همان ستون جایگزین می‌کنیم تا هیچ یک از ویژگی ‌ها مقدار صفر نداشته باشند و دقت تحلیل داده افزایش یابد. سپس، مقادیر Outcome برای هر فرد که مقدار 1 داشته باشد، را به عنوان "دیابتی" در دیتابیس تغییر می‌دهیم و مقادیری که مساوی 0 هستند، را به عنوان "غیر دیابتی" در دیتابیس تعریف می‌کنیم. این کار را به منظور نمایش بهتر نمودار و درک بهتر داده ها انجام میدهیم.

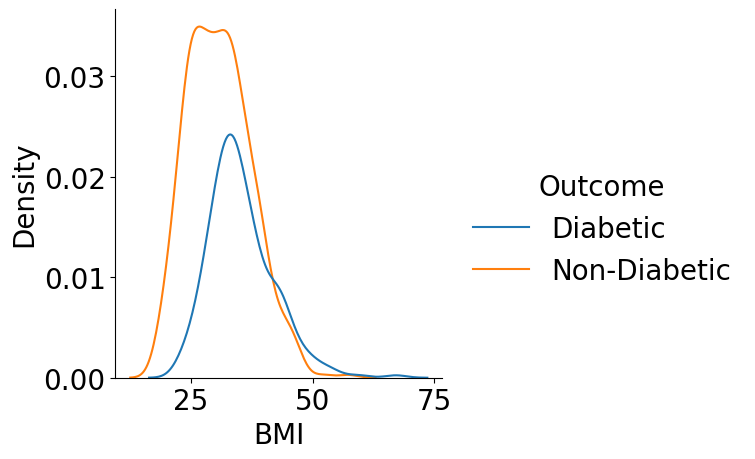
از نمودارهای پایین، یک شکل با دو نمودار فرعی دریافت خواهد شد. نمودار سمت چپ یک نمودار شمارش است که توزیع 'Outcome' در مجموعه داده را با استفاده از نوار ها نمایش می‌دهد و مقادیر بر روی هر نوار نمایش داده شده‌اند. نمودار سمت راست یک نمودار دایره‌ای است که توزیع نتایج را با استفاده از بخش ‌های خود نمایش می‌دهد و درصد مقادیر بر روی هر بخش نمایش داده شده‌است. هر دو نمودار فرعی دارای عنوانی هستند که هدف نمودارها را نشان می‌دهند.



1. همانگونه که تأثیر گلوکز بر دیابت به شدت مشهود است، ما تصمیم گرفتیم یک مطالعه در مورد این ویژگی انجام دهیم. به عبارت دیگر، افرادی که مقدار گلوکز خونشان کمتر از 140 است، دیابت ندارند. افرادی که گلوکز خونشان بین 140 و 199 قرار دارد، دارای وضعیت پیش‌دیابت هستند که این نشان می‌دهد میزان قند خون آنها از حد معمول بالاتر است، اما برای دسته‌بندی به عنوان دیابت، مقدار قند خون آنها به اندازه کافی بالا نیست و در آخر افرادی که گلوکز خونشان بیشتر از 199 است، به عنوان افراد مبتلا به دیابت شناخته می‌شوند.

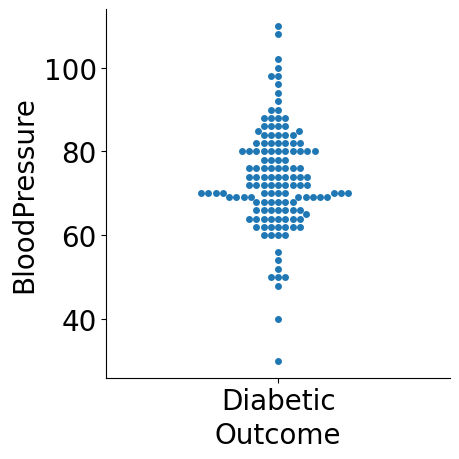
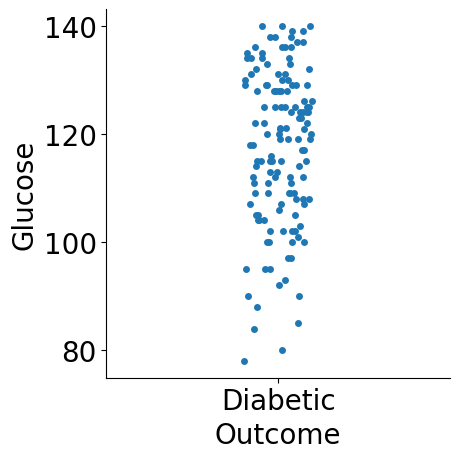
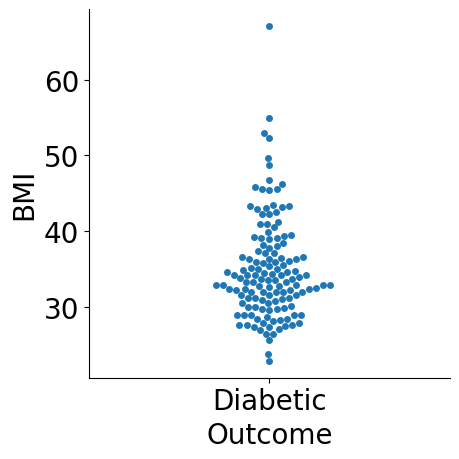
حال آنکه با درنظر گرفتن تنها گلوکز، تعداد افراد دیابتی در داده های ما صفر نفر میشود در حالیکه داده هایی وجود دارد که ستون Outcome آنها دیابتی است و این تناقض نشان دهنده این است که تحلیل، تنها بر اساس ویژگی گلوکز انجام پذیر نیست و باید سایر ویژگی ها نظیر انسولین، فشار خون، BMI و ... را برای تحلیل به کار بگیریم.

1. نمودار چگالی، یک خلاصه تصویری از توزیع و چگالی مقادیر BMI را ارائه می‌دهد و درکی از چگونگی توزیع داده و الگوها یا تفاوت‌های احتمالی بر اساس متغیر Outcome فراهم می‌کند.

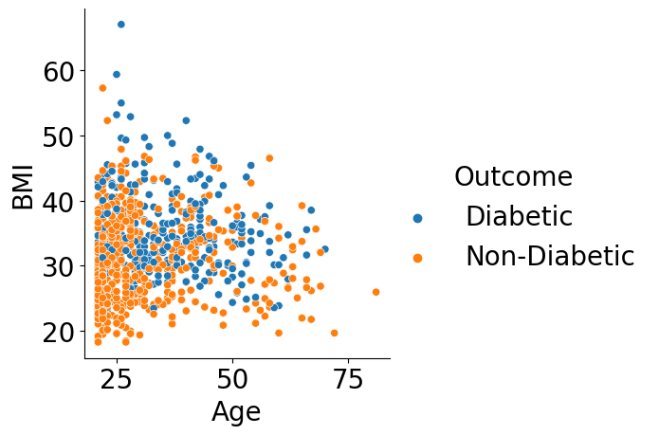
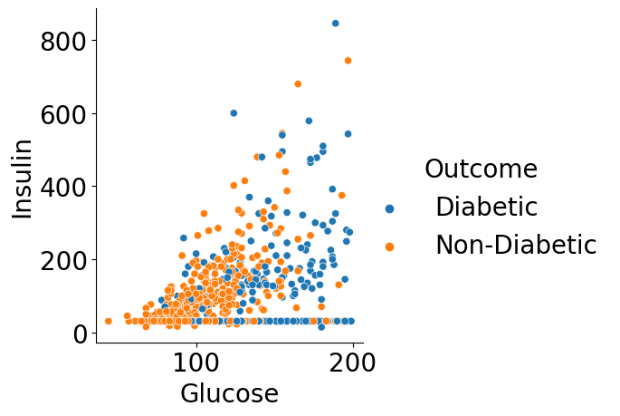


نمودار چگالی یک منحنی را نشان می‌دهد که چگالی تخمین ‌زده ‌شده از مقادیر BMI در مجموعه داده را نمایش می‌دهد. با درج متغیر Outcome، این امکان را فراهم می‌آورد که توزیع‌ چگالی برای Outcome را به صورت تصویری مقایسه کنیم که می‌تواند به شناسایی تفاوت ‌ها یا تشابه ‌ها در توزیع مقادیر BMI بر اساس نتایج مختلف کمک کند.

1. طبق گزارشات پزشکی اگر میزان گلوکز زیر 140 باشد فرد مبتلا به بیماری دیابت نیست که اگر این شرط را با ستون Outcome چک کنیم، مشاهده می‌کنیم افرادی هستند که دیابتی تشخیص داده شده‌اند و این تناقض به دلیل تاثیر چشمگیر ویژگی سن، انسولین و BMI می‌باشد.



1. در این مرحله دو نمودار مشاهده می‌کنیم که اولی ویژگی های گلوکز و انسولین و دومی ویژگی های BMI و سن را برای Outcome بررسی می‌کند.



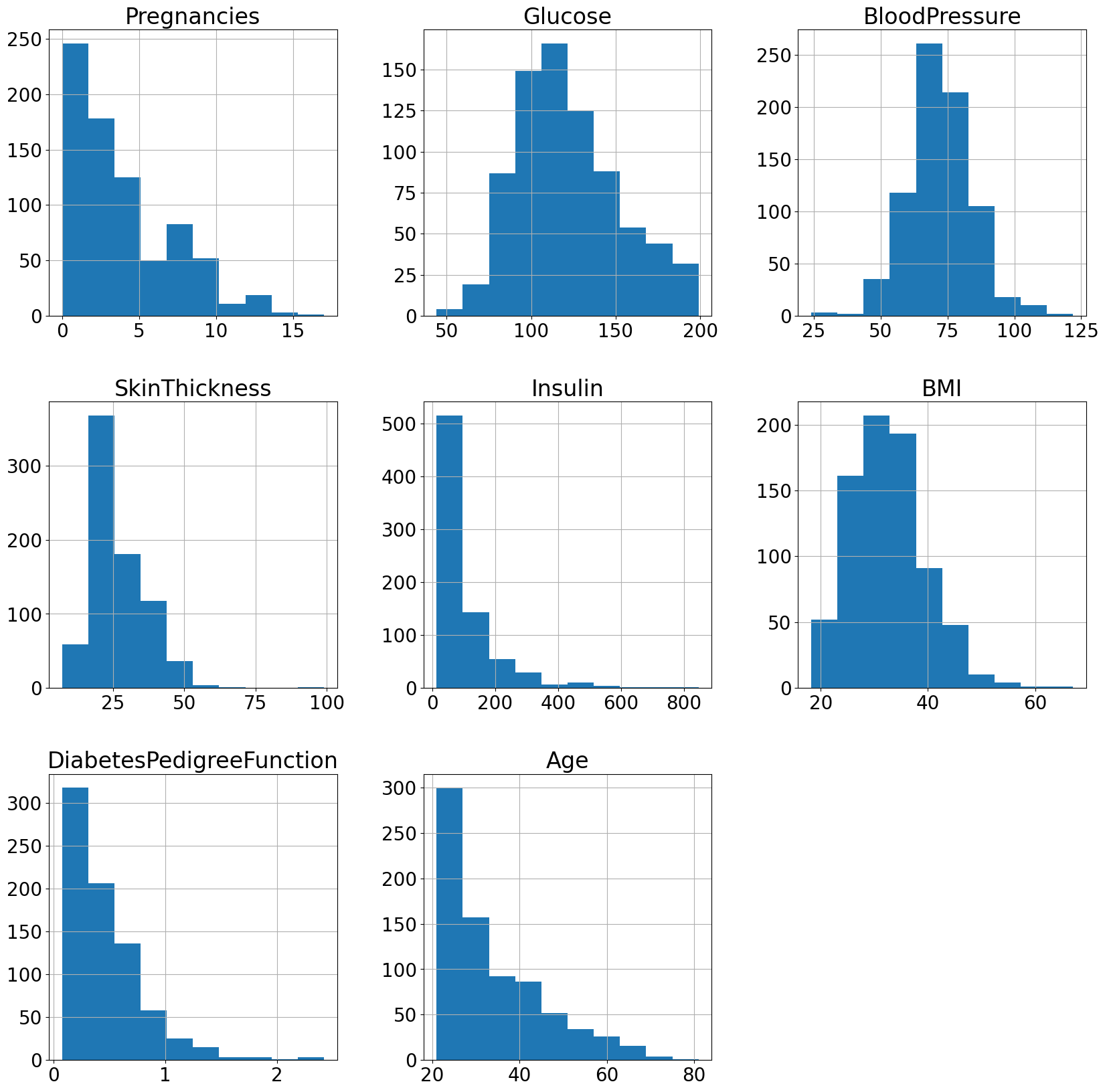
در شکل اول مشاهده می‌شود که با بالا رفتن سن در صورت کنترل وزن احتمال ابتلا به دیابت کاهش پیدا می‌کند.

محدوده :BMI

* زیر 18.5 - شما در محدوده وزن کم هستید.
* بین 18.5 و 24.9 - شما در محدوده وزن سالم هستید.
* بین 25 و 29.9 - شما در محدوده وزن چاق هستید.
* بین 30 و 39.9 - شما در محدوده وزن چاق مفرط هستید.

در شکل دوم مشاهده می‌شود که بالا رفتن انسولین متناسب با بالا رفتن گلوکز، فرد غیر دیابتی تشخیص داده می‌شود اما گاها با توجه به دیابت نوع دوم، ممکن است با بالا رفتن هم گلوکز و هم انسولین فرد دیابتی تشخیص داده شود چراکه تولید انسولین در خون بی تاثیر خواهد بود.

1. در اینجا بعد از مدتی تجزیه و تحلیل داده ها یک نمای کلی از وضعیت مجموعه داده و تغییراتی که اعمال کرده‌ایم را به کمک نمودار هیستوگرام توزیع داده مشاهده می‌کنیم:



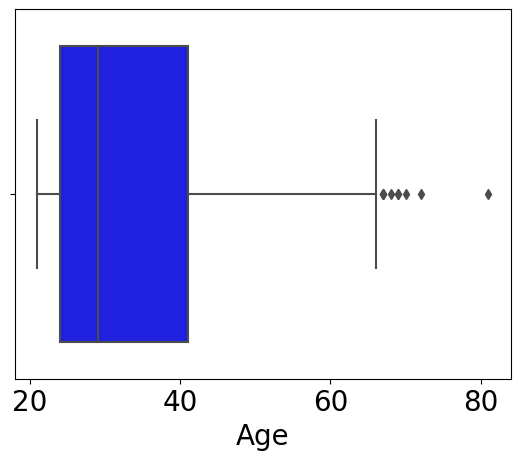
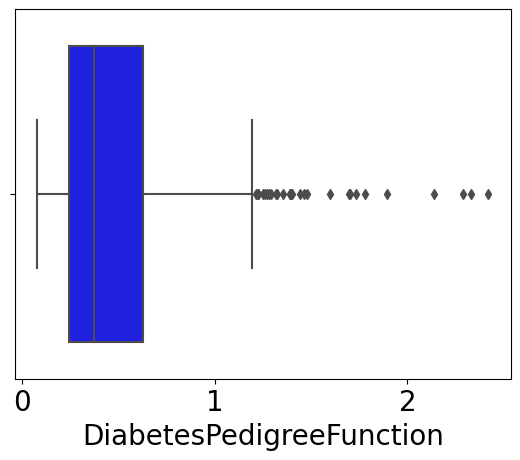
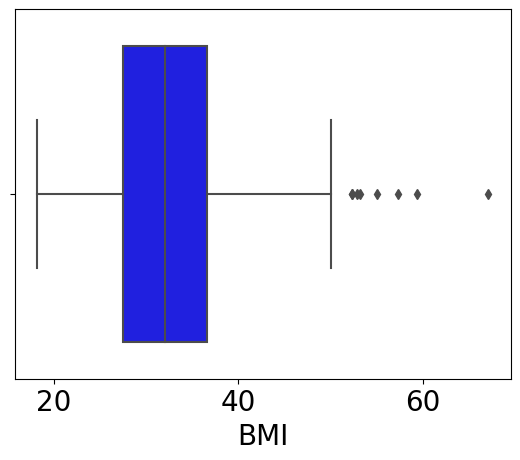
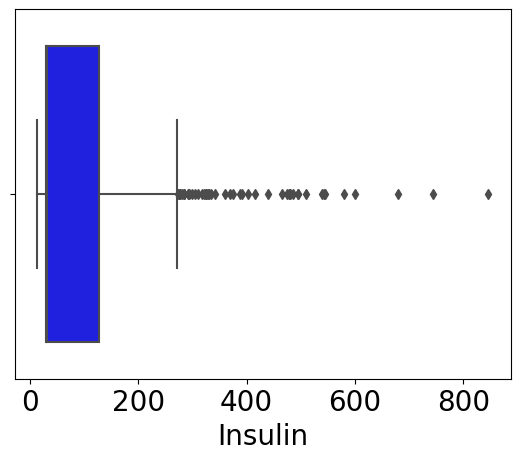
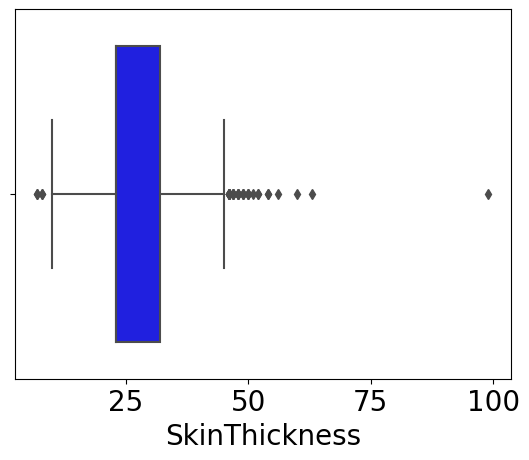
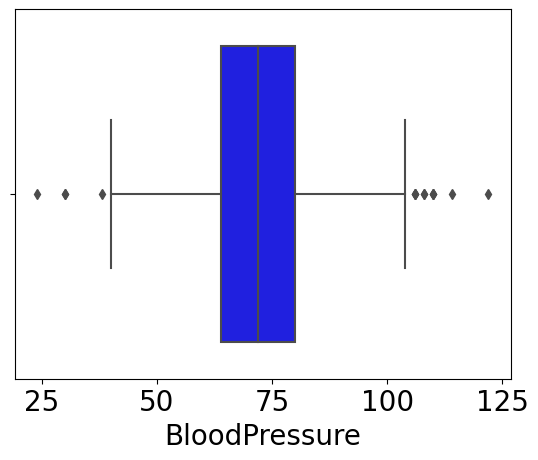
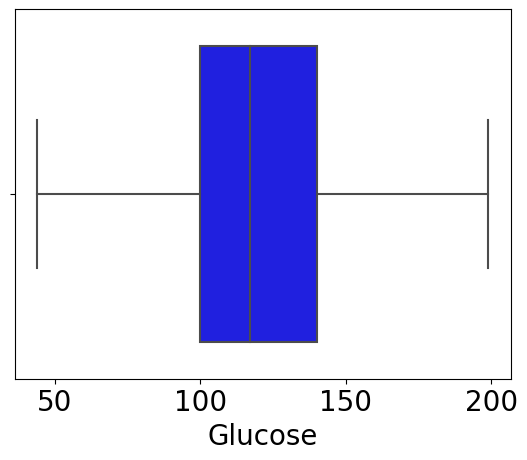
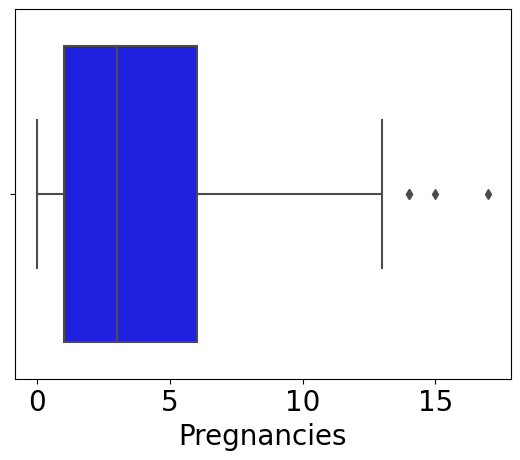
1. رسم Box Plot

Box Plot چیست و چه کاربرد هایی دارد؟

این نمودار به عنوان نمودار جعبه و شاخه (Box-and-Whisker plot) نیز شناخته می‌شود که یک نمایش گرافیکی از توزیع و آمار خلاصه داده‌ هاست. این نمودار یک خلاصه بصری از میانگین مرکزی داده، پراکندگی و هر نقطه غیرعادی (Outlier) احتمالی را ارائه می‌دهد. نمودار جعبه به ویژه برای مقایسه توزیع‌ های چندین مجموعه داده یا شناسایی وجود نقاط نمایی در یک مجموعه داده بسیار مفید است. این نمودار خلاصه بصری از ویژگی‌های آماری کلیدی داده‌ ها ارائه می‌دهد که تفسیر و مقایسه داده‌ ها در میان گروه ‌ها یا دسته ‌های مختلف را آسان ‌تر می‌کند.

در ادامه، تجزیه و تحلیلی از مؤلفه‌ها و اطلاعات انتقالی توسط نمودار جعبه آورده شده است:

* Box: جعبه مرکزی در نمودار نمایانگر دامنه ربع بالا و پایینی (IQR) است، که معادل 50% میانه داده‌هاست. جعبه از چارک اول (Q1) تا چارک سوم (Q3) گسترش می‌یابد و خطی که روی جعبه است نمایانگر میانه (Q2) است.
* Whiskers: خطوطی که از جعبه به بیرون گسترش می‌یابند، به عنوان شاخه‌ها شناخته می‌شوند و محدوده داده‌های خارج از IQR را نشان می‌دهند. معمولاً به حداکثر 1.5 برابر طول IQR از Q1 یا Q3 گسترش می‌یابند. نقاط داده‌ای که خارج از این محدوده قرار دارند به عنوان Outlier احتمالی در نظر گرفته می‌شوند و معمولاً به صورت جداگانه به عنوان نقاط نمایش داده می‌شوند.
* نقاط نمایی: نقاط داده‌ای که خارج از شاخه‌ها قرار دارند، به عنوان نقاط نمایی در نمودار علامت‌گذاری می‌شوند. آنها ممکن است به عنوان نقاط فردی یا نمادها، با توجه به کنوانسیون نمایش، به تصویر کشیده شوند.
* میانه (Q2): خط درون جعبه نمایانگر میانه است که میانه مقدار میانی داده‌ها هنگامی که به ترتیب صعودی مرتب شود. این یک اندازه از میانگین مرکزی است.
* چارک پایین (Q1) و چارک بالا (Q3): چارک پایین (Q1) که معادل چارک 25 (25th percentile) و چارک بالا (Q3) که معادل چارک 75 (75thpercentile) است. این مقادیر به تقسیم داده‌ها به چهار بخش مساوی کمک می‌کنند و اطلاعاتی را در مورد پراکندگی 50% میانه داده‌ها ارائه می‌دهند.

Box Plot توزیع داده ها بر اساس ویژگی های موجود:

Outlier ها در مجموعه داده، مقادیر غیرمعمولی هستند و می‌توانند تحلیل‌ های آماری را نادرست و فرضیات آنها را نقض کنند. بنابراین، کنترل Outlier ها بسیار مهم است. در این مورد، حذف آنها ممکن است منجر به از دست رفتن اطلاعات شود، بنابراین ما باید با استفاده از تکنیک ‌های مقیاس‌ دهی و نرمالایز کردن با آن ها کنار بیاییم.

1. نحوه مقیاس دهی و استاندارد کردن داده:

Quantile Transformer یک تکنیک پیش پردازش داده است که به طور معمول در یادگیری ماشین و آمار استفاده می‌شود. این تکنیک مجموعه داده را به یک مجموعه داده جدید تبدیل می‌کند که مقادیر آن توزیع احتمالی خاصی دارند. بسته به گزینه‌ های مشخص شده Quantile Transformer، داده را به یک توزیع یکنواخت یا Gaussian تبدیل می‌کند. Quantile Transformer به تخمین درصدی (کوانتایل‌ها) از داده‌ های اصلی می‌پردازد و سپس هر نقطه داده را به درصدی متناظر در توزیع مقصدش نگاشت می‌کند. این تبدیل به کاهش تأثیر Outlier ها و به داده‌ ها کمک می‌کند تا برای مدل‌ سازی مناسب‌ تر شوند. در ادامه توضیح داده می‌شود که این تکنیک چه کارهایی انجام می‌دهد:

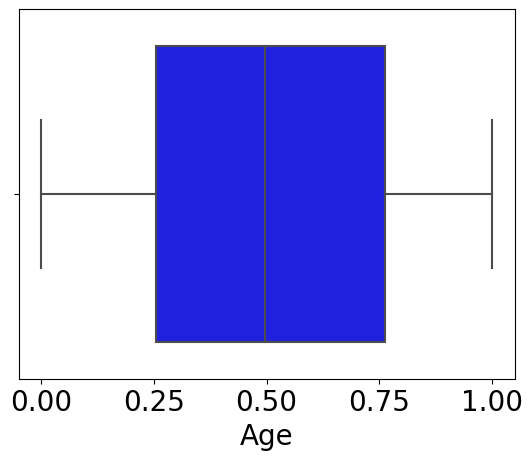
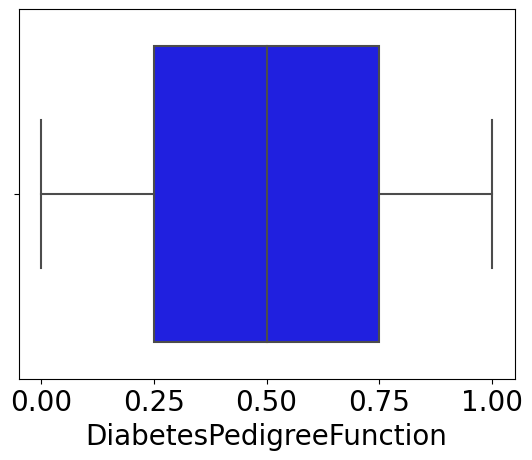
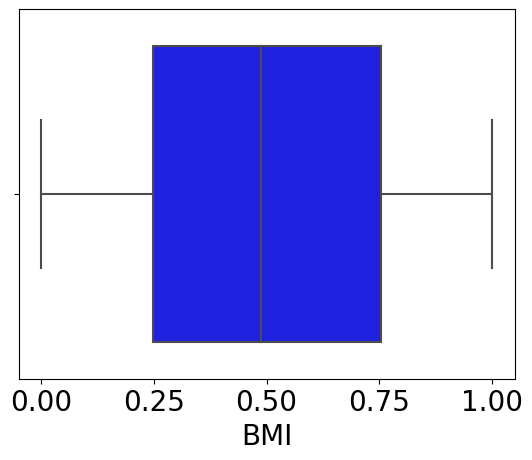
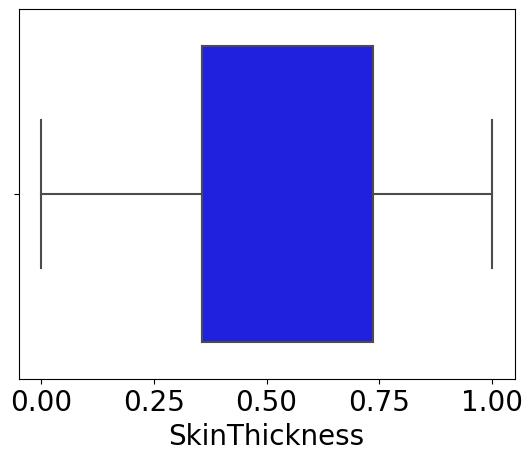
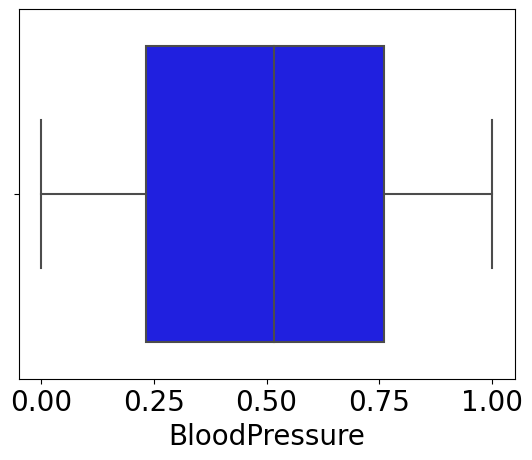
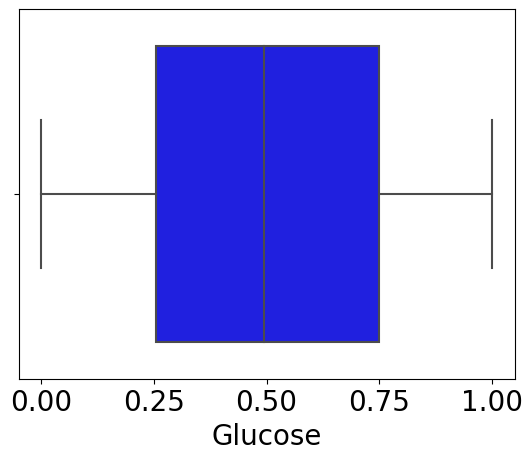
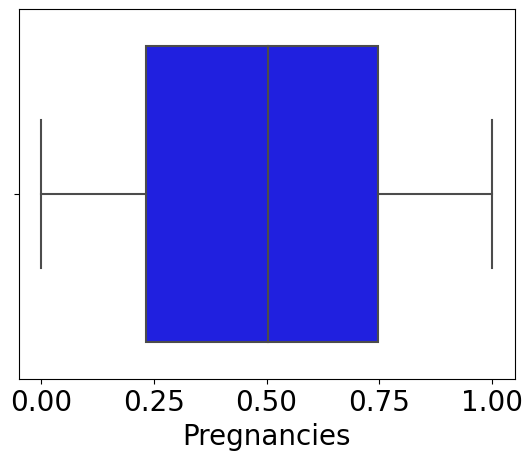
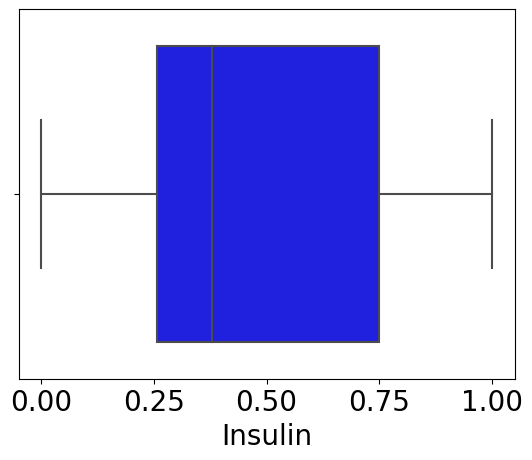
* توزیع یکنواخت: اگر بخواهیم داده ‌های خود را به یک توزیع یکنواخت تبدیل کنیم، Quantile Transformer داده ‌های اصلی را به یک توزیع یکنواخت بین 0 و 1 نگاشت می‌کند. این کار زمانی مفید است که داده‌ ها دنبال یک توزیع خاص نباشند و می‌خواهیم آنها را به توزیعی یکنواخت‌تر تبدیل کنیم که برای برخی از الگوریتم ‌های یادگیری ماشین خاص ممکن است مفید باشد.
* توزیع Gaussian: به عنوان گزینه دیگر، می‌توان تصمیم گرفت داده ‌ها را به یک توزیع Gaussian (نرمال) تبدیل کرد. در این صورت، Quantile Transformer داده‌ها را به یک توزیع نرمال با میانگین 0 و انحراف معیار 1 نگاشت می‌کند. این تبدیل زمانی می‌تواند مفید باشد که داده ‌ها شیب دار یا غیر نرمال هستند و می‌خواهیم آنها را به یک توزیع نرمال ‌تر تبدیل کنیم که ممکن است برای برخی از تکنیک‌ های آماری و مدل ‌های یادگیری ماشین الزامی باشد.

بعضی نکات مهم در استفاده از Quantile Transformer عبارتند از:

* 1. این یک تبدیل غیر خطی است، بنابراین ممکن است ارتباطات بین نقاط داده را از بهم بریزد.
  2. نسبت به وجود مقادیر تکراری در داده حساس است. این می‌تواند برای بهبود عملکرد برخی از الگوریتم ‌های یادگیری ماشین که یک توزیع داده خاص را فرض می‌کنند، مفید باشد.

می‌توان پارامتر هایQuantile Transformer را تنظیم کرد، مانند توزیع خروجی (یکنواخت یا نرمال) و اینکه آیا می‌خواهیم داده‌ های تبدیل شده را به دامنه خاصی مقیاس کنیم یا نه. این به ما امکان را می‌دهد تا تبدیل را به نیاز ها و داده ‌های خاص خودمان تنظیم کنیم.

در شکل های زیر نشان داده می‌شود که به کمک تکنیکQuantile Transformer، توزیع داده به صورت یکنواخت صورت گرفته و با کاهش تأثیر Outlier ها دیتا ما برای مدل‌ سازی مناسب‌ تر آماده می‌شوند:



1. چرا داده را تقسیم می‌کنیم؟ داده Train و Test

تقسیم داده به دلایل مهمی در یادگیری ماشین و تحلیل داده به کار می‌رود. این عمل به دو بخش آموزش و آزمون به ما این امکان را می‌دهد که عملکرد و توانایی عمومی مدل را ارزیابی کنیم و همینطور به ما کمک می‌کند که عملکرد مدل را بر روی داده‌های جدید ارزیابی کرده، از Overfitting جلوگیری کنیم، پارامترهای مدل را بهینه‌سازی کنیم و تخمین‌های عملکرد بی‌انحراف دریافت کنیم. این یک مرحله مهم در فرآیند یادگیری ماشین است تا اطمینان حاصل شود که مدل به خوبی عمل می‌کند و توانایی تعمیم به داده‌های ناشناخته را دارد.

در ادامه دلایل اصلی تقسیم داده به دو بخش را برای شما توضیح می‌دهیم:

* ارزیابی مدل: با تقسیم داده به دو بخش آموزش و آزمون، می‌توانیم ارزیابی کنیم که مدل ما چقدر خوب عمل می‌کند. بخش آموزش برای یادگیری مدل استفاده می‌شود و بخش آزمون برای ارزیابی عملکرد مدل بر روی داده ‌هایی استفاده می‌شود که تاکنون برای مدل ناشناخته بوده‌اند. این ارزیابی به ما کمک می‌کند تا بفهمیم مدل چقدر توانایی تعمیم به داده‌ های جدید و ناشناخته دارد.
* جلوگیری از Overfitting: این اتفاق وقتی رخ می‌دهد که یک مدل به طور زیادی به داده ‌های آموزشی خود پیوسته و روی داده‌های جدید عملکرد ضعیفی داشته باشد. تقسیم داده به دو بخش به ما این امکان را می‌دهد که Overfitting را تشخیص دهیم و از آن جلوگیری کنیم. با ارزیابی عملکرد مدل بر روی بخش آزمون، می‌توانیم تشخیص دهیم که آیا مدل به طور زیادی به داده‌های آموزشی پیوسته و Overfit شده است یا آیا می‌تواند به خوبی به داده‌های جدید تعمیم دهد.

پیاده سازی مدل:

بعد از تجزیه و تحلیل داده ها و بررسی نتایج بدست آماده نوبت به مدل سازی و آموزش (Train) دیتا می‌رسد. پیاده سازی مدل این اجازه را می‌دهد تا کامپیوتر یا ماشین، از داده ‌ها یاد بگیرد و الگوها و اطلاعات مفهومی را استخراج کنند. در اینجا از 4 عدد از مدل های یادگیری استفاده کرده‌ایم که اعم از Decision Tree، Logistic Regression، KNN و SVM است. به شرح مختصری از آنها میپردازیم:

* درخت تصمیم گیری (Decision Tree):

مدل درخت تصمیم گیری یکی از روش‌های مهم در مسائل پیش‌بینی و تصمیم‌گیری در ماشین‌یادگیری است. این مدل بر اساس ساختار درختی، تصمیم ‌ها را اتخاذ می‌کند و به ازای هر نمونه ورودی تصمیم ‌های متوالی اتخاذ می‌کند تا به یک خروجی یا پیش‌بینی برسد. در مدل درخت تصمیم گیری برای پروژه تشخیص دیابت، می‌توان به شرح زیر عمل کرد:

* 1. ساخت درخت تصمیم: مرحله اول در استفاده از مدل درخت تصمیم گیری برای تشخیص دیابت، ساخت درخت است. این درخت از ریشه شروع می‌شود و در هر گام به یکی از ویژگی ‌های ورودی تقسیم می‌شود. به عنوان مثال، از یکی از ویژگی‌های ورودی مثل سطح گلوکز می‌توان به عنوان یک نود تصمیم‌گیری استفاده کرد. سپس از مقدار سطح گلوکز، درخت به دو شاخه تقسیم می‌شود: یک شاخه که سطح گلوکز کمتر از مقداری خاصی است و یک شاخه دیگر که سطح گلوکز بیشتر است.
  2. پیش‌بینی: هر نمونه ورودی به ازای ویژگی‌هایش از ریشه به سمت یکی از گره‌ های پایانی درخت حرکت می‌کند. در هر گره پایانی یک پیش‌بینی انجام می‌شود که معمولاً یک مقدار باینری (بله/خیر) است. در اینجا، پیش‌بینی ما مربوط به تشخیص دیابت یا عدم دیابت است. به عبارت دقیق‌تر، اگر در درخت به یک گره پایانی برسیم که می‌گوید شخص دیابت دارد، آن نمونه به عنوان مثبت دیابت معرفی می‌شود؛ در غیر این صورت، به عنوان منفی دیابت معرفی می‌شود.
  3. آموزش و ارزیابی: برای ساخت یک درخت تصمیم گیری بهینه، از داده‌های آموزشی استفاده می‌شود. الگوریتم تلاش می‌کند تا بهینه ‌ترین تقسیم‌ بندی ‌ها و شاخه ‌ها را بر اساس ویژگی‌ها پیدا کند تا در نهایت دقت پیش‌بینی بالا باشد. سپس از مدل بر روی داده‌ های آزمایشی یا داده‌ های جدید برای ارزیابی و پیش‌بینی استفاده می‌شود.
* رگرسیون لجستیک (Logistic Regression):

چرا از رگرسیون لجستیک استفاده می‌کنیم؟

به طور کلی رگرسیون لجستیک یک تکنیک مدل ‌سازی آماری است که برای دسته ‌بندی دودویی (binary classification) استفاده می‌شود. به خصوص برای مواردی که متغیر وابسته یا نتیجه به صورت دو کلاس ممکن است، مانند "بله/خیر"، "صحیح/غلط" یا در مورد مجموعه داده دیابت، "دیابتی/غیر دیابتی".

در رگرسیون لجستیک، هدف این است که احتمال تعلق یک مشاهده به یک کلاس خاص بر اساس یک یا چند متغیر مستقل (که به عنوان ویژگی‌ها یا پیش ‌بیننده ‌ها هم شناخته می‌شوند) تخمین زده شود. خروجی رگرسیون لجستیک یک مقدار احتمال بین 0 و 1 است که می‌توان آن را به عنوان احتمال تعلق به کلاس مثبت (مثلاً داشتن دیابت) تفسیر کرد.

مدل رگرسیون لجستیک از یک تابع لجستیک (یا تابع سیگموئید) برای تبدیل یک ترکیب خطی از ویژگی‌های ورودی به احتمال پیش‌بینی شده استفاده می‌کند. این تبدیل اطمینان می‌دهد که احتمالات پیش‌بینی شده در محدوده معتبر 0 تا 1 قرار می‌گیرند. تابع لجستیک به صورت زیر تعریف می‌شود:

P(y = 1|X) = 1 / (1 + e^-(b0 + b1x1 + b2x2 + ... + bnxn))

در حالی که:

1. P(y = 1|X) احتمال متغیر وابسته به کلاس مثبت با توجه به ویژگی‌های ورودی X است.
2. e پایه لگاریتم طبیعی (تقریباً 2.718) است.
3. b0، b1، b2، ...، bn ضرایب مدل رگرسیون لجستیک هستند که تأثیر هر ویژگی بر نتیجه را تعیین می‌کنند.
4. x1، x2، ...، xn مقادیر ویژگی متناظر برای یک داده مشاهده شده هستند.

نتایج ارائه شده به شرح زیر است:

1. دقت آموزش: دقت آموزش مدل رگرسیون لجستیک برابر با 75.98% است. این به معنای این است که مدل به درستی برای تقریباً 75.98% از نمونه ‌ها در مجموعه داده آموزش نتیجه را پیش‌بینی کرده است.
2. امتیاز دقت مدل: امتیاز دقت مدل رگرسیون لجستیک بر روی داده‌های آزمون برابر با 77.92% است. این نشان می‌دهد که مدل به درستی برای تقریباً 77.92% از نمونه‌ها در مجموعه داده آزمون نتیجه را پیش‌بینی کرده است.
3. گزارش دسته‌بندی: گزارش دسته‌بندی معیارهای ارزیابی مانند دقت (پریسیژن)، حساسیت (ریکال) و امتیاز F1 را برای هر کلاس (0.0 و 1.0) در داده‌های آزمون ارائه می‌دهد. در ادامه توضیح معیارها آمده است:

- دقت (پریسیژن): دقت نسبت مثبت‌های به درستی پیش‌بینی شده به تمام نمونه‌های پیش‌بینی شده به عنوان مثبت را اندازه می‌گیرد. برای کلاس 0.0، دقت برابر با 0.80 است که نشان می‌دهد 80% از نمونه‌های پیش‌بینی شده به عنوان غیر دیابتی، واقعاً غیر دیابتی هستند. برای کلاس 1.0، دقت برابر با 0.73 است که نشان می‌دهد 73% از نمونه‌های پیش‌بینی شده به عنوان دیابتی، واقعاً دیابتی هستند.

- حساسیت (ریکال): حساسیت (همچنین به نام حساسیت یا نرخ مثبت واقعی) نسبت مثبت‌های به درستی پیش‌بینی شده به تمام مثبت‌های واقعی را اندازه می‌گیرد. برای کلاس 0.0، حساسیت برابر با 0.86 است که نشان می‌دهد 86% از نمونه‌های واقعی غیر دیابتی به درستی به عنوان غیر دیابتی تشخیص داده می‌شوند. برای کلاس 1.0، حساسیت برابر با 0.63 است که نشان می‌دهد 63% از نمونه‌های واقعی دیابتی به درستی به عنوان دیابتی تشخیص داده می‌شوند.

- امتیاز F1: امتیاز F1 میانگین هارمونیک دقت و حساسیت است. این امتیاز اندازه‌گیری متوازنی از عملکرد مدل را با در نظر گرفتن هم دقت و هم حساسیت فراهم می‌کند. برای کلاس 0.0، امتیاز F1 برابر با 0.83 است و برای کلاس 1.0، امتیاز F1 برابر با 0.68 است.

- پشتیبانی: پشتیبانی تعداد نمونه‌های تعلق گرفته به هر کلاس در داده‌های آزمون را نمایش می‌دهد. در اینجا، 147 نمونه از کلاس 0.0 (غیر دیابتی) و 84 نمونه از کلاس 1.0 (دیابتی) وجود دارد.

1. دقت، میانگین ماکرو و میانگین وزنی: دقت (accuracy) کلی مدل، بر روی داده‌های آزمون است که برابر با 78% است. میانگین ماکرو میانگین دقت، حساسیت و امتیاز F1 برای همه کلاس ‌ها را محاسبه می‌کند و وزن یکسانی به هر کلاس می‌دهد. میانگین وزنی تعداد نمونه‌ ها برای هر کلاس را در نظر می‌گیرد و میانگین وزنی بر اساس توزیع کلاس ‌ها ارائه می‌دهد.

بر اساس این نتایج، مدل رگرسیون لجستیک سطح معقولی از عملکرد در پیش‌بینی دیابت در مجموعه داده داده شده را نشان می‌دهد. این مدل در دو مرحله آموزش و آزمون دقت بالایی (بیش از 75%) دارد، که نشان می‌دهد می‌تواند پیش‌بینی ‌های به نسبت دقیقی انجام دهد. با این حال، عملکرد مدل ممکن است برای کلاس ‌های مختلف متغیر باشد و دقت، حساسیت و امتیاز F1 کمی برای کلاس دیابتی نسبت به کلاس غیر دیابتی فرق کند.

* نزدیک ترین همسایه (K-Nearest Neighbors):

مدل نزدیک ترین همسایه یکی دیگر از مدل‌های مهم در ماشین یادگیری ماشینی برای مسائل دسته‌بندی و پیش‌بینی است. این مدل برای مسائل دسته‌بندی و پیش‌بینی بر اساس شباهت میان نمونه ‌ها از داده ‌های آموزشی استفاده می‌کند. در پروژه تشخیص دیابت، می‌توان از مدل KNN استفاده کرد تا بر اساس ویژگی ‌های فردی، دیابت را پیش‌بینی کنیم. توضیحات زیر توضیح دقیق‌تری ارائه می‌دهد:

1. انتخاب تعداد همسایه‌ها (K): در مدل KNN، ابتدا تعداد مشخصی از نزدیک‌ترین همسایه‌ها (نمونه‌های آموزشی) انتخاب می‌شود. این تعداد به عنوان K نمایش داده می‌شود و معمولاً به صورت دستی تنظیم می‌شود. برای مثال، اگر K برابر با 5 باشد، پنج همسایه نزدیک‌تر به نمونه ورودی ما انتخاب می‌شوند.
2. محاسبه فاصله: برای یافتن همسایه‌ها، باید فاصله بین نمونه ورودی و همه نمونه‌های آموزشی محاسبه شود. فاصله معمولاً با استفاده از معیارهایی مانند فاصله اقلیدسی (Euclidean) یا فاصله منهتن (Manhattan) محاسبه می‌شود.
3. رتبه‌بندی همسایه‌ها: پس از محاسبه فاصله بین نمونه ورودی و همه نمونه‌های آموزشی، همسایه‌ها بر اساس فاصله به ترتیب مرتب می‌شوند و K همسایه نزدیکتر به نمونه ورودی انتخاب می‌شوند.
4. تصمیم‌گیری دسته‌بندی: با داشتن K همسایه انتخاب شده، مدل تصمیم می‌گیرد که نمونه ورودی به کدام دسته (بیمار یا عدم بیمار بودن) تعلق دارد. این تصمیم معمولاً با شمارش تعداد همسایه‌های هر دسته و انتخاب دسته‌ای که بیشترین همسایه‌ها را دارد اتخاذ می‌شود.

* ماشین ‌های پشتیبانی (Support Vector Machines):

چرا ما از ماشین‌های بردار پشتیبانی (SVMs) استفاده می‌کنیم؟

هرچند ماشین ‌های بردار پشتیبانی (SVMs) مزایایی در مدیریت داده‌ های غیرخطی، نقاط نادر، و فضا های بعدی بالا دارند، اما لازم به ذکر است که ممکن است محدودیت ‌هایی نیز داشته باشند. SVMs ممکن است در مصرف منابع محاسباتی مطرح باشند و نیاز به تنظیم دقیق پارامترها داشته باشند. علاوه بر این، اگر مجموعه داده نامتوازن باشد یا تداخل زیادی بین کلاس‌ها وجود داشته باشد SVMs ممکن است در مواردی عملکرد بهینه را ارائه ندهند.

SVMs به دلیل توانایی‌اش در مدیریت داده‌های غیرخطی، مقاومت در برابر نقاط نادر، کارآیی در فضا های بعد بالا، عملکرد تعمیمی و قابلیت دسته‌بندی دودویی به عنوان یک انتخاب محبوب برای مجموعه داده دیابت مطرح است.

نتایج ارائه شده برای مدل SVM بر روی مجموعه داده دیابت به شرح زیر هستند:

1. دقت آموزش: مدل SVM دقت آموزشی برابر با 100% دارد. این به این معناست که مدل به درستی پیش‌بینی نتایج تمام نمونه‌ها در مجموعه داده آموزش را انجام داده است. دقت آموزشی 100% نشان‌دهنده این است که مدل داده‌های آموزش را به خوبی حفظ کرده و ممکن است باز‌آموزی داشته باشد.
2. امتیاز دقت مدل: امتیاز دقت مدل SVM بر روی داده‌های آزمون برابر با 69.7% است. این نشان‌دهنده این است که مدل به درستی نتایج تقریباً 69.7% از نمونه‌ها در مجموعه داده آزمون پیش‌بینی کرده است. دقت مدل بر روی داده‌های آزمون کمتر از دقت آموزش است، که ممکن است نشانه باز‌آموزی یا محدودیت‌هایی در تعمیم به داده‌های ناشناخته باشد.
3. گزارش دسته‌بندی: گزارش دسته‌بندی معیارهای ارزیابی مانند دقت، حساسیت و امتیاز F1 برای هر کلاس (0.0 و 1.0) در داده‌های آزمون ارائه می‌دهد. در ادامه توضیحاتی درباره این معیارها آمده است:

- دقت: دقت نسبت تعداد نمونه‌های صحیح پیش‌بینی‌شده به تعداد نمونه‌های پیش‌بینی‌شده به عنوان مثبت را اندازه‌گیری می‌کند. برای کلاس 0.0، دقت برابر با 0.78 است که نشان‌دهنده این است که 78% از نمونه‌های پیش‌بینی‌شده به عنوان غیر دیابتی واقعاً غیر دیابتی هستند. برای کلاس 1.0، دقت برابر با 0.57 است که نشان‌دهنده این است که 57% از نمونه‌های پیش‌بینی‌شده به عنوان دیابتی واقعاً دیابتی هستند.

- حساسیت: حساسیت (همچنین به نام تشخیص یا نرخ مثبت واقعی) نسبت تعداد نمونه‌های مثبت درست پیش‌بینی‌شده به تعداد نمونه‌های مثبت واقعی را اندازه‌گیری می‌کند. برای کلاس 0.0، حساسیت برابر با 0.73 است که به معنای این است که 73% از نمونه‌های واقعی غیر دیابتی به درستی به عنوان غیر دیابتی شناسایی شده‌اند. برای کلاس 1.0، حساسیت برابر با 0.64 است که به معنای این است که 64% از نمونه‌های واقعی دیابتی به درستی به عنوان دیابتی شناسایی شده‌اند.

- امتیاز F1: امتیاز F1 میانگین هارمونیک دقت و حساسیت است. این امتیاز اندازه‌گیری متوازنی از عملکرد مدل با در نظر گرفتن هم دقت و هم حساسیت ارائه می‌دهد. برای کلاس 0.0، امتیاز F1 برابر با 0.75 و برای کلاس 1.0، امتیاز F1 برابر با 0.61 است.

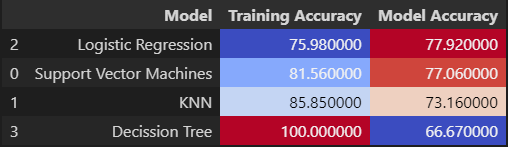
- پشتیبانی: پشتیبانی تعداد نمونه‌های تعلق گرفته به هر کلاس در داده‌های آزمون را نمایش می‌دهد. تعداد نمونه‌های کلاس 0.0 (غیر دیابتی) برابر با 147 و تعداد نمونه‌های کلاس 1.0 (دیابتی) برابر با 84 است.

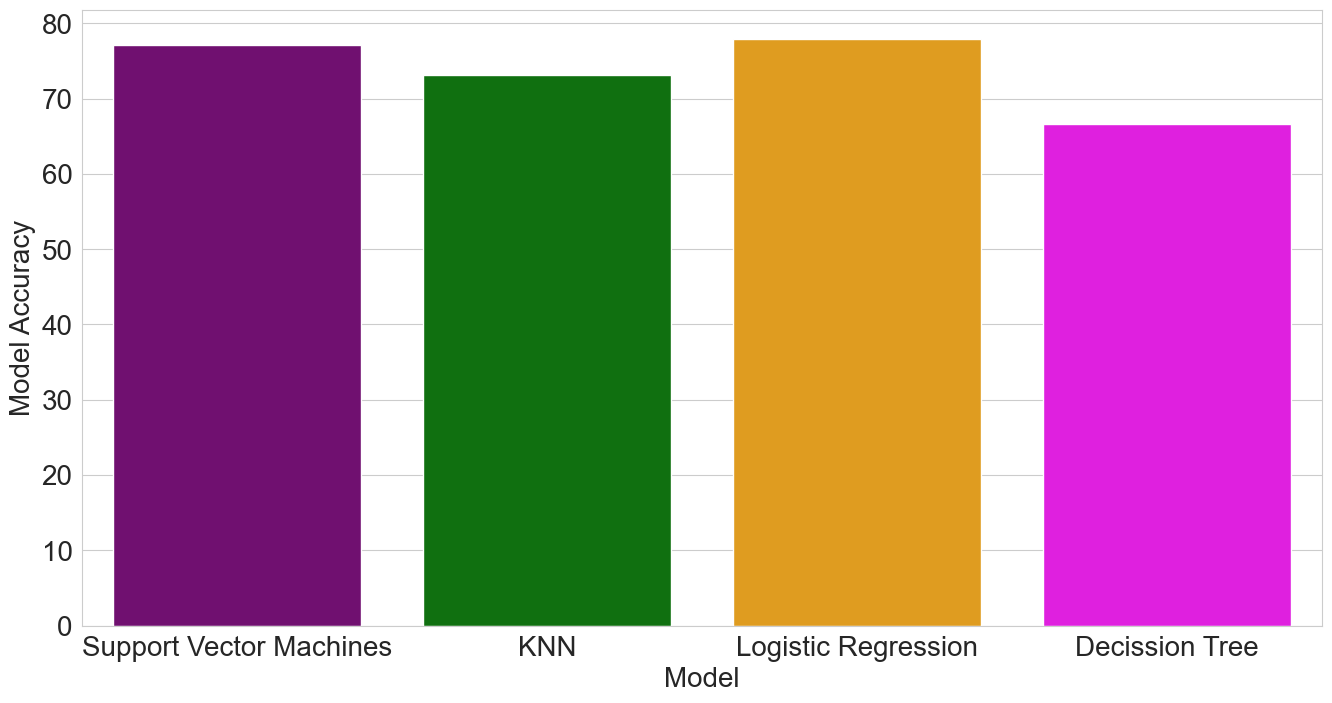
1. دقت، میانگین ماکرو، میانگین وزنی: دقت کلی مدل بر روی داده‌های

آزمون برابر با 70% است. میانگین ماکرو دقت، حساسیت و امتیاز F1 را در تمام کلاس‌ها محاسبه کرده و ارزیابی متوازنی از عملکرد مدل را نشان می‌دهد که امتیاز 0.68 است، نشان‌دهنده تعادل نسبی در عملکرد در کلاس‌ها است. میانگین وزنی که توزیع کلاس‌ها را در نظر می‌گیرد نیز برابر با 0.70 است.

این نتایج نشان می‌دهند که مدل SVM ممکن است به خوبی به داده‌های ناشناخته تعمیم ندهد. برای بهبود عملکرد مدل و رفع مشکل باز‌آموزی، ممکن است نیاز به تنظیم پارامترها یا در نظر گرفتن مدل‌های جایگزین باشد.

* به طور خلاصه جدول و نمودار هیستوگرام زیر میزان دقت هر مدل را نشان می‌دهد:





و همانطور که مشاهده می‌شود، در می‌یابیم که مدل رگرسیون لجستیک با دقت بیشتری پیش‌بینی می‌کند. و با توجه به همین موضوع دیتا یک فرد جدید را در این مدل تست کرده و جواب نهایی به شرح زیر است:

Input data:

Pregnancies : 2

Glucose : 95

Blood Pressure: 90

Skin Thickness : 40

Insulin: 150

BMI : 24

Diabetes Pedigree Function : 0.727000

Age : 20

Prediction:

Outcome: 1

The patient is predicted to have diabetes

**فصل 4: نتیجه گیری**

مدیریت چالش ‌های مرتبط با پیش‌بینی دیابت با استفاده از یادگیری ماشین اهمیت بسیاری دارد. این چالش‌ ها شامل برهم‌کنش‌ های ویژگی‌ها که ممکن است تشخیص دیابت را پیچیده کنند، برای مثال، افزایش سطح انسولین ممکن است تأثیر معکوسی روی سطح گلوکز در خون داشته باشد. جمع ‌آوری داده‌ های کافی و معتبر برای آموزش مدل ‌ها، تنظیم پارامتر های مدل به درستی و توازن بین دقت و تعداد نادرست‌ ها است. با مدیریت موفق این چالش‌ ها، می‌توان به توسعه مدل‌ های یادگیری ماشین با دقت و عملکرد بهتر در تشخیص و مدیریت دیابت کمک کرد.

سیستم‌ پیش‌بینی دیابتی که طراحی کردیم با استفاده از یادگیری ماشین در تشخیص و پیش‌بینی دیابت کمک می‌کند. این سیستم با تحلیل داده ‌های مرتبط با دیابت، مانند سن و سابقه خانوادگی، مدل‌های پیش‌بینی ایجاد می‌کند. این مدل ها به پزشکان در تشخیص زودهنگام دیابت و مدیریت بهتر بیماران کمک می‌کنند. همچنین، می‌تواند افرادی که در خطر ابتلا به دیابت هستند را شناسایی کرده و اقدامات پیشگیرانه‌ای را توصیه کند. استفاده از این ماشین به بهبود سلامت عمومی افراد و کمک به پزشکان در تشخیص به موقع و مدیریت بهتر دیابت کمک می‌کند.

به منظور توسعه تحقیقات آتی در زمینه پیش‌بینی دیابت با استفاده از یادگیری ماشین، پیشنهاد زیر می‌تواند موجب تعمیق دانش و پیشبرد پژوهش‌ ها شود:

* دقت مدل ها مهم ترین مسئله‌ای است که باید به آن پرداخته شود. ماشین‌ های یادگیری قادر به تجزیه و تحلیل داده‌ های متنوع هستند. پژوهش ‌های آتی می‌تواند به ترکیب داده ‌های بالینی، ژنتیکی و سایر منابع اطلاعاتی مانند تصاویر پزشکی بپردازد و با استفاده از تکنولوژی ‌های پیشرفته از تکنولوژی‌ هایی مانند یادگیری عمیق (Deep Learning) و شبکه‌های عصبی مصنوعی برای ارتقاء دقت مدل‌ها استفاده شوند. این تکنولوژی ‌ها قابلیت تشخیص الگوهای پیچیده‌تر را دارند.

**منابع و مراجع**

1. [**https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database**](https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)
2. [**https://www.databricks.com/glossary/machine-learning-models**](https://www.databricks.com/glossary/machine-learning-models)
3. [**https://online.hbs.edu/blog/post/how-to-analyze-datasets**](https://online.hbs.edu/blog/post/how-to-analyze-datasets)
4. [**https://www.simplilearn.com/data-analysis-methods-process-types-article#:~:text=Data%20analysis%20refers%20to%20the,%2C%20and%20support%20decision%2Dmaking**](https://www.simplilearn.com/data-analysis-methods-process-types-article#:~:text=Data%20analysis%20refers%20to%20the,%2C%20and%20support%20decision%2Dmaking)**.**
5. [**https://www.simplilearn.com/tutorials/python-tutorial/data-visualization-in-python#:~:text=Matplotlib%20and%20Seaborn%20are%20python,primarily%20used%20for%20statistical%20graphs**](https://www.simplilearn.com/tutorials/python-tutorial/data-visualization-in-python#:~:text=Matplotlib%20and%20Seaborn%20are%20python,primarily%20used%20for%20statistical%20graphs)**.**
6. [**https://www.geeksforgeeks.org/data-analysis-with-python/**](https://www.geeksforgeeks.org/data-analysis-with-python/)
7. [**https://www.ninisite.com/article/12666/%D8%AF%D8%B1%D8%A8%D8%A7%D8%B1%D9%87-%D8%AF%DB%8C%D8%A7%D8%A8%D8%AA-%D9%86%D9%88%D8%B9-%DB%B1-%DA%86%D9%87-%D9%85%DB%8C%D8%AF%D8%A7%D9%86%DB%8C%D8%AF**](https://www.ninisite.com/article/12666/%D8%AF%D8%B1%D8%A8%D8%A7%D8%B1%D9%87-%D8%AF%DB%8C%D8%A7%D8%A8%D8%AA-%D9%86%D9%88%D8%B9-%DB%B1-%DA%86%D9%87-%D9%85%DB%8C%D8%AF%D8%A7%D9%86%DB%8C%D8%AF)
8. [**https://www.ninisite.com/article/13274/%D8%B1%D8%A7%D9%87%D9%87%D8%A7%DB%8C-%D8%B3%D8%A7%D8%AF%D9%87-%D8%A8%D8%B1%D8%A7%DB%8C-%D9%85%D8%AF%DB%8C%D8%B1%DB%8C%D8%AA-%D8%B9%D9%88%D8%A7%D8%B1%D8%B6-%D8%AF%DB%8C%D8%A7%D8%A8%D8%AA-%D9%86%D9%88%D8%B9-1-%D9%88-2**](https://www.ninisite.com/article/13274/%D8%B1%D8%A7%D9%87%D9%87%D8%A7%DB%8C-%D8%B3%D8%A7%D8%AF%D9%87-%D8%A8%D8%B1%D8%A7%DB%8C-%D9%85%D8%AF%DB%8C%D8%B1%DB%8C%D8%AA-%D8%B9%D9%88%D8%A7%D8%B1%D8%B6-%D8%AF%DB%8C%D8%A7%D8%A8%D8%AA-%D9%86%D9%88%D8%B9-1-%D9%88-2)
9. [**https://kermany.com/what-is-diabetes/#:~:text=%D8%A7%DB%8C%D9%86%20%D8%A8%DB%8C%D9%85%D8%A7%D8%B1%DB%8C%20%D8%B2%D9%85%D8%A7%D9%86%DB%8C%20%D8%A7%D8%AA%D9%81%D8%A7%D9%82%20%D9%85%DB%8C%E2%80%8C%D8%A7%D9%81%D8%AA%D8%AF,%DA%86%D8%B4%D9%85%E2%80%8C%D9%87%D8%A7%20%D9%88%20%D8%A7%D8%B9%D8%B5%D8%A7%D8%A8%20%D8%A2%D8%B3%DB%8C%D8%A8%20%D8%A8%D8%B1%D8%B3%D8%A7%D9%86%D8%AF**](https://kermany.com/what-is-diabetes/#:~:text=%D8%A7%DB%8C%D9%86%20%D8%A8%DB%8C%D9%85%D8%A7%D8%B1%DB%8C%20%D8%B2%D9%85%D8%A7%D9%86%DB%8C%20%D8%A7%D8%AA%D9%81%D8%A7%D9%82%20%D9%85%DB%8C%E2%80%8C%D8%A7%D9%81%D8%AA%D8%AF,%DA%86%D8%B4%D9%85%E2%80%8C%D9%87%D8%A7%20%D9%88%20%D8%A7%D8%B9%D8%B5%D8%A7%D8%A8%20%D8%A2%D8%B3%DB%8C%D8%A8%20%D8%A8%D8%B1%D8%B3%D8%A7%D9%86%D8%AF)**.**
10. [**https://www.geeksforgeeks.org/machine-learning-introduction/**](https://www.geeksforgeeks.org/machine-learning-introduction/)
11. [**https://www.coursera.org/articles/python-machine-learning-library**](https://www.coursera.org/articles/python-machine-learning-library)