Amirkabir University of Technology
(Tehran Polytechnic)

Electrical Engineering Department

Data Mining – Dr.Amir Mazlaghani

Report Exp 2

نام دانشجو: علی بابالو – ۹۸۳۱۳۲۲

ایمیل : alibabaloo@aut.ac.ir

در ابتدا کتابخانه های مورد نیاز و دیتاستمان را ایمپورت می‌کنیم. سپس با استفاده از تابع loc یک ستون جدید به نام population level تشکیل می‌دهیم که مقادیر آن با توجه مقدار اتریبیوتpopulation تعریف می‌شود. یعنی مقدار داخل ستون جدید اضافه شده در یکی از چارک های population قرار بگیرد دارای لیبل های Low, Mid, High, Over می‌شود. برای چارک بندی کردن از تابع qcut استافاده می‌کنیم که ابتدا لیست دیتا ها (در اینجا ستون population)، تعداد تقسیم بندی ها (در اینجا q = 4)و نام لیبل هارا به آن می‌دهیم.

برای مرحله preprocess ابتدا با استفاده از تابع dropna، data object هایی که مقدار NaN دارند را دراپ می‌کنیم. سپس ستون هایی با داده هایی از نوع categorical را با استفاده از تابع LableEncoder، انکود کرده سپس با استفاده از تابع StandardScaler که هردوی آنها در کتابخانه sklearn هستند، نورمالایز می‌کنیم.

سپس برای هر دو قسمت رگرسیون و کلسیفیکیشن داده هارا به نسبت ۸۰ به ۲۰ با استفاده از تابع train_test_split از کتاخانه sklearn تقسیم بندی می‌کنیم. برای قسمت رگرسیون داده های نهایی از ستون population و قسمت کلاسیفیکیشن داده نهایی از ستون population level هستند.

برای قسمت رگرسیون خطی ابتدا یک مدل با استفاده از تابع ()linear_model.LinearRegression تشکیل می‌دهیم. سپس این مدل را با داده آموزش fit می‌کنیم و در نهایت آن مدل اموزش داده شده را برای داده تست predict می‌کنیم. سپس مقدار MSE را بوسیله تابع mean_squared_error از کتابخانه sklearn بدست می‌آوریم. برای قسمت غیر خطی هم تمامی مراحل بالا را طی می‌کنیم اما مدلمان را بصورت غیرخطی با استفاده از تابع PolynomialFeatures با درجه ۲ تشکیل می دهیم.

برای قسمت کلسیفیکیشن ابتدا مدل درخت تصمیم داریم که با استفاده از تابع DecisionTreeClassifier با معیار Entropy درست می‌کنیم و آن را با داده های آموزش فیت کرده و برای داده های تست پردیکت می‌کنیم. سپس با استفاده از تابع های Accuracy_score و Precision_score مقدار دقت و خطا این مدل را برای داده های تست و ترین محاسبه می-کنیم.

برای قسمت رندوم فارست هم مراحل بالا را تکرار می‌کنیم اما در اینجا مدلمان RandomForestClassifier می‌باشد.

در قسمت KNN ابتدا مدلمان را با استفاده از تابع KNeighborsClassifier تعریف می‌کنیم. در هر مرحله باید مقدار پارامتر n_neighbors را بترتیب برابر ۲ ، ۳ و ۵ قرار می‌دهیم. این پارامتر تعداد همسایه هارا نشان می‌دهد. سپس با استفاده از تابع Precision_score و Accuracy_score بترتیب خطا و دقت این مدل هارا بدست می‌آوریم.

در مرحله آخر نیز مدل SVM را با استفاده از تابع SVC تعریف می‌کنیم و آن را برای داده های تست و ترین آموزش می‌دهیم سپس خطا و accuracy آن مدل را با تابع Precision_score و Accuracy_score بدست می‌آوریم.

پایان