به نام خدا

گزارش مینی پروژه اول یادگیری ماشین

ائلدار صمدزاده طریقت ۴۰۰۳۳۰۴۴

مقدمه

این پروژه بر توسعه و ارزیابی مدل‌های یادگیری ماشین برای پیش‌بینی سطح گلوکز خون در دیتاست دیابت بومیان پیما تمرکز دارد .پیش‌بینی دقیق گلوکز خون برای مدیریت دیابت و درک پیشرفت آن حیاتی است .این پروژه مدل‌های رگرسیونی مختلف و تکنیک‌های پیش‌پردازش را برای دستیابی به عملکرد پیش‌بینی بهینه بررسی می‌کند.

اهداف پروژه:

- مقایسه عملکرد مدل‌های رگرسیونی مختلف.
- تحلیل تأثیر ویژگی‌های مختلف بر پیش‌بینی سطح گلوکز خون.
- انتخاب بهترین مدل و تحلیل اثرات تغییرات در پیش‌پردازش داده‌ها.

داده‌ها

دیتاست مورد استفاده Pima Indians Diabetes Database است که یک دیتاست عمومی محسوب می‌شود .این دیتاست شامل چندین متغیر پیش‌بینی‌کننده پزشکی و یک متغیر هدف است .ویژگی‌ها عبارتند از:

- تعداد بارداری‌ها(Pregnancies)
- گلوکز(Glucose)
- فشار خون(BloodPressure)
- ضخامت پوست(SkinThickness)
- انسولین(Insulin)
- BMI (BMI)
- تابع وراثت دیابت(DiabetesPedigreeFunction)
- سن (Age) متغیر هدف برای این پروژه Glucose است که هدف آن پیش‌بینی سطح گلوکز خون است.

ابعاد دیتاست:

- دیتاست شامل 768 سطر و 9 ستون است.

پیش‌پردازش

مدیریت مقادیر گمشده

چندین ستون (Glucose, BloodPressure, SkinThickness, BMI) حاوی مقادیر صفر بودند که در این زمینه به عنوان مقادیر گمشده یا نامعتبر در نظر گرفته شدند .این مقادیر صفر با np.nanجایگزین شدند و سپس با میانه ستون‌های مربوطه پر شدند .پر کردن با میانه در برابر نقاط پرت مقاوم است و به حفظ توزیع داده‌ها کمک می‌کند.

حذف نقاط پرت

نقاط پرت با استفاده از روش Z-score حذف شدند .نقاط داده‌ای با Z-score مطلق بزرگتر از 3 برای هر ویژگی به عنوان نقاط پرت در نظر گرفته شده و از دیتاست حذف شدند .این کار به کاهش تأثیر مقادیر شدید بر آموزش مدل کمک می‌کند.

انتخاب ویژگی

SelectKBestبا استفاده از f_regressionبه عنوان تابع امتیازدهی، برای انتخاب 6 ویژگی برتر از داده‌های پیش‌پردازش شده (به استثنای هدف (Glucoseبه کار گرفته شد f_regression .مقدار F را برای هر ویژگی محاسبه می‌کند که نشان‌دهنده وابستگی خطی بین ویژگی و متغیر هدف است .این مرحله ابعاد را کاهش می‌دهد و به طور بالقوه عملکرد مدل را با تمرکز بر مهمترین پیش‌بینی‌کننده‌ها و کاهش نویز بهبود می‌بخشد.

تقسیم‌بندی داده‌ها

داده‌های پیش‌پردازش شده و دارای ویژگی‌های انتخاب شده (X_selected) با استفاده از train_test_splitبه مجموعه‌های آموزشی و آزمایشی تقسیم شدند.

- **اندازه مجموعه آزمایشی** 20% :از داده‌ها برای مجموعه آزمایشی اختصاص یافت.
- **حالت تصادفی(Random State)** : از یک random_stateبرابر با 42 برای اطمینان از قابلیت بازتولید تقسیم‌بندی استفاده شد.

مدل‌های رگرسیونی

نه مدل رگرسیونی مختلف برای ارزیابی انتخاب شدند که هر یک به عنوان یک پایپ‌لاین Scikit-learn پیاده‌سازی شده‌اند تا مراحل پیش‌پردازش (مانند StandardScaler)را در تعریف مدل کپسوله کنند .این کار اطمینان از مقیاس‌بندی یکنواخت در آموزش و پیش‌بینی را فراهم می‌کند.

مدل‌ها عبارتند از:

- **رگرسیون خطی(Linear Regression)** : یک مدل خطی پایه.
- **رگرسیون ریج(Ridge Regression)** : رگرسیون خطی با تنظیم‌کننده L2 برای جلوگیری از بیش‌برازش.
- **رگرسیون لسو(Lasso Regression)** : رگرسیون خطی با تنظیم‌کننده L1 برای انتخاب ویژگی و جلوگیری از بیش‌برازش.
- **رگرسیون چندجمله‌ای(Polynomial Regression)** : رگرسیون خطی را با اضافه کردن ویژگی‌های چندجمله‌ای گسترش می‌دهد و به آن امکان می‌دهد روابط غیرخطی را تشخیص دهد.
- **رگرسیون نزدیکترین همسایه(KNN Regression)** : یک روش غیرپارامتری که بر اساس میانگین k نزدیکترین همسایه خود پیش‌بینی می‌کند.

- **رگرسور بردار پشتیبان(SVR)** : یک مدل قدرتمند که می‌تواند روابط خطی و غیرخطی را مدیریت کند و از یک ترفند کرنل استفاده می‌کند.
- **درخت تصمیم(Decision Tree Regressor)** : یک مدل مبتنی بر درخت که داده‌ها را بر اساس مقادیر ویژگی‌ها تقسیم می‌کند.
- **جنگل تصادفی(Random Forest Regressor)** : یک روش ترکیبی است که چندین درخت تصمیم را برای بهبود دقت و کاهش بیش‌برازش ترکیب می‌کند.
- **رگرسیون خطی بیزی(Bayesian Linear Regression)** : یک رویکرد احتمالی برای رگرسیون خطی که توزیعی از مقادیر پارامترهای ممکن را ارائه می‌دهد.

ارزیابی

هر مدل رگرسیونی تعریف شده بر روی مجموعه داده‌های X_train و y_train آموزش داده شد و سپس بر روی مجموعه X_test(داده‌های دیده نشده) ارزیابی گردید .عملکرد هر مدل با استفاده از دو معیار رگرسیون رایج ارزیابی شد:

- **میانگین خطای مربعات(MSE)** : میانگین مربع تفاوت بین مقادیر تخمین زده شده و مقدار واقعی را اندازه‌گیری می‌کند MSE .کمتر نشان‌دهنده برازش بهتر است.
- **امتیاز R2 (ضریب تعیین)** :نسبت واریانس در متغیر وابسته را نشان می‌دهد که قابل پیش‌بینی از متغیرهای مستقل است .امتیاز R2 بالاتر نشان‌دهنده برازش بهتر است.

نتایج در یک DataFrame پانداس گردآوری شده و بر اساس MSE به ترتیب صعودی مرتب شده‌اند تا مقایسه عملکرد مدل‌ها آسان باشد.

نتایج و بحث

مشاهدات کلیدی:

- MSE و R2 Score هر مدل را برای شناسایی بهترین الگوریتم‌ها برای پیش‌بینی گلوکز خون در این دیتاست تحلیل کنید.
- مشاهده کنید که آیا انواع خاصی از مدل‌ها (مانند روش‌های ترکیبی مانند جنگل تصادفی، یا مدل‌های دارای تنظیم‌کننده مانند ریج/لسو) عملکرد بهتری نسبت به مدل‌های خطی ساده‌تر نشان می‌دهند.
- تأثیر StandardScaler را در پایپ‌لاین‌ها، به ویژه برای مدل‌های مبتنی بر فاصله مانند KNN و SVR ، و مدل‌های تنظیم‌کننده در نظر بگیرید.

نتیجه‌گیری

این پروژه با موفقیت مدل‌های رگرسیونی مختلفی را برای پیش‌بینی سطح گلوکز خون در دیتاست دیابت بومیان پیما پیاده‌سازی و ارزیابی کرد. اما از انجایی که مقدار mse ۴ ۵ برابر مقدار ماکس گلوکز است، این نتیجه اصلا قابل قبول نیست و r2score ۰.۳۲ حتی از شیر و خط انداختن هم بدتر است! در نتیجه regression برای این پیشبینی اصلا پیشنهاد نمیشود.