به نام خدا

گزارش مینی پروژه دوم یادگیری ماشین

ائلدار صمدزاده طریقت ۴۰۰۳۳۰۴۴

مقدمه

هدف این پروژه، توسعه و ارزیابی مدل‌های یادگیری ماشین برای تشخیص سرطان سینه با استفاده از دیتاست سرطان سینه ویسکانسین (تشخیصی) است. تشخیص زودهنگام و دقیق سرطان سینه برای درمان موثر و بهبود نتایج بیماران بسیار حیاتی است. این پروژه به بررسی تأثیر روش‌های مختلف انتخاب ویژگی و کاهش ابعاد بر عملکرد الگوریتم‌های طبقه‌بندی مختلف می‌پردازد.

اهداف پروژه:

- مقایسه عملکرد مدل‌های طبقه‌بندی مختلف.
- انتخاب ویژگی‌های مناسب.
- بررسی تأثیر ویژگی‌های مختلف بر عملکرد مدل.
- استفاده از تکنیک‌های اعتبارسنجی متقابل برای ارزیابی مدل.

توصیف دیتاست

دیتاست مورد استفاده، **دیتاست سرطان سینه ویسکانسین (تشخیصی)** است. این دیتاست شامل ویژگی‌هایی است که از تصاویر دیجیتالی آسپیراسیون سوزن ظریف (FNA) توده پستان محاسبه شده‌اند. این ویژگی‌ها، خصوصیات هسته سلول را توصیف می‌کنند، مانند:

- اندازه تومور (مانند mean radius، mean perimeter، mean area)
- بافت (mean texture، worst texture)
- تراکم، فرورفتگی، نقاط مقعر، تقارن و ابعاد فرکتال .هر ویژگی دارای مقادیر میانگین، خطای استاندارد و "بدترین" (بزرگترین) مقدار است .متغیر هدف، diagnosisاست که نشان می‌دهد تومور **بدخیم (0)** است یا **خوش‌خیم**.**(1)**

ابعاد دیتاست:

- **ویژگی‌ها** 569 :**(X_full)** نمونه، 30 ویژگی.
- **هدف** 569 :**(y_full)** نمونه.

پیش‌پردازش

تشخیص نقاط پرت (IQR و Z-score)

نقاط پرت (Outliers) در دیتاست با استفاده از دو روش رایج بررسی شدند:

- **دامنه بین چارکی (IQR):** IQR برای هر ویژگی محاسبه شد .نمایانگر دامنه بین صدک 25 (Q1) و صدک 75 (Q3)داده‌ها است. نقاط پرت معمولاً به عنوان نقاط داده‌ای که کمتر از Q1- 1.5 * IQR یا بیشتر از Q3 + 1.5 * IQRقرار می‌گیرند، شناسایی می‌شوند.
- **امتیاز Z (Z-score):** نقاط پرت همچنین به عنوان نقاط داده‌ای با امتیاز Z مطلق بزرگتر از 3 شناسایی شدند، که نشان‌دهنده مقادیری است که بیش از 3 انحراف معیار از میانگین فاصله دارند.

مشاهده :خروجی محاسبات IQR وZ-score ، وجود نقاط پرت در بسیاری از ویژگی‌ها، به ویژه در خصوصیات مانند mean radius، mean texture، mean perimeter، mean areaو نسخه‌های "بدترین" این ویژگی‌ها را تأیید می‌کند. این ویژگی‌ها اغلب دارای چولگی مثبت هستند که منجر به نقاط پرت بیشتر در انتهای بالایی می‌شود که از نظر بیولوژیکی برای اندازه‌گیری تومور منطقی است.

بصری‌سازی (نمودارهای جعبه‌ای)

نمودارهای جعبه‌ای (Box Plots) برای تمامی ویژگی‌ها، گروه‌بندی شده بر اساس تشخیص (بدخیم در مقابل خوش‌خیم)، تولید شدند. این بصری‌سازی برای موارد زیر حیاتی است:

- **شناسایی نقاط پرت :**نقاط پرت به وضوح به صورت نقاط مجزا که از خطوط whiskers نمودارهای جعبه‌ای فراتر می‌روند، قابل مشاهده هستند.
- **درک توزیع ویژگی :**چولگی و پراکندگی داده‌ها را برای هر تشخیص نشان می‌دهد.
- **ارزیابی قابلیت جداسازی ویژگی‌ها :**به طور بصری نشان می‌دهد کدام ویژگی‌ها جداسازی واضحی بین کلاس‌های بدخیم و خوش‌خیم نشان می‌دهند. به عنوان مثال، mean radius، mean perimeter، mean area، worst radius، worst perimeterو worst areaتفاوت‌های قابل توجهی بین دو گروه تشخیصی نشان می‌دهند که تومورهای بدخیم عموماً مقادیر بالاتری دارند.

مقیاس‌بندی ویژگی‌ها (استانداردسازی)

قبل از اعمال اکثر الگوریتم‌های یادگیری ماشین و تکنیک‌های کاهش ابعاد (مانند PCA وLDA)، از StandardScalerبرای استانداردسازی ویژگی‌ها استفاده شد. این کار داده‌ها را به گونه‌ای تبدیل می‌کند که میانگین 0 و انحراف معیار 1 داشته باشند. مقیاس‌بندی ضروری است زیرا الگوریتم‌هایی که به مقیاس ویژگی‌ها حساس هستند (مانندSVM ، رگرسیون لجستیک، PCA، LDA) در غیر این صورت به ویژگی‌هایی با دامنه‌های عددی بزرگتر اهمیت نامتناسبی می‌دهند.

کاهش ابعاد (PCAوLDA)

- **تحلیل مؤلفه‌های اصلی PCA (PCA):** بر روی دیتاست کامل scaledاعمال شد تا ابعاد را کاهش دهد و در عین حال بیشترین واریانس ممکن را حفظ کند. دو مؤلفه اصلی (n_components=2) برای بصری‌سازی و استفاده احتمالی به عنوان مجموعه ویژگی تولید شدند.
- **تحلیل تفکیک خطی (LDA) برای کاهش ابعاد LDA (LDA_DR):** بر روی دیتاست کامل scaled، به طور خاص برای کاهش ابعاد، اعمال شد و یک مؤلفه (n_components=1) تولید کرد، زیرا یک مسئله طبقه‌بندی دوتایی است LDA .بر روی به حداکثر رساندن قابلیت جداسازی کلاس‌ها تمرکز دارد.

انتخاب ویژگی

چهار روش متمایز انتخاب ویژگی برای شناسایی مرتبط‌ترین ویژگی‌ها به کار گرفته شد، با هدف کاهش نویز، بهبود کارایی مدل و به طور بالقوه افزایش عملکرد. برای هر روش، 10 ویژگی برتر انتخاب شدند.

- **اطلاعات متقابل :(MI)** این روش وابستگی آماری بین هر ویژگی و متغیر هدف را اندازه‌گیری می‌کند. نمرات MI بالاتر نشان‌دهنده وابستگی بیشتر است. این روش برای کشف روابط خطی و غیرخطی موثر است.
- **حذف ویژگی بازگشتی** RFE **:(RFE)** یک روش Wrapper است که با برازش مکرر یک مدل (در این مورد رگرسیون لجستیک) و حذف ضعیف‌ترین ویژگی‌ها (یا مجموعه‌ای از ویژگی‌ها) تا رسیدن به تعداد ویژگی‌های مورد نظر (10) کار می‌کند. این روش ویژگی‌هایی را انتخاب می‌کند که برای مدل انتخابی مهم هستند.
- **کای-دو :(chi2)** این آزمون آماری وابستگی بین متغیرهای تصادفی را اندازه‌گیری می‌کند. برای ویژگی‌های عددی غیر منفی و یک هدف طبقه‌بندی مناسب است و ارزیابی می‌کند که آیا یک ویژگی مستقل از کلاس است یا خیر. مقادیر کای-دو بالاتر نشان‌دهنده وابستگی بیشتر است.
- **SelectKBest(با امتیازدهی chi2) :**این یک روش فیلتر است که 10 ویژگی برتر را بر اساس یک تابع امتیازدهی مشخص انتخاب می‌کند. در اینجا، chi2به عنوان تابع امتیازدهی استفاده شد، به این معنی که نتایج با روش مستقیم کای-دو یکسان هستند.

طبقه‌بندی و ارزیابی

بخش اصلی پروژه شامل ارزیابی عملکرد شش طبقه‌بند کننده یادگیری ماشین مختلف در مجموعه‌های ویژگی متفاوت با استفاده از **اعتبارسنجی متقابل 5-فولد طبقه‌بندی‌شده (Stratified K-Fold Cross-Validation)** است-Stratified K . Foldتضمین می‌کند که هر فولد همان نسبت توزیع کلاس را که در دیتاست اصلی وجود دارد، حفظ می‌کند و تخمین‌های قابل اطمینان‌تری از عملکرد مدل ارائه می‌دهد.

طبقه‌بندکننده‌های مورد استفاده:

- درخت تصمیم(Decision Tree Classifier)
- نایف بیز(Naive Bayes / GaussianNB)
- ماشین بردار پشتیبان(SVM / SVC)
- تحلیل تفکیک خطی -LDA) به عنوان طبقه‌بند کننده(
- جنگل تصادفی(Random Forest Classifier)
- رگرسیون لجستیک(Logistic Regression)
- بگینگ(Bagging Classifier)

مجموعه‌های ویژگی ارزیابی‌شده:

- **Original**: تمامی 30 ویژگی اصلی.

- **Scaled:** تمامی 30 ویژگی پس از تبدیل با.StandardScaler
- **MI:** 10ویژگی برتر انتخاب‌شده توسط اطلاعات متقابل.
- **RFE:** 10 ویژگی برتر انتخاب‌شده توسط حذف ویژگی بازگشتی.
- **Chi2:** 10 ویژگی برتر انتخاب‌شده توسط آزمون کای-دو.
- **KBest:** 10 ویژگی برتر انتخاب‌شده توسط) SelectKBest با استفاده از.(chi2
- **PCA:** 2 مؤلفه اصلی.
- **LDA_DR:** 1 مؤلفه تفکیک خطی (از کاهش ابعاد LDA).

معیارهای ارزیابی :برای هر ترکیب مدل و مجموعه ویژگی، معیارهای زیر محاسبه و میانگین‌گیری شدند (میانگین‌گیری بر روی 5 فولد اعتبارسنجی متقابل):

- **دقت(Accuracy) :** صحت کلی پیش‌بینی‌ها.
- **دقت مثبت(Precision) :** نسبت پیش‌بینی‌های مثبت که واقعاً صحیح بوده‌اند.
- **فراخوانی(Recall / Sensitivity) :** نسبت موارد مثبت واقعی که توسط مدل به درستی شناسایی شده‌اند .**این معیار به ویژه در پیش‌بینی سرطان برای به حداقل رساندن False Negatives (تشخیص‌های از دست رفته) حیاتی است.**
- **امتیاز(F1-Score) F1 :** میانگین هارمونیک دقت و فراخوانی، که تعادلی بین این دو ارائه می‌دهد.

.6نتایج و بحث

نتایج اعتبارسنجی متقابل برای تمامی ترکیبات مدل و مجموعه ویژگی در جدول زیر خلاصه شده است، که ابتدا بر اساس فراخوانی (نزولی) و سپس بر اساس دقت (نزولی) مرتب شده‌اند، زیرا در پیش‌بینی سرطان، به حداقل رساندن False Negativesدر اولویت است.

مشاهدات و بینش‌های کلیدی:

- **تأثیر مهندسی ویژگی :**مشاهده کنید که چگونه مجموعه‌های ویژگی مختلف) مقیاس‌بندی‌شده، MI، RFE، Chi2، PCA، (LDA_DRبر عملکرد مدل‌های مختلف تأثیر می‌گذارند. کاهش ابعادPCA) ، (LDA_DRیا انتخاب ویژگی اغلب منجر به بهبود در برخی معیارها می‌شود، به ویژه برای مدل‌هایی که مستعد overfitting در داده‌های با ابعاد بالا هستند.
- **تغییرپذیری عملکرد مدل :**مدل‌های مختلف با مجموعه‌های ویژگی متفاوت، عملکرد بهتری دارند. به عنوان مثال، برخی مدل‌ها ممکن است بهترین عملکرد را بر روی داده‌های مقیاس‌بندی‌شده داشته باشند، در حالی که برخی دیگر ممکن است از مجموعه ویژگی کاهش‌یافته مانند LDA_DR یا RFE بهره بیشتری ببرند.
- **فراخوانی به عنوان معیار اصلی :**در زمینه پیش‌بینی سرطان، مدل‌هایی با فراخوانی بالا مطلوب هستند، حتی اگر به معنای کاهش جزئی در دقت باشد. این امر خطر نتایج منفی کاذب را به حداقل می‌رساند، جایی که به یک بیمار مبتلا به سرطان به اشتباه گفته می‌شود که سالم است که می‌تواند عواقب تهدیدکننده زندگی داشته باشد.

- **معاوضه‌ها** :معاوضه‌های بین دقت و فراخوانی را تحلیل کنید. مدلی با فراخوانی بسیار بالا اما دقت بسیار پایین ممکن است هشدارهای کاذب زیادی تولید کند که منجر به استرس و اقدامات غیرضروری برای افراد سالم می‌شود.

نتیجه‌گیری و کار آتی

این پروژه یک روش‌شناسی قوی برای طبقه‌بندی تشخیص سرطان سینه را نشان می‌دهد، که بر اهمیت پیش‌پردازش، مهندسی ویژگی و ارزیابی جامع مدل با استفاده از اعتبارسنجی متقابل تأکید دارد. مقایسه سیستماتیک در مجموعه‌های ویژگی و مدل‌های مختلف، بینش‌هایی را در مورد اینکه کدام ترکیب‌ها برای این مسئله خاص بهترین عملکرد را دارند، با تمرکز بر به حداقل رساندن منفی‌های کاذب، ارائه می‌دهد.