

دانشگاه صنعتي امیرکبیر  
(پلی­تکنیک تهران)

دانشكده ریاضی و علوم کامپیوتر

درس هوش مصنوعی و کارگاه

**پروژه پنجم**

نگارش

فاطمه گل محمدی

40113426

استاد اول

دکتر مهدی قطعي

استاد دوم

بهنام یوسفي مهر

اذر 1403

# چكيده

در این مقاله به بررسی پیش پردازش و تحلیل اکتشافی داده ها میپردازیم و از الگوریتم یادگیری ماشین برای شناسایی الگوهای موجود در داده‌ها استفاده می‌کند. در ابتدا، داده‌ها بارگذاری و پیش‌پردازش می‌شوند. این مرحله شامل شناسایی و پر کردن مقادیر گمشده، تبدیل ویژگی‌های غیر عددی به عددی، و استانداردسازی داده‌ها است. سپس، با استفاده از الگوریتم ها ، مدل آموزش داده می‌شود و عملکرد آن با استفاده از معیارهای مختلف ارزیابی می‌گردد.

واژه های کلیدی:

پیش پردازش ، تحلیل اکتشافی ، داده ، ارزیابی

|  |  |
| --- | --- |
| فهرست مطالب | صفحه |

[چكيده ‌أ](#_Toc148112500)

[۱. فصل اول مقدمه 3](#_Toc148112501)

[۲. فصل دوم پیش پردازش داده 7](#_Toc148112504)

[۱-۲- بررسی و رسیدگی به مقادیر گمشده 8](#_Toc148112505)

[۲-۲- بررسی داده ها و نرمال سازی 9](#_Toc148112506)

[۳. فصل سوم تحلیل اکتشافی داده ها 12](#_Toc148112509)

[۱-۳- الگو ها و ارتباط بین ویژگی ها 13](#_Toc148112510)

[۳-۲- شناسایی داده های پرت 14](#_Toc148112511)

[۴. فصل چهارم کد 18](#_Toc148112513)

[۱-۴- شرح و بررسی پیاده سازی کد 19](#_Toc148112514)

[۵. فصل پنجم جمع‌بندی و نتیجه‌گیری 23](#_Toc148112517)

[منابع و مراجع 25](#_Toc148112518)

# ۱. فصل اول مقدمه

**مقدمه**

امروزه با پیشرفت فناوری و افزایش آگاهی عمومی نسبت به اهمیت سلامت و تندرستی ردیاب ‌های تناسب اندام به یکی از ابزارهای مهم برای پیگیری فعالیت‌های روزانه تبدیل شده‌اند. این دستگاه‌ها با استفاده از حسگرهایی مانند شتاب‌سنج‌ها و ژیروسکوپ‌ها داده ‌های حرکتی دقیقی را جمع‌آوری می‌کنند که می‌تواند به کاربران در درک بهتر وضعیت جسمانی و بهبود عادات روزمره انها کمک کند. با توجه به اینکه این ردیاب‌ ها همواره اطلاعاتی مانند تعداد قدم‌ها، مسافت طی شده، کالری سوزانده شده و حتی کیفیت خواب را ثبت می‌کنند تحلیل این داده‌ها به یک حوزه تحقیقاتی جذاب و کاربردی تبدیل شده است.

برای تحلیل داده‌های جمع‌آوری شده اولین قدم بارگذاری و پیش‌پردازش داده ‌ها است. در این مرحله با استفاده از کتابخانه‌هایی مانند Pandas در زبان Python داده‌ها بارگذاری شده و هر گونه داده گم ‌شده یا نادرست شناسایی میشود و اصلاح می‌شود. بعد نرمال‌سازی یا استانداردسازی داده‌ ها انجام می‌شود تا مدل‌های یادگیری ماشین بتوانند به درستی بر روی آن‌ ها کار کنند. این فرآیند به بهبود دقت مدل کمک می‌کند.

پس از پیش‌پردازش داده ‌ها مرحله تحلیل اکتشافی داده‌ها (EDA) آغاز می‌شود. در این مرحله با استفاده از ابزارهای Matplotlib و Seaborn الگو های موجود در داده‌ها بررسی می‌شود. این تحلیل می‌تواند شامل شناسایی نقاط دورافتاده ، بررسی توزیع داده‌ها و تحلیل همبستگی بین متغیرها باشد. در اخر برای انتخاب مدل مناسب از کتابخانه Lazypredict استفاده میکنیم که به مقایسه سریع چندین مدل یادگیری ماشین می‌پردازد و بهترین گزینه را برای داده‌های موجود پیشنهاد می‌دهد.

# ۲. فصل دوم پیش پردازش داده

## ۱-۲- بررسی و رسیدگی به مقادیر گمشده

پیش‌پردازش داده‌ها یکی از مراحل کلیدی در تحلیل داده‌ها و یادگیری ماشین است که به منظور آماده‌سازی داده‌ها برای مدل‌سازی و تحلیل انجام می‌شود. این مرحله شامل مجموعه‌ای از تکنیک‌ها و فرآیندها است که به بهبود کیفیت داده‌ها و افزایش دقت مدل‌های پیش‌بینی کمک می‌کند. یکی از مهم‌ترین جنبه‌های پیش‌پردازش، شناسایی و مدیریت داده‌های گمشده است. داده‌های گمشده می‌توانند به دلایل مختلفی ایجاد شوند، از جمله خطاهای ثبت اطلاعات یا عدم دسترسی به برخی از ویژگی‌ها. در این مرحله، معمولاً از روش‌هایی مانند پر کردن مقادیر گمشده با میانگین یا میانه ویژگی‌ها استفاده می‌شود تا تأثیر منفی آن‌ها بر روی مدل کاهش یابد.

علاوه بر مدیریت داده‌های گمشده، تبدیل داده‌های غیر عددی به عددی نیز جزء مهمی از پیش‌پردازش است. بسیاری از الگوریتم‌های یادگیری ماشین تنها با داده‌های عددی کار می‌کنند، بنابراین لازم است که ویژگی‌های غیر عددی مانند دسته‌بندی‌ها به فرمت عددی تبدیل شوند. این کار معمولاً با استفاده از تکنیک‌هایی مانند One-Hot Encoding یا Label Encoding انجام می‌شود. این روش‌ها به ما این امکان را می‌دهند که اطلاعات مفید موجود در ویژگی‌های غیر عددی را حفظ کنیم و آن‌ها را برای مدل‌های یادگیری ماشین قابل استفاده کنیم.

## ۲-۲- بررسی داده ها و نرمال سازی

استانداردسازی داده‌ها نیز یکی دیگر از مراحل مهم پیش‌پردازش است. در بسیاری از الگوریتم‌های یادگیری ماشین، مقیاس ویژگی‌ها می‌تواند تأثیر زیادی بر عملکرد مدل داشته باشد. به همین دلیل، مقیاس‌گذاری ویژگی‌ها به یک مقیاس مشترک، مثلاً با استفاده از روش‌هایی مانند StandardScaler، می‌تواند به بهبود دقت مدل کمک کند. با این کار، ویژگی‌هایی که دارای مقادیر بزرگ‌تری هستند، تأثیر بیشتری بر روی مدل نخواهند داشت و این امر باعث می‌شود که مدل به طور عادلانه‌تر به همه ویژگی‌ها توجه کند.

در نهایت، تجزیه و تحلیل اکتشافی داده‌ها (EDA) نیز بخشی از پیش‌پردازش است که به ما کمک می‌کند تا درک بهتری از ساختار و الگوهای موجود در داده‌ها پیدا کنیم.

## 

# ۳. فصل سوم تحلیل اکتشافی داده ها

## ۱-۳- الگو ها و ارتباط بین ویژگی ها

تحلیل اکتشافی داده‌ها (EDA) یک مرحله حیاتی در فرآیند تحلیل داده‌ها است که به محققان و تحلیلگران کمک می‌کند تا درک عمیق‌تری از داده‌های خود پیدا کنند. هدف اصلی EDA شناسایی الگوها، روندها، و روابط موجود در داده‌ها است که می‌تواند به تصمیم‌گیری‌های بهتر و انتخاب مدل‌های مناسب‌تر منجر شود. این مرحله معمولاً شامل تجزیه و تحلیل توصیفی، بصری‌سازی داده‌ها، و بررسی ویژگی‌های مختلف است.

یکی از اولین گام‌ها در EDA، استفاده از آمار توصیفی برای خلاصه‌سازی ویژگی‌های کلیدی داده‌ها است. این شامل محاسبه معیارهای مختلفی مانند میانگین، میانه، و انحراف معیار برای متغیرهای عددی و توزیع فراوانی برای متغیرهای دسته‌ای می‌شود. این اطلاعات می‌تواند به ما کمک کند تا بفهمیم داده‌ها به چه صورت توزیع شده‌اند و آیا ناهنجاری یا نقاط دورافتاده‌ای وجود دارد یا خیر. به عنوان مثال، اگر میانگین و میانه یک ویژگی به طور قابل توجهی متفاوت باشند، این می‌تواند نشانه‌ای از وجود نقاط دورافتاده باشد که نیاز به بررسی بیشتری دارند.

در طول EDA، تحلیلگران به دنبال شناسایی روابط بین متغیرها هستند. این روابط می‌توانند خطی یا غیرخطی باشند و ممکن است به ما کمک کنند تا پیش‌بینی‌های بهتری انجام دهیم. برای مثال، اگر یک ویژگی خاص با برچسب هدف همبستگی بالایی داشته باشد، می‌توانیم آن را به عنوان یک ویژگی کلیدی در مدل‌سازی در نظر بگیریم. همچنین، شناسایی گروه‌های مختلف در داده‌ها می‌تواند به ما کمک کند تا مدل‌های متفاوتی برای هر گروه ایجاد کنیم.

## ۳-۲- شناسایی داده های پرت

در EDA، همچنین باید به بررسی داده‌های گمشده و ناهنجاری‌ها توجه کنیم. شناسایی و مدیریت داده‌های گمشده از اهمیت بالایی برخوردار است، زیرا این داده‌ها می‌توانند تأثیر قابل توجهی بر روی نتایج مدل‌ها داشته باشند. همچنین، ناهنجاری‌ها (outliers) باید مورد بررسی قرار گیرند، زیرا ممکن است نشان‌دهنده مشکلات در جمع‌آوری داده‌ها یا ویژگی‌های خاصی از نمونه‌ها باشند که نیاز به توجه ویژه دارند.

در نهایت، تحلیل اکتشافی داده‌ها به ما کمک می‌کند تا تصمیمات بهتری در مورد مراحل بعدی تحلیل و مدل‌سازی بگیریم. این مرحله می‌تواند به ما نشان دهد که کدام ویژگی‌ها باید در مدل‌سازی گنجانده شوند، کدام روش‌های پیش‌پردازش باید انجام شود، و چه نوع مدل‌هایی ممکن است بهترین عملکرد را داشته باشند. با استفاده از بینش‌های به دست آمده از EDA، می‌توانیم به سمت ایجاد مدل‌های پیش‌بینی دقیق‌تر و مؤثرتر حرکت کنیم.

# ۴. فصل چهارم کد

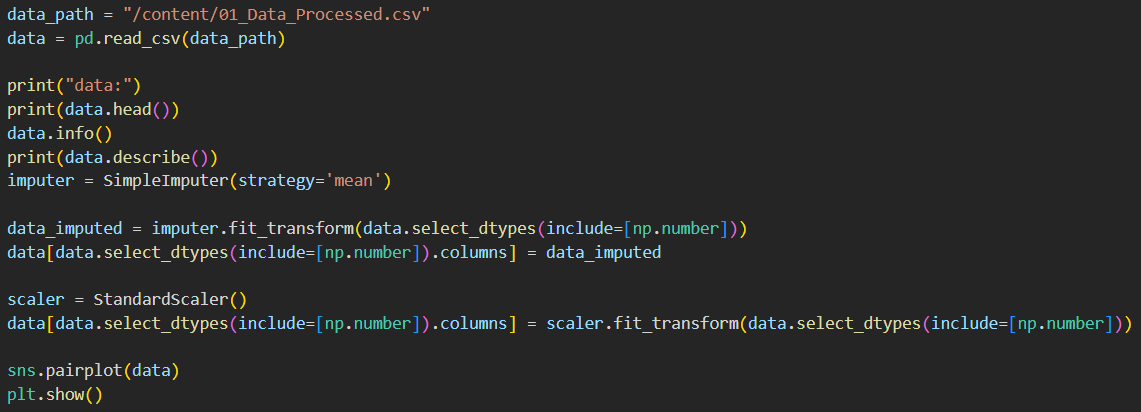
این کد برای آماده‌سازی و تحلیل اولیه‌ی داده‌ها طراحی شده است. ابتدا داده‌ها از یک فایل CSV بارگذاری شدند و ساختار آن‌ها بررسی شد تا اطلاعاتی مانند تعداد ستون‌ها، نوع داده‌ها، و مقدار داده‌های گمشده مشخص شود.

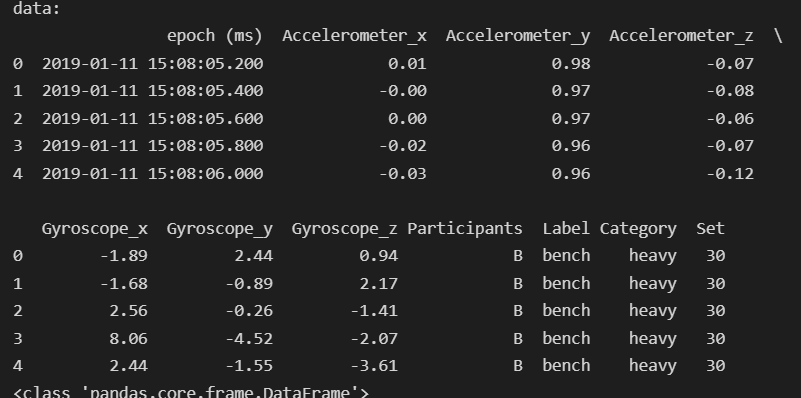
برای حل مشکل مقادیر گمشده در ستون‌های عددی، از روش جایگزینی با میانگین استفاده شد. این کار تضمین می‌کند که داده‌ها برای تحلیل و مدل‌سازی کامل و بدون خطا باشند.

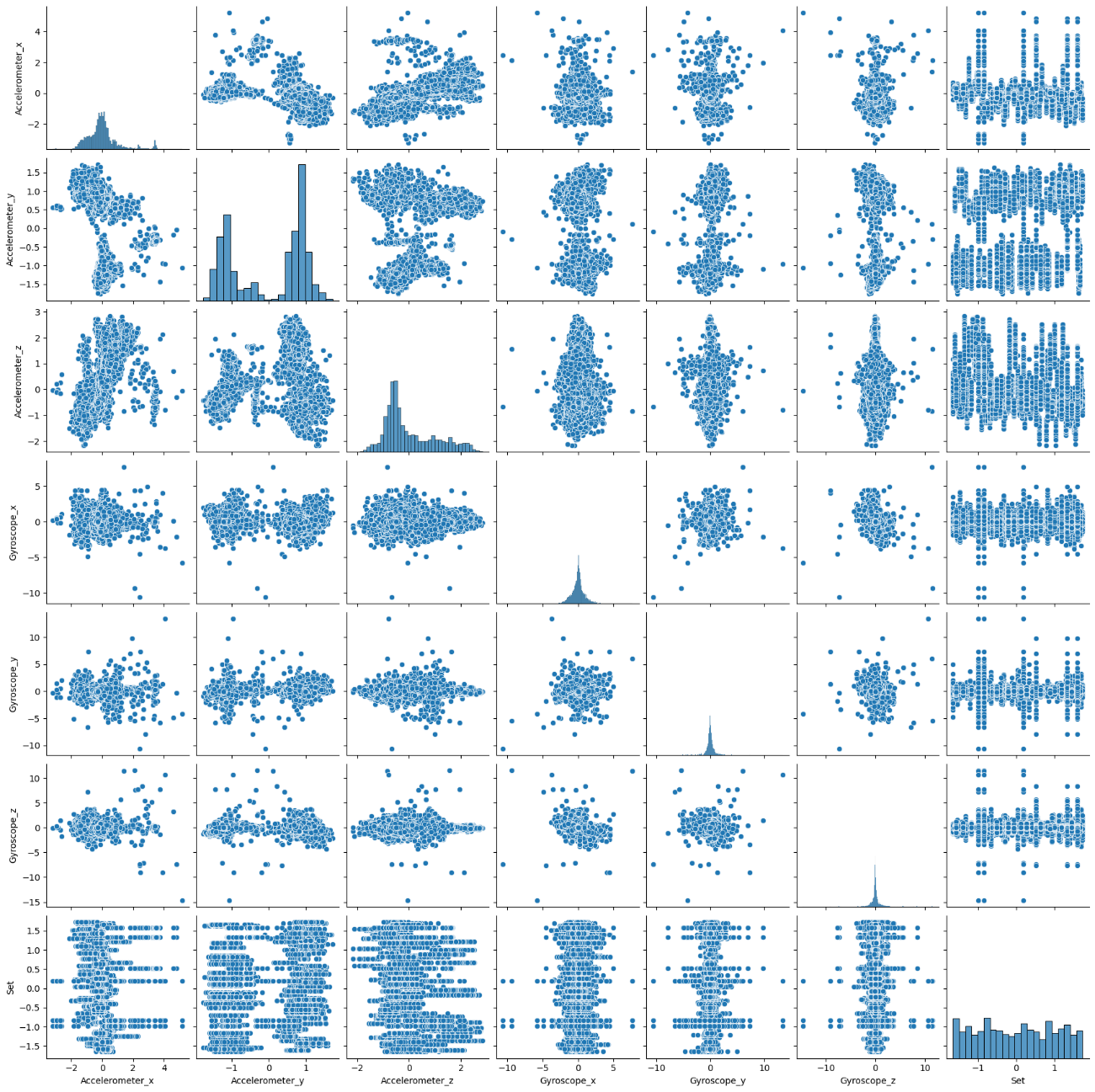
داده‌ها استانداردسازی شدند تا مقادیر عددی با میانگین صفر و انحراف معیار یک تنظیم شوند. این مرحله برای مدل‌های یادگیری ماشین بسیار اهمیت دارد زیرا مقیاس‌های متفاوت بین ویژگی‌ها ممکن است به مدل‌سازی ضعیف منجر شود.

پس از پیش‌پردازش داده‌ها، تحلیل اکتشافی داده‌ها انجام شد. با استفاده از نمودار پراکندگی جفت‌ستونی (pairplot) روابط بین متغیرهای عددی و توزیع آن‌ها بررسی شد. این ابزار به کشف الگوها، همبستگی‌ها، و نقاط پرت کمک می‌کند.

این مراحل به تحلیل بهتر داده‌ها و درک ویژگی‌های اصلی آن‌ها پیش از ورود به مراحل پیشرفته‌تر مدل‌سازی کمک می‌کنند.







اول: مونه‌ای از ۵ ردیف اول داده‌ها را نشان می‌دهد که برای شناسایی سریع ساختار و نوع داده‌ها مفید است.

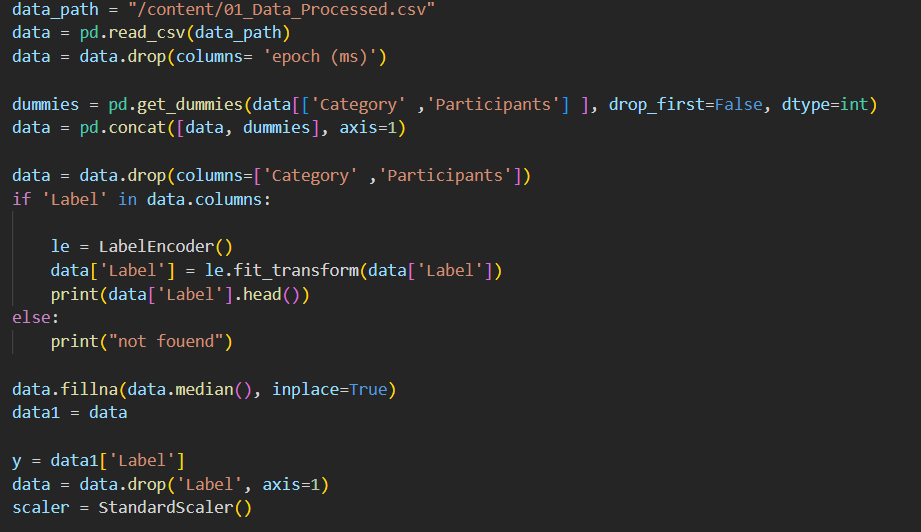
دوم: این نمودار نشان‌دهنده روابط دوبه‌دوی میان تمام ستون‌های عددی است و پراکندگی را نشان میدهد.

این کد یک فرآیند کامل پیش‌پردازش داده‌ها، مهندسی ویژگی‌ها، و مدل‌سازی را انجام می‌دهد. ابتدا داده‌ها بارگذاری و یک ستون غیرضروری حذف شد.

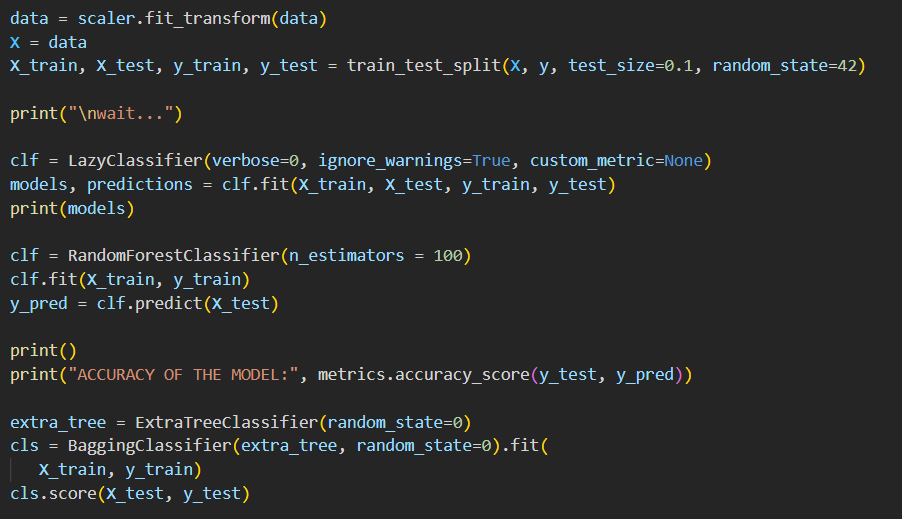
سپس متغیرهای اسمی به مقادیر عددی دودویی تبدیل شدند و ستون‌های اصلی آن‌ها حذف شدند.

متغیر هدف (Label) اگر موجود باشد، مقادیر متنی به اعداد تبدیل شوند.

مقادیر گمشده با مقدار میانه‌ی ستون‌ها جایگزین شد و سپس داده‌ها استانداردسازی شدند تا مقیاس ویژگی‌های عددی یکنواخت شود. مجموعه داده به دو بخش آموزشی و آزمایشی تقسیم شد تا داده‌ها برای مدل‌سازی آماده شوند.



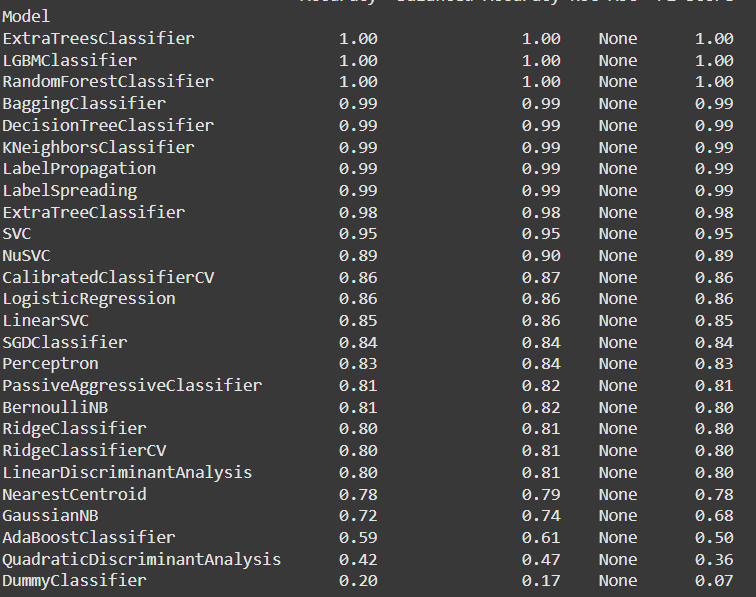
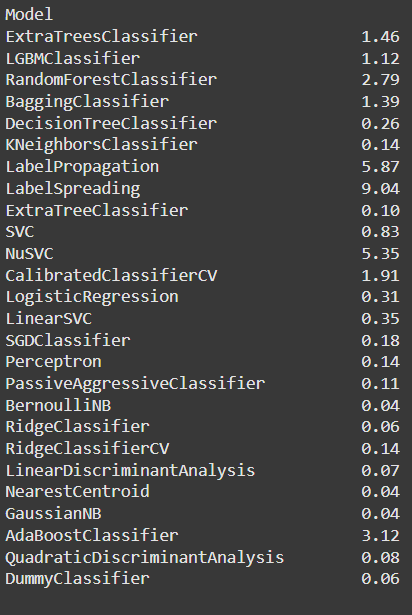
در مرحله‌ی مدل‌سازی، ابتدا با استفاده از ابزار لیزی کلسیفایر عملکرد مدل‌های مختلف یادگیری ماشین روی داده‌ها مقایسه شد.

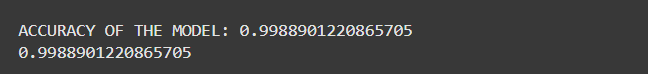


سپس مد(Random Forest) با ۱۰۰ درخت تصمیم آموزش داده شد و دقت مدل روی داده‌های آزمایشی محاسبه شد.

در اخرعملکرد این مدل‌ها با معیارهایی مانند دقت (Accuracy) ارزیابی شد.

این فرآیند نشان می‌دهد که چگونه می‌توان داده‌ها را از پیش‌پردازش تا مدل‌سازی هدایت کرد و بهترین مدل را انتخاب کرد.





# ۵. فصل پنجم جمع‌بندی و نتیجه‌گیری

**جمع‌بندی و نتیجه‌گیری**

در نهایت میتوانیم بگوییم که ردیاب‌های تناسب اندام و تحلیل داده‌های آن‌ها ابزارهای قدرتمندی برای بهبود سلامت و کیفیت زندگی افراد هستند. با استفاده از کدهای تحلیلی و الگوریتم‌های یادگیری ماشین، می‌توان به بینش‌های ارزشمندی دست یافت که به کاربران در دستیابی به اهداف سلامتی و تناسب اندام کمک می‌کند. با این حال، لازم است چالش‌ها و محدودیت‌های این فناوری‌ها نیز مورد توجه قرار گیرد تا از حداکثر پتانسیل آن‌ها بهره ‌برداری شود.

**منابع و مراجع**

["faradars"](file:///C:\Users\Asus\AppData\Roaming\Microsoft\Word\AI_HW1.docx)

["geeksforgeeks"](file:///C:\Users\Asus\AppData\Roaming\Microsoft\Word\AI_HW1.docx)

["researchgate"](file:///C:\Users\Asus\AppData\Roaming\Microsoft\Word\AI_HW1.docx)

["logistic"](file:///C:\Users\Asus\AppData\Roaming\Microsoft\Word\AI_HW1.docx)

["gardner"](file:///C:\Users\Asus\AppData\Roaming\Microsoft\Word\AI_HW1.docx)

["wikipedia"](file:///C:\Users\Asus\AppData\Roaming\Microsoft\Word\AI_HW1.docx)

["smart"](file:///C:\Users\Asus\AppData\Roaming\Microsoft\Word\AI_HW1.docx)

["ijert"](file:///C:\Users\Asus\AppData\Roaming\Microsoft\Word\AI_HW1.docx)