# گزارش بررسی مدل‌های رگرسیون بر روی داده‌های دیابت

## ۱. مقدمه

در این تمرین، با استفاده از داده‌های دیابت و مدل‌های مختلف رگرسیون، به بررسی دقت پیش‌بینی و تأثیر هایپرپارامترهای هر مدل پرداخته شد. هدف اصلی این است که بهینه‌ترین مدل را برای پیش‌بینی سطح دیابت با استفاده از ویژگی‌های مختلف بیماران پیدا کنیم و تأثیر تنظیمات مختلف (هایپرپارامترها) را در عملکرد مدل‌ها بررسی کنیم.

## ۲. مدل‌های بررسی‌شده

مدل‌های رگرسیون بررسی‌شده شامل موارد زیر است:

* **رگرسیون خطی ساده (Linear Regression)**
* **رگرسیون Ridge**
* **رگرسیون Lasso**
* **Elastic Net**
* **رگرسیون چندجمله‌ای (Polynomial Regression)**
* **...**

## ۳. دقت مدل‌ها

دقت هر مدل با استفاده از معیار **R² (R-squared)** محاسبه شد. به‌طور کلی، این معیار مقدار واریانس توضیح داده شده توسط مدل را نمایش می‌دهد. دقت‌های به‌دست‌آمده در هر مدل بخاطر محدودیت در حجم دیتا کمتر از ۵۰ درصد شد و توضیحات رگرسیون و هایپر پارامتر های هر رگرسیون به صورت زیر است:

* **رگرسیون خطی ساده**: در این مدل، دقت متوسطی به دست آمد و برخی از ویژگی‌ها به دلیل همبستگی زیاد با یکدیگر، باعث ناپایداری مدل شدند.
* **رگرسیون Ridge**: این مدل با اعمال پنالتی، تأثیر ویژگی‌های با همبستگی بالا را کاهش داد و دقت بیشتری نسبت به رگرسیون خطی ساده ارائه کرد. هایپرپارامتر alpha (ضریب پنالتی) در این مدل تأثیر چشم‌گیری بر روی دقت مدل داشت.
* **رگرسیون Lasso**: با اعمال پنالتی L1، برخی از ویژگی‌ها به صفر رسانده شد. این روش برای کاهش بعد داده‌ها مناسب است. انتخاب مقدار alpha تأثیر زیادی داشت؛ با افزایش alpha، ویژگی‌های بیشتری حذف شدند.
* **Elastic Net**: این مدل ترکیبی از Ridge و Lasso بود و تعادلی بین ویژگی‌های انتخاب شده و ویژگی‌های حذف شده فراهم کرد. هایپرپارامترهای alpha و l1\_ratio در این مدل نقش مهمی داشتند. بهترین نتایج در مقادیر متعادل l1\_ratio حاصل شد.
* **رگرسیون چندجمله‌ای (Polynomial Regression)**: این مدل به دلیل افزایش تعداد ویژگی‌های چندجمله‌ای، قدرت مدل را افزایش داد و بهبود دقت قابل توجهی داشت. هایپرپارامتر degree (درجه چندجمله‌ای) در این مدل نقش کلیدی داشت و افزایش آن باعث پیچیده‌تر شدن مدل شد.

## ۴. تأثیر هایپرپارامترها

* **alpha در Ridge و Lasso**: در مدل‌های Ridge و Lasso، افزایش alpha باعث کاهش تأثیر ویژگی‌های با اهمیت کمتر شد. در Ridge، دقت مدل بهبود یافت اما در Lasso، برخی ویژگی‌ها حذف شدند و باعث سادگی مدل شد.
* **l1\_ratio در Elastic Net**: این پارامتر تعیین‌کننده نسبت تأثیر پنالتی L1 به L2 است. تغییر این پارامتر تأثیر زیادی روی انتخاب ویژگی‌ها و دقت مدل داشت.
* **degree در Polynomial Regression**: افزایش این پارامتر باعث پیچیده‌تر شدن مدل شد و دقت را بهبود بخشید اما در درجات بالا منجر به بیش‌برازش شد.

## ۵. نتیجه‌گیری

مدل‌های پیچیده‌تر مانند **Polynomial Regression** در درجات پایین و **Elastic Net** با تنظیمات متعادل، نتایج بهتری ارائه دادند. در حالی که رگرسیون‌های Ridge و Lasso به دلیل استفاده از پنالتی، ساده‌سازی داده و کاهش پیچیدگی مدل‌ها، می‌توانند در مواقعی که داده‌ها دارای همبستگی بالایی هستند مناسب‌تر باشند.

## ۶. نگاهی دیگر (توضیحات مربوط به classification بخش آخر کد)

داده های دیابت با استفاده از تبدیل مسئله رگرسیونی به کلاسیفیکیشن، مورد تحلیل قرار گرفت. به طور خاص، مقدار ستون Result که در ابتدا به صورت عددی و پیوسته بود، به سه دسته مختلف تبدیل شد. هدف این بود که با استفاده از این دسته‌بندی جدید، مدل‌های مختلف کلاسیفیکیشن مانند KNN، درخت تصمیم (Decision Tree)، رگرسیون لجستیک (Logistic Regression) و ماشین بردار پشتیبان (SVM) را برای پیش‌بینی دسته‌بندی‌ها آزمایش کنیم.

(با تحقیق در مورد اینکه پزشکان معمولا با مشاهده نتیجه گزارش دیتابت به بیمار چه feedback میدهند ستون نتیجه را دسته بندی کردم)

در ابتدا ستون Result به دسته‌هایی تقسیم شد تا مسئله رگرسیونی به یک مسئله کلاسیفیکیشن تبدیل شود:

* **دسته ۰**: برای مقادیر Result بین ۰ تا ۵۰
* **دسته ۱**: برای مقادیر Result بین ۵۰ تا ۲۰۰
* **دسته ۲**: برای مقادیر Result بزرگتر از ۲۰۰

این دسته‌بندی به ما این امکان را می‌دهد که از الگوریتم‌های کلاسیفیکیشن برای پیش‌بینی گروه مربوط به هر نمونه استفاده کنیم.

با تبدیل مسئله به یک کلاسیفیکیشن و استفاده از مدل‌های مختلف، توانستیم دقت‌های قابل قبولی برای پیش‌بینی دسته‌بندی‌های مختلف ستون Result بدست آوریم. از میان مدل‌های مختلف، **SVM** و **Decision Tree** با تنظیم مناسب هایپرپارامترها دقت بالاتری ارائه دادند. این در حالی است که **Logistic Regression** به دلیل ساده بودن مدل، دقت کمتری داشت و بیشتر برای داده‌های ساده و خطی مناسب است.