

شرح تمرین

یک مجموعه داده مربوط به تشخیص بدافزار (ویروس) نرم افزاری در پوشه دیتاست موجود است. این مجموعه داده حاصل تحقیقات مرکز آپای دانشگاه محقق اردبیلی در سال ۱۳۹۸ و ۱۳۹۹ بوده و با هدف طراحی و تولید نرم افزار بومی شناسایی بدافزار جمع آوری شده است.

الف) این مجموعه داده را خوانده و پیش پردازش های لازم را اعمال نموده و آن را به دو مجموعه داده ی آزمون و آموزش به نسبت ۲۰:۸۰ تقسیم کنید.

ب) یک مدل رگرسیون لجستیکی با هدف شناسایی بدافزار (ستون *OUTPUT*) آموزش داده و ماتریس درهم ریختگی آن رسم کنید.

ج) یک مدل بیز با هدف شناسایی بدافزار (ستون *OUTPUT*) آموزش داده و ماتریس درهم ریختگی آن را گزارش کنید.

د) بر اساس ماتریس های درهم ریختگی حاصل، کدام یک از دو مدل عملکرد مناسب تری نشان می دهد؟ توضیح دهید.