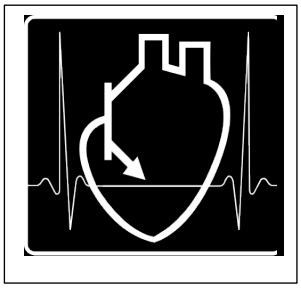
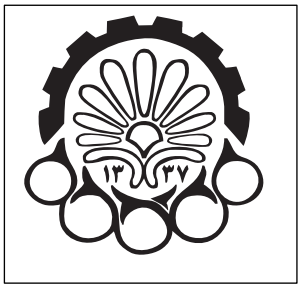




پروژه میانی ۱ درس یادگیری ماشین

استاد درس: دکتر بغدادی

آریان افشار

۴۰۱۳۳۰۰۴

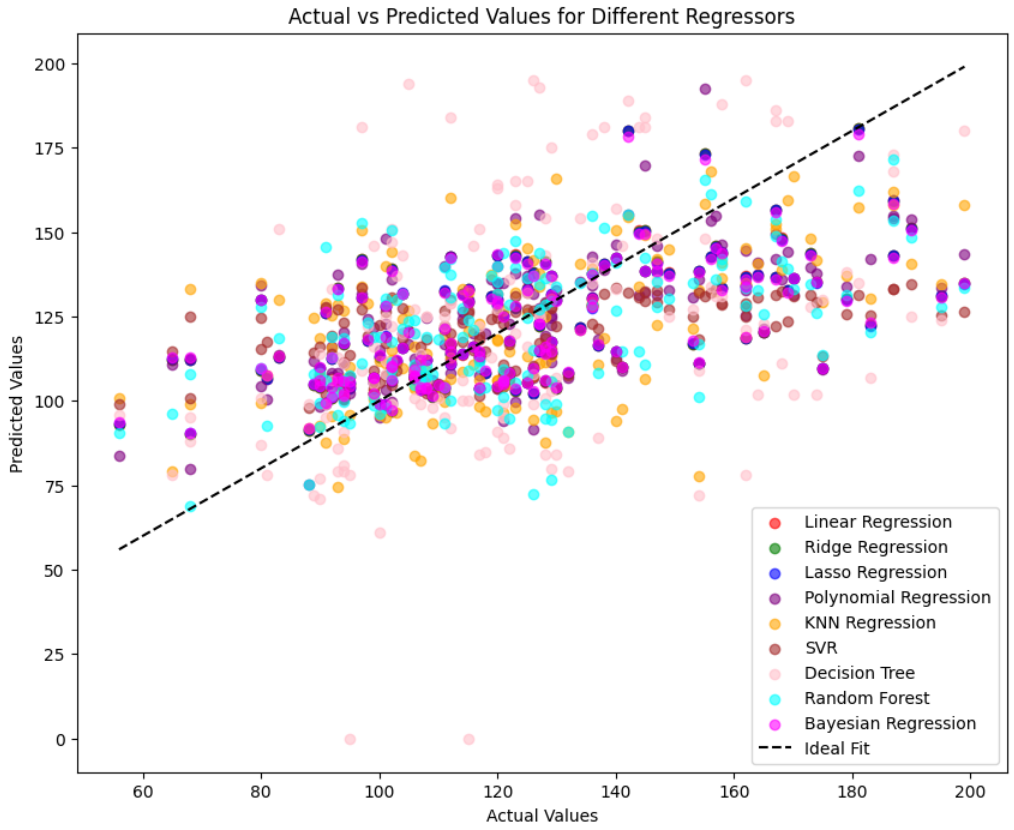
مقدمه:

در این پروژه، با استفاده از دیتاست Pima Indians Diabetes Database ، که شامل ویژگی های پزشکی مختلفی از جمله سن، فشار خون، و سطح گلوکز است، به بررسی و مقایسه عملکرد مدل های مختلف رگرسیونی برای پیش بینی مقدار پیوسته ی سطح گلوکز خون پرداختم. هدف اصلی، ارزیابی توانایی مدل های مختلف رگرسیون در یادگیری الگوی ارتباط بین ویژگی های ورودی و سطح گلوکز و انتخاب بهترین مدل برای این منظور است.

توضیحات مربوط به نوشتن کد در خود کد به صورت تکست آورده شده و در این گزارش کار من فقط به بخش تحلیل نمودار ها و انتظارات خواسته شده پرداختم. همچنین با اجازه استاد از کد های تمرین ۳ برای رسم نمودار ها و پیاده سازی مدل ها استفاده کردم.

۱ . مقایسه عملکرد مدل های مختلف رگرسیونی:

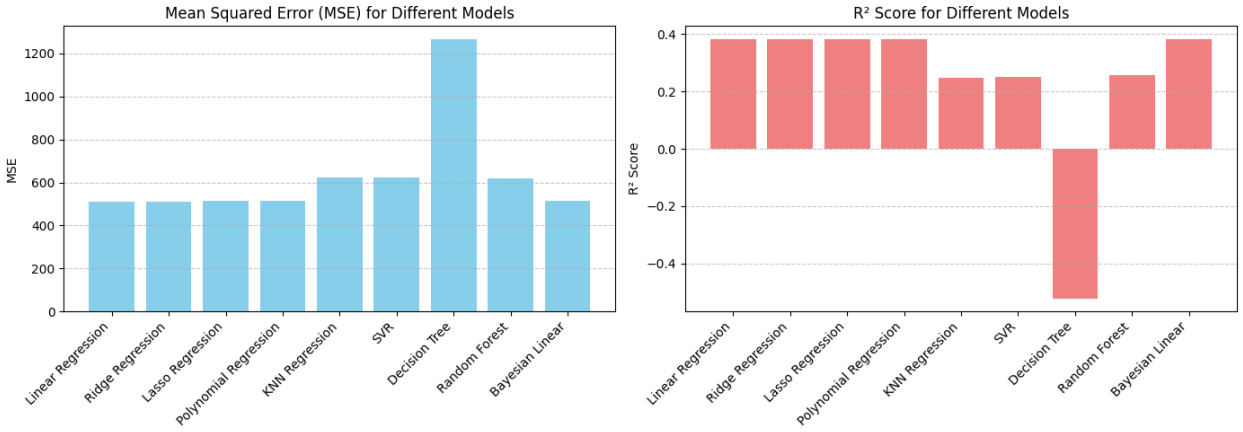
برای ارزیابی عملکرد مدل های مختلف رگرسیونی، از معیارهای رایج مانند Mean Squared Error (MSE) و R-squared Score (R^2) استفاده شده است. همچنین، نمودار مقادیر واقعی در برابر مقادیر پیش بینی شده برای هر مدل ترسیم شده است تا به صورت بصری عملکرد آن ها را مقایسه کنم.



شکل ۱: نمودار Actual vs Predicted Values برای مدل‌های مختلف رگرسیونی

این نمودار مقادیر واقعی (Actual Values) را در برابر مقادیر پیش بینی شده (Predicted Values) برای هر یک از مدل های رگرسیونی نشان می‌دهد. خط چین سیاه نیز "Ideal Fit" را نشان می دهد که در آن مقادیر واقعی و پیش بینی شده برابر هستند.

با بررسی شکل ۱، می توان مشاهده کرد که مدل هایی که نقاط داده ها آن‌ها بیشتر در نزدیکی خط چین سیاه تجمع جمع شده اند، عملکرد بهتری در پیش بینی سطح گلوکز داشته اند. در نگاه اول، به نظر می رسد مدل‌های Linear Regression ، Ridge Regression، Lasso Regression، Polynomial Regression و Bayesian Regression نقاطی دارند که به خط ایده آل نزدیک تر هستند نسبت به مدل هایی مانند Decision Tree و SVR مدل Random Forest نیز پراکندگی قابل قبولی از نقاط را نشان می دهد. مدل Decision Tree بیشترین پراکندگی نقاط را دارد که نشان دهنده دقت پایین تر آن در پیش بینی است.



شکل ۲: معیارهای عملکرد مدل‌های مختلف رگرسیونی MSE وR^2 Score

این نمودار مقدار ضریب تعیین (R^2 Score) را برای هر یک از مدل ها نشان می دهد . R^2 Score نشان می دهد که چه نسبتی از واریانس متغیر هدف (در اینجا سطح گلوکز خون) توسط مدل توضیح داده می شود. مقادیر بالاتر R^2 (نزدیک به ۱) نشان دهنده عملکرد بهتر است. مقادیر منفی R^2 نشان دهنده این است که مدل بدتر از پیش بینی میانگین عمل کرده است. همچنین مقدار خطای میانگین مربعات (MSE) را برای هر یک از مدل ها نیز نشان می‌دهد . MSE یک معیار رایج برای ارزیابی عملکرد مدل های رگرسیونی است و مقادیر کمتر MSE نشان دهنده عملکرد بهتر است.

تحلیل Mean Squared Error (MSE)

بر اساس نمودار MSE در شکل ۲، مقادیر MSE برای مدل های مختلف به شرح زیر است:

Linear Regression تقریباً ۵۰۰

Ridge Regression تقریباً ۵۰۰

Lasso Regression تقریباً ۵۰۰

Polynomial Regression تقریباً ۵۰۰

KNN Regression تقریباً ۶۰۰

SVR تقریباً ۶۰۰

Decision Tree تقریباً ۱۲۵۰ (بالاترین MSE)

Random Forest تقریباً ۶۰۰

Bayesian Regression تقریباً ۵۰۰ (کمترین MSE)

البته که مقدار دقیق ان ها در کد چاپ شده است:

```
Linear Regression: MSE=511.9708, R2=0.3832
Ridge Regression: MSE=512.0028, R2=0.3831
Lasso Regression: MSE=512.6320, R2=0.3824
Polynomial Regression: MSE=513.6550, R2=0.3812
KNN Regression: MSE=624.5678, R2=0.2475
SVR: MSE=622.5677, R2=0.2499
Decision Tree Regression: MSE=1263.9156, R2=-0.5227
Random Forest Regression: MSE=617.0863, R2=0.2565
Bayesian Linear Regression: MSE=512.9068, R2=0.3821
```

همانطور که مشاهده می‌شود، مدل linear Regression با کمترین مقدار MSE تقریباً (511) بهترین عملکرد را از نظر حداقل کردن خطای پیش بینی نشان می دهد. مدل‌هایbayesian Regression ، Ridge Regression، Lasso Regression و Polynomial Regression نیز MSE مشابه و پایینی دارند. در مقابل، مدل Decision Tree با MSE بسیار بالا، عملکرد ضعیف تری دارد.

تحلیل:R-squared Score (R^2)

بر اساس نمودار R^2 Score در شکل ۲، مقادیر R^2 برای مدل های مختلف به شرح زیر است:

Linear Regression تقریباً ۰.۴

Ridge Regression تقریباً ۰.۴

Lasso Regression تقریباً ۰.۴

Polynomial Regression تقریباً ۰.۴

KNN Regression تقریباً ۰.۲۵

SVR تقریباً ۰.۲۵

Decision Tree تقریباً -۰.۵ (مقدار منفی)

Random Forest تقریباً ۰.۲۵

Bayesian Linear تقریباً ۰.۴

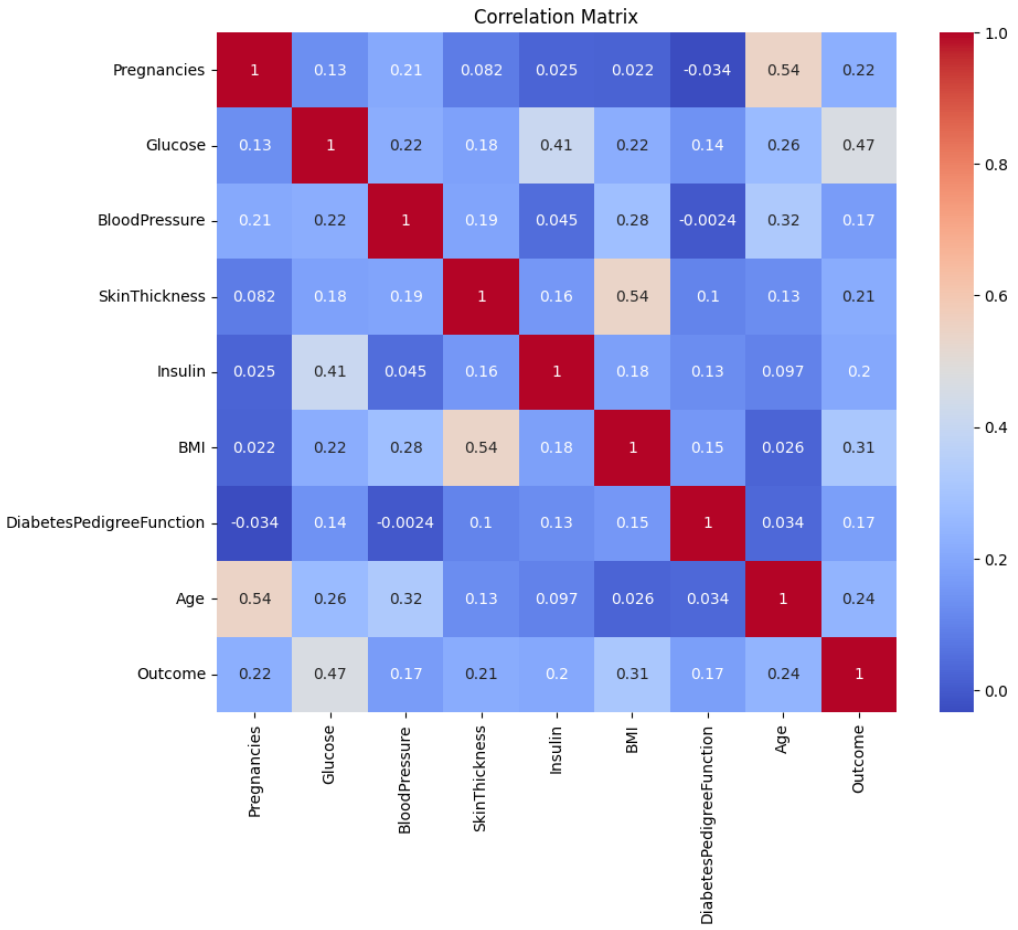
مدل‌های Linear Regression ، Ridge Regression، Lasso Regression، Polynomial Regression و Bayesian Linear با R^2 Score تقریباً ۰.۴، بیشترین نسبت از واریانس سطح گلوکز خون را نشان می دهند. این نشان دهنده توانایی نسبتاً خوب این مدل ها در کپچر کردن رابطه بین ویژگی های ورودی و متغیر هدف است. مدل‌های KNN Regression ، SVR و Random Forest R^2 Score کمتری دارند (تقریباً ۰.۲۵). قابل توجه است که مدل Decision Tree دارای R^2 Score منفی است که نشان دهنده عملکرد ضعیف تر آن نسبت به پیش بینی صرفاً میانگین سطح گلوکز خون است.

خلاصه مقایسه عملکرد:

با توجه به تحلیل MSE وR^2 Score ، مدل‌های Linear Regression ، Ridge Regression، Lasso Regression، Polynomial Regression و Bayesian Regression عملکرد بهتری در پیش بینی سطح گلوکز خون در این دیتاست از خود نشان دادند. در میان آن ها، **linear Regression** با کمترین MSE، عملکرد اندکی بهتر دارد. مدل **Decision Tree ضعیف ترین** عملکرد را داشته است.

۲ . تحلیل تاثیر ویژگی‌های مختلف بر پیش‌بینی سطح گلوکز خون:

تحلیل تاثیر هر ویژگی بر پیش بینی سطح گلوکز خون می تواند با بررسی ضرایب مدل های خطی (Linear, Ridge, Lasso, Bayesian) یا با استفاده از روش های تحلیل همبستگی انجام شود. با توجه به ماهیت دیتاست Pima Indians و همچنین رسم correlation matrix که ماتریس همبستگی مقدار سطح همبستگی میان هر جفت ویژگی ها را نشان می دهد که این مقادیر در بازه [-1, 1] هستند. مقادیر نزدیک به 1 نشان دهنده همبستگی مثبت قوی، مقادیر نزدیک به -1 نشان دهنده همبستگی منفی قوی، و مقادیر نزدیک به 0 نشان دهنده عدم وجود ارتباط مشخص هستند که این کار را درآخر کد انجام داده ام.



تحلیل ویژگی‌ها:

1. **Glucose (میزان گلوکز)**:

با ضریب همبستگی ۱ (مقدار هدف)، این ویژگی به صورت مستقیم بیشترین ارتباط را با مقدار هدف (Glucose) دارد و انتظار می رود بیشترین تأثیر را بر پیش بینی داشته باشد. این ویژگی اصلی ترین شاخص برای پیش بینی سطح گلوکز خون است.

2. **Insulin (میزان انسولین سرم)**:

با ضریب همبستگی **0.41**، میزان انسولین دومین ویژگی مهم در پیش بینی سطح گلوکز است. انسولین نقش کلیدی در تنظیم سطح قند خون دارد و رابطه مثبت با سطح گلوکز نشان دهنده اهمیت آن است.

3. **BMI (شاخص توده بدنی)**:

ضریب همبستگی **0.22** نشان دهنده تأثیر مثبت BMI بر سطح گلوکز خون است. با توجه به ارتباط چاقی و اضافه وزن با دیابت، این ارتباط طبیعی به نظر می‌رسد.

4. **Age (سن)**:

با ضریب همبستگی **0.26**، افزایش سن ارتباط مثبت با سطح گلوکز دارد. این ارتباط نشان دهنده این موضوع است که افراد مسن تر ممکن است مستعد افزایش سطح گلوکز و خطر ابتلا به دیابت باشند.

5. **BloodPressure (فشار خون)**:

با ضریب همبستگی **0.22**، فشار خون نیز تأثیر نسبی بر سطح گلوکز دارد.

6. **SkinThickness (ضخامت پوست)**:

ارتباط مثبت و اندک (ضریب همبستگی **0.18**) نشان‌دهنده تاثیر محدود این ویژگی در پیش‌بینی سطح گلوکز است.

7. **DiabetesPedigreeFunction (تابع سابقه دیابت در خانواده)**:

با ضریب همبستگی **0.14**، این ویژگی نشان دهنده اثرات سابقه دیابت خانوادگی بر سطح گلوکز خون است. این ویژگی ممکن است در ترکیب با سایر ویژگی‌ها، اهمیت بیشتری پیدا کند.

8. **Pregnancies (تعداد بارداری‌ها)**:

ضریب همبستگی **0.13** نشان‌دهنده ارتباط کم و مثبت تعداد بارداری ها با سطح گلوکز است. این ویژگی احتمالاً اثرات غیرمستقیم بر سطح گلوکز دارد.

تحلیل همبستگی نشان داد که ویژگی های **Glucose**, **Insulin**, **Age** و **BMI** بالاترین همبستگی را با هدف پیش‌بینی دارند و از اهمیت بیشتری برخوردارند. ویژگی‌هایی مانند **BloodPressure** و **SkinThickness** تأثیرات کمتری دارند، اما همچنان ممکن است در مدل‌سازی ترکیبی نقش ایفا کنند. ویژگی هایی مانند **Pregnancies** و **DiabetesPedigreeFunction** در این ماتریس ارتباط کمتری دارند و ممکن است با تنظیم دقیق مدل اهمیت بیشتری پیدا کنند.

۳ . انتخاب بهترین مدل و بررسی اثرات تغییرات در پیش‌پردازش داده‌ها:

انتخاب بهترین مدل:

بر اساس نتایج تحلیل عملکرد MSE وR^2 Score مدل Linear Regression به عنوان بهترین مدل برای پیش بینی سطح گلوکز خون در این پروژه انتخاب می شود. این مدل با کمترین MSE و R^2 Score مشابه با سایر مدل های خطی برتر، عملکرد مناسبی را از خود نشان داده است. مدل های Bayesian Regression ، Ridge Regression، Lasso Regression و Polynomial Regression نیز عملکرد خوبی دارند و می توانند به عنوان گزینه های جایگزین در نظر گرفته شوند.

بررسی اثرات تغییرات در پیش پردازش داده ها:

من در کد نتایج بالا را با استفاده از پیش پردازش standard scaler و حذف ۰ ها و feature selection بر اساس تحلیل ویژگی های مرحله قبل انجام داده ام و در پایین تاثیر آن ها را ذکر کردم , پس تمام کد و نتایج با استفاده از این روش ها بود ولی راه های دیگری نیز وجود دارد که در پایین ذکر کردم.

۱ .استاندارد سازی(StandardScaler)

استاندارد سازی باعث می شود تمام ویژگی ها در یک مقیاس قرار گیرند (میانگین صفر و انحراف معیار یک). اگر یک ویژگی مانند سن مقادیر کوچکی داشته باشد ولی ویژگی دیگری مانند انسولین مقادیر بزرگی داشته باشد، مدل ممکن است به‌طور unbalanced وزن بیشتری به ویژگی بزرگ تر بدهد. با استاندارد سازی، مدل بین ویژگی ها تعادل برقرار می کند و در نتیجه، عملکرد و پایداری آن افزایش می یابد.

۲ . حذف مقادیر صفر(Zero Removal)

در دیتاست Pima ، برخی ویژگی ها مثل فشار خون یا BMI دارای مقادیر صفر هستند، در حالی که در واقع این مقادیر نمی توانند صفر باشند (مثلاً هیچ فرد زنده ای نمی‌تواند فشار خون صفر داشته باشد). بنابراین، این صفرها به‌عنوان مقادیر گم شده تلقی می شوند و می توانند مدل را گمراه کنند. با حذف ردیف هایی که شامل این داده ها هستند، داده‌های واقعی تری در اختیار مدل قرار می گیرند. این کار معمولاً باعث افزایش دقت مدل می شود، هرچند که کاهش حجم داده ها نیز می تواند ریسک کمبود اطلاعات را به‌همراه داشته باشد.

۳ .انتخاب ویژگی(Feature Selection)

تمام ویژگی های موجود در داده‌ها الزاماً برای پیش بینی هدف مفید نیستند. برخی از آن‌ها ممکن است نویز ایجاد کنند یا همبستگی کمی با خروجی داشته باشند. استفاده از روش‌های feature selection کمک می کند تنها ویژگی های مهم و مؤثر برای پیش بینی انتخاب شوند. این کار موجب ساده تر شدن مدل، کاهش احتمال بیش‌برازش (overfitting) ، و در بسیاری از موارد بهبود دقت پیش بینی می شود.

۴ . مقیاس‌بندی ویژگی‌ها:

مدل‌هایی مانند SVR و KNN به مقیاس ویژگی ها بسیار حساس هستند. اعمال روش های مقیاس بندی مانند Standard Scalerیا Min-Max Scaler می تواند دامنه مقادیر ویژگی ها را یکسان کرده و عملکرد این مدل ها را به طور قابل توجهی بهبود بخشد. در این پروژه، مقیاس بندی انجام شده است که عملکرد مدل‌های SVR وKNN را به سطحی قابل قبول رسانده است، هرچند نسبت به مدل های خطی ضعیف تر هستند.

مدیریت داده‌های پرت (Outliers):

وجود داده های پرت می تواند مدل های خطی را تحت تاثیر قرار داده و باعث افزایش خطای آن‌ها شود. روش های شناسایی و حذف یا تبدیل داده‌های پرت می توانند عملکرد این مدل ها را بهبود بخشند.

مهندسی ویژگی :(Feature Engineering)

ایجاد ویژگی های جدید از ویژگی های موجود (مثلاً تعامل بین دو ویژگی یا تبدیل های غیرخطی) می‌تواند اطلاعات جدیدی را برای مدل فراهم کند و عملکرد آن را بهبود بخشد.

در این پروژه، با توجه به تفاوت عملکرد مدل های مختلف، می توان نتیجه گرفت که انتخاب روش های پیش پردازش به کار گرفته شده تاثیر مشخصی بر نتایج داشته است. برای مثال، اگر standard scaling انجام نمی شد، احتمالاً عملکرد مدل‌های SVR و KNN ضعیف تر از آنچه مشاهده می شود، بود.

۴. نتیجه‌گیری:

در این پروژه، با هدف پیش بینی سطح گلوکز خون در دیتاست Pima Indians Diabetes ، عملکرد چندین مدل رگرسیونی مورد ارزیابی قرار گرفت. نتایج نشان داد که مدل های Linear Regression ، Ridge Regression، Lasso Regression، Polynomial Regressionو Bayesian Regression عملکرد بهتری نسبت به مدل های KNN Regression، SVR، Random Forestو Decision Tree دارند. مدل Linear Regression با کمترین MSE، به عنوان بهترین مدل انتخاب شد. تحلیل تاثیر ویژگی ها نشان می دهد که ویژگی هایی مانند میزان گلوکز، BMI، و سن نقش مهمی در پیش‌بینی سطح گلوکز خون دارند. همچنین، اهمیت پیش پردازش داده‌ها در دستیابی به عملکرد مطلوب مدل‌ها مورد بررسی قرار گرفت.