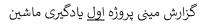
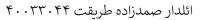


# به نام خدا







#### 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 يروژه:

- 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 می شود .این دیتاست شامل چندین متغیر پیش بینی کننده پزشکی و یک متغیر هدف است .ویژگی ها عبارتند از:

- تعداد بارداریها(Pregnancies)
  - گلوکز(Glucose)
  - فشارخون(BloodPressure)
- ضخامت یوست(SkinThickness)
  - انسولين(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.nan جایگزین 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 می کند.

### حذف نقاط يرت

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 می کند.

## انتخاب ويژگي

SelectKBest با استفاده از regression\_fبه عنوان تابع امتیازدهی، برای انتخاب 6 ویژگی برتر از دادههای پیش پردازش شده (به استثنای هدف (Glucose)به کار گرفته شد f\_regression .مقدار F را برای هر ویژگی محاسبه می کند که نشاندهنده وابستگی خطی بین ویژگی و متغیر هدف است .این مرحله ابعاد را کاهش می دهد و به طور بالقوه عملکرد مدل را با تمرکز بر مهمترین پیش بینی کننده ها و کاهش نویز بهبود می بخشد.

#### تقسيمبندي دادهها

دادههای پیش پردازش شده و دارای ویژگیهای انتخاب شده (X\_selected) با استفاده از train\_test\_split به مجموعههای آموزشی و آزمایشی تقسیم شدند.

- اندازه مجموعه آزمایشی %20 :از دادهها برای مجموعه آزمایشی اختصاص یافت.
- حالت تصادق(Random State): از یک random\_state): از یک تصادق(Random State) از یک تقسیم بندی 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) StandardScalerرا در تعریف مدل کپسوله کنند. این کار اطمینان از مقیاسبندی یکنواخت در آموزش و پیشبینی را فراهم می کند.

#### مدلها عبارتند از:

- رگرسیون خطی(Linear Regression): یک مدل خطی پایه.
- رگرسیون ریج(Ridge Regression): رگرسیون خطی با تنظیم کننده L2 برای جلوگیری از بیشبرازش.
- رگرسیون لسو (Lasso Regression) : رگرسیون خطی با تنظیم کننده 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 پیش بینی می کند.

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

#### ارزيايي

هر مدل رگرسیونی تعریف شده بر روی مجموعه دادههای  $Y_train = X_train$  و سپس بر روی مجموعه  $X_train = X_train$  (گرسیون رایج ارزیایی شد:  $X_train = X_train$  (گرسیون رایج ارزیایی شد:

- میانگین خطای مربعات(MSE): میانگین مربع تفاوت بین مقادیر تخمین زده شده و مقدار واقعی را اندازه گیری می کند MSE. کمتر نشان دهنده 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.

### نتايج و بحث

#### مشاهدات كليدى:

- 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 تنظیم کننده مانند ریج/لسو) عملکرد بهتری نسبت به مدلهای خطی سادهتر نشان میدهند.
- تأثیر StandardScalerرا در پایپلاینها، به ویژه برای مدلهای مبتنی بر فاصله مانند KNN و SVR ، و مدلهای تنظیم کننده 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 انجابی که مقدار ۴ mse برابر مقدار ماکس گلوکز است، این نتیجه اصلا قابل قبول نیست و regression برای این پیشبینی اصلا پیشنهاد نمیشود.