دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

گزارش تمرین دوم داده‌کاوی

استخراج ویژگی مبتنی‌بر استقلال خطی و تحلیل تأثیر آن بر فرآیند بهینه‌سازی توابع زیان

نگارش
زهرا براتی

استاد درس
دکتر مهدی قطعی

تدریسیار
آقای بهنام یوسفی‌مهر

آبان ۱۴۰۴

چکیده

در این مطالعه، سه مسئله‌ی اصلی در یادگیری ماشین شامل رگرسیون، خوشه‌بندی و طبقه‌بندی با هدف بررسی تأثیر روش‌های کاهش بعد و انتخاب ویژگی بر عملکرد مدل‌ها مورد تحلیل قرار گرفتند.

در مسئله‌ی رگرسیون، نتایج نشان داد که کاهش بعد از طریق PCA موجب افزایش پایداری ضرایب در مدل Linear Regression و بهبود سرعت همگرایی در مدل SGDRegressor می‌شود. در مسئله‌ی خوشه‌بندی با الگوریتم KMeans، استفاده از داده‌های کاهش‌یافته ضمن کاهش زمان محاسبه، منجر به بهبود شاخص‌های کیفیت مانند سیلوییت گردید. در مسئله‌ی طبقه‌بندی نیز، مدل KNeighborsClassifier با داده‌های PCA زمان پیش‌بینی کمتری داشت، در حالی که در مدل RandomForestClassifier، روش انتخاب ویژگی عملکرد مشابه یا اندکی بهتر از PCA ارائه داد.

در مجموع، نتایج نشان داد که هیچ‌یک از دو رویکرد بر دیگری برتری مطلق ندارند و کارایی آن‌ها به نوع مدل و ساختار داده بستگی دارد. روش PCA برای مدل‌های خطی و مبتنی بر فاصله مناسب‌تر است، درحالی‌که روش SelectKBest در مدل‌های درختی و غیرخطی عملکرد بهتری دارد.

ترکیب این دو رویکرد می‌تواند داده‌هایی بهینه، پایدار و کم‌بعد فراهم کند و دقت و کارایی الگوریتم‌های یادگیری را به شکل محسوسی افزایش دهد.

واژه‌های کلیدی:

کاهش بعد، استخراج ویژگی، انتخاب ویژگی، تحلیل مؤلفه‌های اصلی، انتخاب ویژگی

فهرست مطالب

صفحه

فصل اول

مقدمه

مقدمه

در این پروژه، به بررسی نقش استقلال خطی[1] و هم‌خطی[2] در عملکرد مدل‌های یادگیری ماشین و فرآیندهای بهینه‌سازی پرداخته می‌شود.

هدف اصلی، تحلیل تأثیر وجود هم‌بستگی یا وابستگی خطی میان ویژگی‌های ورودی بر پایداری ضرایب مدل، سرعت همگرایی الگوریتم‌های مبتنی بر گرادیان و دقت نهایی مدل‌های پیش‌بینی و طبقه‌بندی است.

از سوی دیگر، با بهره‌گیری از روش‌های استخراج ویژگی[3] و انتخاب ویژگی[4]، تلاش می‌شود مجموعه‌ای از ویژگی‌های بهینه و مستقل‌تر ایجاد شود تا مدل نهایی ضمن کاهش پیچیدگی، از نظر پایداری و سرعت آموزش نیز عملکرد بهتری داشته باشد.

[1] Linear Independence

[2] Collinearity

[3] Feature Extraction

[4] Feature Selection

فصل دوم

انتخاب سه مسئله داده‌کاوی

انتخاب سه مسئله داده‌کاوی

در این بخش، سه مجموعه داده‌ی پیشنهادی بارگیری شده تا فرآیندهای تحلیل، استخراج و انتخاب ویژگی بر روی آن‌ها انجام شود.

هر مجموعه‌داده به‌گونه‌ای انتخاب شده است که یکی از سه نوع مسأله‌ی اصلی در داده‌کاوی را پوشش دهد: رگرسیون[5]، طبقه‌بندی[6] و خوشه‌بندی[7]

در ادامه، هر مجموعه داده معرفی و ویژگی‌های آماری آن بررسی می‌شود.

۲-۱- مجموعه داده‌ی Wisconsin Breast Cancer

در این مجموعه داده پس از حذف مقادیر گمشده، تعداد کل نمونه‌ها به ۶۸۳ ردیف کاهش یافته است. این داده شامل ۹ ویژگی عددی اصلی و یک برچسب (Class) است که دو مقدار ۲ (خوش‌خیم) و ۴ (بدخیم) را می‌پذیرد. ویژگی‌های استخراج‌شده حاصل از تصاویر سلول‌های بافت پستان است که برای تشخیص خوش‌خیم یا بدخیم بودن تومور استفاده می‌شود.

هر نمونه نمایانگر یک بیمار و هر ویژگی نشان‌دهنده‌ی یکی از خصوصیات بافت سلولی (مانند ضخامت کلوخه، یکنواختی اندازه و شکل سلول‌ها و غیره) است.

خلاصه‌ی آماری این مجموعه داده نشان می‌دهد که هر ویژگی در مقیاس عددی ۱ تا ۱۰ مقداردهی شده است. میانگین اغلب ویژگی‌ها در حدود ۳ تا ۴ قرار دارد که بیانگر تمرکز بیشتر داده‌ها در نواحی پایین بازه است.

انحراف معیار بالا در ویژگی‌هایی نظیر Clump Thickness، Uniformity of Cell Size و Uniformity of Cell Shape نشان‌دهنده‌ی تنوع زیاد در مقادیر این متغیرها میان بیماران مختلف است. از طرفی، ویژگی‌هایی مانند Mitoses و Normal Nucleoli دارای میانگین نزدیک به مقدار حداقل هستند، که نشان می‌دهد اکثر نمونه‌ها مقادیر کوچکی در این خصوصیات دارند و تنها درصد کمی از بیماران

[5] Regression

[6] Classification

[7] Clustering

دارای مقادیر بزرگ‌ترند. همچنین ویژگی Bare Nuclei دارای بیشترین انحراف معیار (۳.۶۴) است و می‌تواند نقش مهمی در تفکیک داده‌های خوش‌خیم و بدخیم داشته باشد.

با توجه به مقادیر حداقل و حداکثر (۱ تا ۱۰)، می‌توان نتیجه گرفت که همه‌ی ویژگی‌ها در یک مقیاس یکسان قرار دارند و نیازی به نرمال‌سازی شدید وجود ندارد، اما بررسی هم‌بستگی بین آن‌ها ضروری است، زیرا شباهت زیاد بین ویژگی‌هایی مانند Uniformity of Cell Size و Uniformity of Cell Shape احتمال وجود هم‌خطی را بالا می‌برد.

در ستون برچسب (Class) میانگین برابر با ۲.۷ است که با توجه به مقادیر مجاز (۲ = خوش‌خیم، ۴ = بدخیم)، نشان‌دهنده‌ی تعداد بیشتر موارد خوش‌خیم در مقایسه با بدخیم است. بنابراین، داده‌ها نامتوازن[8] هستند و باید این نکته در آموزش مدل طبقه‌بندی در نظر گرفته شود.

۲-۲- مجموعه داده‌ی Boston Housing

این مجموعه داده با ۵۰۶ نمونه اولیه یکی از شناخته‌شده‌ترین داده‌های حوزه‌ی یادگیری ماشین است و برای پیش‌بینی قیمت متوسط خانه‌ها در مناطق مختلف شهر بوستون به‌کار می‌رود.

لازم به ذکر است تعداد نمونه‌ها پس از پاک‌سازی به ۳۹۴ ردیف کاهش یافته است. این مجموعه شامل ۱۳ ویژگی عددی به‌همراه برچسب (MEDV) است که بیانگر قیمت متوسط خانه‌ها (به هزار دلار) است.

ویژگی‌ها شامل شاخص‌های اجتماعی، اقتصادی و زیست‌محیطی هستند (مانند نسبت جمعیت دانش‌آموزان، نرخ جرم، فاصله تا مراکز شغلی، و غیره).

در این مجموعه داده، برچسب (MEDV) میانگین ۲۲.۳۶ هزار دلار دارد و دامنه‌ی تغییرات آن از ۵ تا ۵۰ هزار دلار است. انحراف معیار ۹.۱۴ نشان می‌دهد که اختلاف قابل‌توجهی میان مناطق مختلف شهر وجود دارد.

در بین ویژگی‌ها، میانگین تعداد اتاق‌ها (RM) دارای میانگین ۶.۲۸ و انحراف معیار ۰.۷ است که پراکندگی نسبتاً کمی دارد. درصد جمعیت کم‌درآمد (LSTAT) دارای میانگین ۱۲.۷۷ و انحراف معیار

[8] imbalanced

۷.۳۱ است که نشان‌دهنده‌ی تنوع زیاد در وضعیت اقتصادی مناطق است.

ویژگی غلظت اکسید نیتروژن (NOX) میانگین ۰.۵۵ دارد که محدوده‌ی کوچکی را پوشش می‌دهد، درحالی‌که نرخ جرم (CRIM) از ۰.۰۱ تا حدود ۸۹ تغییر می‌کند و دارای انحراف معیار بسیار بزرگ (۹.۲) است. این موضوع وجود پراکندگی زیاد و داده‌های پرت[9] را در این ویژگی تأیید می‌کند.

همچنین مشاهده می‌شود که برخی ویژگی‌ها مانند فاصله تا مراکز شغلی (DIS) و دسترسی به بزرگراه‌ها (RAD) توزیع‌های ناهمگون دارند و ممکن است بر هم‌خطی تأثیر بگذارند.

به‌طور کلی، داده‌ها نیاز به نرمال‌سازی دارند تا تأثیر مقیاس‌های متفاوت بر مدل کاهش یابد.

۲-۳- مجموعه داده‌ی UCI Iris

مجموعه داده‌ی معروف Iris با ۱۵۰ نمونه شامل اندازه‌گیری‌های گلبرگ و کاسبرگ سه گونه‌ی مختلف گل زنبق (Setosa، Versicolor و Virginica) است.

در این پروژه، از آن برای آزمایش روش‌های کاهش بعد و تحلیل خوشه‌بندی بدون استفاده از برچسب‌ها استفاده می‌شود.

میانگین طول و عرض کاسبرگ و گلبرگ در محدوده‌های نزدیک به یکدیگر قرار دارد، اما انحراف معیار PetalLengthCm (۱.۷۶) نسبت به سایر ویژگی‌ها بیشتر است و نشان‌دهنده‌ی پراکندگی بالاتر این متغیر است.

در مقابل، ویژگی SepalWidthCm با انحراف معیار ۰.۴۳ پراکندگی کمتری دارد.

با توجه به دامنه‌ی مقادیر (حداقل تا حداکثر)، مشخص است که داده‌ها مقیاس‌های متفاوتی دارند (برای مثال طول گلبرگ از ۱ تا ۶.۹ و عرض آن از ۰.۱ تا ۲.۵). بنابراین، پیش از اجرای الگوریتم‌های مبتنی بر فاصله نظیر KNN، استانداردسازی داده‌ها ضروری است تا ویژگی‌ها وزن مساوی در محاسبات داشته باشند.

از آنجا که در این مجموعه‌داده هر سه گونه‌ی گل دارای ویژگی‌های نزدیک اما قابل تفکیک‌اند، انتظار می‌رود روش‌هایی مانند PCA بتوانند جدایی نسبی گونه‌ها را در فضای دوبعدی به‌خوبی نمایش دهند.

[9] Outlier

۲-۴- جمع‌بندی

به‌طور کلی می‌توان گفت داده‌های Breast Cancer دارای مقیاس یکسان ولی هم‌بستگی بالا بین برخی ویژگی‌ها است. داده‌های Boston Housing دامنه‌ی مقادیر بسیار متفاوت و داده‌های پرت دارند و نیاز به نرمال‌سازی دارد. داده‌های UCI Iris توزیع‌های منظم‌تری دارد ولی اختلاف مقیاس بین ویژگی‌ها قابل توجه است.

فصل سوم

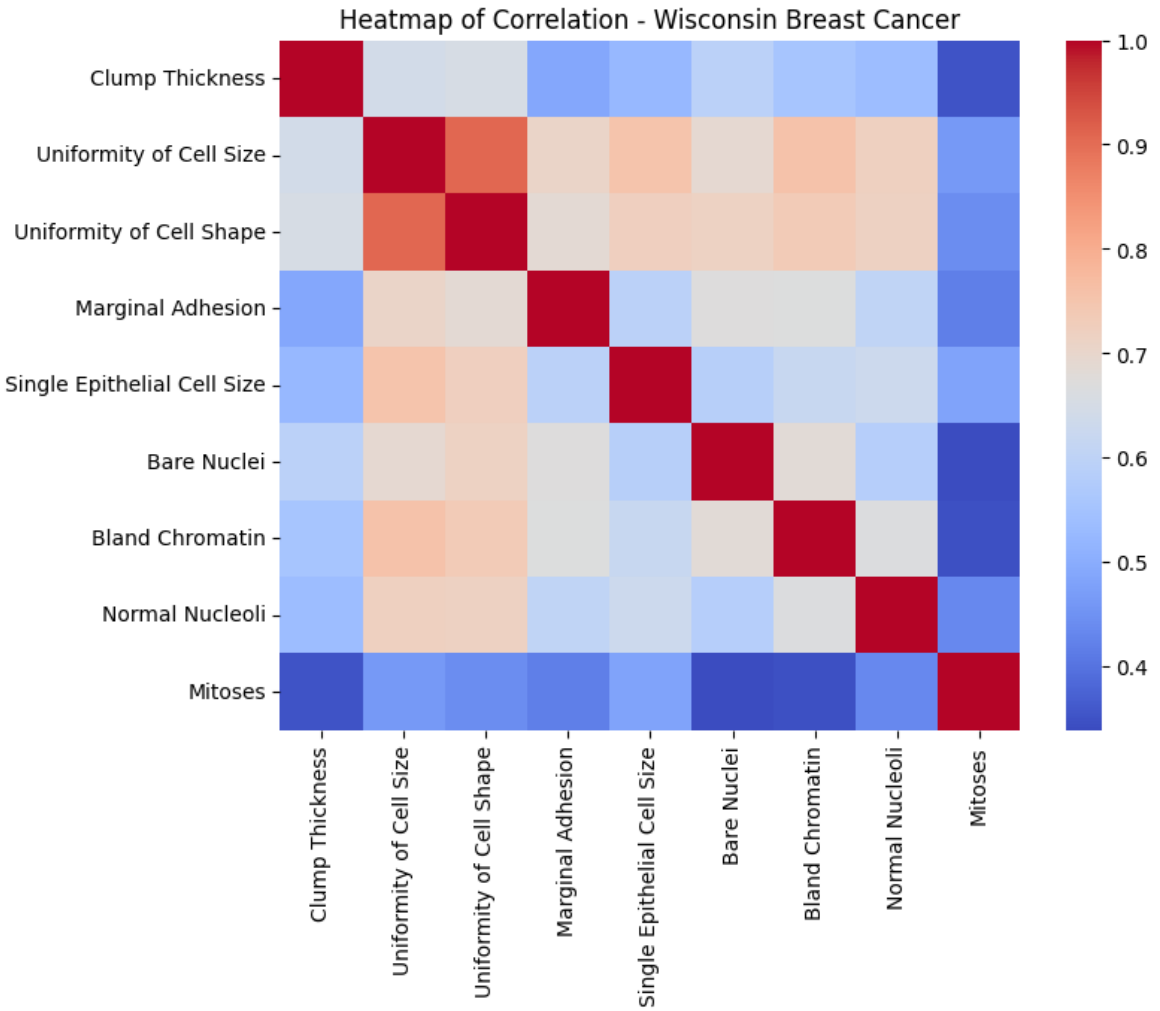
بررسی هم خطی اولیه داده‌ها

بررسی هم خطی اولیه داده‌ها

در این بخش، برای هر مجموعه‌داده، ماتریس همبستگی[10] بین ویژگی‌ها محاسبه و به‌صورت Heatmap نمایش داده شده است.

هدف از این تحلیل، شناسایی میزان وابستگی خطی بین متغیرها و تشخیص وجود هم‌خطی است؛ چرا که وجود هم‌خطی بالا می‌تواند موجب ناپایداری در مدل‌سازی و افزایش واریانس ضرایب شود.

۳-۱- مجموعه داده‌ی Wisconsin Breast Cancer



[10] Correlation Matrix

نمودار Heatmap مربوط به این داده‌ها نشان می‌دهد که برخی از ویژگی‌ها دارای همبستگی بالا (بیش از ۰.۷) هستند.

بین ویژگی‌های زیر وابستگی بسیار قوی مشاهده می‌شود:

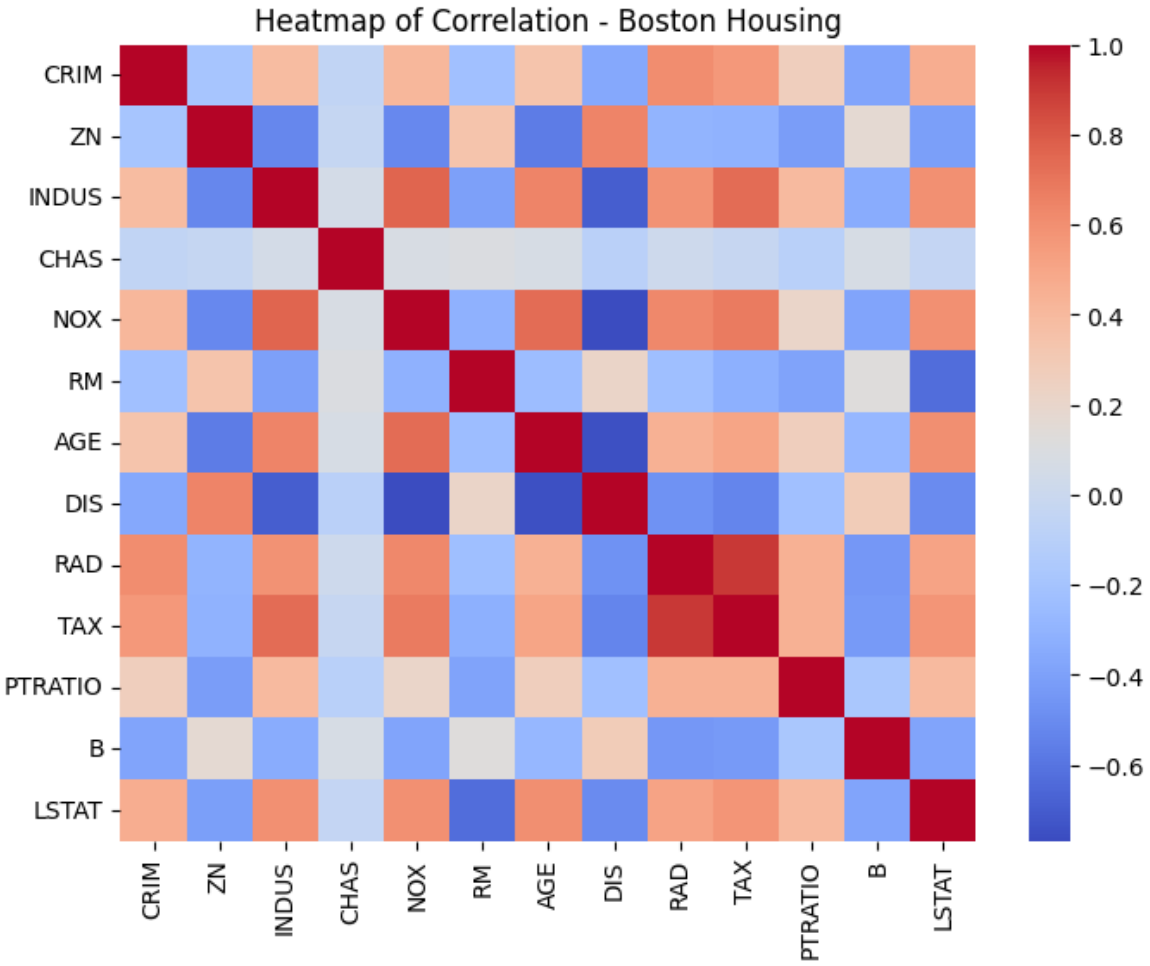
- Uniformity of Cell Size و Uniformity of Cell Shape با ضریب همبستگی حدود ۰.۹۱
- Uniformity of Cell Size و Bland Chromatin با ضریب ۰.۷۶
- Bare Nuclei و Bland Chromatin با ضریب ۰.۶۸

وجود چنین مقادیر بالایی از همبستگی نشان می‌دهد که بسیاری از ویژگی‌ها اطلاعات مشابهی دارند و ممکن است باعث افزونگی[11] در داده شوند.

به‌طور کلی، داده‌های Breast Cancer به‌ویژه در گروه ویژگی‌های مرتبط با یکنواختی و ساختار سلولی مستعد چندهم‌خطی هستند.

[11] Redundancy

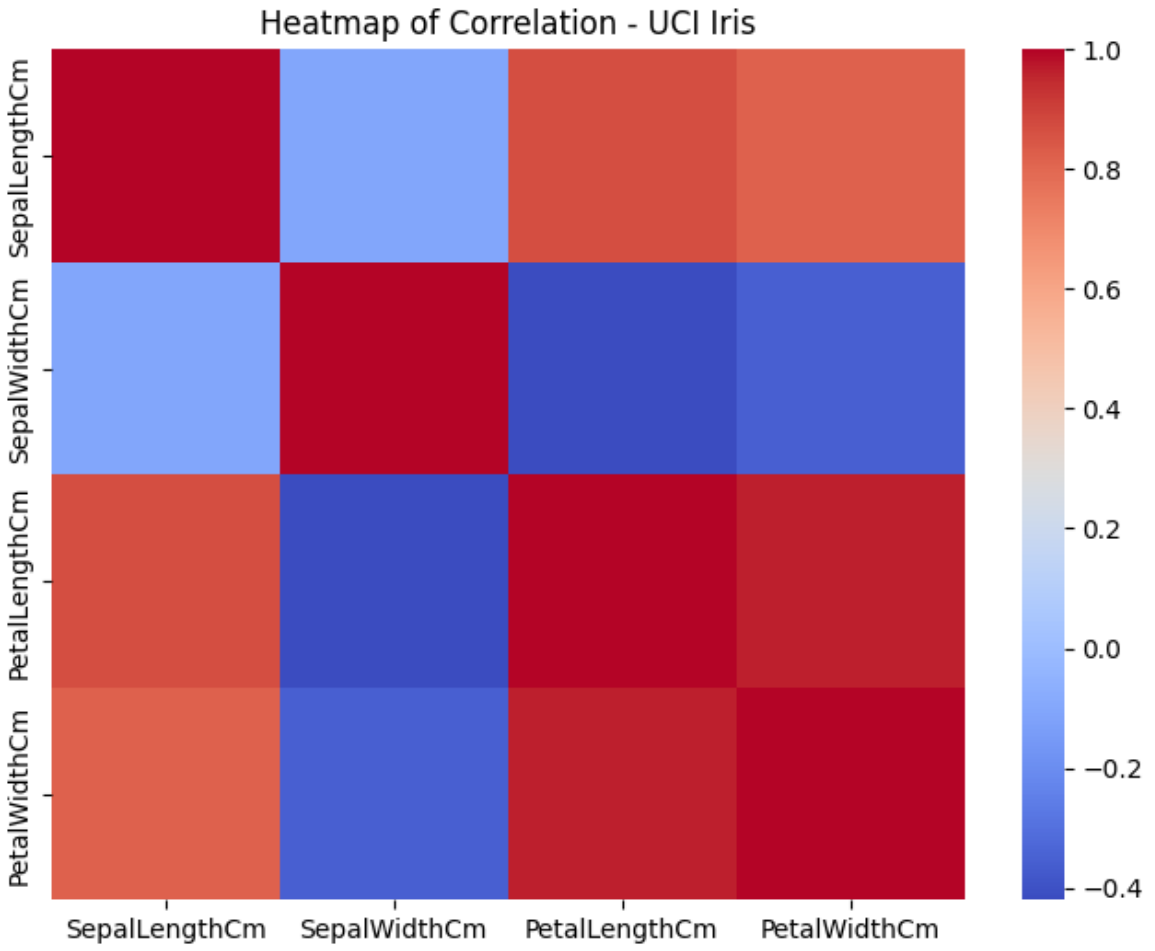
۳-۲- مجموعه داده‌ی Boston Housing



در Heatmap مربوط به داده‌های Boston Housing، چندین رابطه قوی میان متغیرها مشاهده می‌شود:

- INDUS با NOX (ضریب ۰.۷۶) و TAX (ضریب ۰.۷۳) همبستگی بالایی دارد.
- RAD با TAX دارای همبستگی بسیار بالا (۰.۹) است، که نشان‌دهنده‌ی هم‌پوشانی اطلاعات بین این دو متغیر است.
- AGE با NOX و INDUS همبستگی مثبت نسبتاً بالا (حدود ۰.۷) دارد.
- از سوی دیگر، RM با LSTAT دارای همبستگی منفی قوی (حدود۰.۶۴-) است؛ یعنی هرچه تعداد اتاق‌ها بیشتر باشد، درصد جمعیت کم‌درآمد کمتر است.

۳-۳- مجموعه داده‌ی UCI Iris



در نمودار همبستگی داده‌های UCI Iris ، ساختار هم‌خطی ساده‌تر ولی مشخص است.

- بین PetalLengthCm و PetalWidthCm ضریب همبستگی بسیار بالا (۰.۹۶) وجود دارد، که بیانگر وابستگی شدید این دو ویژگی است.
- SepalLengthCm نیز با PetalLengthCm و PetalWidthCm همبستگی مثبت بالا (حدود ۰.۸۷ و ۰.۸۲) دارد.
- در مقابل SepalWidthCm با سایر ویژگی‌ها همبستگی منفی دارد (در حدود ۰.۳- تا ۰.۴-).

نتیجه‌ی این تحلیل نشان می‌دهد که اگرچه داده‌های UCI Iris از نظر تعداد ویژگی‌ها کوچک هستند، اما بین برخی از متغیرها هم‌پوشانی قابل توجهی وجود دارد.

۳-۴- جمع‌بندی

بر اساس نتایج به‌دست‌آمده از ماتریس همبستگی و نمودارهای Heatmap، هر سه مجموعه‌داده دارای درجات متفاوتی از وابستگی خطی[۱۲] میان ویژگی‌ها هستند.

در مجموعه داده‌ی Breast Cancer، وابستگی‌های درونی بسیار قوی بین ویژگی‌های مرتبط با یکنواختی سلول‌ها مشاهده شد. این موضوع نشان می‌دهد که ویژگی‌ها اطلاعات تکراری و هم‌پوشان دارند. چنین ساختاری می‌تواند در مدل‌های رگرسیونی منجر به چندهم‌خطی و در نتیجه ناپایداری ضرایب شود.

در مجموعه داده‌ی Boston Housing، میزان هم‌خطی از سایر مجموعه‌ها بیشتر است. متغیرهایی مانند RAD و TAX با ضریب همبستگی بسیار بالا (۰.۹) به‌شدت با یکدیگر هم‌بسته‌اند. همچنین، INDUS و NOX نیز با ضریب حدود ۰.۷۵ ارتباط قوی دارند. چنین وابستگی‌هایی نشان می‌دهد که چندین ویژگی در واقع بیانگر جنبه‌های مشابهی از ساختار شهری‌اند. در نتیجه، برای مدل‌سازی دقیق‌تر، نیاز به کاهش بعد و حذف افزونگی اطلاعاتی وجود دارد.

در مجموعه داده‌ی Iris نیز گرچه ابعاد داده کم‌تر است، اما دو ویژگی PetalLengthCm و PetalWidthCm ضریب همبستگی بسیار بالا (۰.۹۶) دارند. بنابراین، حتی در این داده‌ی ساده نیز هم‌پوشانی اطلاعات دیده می‌شود و کاهش بعد می‌تواند به بهبود نمایش داده‌ها در فضای دوبعدی کمک کند.

[12] Linear Dependence

فصل چهارم

استخراج ویژگی‌های مستقل

استخراج ویژگی های مستقل

در این مرحله داده‌ها پس از استانداردسازی با سه روش خواسته شده بررسی شدند.

- PCA: حداکثرسازی واریانسِ توضیح‌داده‌شده و ایجاد مؤلفه‌های متعامد؛
- ICA: استخراج مؤلفه‌های مستقل آماری؛
- SVD: تجزیه‌ی مقدار تکین و تقریب کم‌رتبه.

معیار انتخاب تعداد مؤلفه‌ها در PCA رسیدن به حداقل ۹۵٪ واریانس تجمعی بود؛ در ICA و SVD نیز برای امکان مقایسه، همان تعداد مؤلفه استفاده شد.

در اپامه نتایج حاصل بررسی خواهد شد.

۴-۱- مجموعه داده‌ی Wisconsin Breast Cancer

- PCA: با ۷ مؤلفه، حدود ۹۶/۱۲٪ از واریانس کل توضیح داده شده.
- ICA: خروجی با ابعاد (۶۸۳,۷) و ۷ مؤلفه‌ی مستقل آماری.
- SVD: خروجی با ابعاد (۶۸۳,۷) و هم‌اندازه با PCA بر مبنای مقادیر تکین.

ساختار داده نسبتاً دارای بعد بالا است و اطلاعات معنادار در بیش از چند محور فشرده شده است؛ این مسئله با هم‌بستگی‌های متعدد بین ویژگی‌های «Uniformity» سازگار است و نشان می‌دهد کاهش بعد شدید (مثلاً ۲ یا ۳ بعد) می‌تواند اطلاعات مهم را حذف کند.

۴-۲- مجموعه داده‌ی Boston Housing

- PCA: با ۹ مؤلفه، حدود ۹۵/۱۸٪ از واریانس کل توضیح داده شده.
- ICA: خروجی با ابعاد (۳۹۴,۹) و ۹ مؤلفه‌ی مستقل آماری.
- SVD: خروجی با ابعاد (۳۹۴,۹) و هم‌اندازه با PCA بر مبنای مقادیر تکین.

این مجموعه داده بعد بالاتری نسبت به دو مجموعه‌ی دیگر دارد؛ پراکندگی بالا و هم‌بستگی‌های قوی، همان‌طور که پیش‌تر توضیح داده شد، باعث می‌شود برای حفظ اطلاعات لازم باشد تعداد مؤلفه‌ها بیشتر نگه داشته شود.

۴-۳- مجموعه داده‌ی UCI Iris

- PCA: با ۲ مؤلفه، حدود ۹۵/۸٪ از واریانس کل توضیح داده شده.
- ICA: خروجی با ابعاد (۱۵۰,۲) و ۲ مؤلفه‌ی مستقل آماری.
- SVD: خروجی با ابعاد (۱۵۰,۲) و هم‌اندازه با PCA بر مبنای مقادیر تکین.

ساختار داده تقریباً دوبعدی است؛ دو مؤلفه‌ی اول تقریباً تمام تنوع را پوشش می‌دهند. این نتیجه با همبستگی بسیار بالای PetalLength–PetalWidth و تفکیک کلاس‌ها در فضاهای دوبعدی سازگار است.

۴-۴- جمع‌بندی

نتایج حاصل از اجرای الگوریتم‌های PCA، ICA و SVD بر سه مجموعه‌داده نشان داد که ساختار درونی و بعد مؤثر داده‌ها با یکدیگر تفاوت چشمگیری دارد.

در مجموعه داده‌ی Wisconsin Breast Cancer، وجود ویژگی‌های متعدد و هم‌بسته باعث شد که برای توضیح بیش از ۹۵٪ از واریانس، به ۷ مؤلفه اصلی نیاز باشد. این نتیجه تأیید می‌کند که داده، ساختاری نسبتاً پیچیده دارد و اطلاعات در ابعاد متعددی پخش شده است. به همین دلیل، کاهش بعد شدید می‌تواند موجب از دست رفتن بخشی از اطلاعات مؤثر شود.

مجموعه داده‌ی Boston Housing برای رسیدن به همین سطح از واریانس به ۹ مؤلفه نیاز داشت. این امر نشان می‌دهد که این داده از نظر آماری حتی وابستگی‌های درونی بیشتری دارد و برای نمایش کامل روابط بین متغیرهای اقتصادی، اجتماعی و زیست‌محیطی باید ابعاد بیشتری حفظ شوند. در نتیجه، کاهش بعد تنها تا حدود متوسط (۸ تا ۱۰ بعد) منطقی است.

در مجموعه داده‌ی UCI Iris، تنها ۲ مؤلفه توانستند بیش از ۹۵٪ از واریانس را توضیح دهند؛ به بیان

دیگر، ساختار داده به‌صورت طبیعی دوبعدی است و همین دو مؤلفه برای تفکیک گونه‌های گل کافی‌اند. این رفتار، سادگی و نظم درونی داده را نسبت به دو مجموعه‌ی دیگر نشان می‌دهد.

فصل پنجم

اعمال روش‌های پایه انتخاب ویژگی

اعمال روش های پایه انتخاب ویژگی

در این مرحله، هدف انتخاب زیرمجموعه‌ای از مهم‌ترین ویژگی‌ها است تا ضمن کاهش ابعاد، دقت مدل افزایش و پیچیدگی آن کاهش یابد.

برای این منظور دو روش مورد استفاده قرار گرفت:

۱. SelectKBest بر اساس آزمون آماری f_regression برای سنجش قدرت ارتباط هر ویژگی با متغیر هدف

۲. RFE (Recursive Feature Elimination) با مدل پایه‌ی LinearRegression برای حذف تدریجی ویژگی‌های کم‌اهمیت

در هر مجموعه داده، تا سقف سه ویژگی برتر انتخاب شد تا مقایسه‌ی نتایج بین روش‌ها ساده‌تر شود.

در ادامه نتایج هر مجموعه داده را بررسی خواهیم کرد.

۵-۱- مجموعه داده‌ی Wisconsin Breast Cancer

در نتایج حاصل برای این مجموعه داده داریم:

- SelectKBest:
  ['Uniformity of Cell Size', 'Uniformity of Cell Shape', 'Bare Nuclei']
- RFE:
  ['Clump Thickness', 'Uniformity of Cell Size', 'Bare Nuclei']

ویژگی‌های منتخب در هر دو روش، مربوط به یکنواختی و ساختار سلولی هستند؛ این نتایج با یافته‌های مرحله‌ی همبستگی سازگار است.

حضور مکرر متغیرهای «Uniformity of Cell Size» و «Bare Nuclei» نشان می‌دهد این دو عامل بیشترین قدرت تمایز بین نمونه‌های خوش‌خیم و بدخیم را دارند.

۵-۲- مجموعه داده‌ی Boston Housing

در نتایج حاصل برای این مجموعه داده داریم:

- SelectKBest:
  ['RM', 'PTRATIO', 'LSTAT']
- RFE:
  ['CHAS', 'NOX', 'RM']

هر دو روش ویژگی RM (میانگین تعداد اتاق‌ها) را به‌طور مشترک انتخاب کرده‌اند، که در پیش‌بینی قیمت خانه دور از انتظار نیست.

ویژگی LSTAT (درصد جمعیت کم‌درآمد) نیز در روش آماری SelectKBest امتیاز بالایی کسب کرده، در حالی‌که RFE ویژگی‌های زیست‌محیطی (NOX و CHAS) را برگزیده است.

بنابراین، ترکیب این دو رویکرد تصویری متوازن از اهمیت متغیرهای اقتصادی و محیطی ارائه می‌دهد.

۵-۳- مجموعه داده‌ی UCI Iris

در نتایج حاصل برای این مجموعه داده داریم:

- SelectKBest:
  ['SepalLengthCm', 'PetalLengthCm', 'PetalWidthCm']
- RFE:
  ['SepalLengthCm', 'PetalLengthCm', 'PetalWidthCm']

هر دو روش دقیقاً سه ویژگی مشابه را انتخاب کرده‌اند.

این هم‌پوشانی کامل نشان می‌دهد که ویژگی‌های مرتبط با گلبرگ‌ها دارای بیشترین قدرت تفکیک بین گونه‌های مختلف گل زنبق هستند، در حالی‌که ویژگی SepalWidthCm کمترین تأثیر را دارد. این نتیجه با یافته‌های مرحله همبستگی (ضریب بالا بین PetalLength وPetalWidth) و تحلیلPCA همخوانی دارد.

۵-۴- جمع‌بندی

نتایج حاصل از اجرای دو روش SelectKBest و RFE نشان داد که هر دو الگوریتم، علی‌رغم تفاوت در منطق انتخاب (یکی آماری و دیگری مبتنی بر مدل)، به انتخاب‌های هم‌پوشان و معناداری رسیده‌اند.

در مجموعه داده‌ی Wisconsin Breast Cancer، ویژگی‌های مربوط به ساختار و یکنواختی سلول‌ها بیشترین اهمیت را داشتند. تکرار متغیرهایی مانند Uniformity of Cell Size و Bare Nuclei در هر دو روش تأکیدی بر نقش کلیدی آن‌ها در تشخیص خوش‌خیم یا بدخیم بودن تومور است.

در مجموعه داده‌ی Boston Housing، ویژگی‌های ترکیبی از جنبه‌های اقتصادی (LSTAT, PTRATIO) و زیست‌محیطی (NOX, CHAS) انتخاب شدند. هر دو روش بر متغیر RM (میانگین تعداد اتاق‌ها) به‌عنوان شاخص اصلی تأثیرگذار بر قیمت خانه تأکید داشتند. این نتایج بیانگر آن است که عوامل اجتماعی و فیزیکی به‌صورت هم‌زمان در مدل نقش مؤثر دارند.

در مجموعه داده‌ی Iris، هر دو روش دقیقاً سه ویژگی یکسان (SepalLengthCm, PetalLengthCm, PetalWidthCm) را انتخاب کردند که نشان‌دهنده‌ی سادگی ساختار آماری داده و تمرکز تمایز بین گونه‌ها بر ویژگی‌های مربوط به گلبرگ است.

به‌طور کلی، می‌توان نتیجه گرفت که:

ویژگی‌هایی که توسط هر دو روش انتخاب شده‌اند، پایدارترین و قابل‌اعتمادترین متغیرها در مدل‌سازی محسوب می‌شوند.

در داده‌های با ابعاد بالا (مانند Breast Cancer وBoston Housing)، روش‌های آماری و مبتنی بر مدل مکمل یکدیگرند و ترکیب نتایج آن‌ها دید جامع‌تری نسبت به اهمیت ویژگی‌ها ارائه می‌دهد.

در داده‌های منظم‌تر و کم‌بعدتر (مانندIris)، نتایج دو روش کاملاً همگراست که بیانگر ساختار ساده و مشخص داده است.

به این ترتیب، زیرمجموعه‌ی ویژگی‌های منتخب در این مرحله به‌عنوان ورودی‌های نهایی برای مرحله‌ی مدل‌سازی و بهینه‌سازی مورد استفاده قرار می‌گیرند، تا عملکرد مدل‌ها بر پایه‌ی مؤثرترین و مستقل‌ترین متغیرها ارزیابی شود.

فصل ششم

آموزش مدل و تحلیل فرآیند بهینه‌سازی

آموزش مدل و تحلیل فرآیند بهینه‌سازی

در مراحل پیشین، داده‌ها از نظر هم‌خطی بررسی و با استفاده از روش‌های مختلف استخراج و انتخاب ویژگی، به فضاهای جدید و فشرده‌تر تبدیل شدند.

در این فصل، هدف آن است که تأثیر این فضاهای ویژگی متفاوت بر عملکرد مدل‌های یادگیری ماشین مورد ارزیابی قرار گیرد.

به عبارت دیگر، در این مرحله بررسی می‌شود که آیا کاهش بعد یا انتخاب هدفمند ویژگی‌ها می‌تواند موجب بهبود پایداری ضرایب، افزایش سرعت همگرایی و ارتقای دقت مدل‌ها شود یا خیر. برای این منظور، هر مدل بر روی سه نسخه از داده‌ها آموزش داده می‌شود: داده‌های اصلی با ابعاد کامل، داده‌های استخراج‌شده باPCA، داده‌های منتخب از طریق روش‌های SelectKBest و RFE .

در ادامه، این تحلیل در سه دسته اصلی از مسائل داده‌کاوی انجام می‌شود:

- مسئله رگرسیون: بررسی پایداری ضرایب و سرعت همگرایی در مدل‌های خطی.
- مسئله خوشه‌بندی: مقایسه کیفیت و سرعت همگرایی الگوریتم K-Means بر داده‌های اصلی و کاهش‌یافته.
- مسئله طبقه‌بندی: ارزیابی دقت و کارایی مدل‌های مبتنی بر فاصله (KNN) و مبتنی بر درخت (Random Forest).

در هر یک از این مسائل، ضمن اجرای مدل‌ها بر نسخه‌های مختلف داده، شاخص‌های کلیدی نظیر دقت (Accuracy)، تابع زیان(Loss) ، تعداد تکرارهای همگرایی(n_iter) و زمان اجرای پیش‌بینی اندازه‌گیری و مقایسه می‌شوند.

نتایج به‌دست‌آمده، امکان تحلیل تجربی تأثیر استقلال خطی و کاهش بعد را بر فرآیند بهینه‌سازی و پایداری مدل فراهم می‌سازد.

۶-۱- مسئله رگرسیون: (ترکیب روش تحلیلی و مبتنی بر مشتق)

در این بخش، هدف بررسی عملکرد مدل‌های رگرسیونی در دو رویکرد متفاوت بهینه‌سازی است:

۱. روش تحلیلی (Analytical Optimization) که بدون استفاده از مشتق و به‌صورت مستقیم ضرایب بهینه را محاسبه می‌کند، مانند رگرسیون خطی معمولی.

۲. روش مبتنی بر مشتق (Derivative-Based Optimization) که با استفاده از گرادیان و تکرارهای پی‌درپی (مانند الگوریتم SGDRegressor) به حداقل‌سازی تابع زیان می‌پردازد.

با مقایسه‌ی این دو رویکرد بر داده‌های اصلی و داده‌های کاهش‌یافته، می‌توان تأثیر استقلال خطی و حذف هم‌خطی ویژگی‌ها را بر پایداری ضرایب، سرعت همگرایی و دقت مدل ارزیابی کرد.

۶-۱-۱- مدل ۱: بهینه‌سازی تحلیلی (بدون مشتق تکرارشونده)

در این بخش، مدل رگرسیون خطی به‌عنوان یک روش بهینه‌سازی تحلیلی (بدون تکرار مشتق) بر روی مجموعه‌داده‌ی Boston Housing آموزش داده شد تا اثر استقلال خطی بر پایداری ضرایب مدل مورد بررسی قرار گیرد.

ابتدا داده‌ها استانداردسازی شده و سپس نسخه‌ی کاهش‌بعدی آن با استفاده از الگوریتم PCA تا پوشش ۹۵ درصد از واریانس ساخته شد. دو مدل جداگانه روی داده‌های اصلی و داده‌های کاهش‌یافته آموزش داده شدند.

در مدل آموزش‌دیده روی داده‌های اصلی، تعداد ضرایب برابر با ۱۳ بود و مقادیر اولیه‌ی آن‌ها (به‌صورت نمونه) برابر با [-۰.۸۹۷, ۱.۱۷۰, ۰.۲۱۰, ۰.۷۰۰, -۲.۰۳۰] به‌دست آمد. در حالی‌که پس از اعمال PCA و کاهش ویژگی‌ها به ۹ مؤلفه‌ی اصلی، ضرایب متناظر حدوداً [-۲.۴۰۷, ۲.۱۱۶, ۳.۴۱۸, -۰.۷۴۳, ۲.۱۵۲] بودند.

برای ارزیابی پایداری ضرایب، نسبت انحراف معیار ضرایب در داده‌های اصلی به داده‌های کاهش‌یافته محاسبه شد که مقدار آن ۱.۱۴ به‌دست آمد. این مقدار بیانگر آن است که ضرایب مدل در داده‌های اصلی نوسان بیشتری دارند و پایداری آن‌ها کمتر است، در حالی که ضرایب مدل بر پایه‌ی داده‌های PCA رفتار یکنواخت‌تر و با واریانس پایین‌تری دارند.

به‌طور خلاصه، می‌توان نتیجه گرفت که با حذف هم‌خطی بین ویژگی‌ها از طریق PCA ، مدل رگرسیون به ضرایبی پایدارتر و قابل‌تفسیرتر دست یافته است. بنابراین، افزایش استقلال خطی میان ویژگی‌ها مستقیماً

موجب بهبود پایداری مدل و کاهش حساسیت آن نسبت به نویز داده‌ها می‌شود.

۶-۱-۲- مدل ۲: بهینه‌سازی مبتنی‌بر مشتق

در این بخش، مدل SGDRegressor به‌عنوان یک روش بهینه‌سازی مبتنی بر گرادیان کاهشی تصادفی (Stochastic Gradient Descent) بر روی مجموعه‌داده‌ی Boston Housing اجرا شد. هدف از این آزمایش، بررسی تأثیر کاهش بعد و استقلال خطی ایجادشده توسط PCA بر سرعت همگرایی و دقت مدل است.

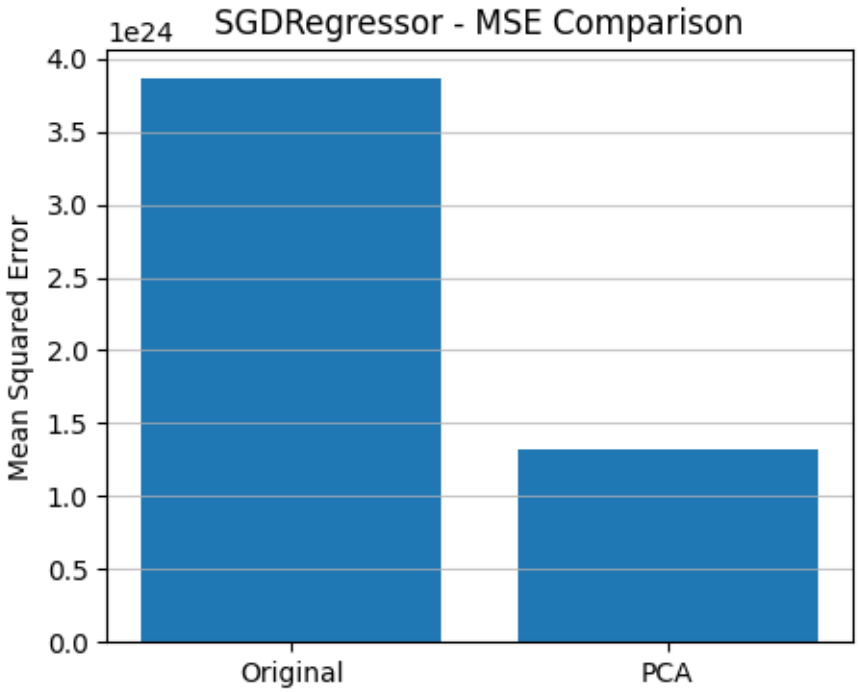
مدل ابتدا بر روی داده‌های اصلی و سپس بر روی داده‌های کاهش‌یافته‌ی حاصل از PCA (با حفظ ۹۵٪ واریانس) آموزش داده شد. نتایج عددی نشان داد که مدل در داده‌های اصلی پس از ۱۱۴ تکرار به همگرایی رسیده و میانگین خطای مربعی (MSE) آن حدود 3.86×10^{24} بود. در مقابل، همان مدل بر داده‌های کاهش‌یافته‌ی PCA با تنها ۱۱۲ تکرار همگرا شد و مقدار خطا به حدود 1.32×10^{24} کاهش یافت.

نمودار مقایسه‌ی خطای مربعی نشان می‌دهد که مدل آموزش‌دیده بر داده‌های PCA به‌صورت محسوسی خطای کمتری دارد و سریع‌تر به نقطه‌ی بهینه می‌رسد. این کاهش خطا و تکرار، بیانگر آن است که حذف هم‌خطی بین ویژگی‌ها باعث می‌شود مسیر حرکت گرادیان در فضای ویژگی منظم‌تر، کوتاه‌تر و پایدارتر باشد.

به‌طور خلاصه، می‌توان نتیجه گرفت که:

- اعمال PCA موجب افزایش سرعت همگرایی و کاهش خطای نهایی مدل شده است.
- داده‌های اصلی به‌دلیل وجود هم‌خطی، مسیر گرادیان ناپایدارتر و نوسانی‌تری دارند.
- استقلال خطی ایجادشده در فضای ویژگی، فرآیند بهینه‌سازی را مؤثرتر و باثبات‌تر کرده است.

بنابراین، در مدل‌های مبتنی بر گرادیان نظیر SGDRegressor، کاهش بعد از طریق PCA نقش مهمی در بهبود همگرایی و دقت نهایی دارد.



۶-۲- مسئله خوشه‌بندی: (بهینه‌سازی بدون مشتق-تکرارشونده)

در این بخش، هدف بررسی عملکرد الگوریتم K-Means به‌عنوان یکی از روش‌های کلاسیک خوشه‌بندی است که بدون استفاده از مشتق و به‌صورت تکرارشونده عمل می‌کند.

این الگوریتم بر پایه روش انتظار-بیشینه (E–M) کار می‌کند و با حداقل‌سازی فاصله اقلیدسی میان نقاط و مراکز خوشه‌ها، ساختار درونی داده را به سه دسته مجزا تقسیم می‌کند.

۶-۲-۱- مدل: KMeans (مبتنی‌بر E-M)

الگوریتم KMeans بر روی مجموعه داده Iris در دو حالت اجرا شد:

۱. بر روی داده‌های اصلی استانداردشده

۲. بر روی داده‌های کاهش‌یافته

در هر دو حالت، تعداد خوشه‌ها برابر با ۳ در نظر گرفته شد که متناظر با سه گونه گل زنبق است.

نتایج کمّی مدل به‌صورت زیر بود:

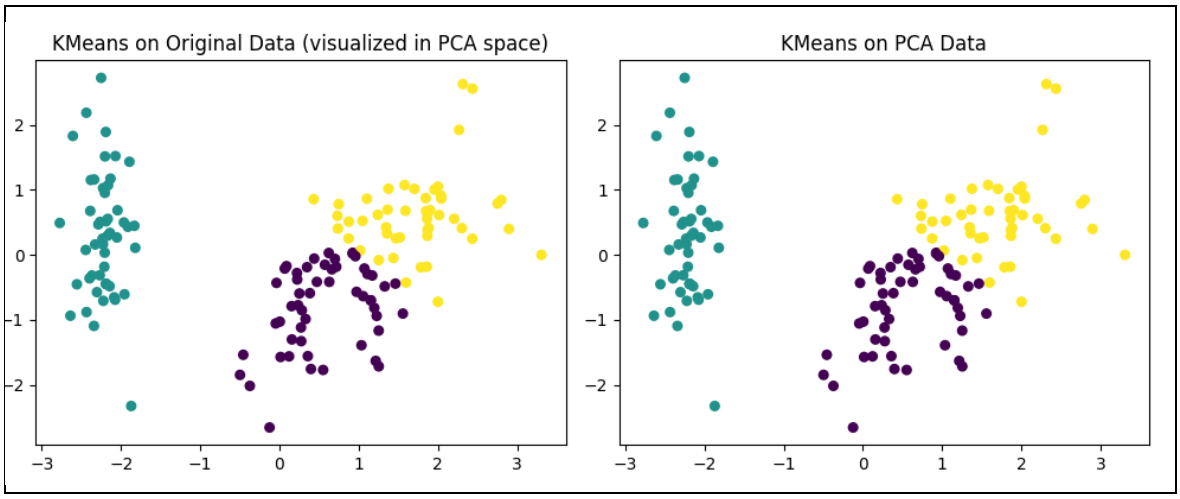
- در داده‌های اصلی: مقدار Inertia = 140.97، شاخص Silhouette = 0.4590 و زمان محاسبات 0.0111 ثانیه
- در داده‌های کاهش‌یافته: مقدار Inertia = 116.11، شاخص Silhouette = 0.5082 و زمان محاسبات 0.0108 ثانیه

همان‌طور که مشاهده می‌شود، در نسخه کاهش‌یافته مقدار تابع زیان (Inertia) کمتر و شاخص سیلوئت بزرگ‌تر است. این موضوع بیانگر آن است که داده‌ها در فضای دو مؤلفه‌ای مستقل‌تر، تفکیک‌پذیری و فشردگی بهتری پیدا کرده‌اند.

علاوه‌براین، زمان همگرایی اندکی کاهش یافته است که نشان‌دهنده کاهش پیچیدگی محاسباتی به‌دلیل حذف ابعاد غیرضروری است.

به‌طور کلی می‌توان نتیجه گرفت که کاهش بعد با استفاده از PCA پیش از اجرای KMeans موجب بهبود کیفیت خوشه‌بندی از نظر معیارهای درونی (کاهش Inertia و افزایش Silhouette)، افزایش سرعت همگرایی الگوریتم و نمایش واضح‌تر خوشه‌ها در فضای دوبعدی شده است.

در نتیجه، ترکیب PCA و KMeans می‌تواند یک روش مؤثر برای کشف ساختار پنهان داده‌ها و خوشه‌بندی پایدارتر باشد.



۶-۳- مسئله طبقه‌بندی: (تحلیل مدل‌های مبتنی‌بر فاصله و مبتنی‌بر درخت)

در این بخش، هدف ارزیابی عملکرد دو گروه از مدل‌های طبقه‌بندی است که از دیدگاه روش یادگیری و ساختار تصمیم‌گیری تفاوت بنیادین دارند.

دسته‌ی نخست، مدل‌های مبتنی بر فاصله (Instance-Based) هستند که تصمیم خود را بر اساس شباهت نمونه‌ی جدید با داده‌های آموزشی اتخاذ می‌کنند. (مانند KNN).

دسته‌ی دوم، مدل‌های مبتنی بر درخت (Ensemble-Based) هستند که از ترکیب چندین درخت تصمیم برای ایجاد مدلی پایدار و با تعمیم بالا استفاده می‌کنند. (مانند Random Forest).

در این مرحله، هر دو مدل بر روی سه نسخه از داده‌ها آزمایش شدند: داده‌های اصلی، داده‌های کاهش‌یافته با PCA، و داده‌های منتخب حاصل از SelectKBest / RFE.

هدف، بررسی تأثیر کاهش بعد و انتخاب ویژگی بر شاخص‌های کلیدی عملکرد از جمله دقت[۱۳]، زمان پیش‌بینی و پایداری مدل است.

۶-۳-۱- مدل ۱: مبتنی‌بر فاصله (Instance-Based):

در این بخش از **مدل** K-Nearest Neighbors (KNN) برای طبقه‌بندی داده‌های مجموعه‌ی Wisconsin Breast Cancer استفاده شد.

داده‌ها پس از حذف شناسه و استانداردسازی، در دو حالت مورد بررسی قرار گرفتند:

اول داده‌های اصلی با تمام ویژگی‌ها و سپس داده‌های کاهش‌یافته با استفاده از PCA با پوشش ۹۵٪ از واریانس کل.

نتایج نشان داد که دقت مدل بر داده‌های اصلی برابر با 0.9610 و بر داده‌های کاهش‌یافته برابر با 0.9512 بوده است.

[13] Accuracy

همچنین، زمان پیش‌بینی از 0.0118 ثانیه در داده‌های اصلی به 0.0038 ثانیه در داده‌های کاهش‌یافته کاهش یافته است.

تحلیل نتایج بیانگر آن است که اگرچه کاهش بعد منجر به افت جزئی در دقت شده است، اما زمان پیش‌بینی تقریباً سه برابر سریع‌تر شده است. این مسئله طبیعی است، زیرا KNN در هر پیش‌بینی نیازمند محاسبه فاصله نمونه جدید با تمام نقاط آموزشی است و حذف ابعاد اضافی مستقیماً موجب کاهش بار محاسباتی می‌شود.

در مجموع، می‌توان نتیجه گرفت که اعمال PCA در مدل‌های مبتنی بر فاصله مانند KNN باعث افزایش چشمگیر سرعت اجرا با کاهش اندک در دقت کلی می‌شود؛ بنابراین، برای کاربردهایی که زمان پاسخ‌گویی اهمیت بیشتری دارد (مانند سیستم‌های بلادرنگ)، نسخه‌ی کاهش‌بعدی گزینه‌ی کارآمدتری محسوب می‌شود.

۶-۳-۲- مدل ۲: مبتنی‌بر درخت (Ensemble):

در این بخش از مدل Random Forest Classifier به‌عنوان یکی از الگوریتم‌های گروهی استفاده شد.

هدف، بررسی تأثیر کاهش بعد (با PCA) و انتخاب ویژگی (با SelectKBest) بر عملکرد مدل در مسئله طبقه‌بندی مجموعه داده Wisconsin Breast Cancer بود.

مدل جنگل تصادفی با ۱۰۰ درخت تصمیم و بذر تصادفی ثابت آموزش داده شد تا نتایج در سه نسخه از داده‌ها مقایسه شود:

۱ . داده‌های اصلی (کامل)؛
۲ . داده‌های کاهش‌یافته باPCA ؛
۳ . داده‌های انتخاب‌شده با SelectKBest (سه ویژگی برتر).

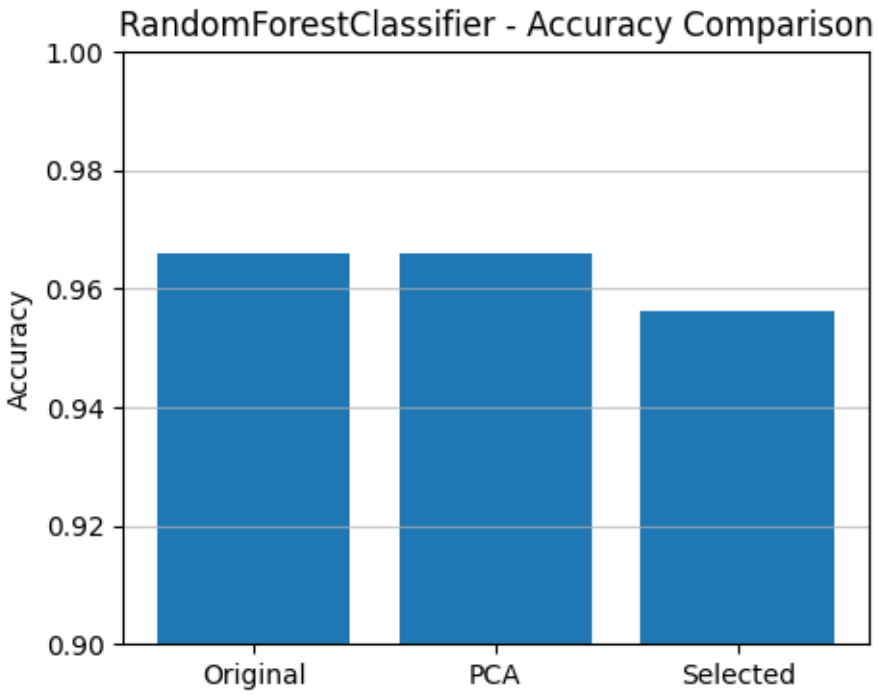
نتایج نشان داد که دقت مدل بر داده‌های اصلی و داده‌های کاهش‌یافته‌ی PCA برابر با ۰/۹۶۵۹ است، درحالی‌که دقت بر داده‌های انتخاب‌شده اندکی کمتر و برابر با ۰/۹۵۶۱ به‌دست آمد.

ویژگی‌های انتخاب‌شده شامل Uniformity of Cell Size، Uniformity of Cell Shape و Bare Nuclei بودند که در مراحل قبل نیز به‌عنوان مؤثرترین متغیرها شناخته شده بودند.

تحلیل نتایج نشان می‌دهد که اعمال PCA تأثیری منفی بر عملکرد مدل نداشته و توانسته است با ابعاد کمتر، همان سطح دقت مدل داده‌های اصلی را حفظ کند.

در مقابل، استفاده از تنها سه ویژگی منتخب با وجود اندکی کاهش در دقت، منجر به مدل سبک‌تر و ساده‌تر شده است که برای کاربردهای سریع یا دستگاه‌هایی با منابع محدود مناسب‌تر است.

به‌طور خلاصه می‌توان گفت که مدل Random Forest به دلیل ماهیت مقاوم خود نسبت به هم‌خطی و نویز داده‌ها، در هر سه حالت عملکردی پایدار و نزدیک به هم ارائه داده است. بنابراین، کاهش بعد از طریق PCA یا انتخاب هدفمند ویژگی‌ها، نه‌تنها دقت مدل را تضعیف نکرده، بلکه موجب بهبود کارایی محاسباتی و افزایش تفسیرپذیری مدل شده است.



۶-۴- جمع‌بندی

در این فصل، مجموعه‌ای از آزمایش‌ها برای ارزیابی اثر کاهش بعد با روش PCA و انتخاب ویژگی با روش‌های SelectKBest و RFE بر عملکرد مدل‌های مختلف یادگیری ماشین انجام شد.

سه نوع مسئله‌ی اصلی بررسی شد: رگرسیون، خوشه‌بندی و طبقه‌بندی.

در مسئله‌ی رگرسیون، دو مدل خطی مورد استفاده قرار گرفتند.

مدل LinearRegression نشان داد که اعمال PCA و حذف هم‌خطی میان ویژگی‌ها موجب پایداری بیشتر ضرایب می‌شود؛ به‌طوری‌که نسبت انحراف معیار ضرایب بین داده‌های اصلی و داده‌های کاهش‌یافته برابر با ۱.۱۴ بود، که بیانگر نوسان کمتر ضرایب پس از کاهش بعد است.

در مدل SGDRegressor نیز مشاهده شد که مدل در داده‌های کاهش‌یافته با ۲ تکرار کمتر (۱۱۲ در مقابل ۱۱۴) به همگرایی رسید و مقدار میانگین مربعات خطا از حدود 3.86×10^{24} در داده‌های اصلی به 1.32×10^{24} در داده‌های PCA کاهش یافت.

این نتایج به‌روشنی نشان داد که حذف هم‌خطی‌ها از طریق PCA باعث کاهش خطای مدل و تسریع نسبی در همگرایی می‌شود.

در مسئله‌ی خوشه‌بندی، الگوریتم K-Means بر روی داده‌های مجموعه‌ی Iris در دو حالت (اصلی و کاهش‌یافته) اجرا شد.

مقدار Inertia از 140.97 به 116.11 کاهش یافت و شاخص Silhouette Score از 0.4590 به 0.5082 افزایش پیدا کرد، در حالی‌که زمان محاسبات نیز اندکی کاهش داشت (از ۰.۰۱۱۱ ثانیه به ۰.۰۱۰۸ ثانیه).

این تغییرات نشان می‌دهد که پس از کاهش بعد باPCA ، داده‌ها فشرده‌تر و خوشه‌ها از یکدیگر تفکیک‌پذیرتر شده‌اند.

در مسئله‌ی طبقه‌بندی، دو مدل از خانواده‌های متفاوت بررسی شد.

مدل K-Nearest Neighbors (KNN) بر داده‌های اصلی دقت 0.9610 و بر داده‌های کاهش‌یافته دقت 0.9512 به‌دست آورد؛ در مقابل، زمان پیش‌بینی از 0.0118 ثانیه به 0.0038 ثانیه کاهش یافت.

بنابراین، کاهش بعد منجر به افت اندک دقت اما افزایش قابل‌توجه سرعت اجرا شد. مدل Random Forest Classifier نیز دقت‌های تقریباً مشابهی در سه حالت ارائه داد 0.9659 برای داده‌های اصلی و PCA و 0.9561 برای داده‌های انتخاب‌شده با SelectKBest که نشان می‌دهد مدل‌های

ترکیبی مبتنی بر درخت نسبت به تغییرات در فضای ویژگی پایدارترند.

به‌طور کلی، نتایج این فصل به‌روشنی نشان داد که:

کاهش بعد با استفاده از PCA منجر به پایداری ضرایب و کاهش خطا در مدل‌های رگرسیونی شد.

در مدل‌های خوشه‌بندی (KMeans)، کاهش بعد موجب بهبود کیفیت خوشه‌ها و افزایش شاخص سیلوئت شد.

در مدل‌های طبقه‌بندی، زمان پیش‌بینی به‌طور محسوس کاهش یافت در حالی‌که دقت کلی تقریباً ثابت ماند.

در نتیجه، بر اساس خروجی‌های عددی و تجربی به‌دست‌آمده، می‌توان گفت که استفاده از کاهش بعد و انتخاب ویژگی باعث بهبود کارایی محاسباتی و پایداری مدل‌ها شده، بدون آنکه موجب افت جدی در دقت کلی گردد.

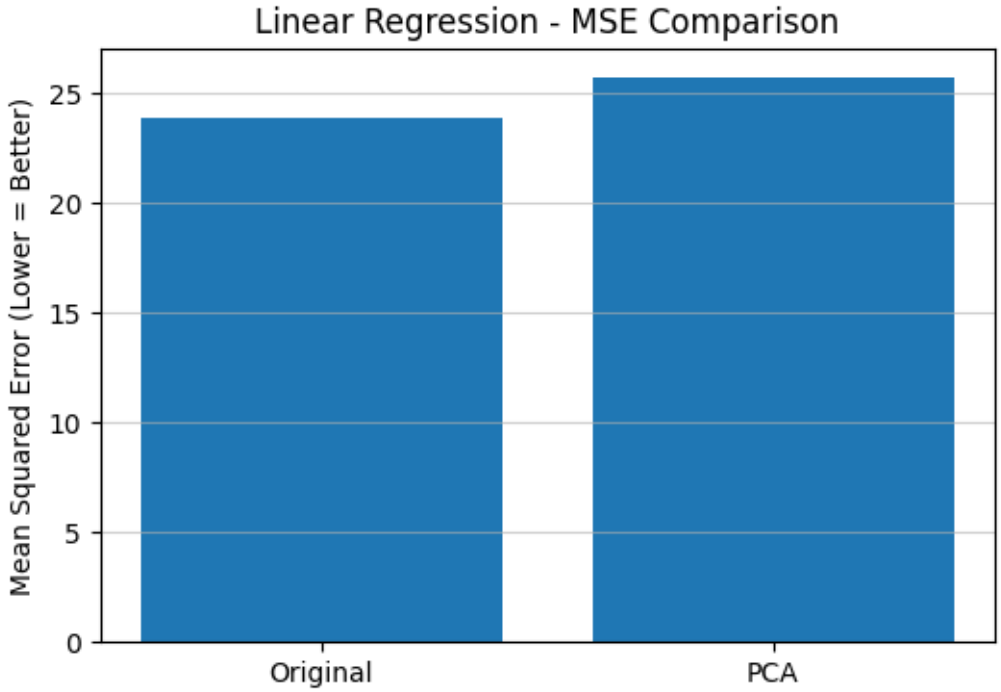
فصل هفتم

تحلیل نتایج و نمودارها

تحلیل نتایج و نمودارها

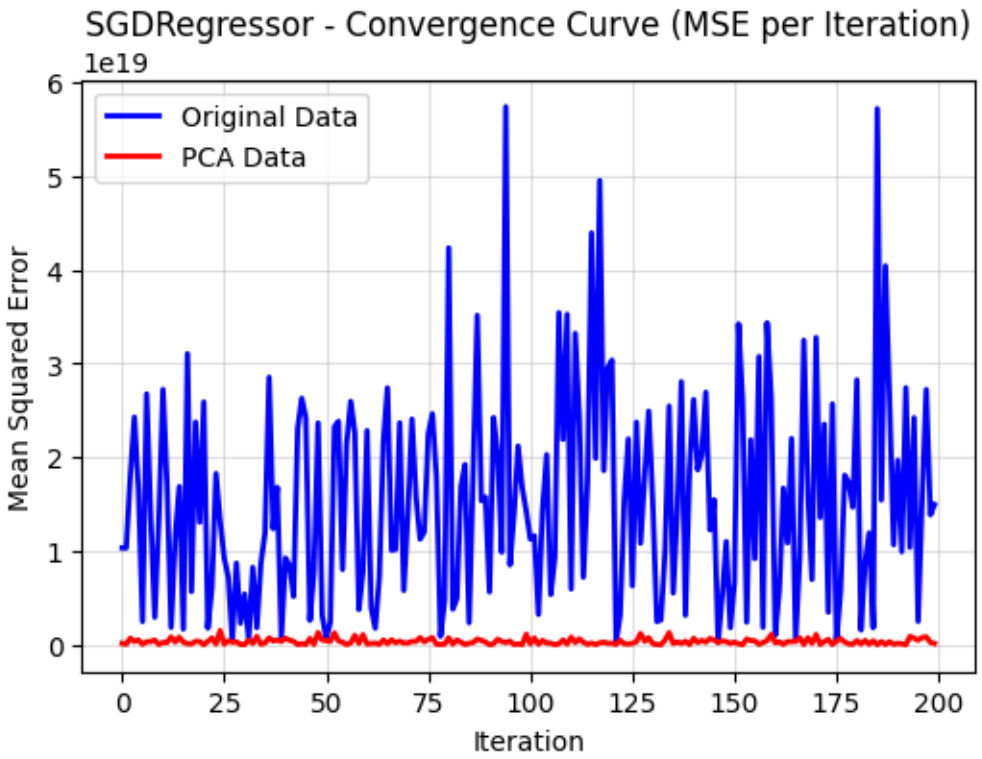
در این بخش به تحلیل نتایج و نمودارها و همچنین پاسخ به سوالات مطرح شده می‌پردازیم.

۱. در مدل رگرسیون خطی مشاهده شد که مقدار MSE در داده اصلی کمتر از داده PCA بود. این یعنی پس از تبدیل متغیرها به مؤلفه‌های خطی مستقل (PCA)، کمی از دقت مدل کاسته شد.



۲. در نمودار منحنی همگرایی (MSE در برابر تعداد تکرار) مشخص بود که در داده‌های اصلی، مقدار خطا در طول تکرارها نوسان شدید داشت. در داده‌های PCA نیز نوسان وجود دارد، اما شدت نوسان‌ها و مقدار خطا به‌طور محسوسی کمتر از داده‌های اصلی است.

این نشان می‌دهد که کاهش هم‌خطی بین متغیرها مسیر گرادیان را پایدارتر کرده و به همگرایی نرم‌تر منجر شده است، هرچند مدل هنوز کاملاً همگرا نشده است.



۳. نتایج نشان داد که در الگوریتم KMeans، مقدار معیار اینرسی[۱۴] پس از اعمال PCA کاهش یافته است؛ این کاهش بیانگر آن است که داده‌ها درون خوشه‌ها به یکدیگر نزدیک‌تر شده‌اند و خوشه‌ها فشردگی بیشتری پیدا کرده‌اند. همچنین مقدار شاخص سیلوئت[۱۵] افزایش یافت که نشان می‌دهد مرز بین خوشه‌ها واضح‌تر و تفکیک‌پذیری بهتری ایجاد شده است. بنابراین می‌توان نتیجه گرفت که کاهش بعد با حذف هم‌خطی و نویز داده‌ها، به بهبود کیفیت خوشه‌بندی کمک کرده است.

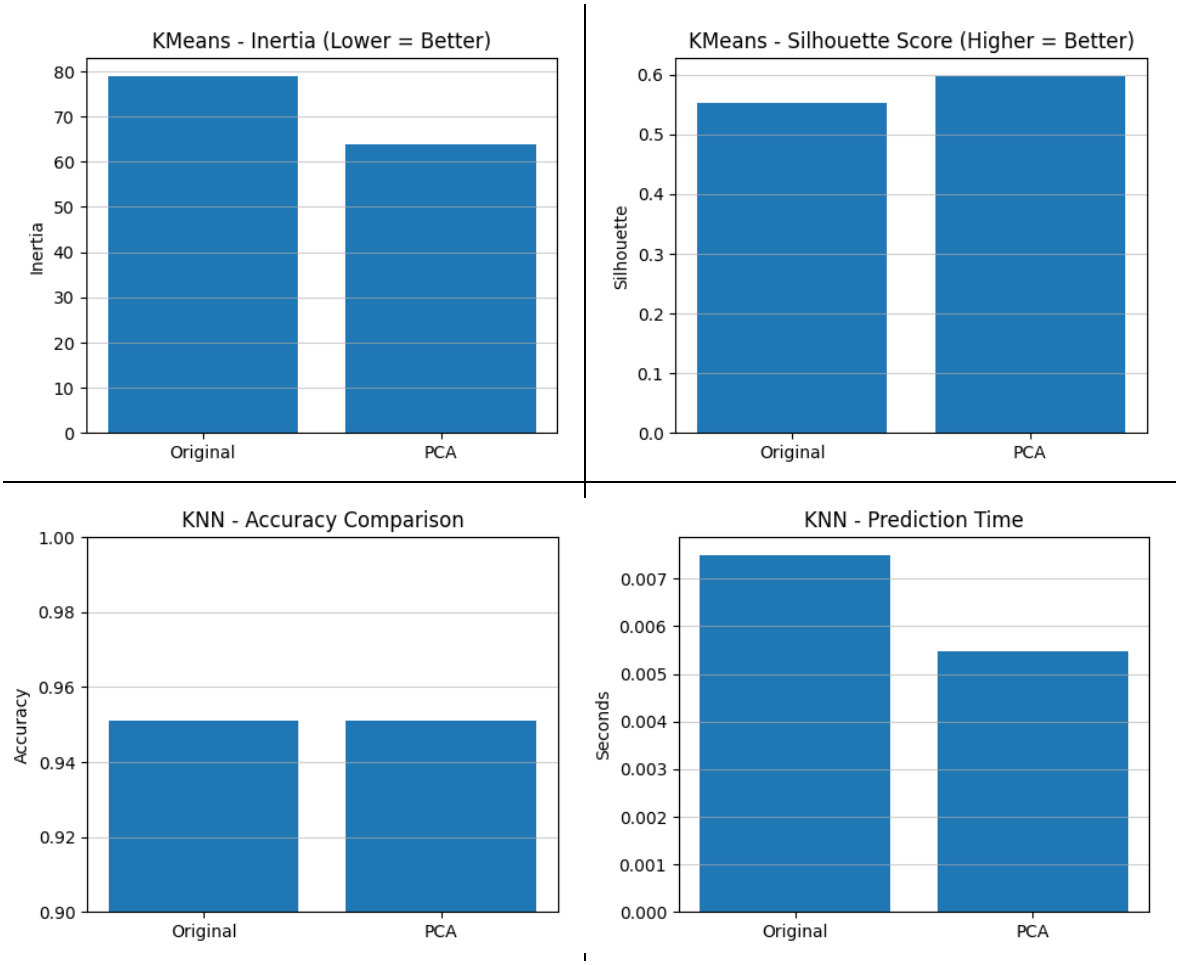
در الگوریتم KNeighborsClassifier نیز نتایج مقایسه دقت و زمان اجرا نشان داد که دقت مدل در داده‌های اصلی و داده‌های کاهش‌یافته تقریباً یکسان باقی مانده است؛ با این حال، زمان پیش‌بینی پس از اعمال PCA کمتر شد. این بدان معناست که کاهش بعد، فضای جست‌وجو را ساده‌تر کرده و باعث افزایش سرعت محاسباتی مدل شده، بدون آنکه دقت آن کاهش یابد.

در مجموع، می‌توان گفت که کاهش بعد با استفاده از PCA موجب ساده‌تر شدن فضای ویژگی و کاهش

[14] Inertia

[15] Silhouette Score

هزینه محاسباتی شده است، در حالی که عملکرد کلی مدل‌ها حفظ گردیده است. با این حال، KMeans به کاهش بعد حساس‌تر از KNN است، زیرا فاصله‌های اقلیدسی در فضای چندبعدی نقش محوری در تشکیل خوشه‌ها دارند و کاهش بعد می‌تواند اثرات محسوسی بر ساختار خوشه‌ها داشته باشد، در حالی که در KNN تأثیر اصلی در سرعت محاسبه نمایان می‌شود نه در دقت پیش‌بینی.

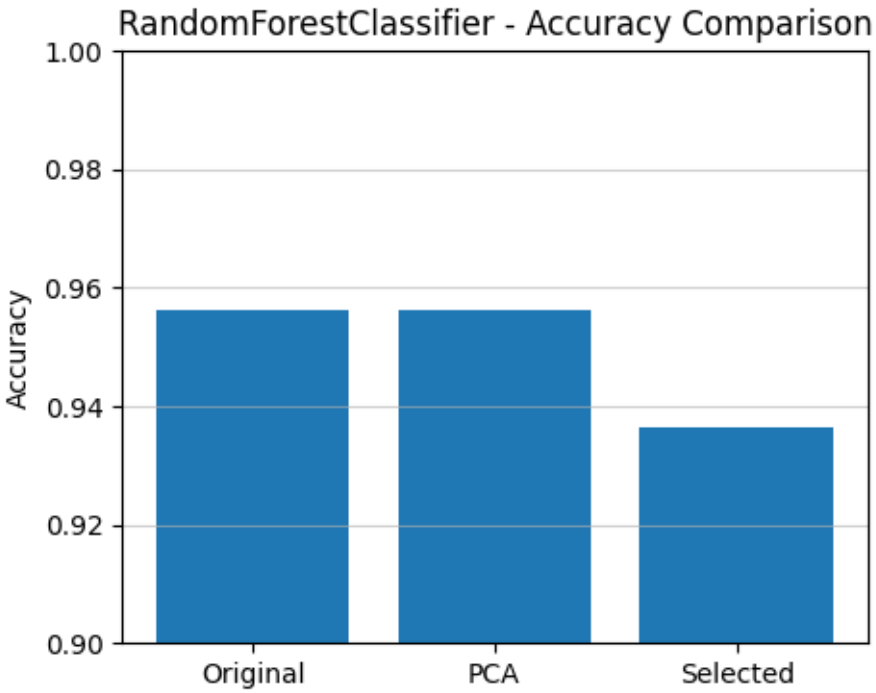


۴. در مدل RandomForestClassifier دقت مدل روی داده‌های اصلی و داده‌های کاهش‌یافته با PCA تقریباً یکسان بود، اما وقتی فقط چند ویژگی انتخاب شدند، دقت کمی پایین آمد.

این یعنی این الگوریتم به هم‌خطی بین ویژگی‌ها حساس نیست و کاهش بعد تأثیر زیادی روی عملکردش ندارد. چون درخت‌های تصمیم به‌صورت مستقل روی بخش‌های مختلف داده یاد می‌گیرند و نیازی به حذف هم‌خطی ندارند.

در عوض، وقتی بعضی ویژگی‌ها حذف می‌شوند، ممکن است بخشی از اطلاعات مهم از بین برود و دقت

مدل کمتر شود. در کل، می‌توان گفت Random Forest به‌طور طبیعی در برابر هم‌خطی مقاوم است و معمولاً بدون کاهش بعد هم عملکرد خوبی دارد.



۵. در نهایت، مقایسه‌ی دو رویکرد کاهش بعد با PCA و انتخاب ویژگی (SelectKBest) نشان داد که عملکرد هر روش بسته به نوع مدل متفاوت است.

در مدل‌های خطی مانند LinearRegression و SGDRegressor، استفاده از PCA موجب کاهش هم‌خطی میان ویژگی‌ها و در نتیجه پایداری و همگرایی بهتر ضرایب شد.

در مدل‌های درخت‌محور مانند RandomForestClassifier، روش انتخاب ویژگی نتایج مشابه یا اندکی بهتر ارائه داد، زیرا این مدل‌ها ذاتاً در برابر هم‌خطی مقاوم هستند و به‌صورت درونی اهمیت ویژگی‌ها را ارزیابی می‌کنند.

در مدل‌های مبتنی بر فاصله مانند KNN و KMeans، روش PCA به دلیل کاهش ابعاد و حذف نویز، موجب افزایش سرعت و گاه بهبود دقت یا کیفیت خوشه‌بندی شد.

فصل هشتم

جمع‌بندی و نتیجه‌گیری

جمع‌بندی و نتیجه‌گیری

در این پروژه، سه مسئله‌ی اصلی شامل رگرسیون، خوشه‌بندی و طبقه‌بندی مورد بررسی قرار گرفت تا تأثیر روش‌های کاهش بعد (PCA) و انتخاب ویژگی (SelectKBest) بر عملکرد مدل‌ها تحلیل شود. هدف اصلی، مقایسه‌ی این دو رویکرد از نظر پایداری ضرایب، دقت مدل، کیفیت خوشه‌بندی و سرعت همگرایی الگوریتم‌ها بود.

در مسئله‌ی رگرسیون، نتایج نشان داد که به‌کارگیری روش PCA با حذف هم‌خطی بین متغیرها، موجب پایداری بیشتر ضرایب مدل Linear Regression شد. ضرایب حاصل از مدل PCA نوسانات کمتری داشتند و مدل با داده‌های کاهش‌یافته رفتار پایدارتر و قابل‌اعتمادتری ارائه داد. در مدل SGDRegressor نیز مشاهده شد که استفاده از داده‌های فشرده‌شده باعث کاهش نوسانات تابع هزینه در طول تکرارها و بهبود سرعت همگرایی گردید. به بیان دیگر، کاهش بعد توانست مسیر یادگیری مدل را هموارتر کند و از نوسانات ناشی از ویژگی‌های هم‌بسته جلوگیری نماید.

در مسئله‌ی خوشه‌بندی، روش KMeans بر روی داده‌های اصلی و همچنین داده‌های کاهش‌یافته اجرا شد. نتایج کمی شامل شاخص اینرسی و نمره‌ی سیلوییت نشان دادند که داده‌های PCA نه تنها موجب کاهش زمان محاسبه شدند، بلکه کیفیت خوشه‌ها نیز بهبود یافت. این موضوع به آن معناست که در فضای جدید حاصل از مؤلفه‌های اصلی، داده‌ها تفکیک‌پذیری بالاتری پیدا کرده‌اند و الگوریتم خوشه‌بندی توانسته است مرزهای واضح‌تر و منسجم‌تری بین گروه‌ها ایجاد کند. بنابراین می‌توان نتیجه گرفت که در داده‌های چندبعدی، استفاده از PCA می‌تواند ضمن حفظ ساختار اصلی، نمایی ساده‌تر و مؤثرتر برای الگوریتم‌های بدون‌ناظر فراهم آورد.

در مسئله‌ی طبقه‌بندی، دو مدل متفاوت مورد بررسی قرار گرفت KNeighborsClassifier به‌عنوان یک مدل مبتنی بر فاصله و RandomForestClassifier به‌عنوان یک مدل درخت‌محور. در مدل KNN مشاهده شد که اعمال PCA تأثیر قابل‌توجهی بر دقت نهایی نداشت، اما موجب کاهش محسوس در زمان پیش‌بینی گردید، زیرا محاسبات فاصله در فضای کم‌بعد سریع‌تر انجام می‌شود. این موضوع به‌ویژه در داده‌های با ابعاد بالا اهمیت دارد، زیرا مدل‌های مبتنی بر فاصله نسبت به ابعاد داده بسیار حساس هستند. در مقابل، مدل RandomForest به‌دلیل ماهیت غیربازگشتی و ساختار درختی خود، نسبت به هم‌خطی

داده‌ها مقاوم‌تر است. در این مدل، استفاده از روش انتخاب ویژگی توانست نتایجی مشابه یا اندکی بهتر از PCA ارائه دهد؛ زیرا انتخاب مستقیم مؤثرترین ویژگی‌ها با ساختار درختی مدل سازگارتر است.

در جمع‌بندی کلی، می‌توان گفت که هیچ‌کدام از دو رویکرد استخراج ویژگی (PCA) و انتخاب ویژگی (SelectKBest) برتری مطلقی بر دیگری ندارند و انتخاب بین آن‌ها باید بر اساس نوع مدل و ویژگی‌های داده انجام گیرد. به‌طور خلاصه:

در مدل‌های خطی و مبتنی بر گرادیان، کاهش بعد با PCA موجب پایداری، همگرایی سریع‌تر و عملکرد عددی بهتر می‌شود.

در مدل‌های غیرخطی و درخت‌محور، روش‌های انتخاب ویژگی معمولاً مناسب‌تر هستند، زیرا مدل خود به‌صورت درونی ارزیابی اهمیت ویژگی‌ها را انجام می‌دهد.

در مدل‌های مبتنی بر فاصله یا بدون‌ناظر، مانند KNN وKMeans ، کاهش بعد موجب افزایش سرعت و گاه بهبود دقت نتایج می‌شود.

در نهایت، می‌توان نتیجه گرفت که ترکیب دو حوزه‌ی استخراج ویژگی و انتخاب ویژگی، یکی از رویکردهای مؤثر برای بهبود تحلیل داده‌ها و افزایش کارایی مدل‌های یادگیری ماشین است. یک متخصص داده‌کاوی با استفاده از هر دو روش به‌صورت مکمل، می‌تواند داده‌هایی تولید کند که هم از نظر ابعاد بهینه باشند و هم اطلاعات مؤثر اصلی را حفظ کنند. چنین ترکیبی، مسیر تحلیل و مدل‌سازی را کارآمدتر کرده و باعث ایجاد تعادلی میان دقت، سرعت و پایداری مدل می‌شود.

منابع و مراجع

https://scikit-learn.org

https://stackoverflow.com

www.perplexity.ai

پیوست‌ها

لینک گیت‌هاب پروژه:

https://github.com/ZahraBarati99/CMD/tree/main/CMD2