نام و نام‌خانوادگی: پریسا عمادی

شماره دانشجویی: ۴۰۱۳۶۲۴۰۲۰

لینک گیت: https://github.com/Dandelion07/BigData_HW6.git

این تمرین با استفاده از کتابخانه pyspark و در محیط colab انجام شده است. لازم به ذکر است قبل از اجرای کد فایل داده باید به فرمت csv. ذخیره شده و همراه فایل کد در یک پوشه قرار گیرند.

برای انجام تمرین ابتدا یک نشست ایجاد شده و در این نشست فایل داده را می خوانیم.

مراحل پیش پردازش شامل پر کردن مقادیر ازدست رفته (NaN) در دیتا فریم و سپس کدگذاری مقادیر غیر عددی برای پردازش ویژگی‌ها با استفاده با الگوریتم ها می باشد. پس از این پیش پردازش ۲۰ درصد داده ها برای ارزیابی الگوریتم ها جدا شده و ۸۰ درصد از داده نیز برای آموزش استفاده شده است.

در ادامه مقادیر Accuracy، Precision و Recall را برای هر الگوریتم محاسبه می کنید. نتایج مطابق جدول زیر است.

#	Algorithm	Accuracy	Precision	Recall
۱	Logistic Regression	0.70	0.67	0.42
۲	Support Vector Machine	0.59	0.57	0.19
۳	Decision Tree	0.75	0.64	0.54
۴	Random Forest	0.61	0.75	0.23

مقایسه الگوریتم‌های مختلف با معیارهای مختلف طبق جدول فوق نتایج متفاوتی در پی دارد. با این حال دقت الگوریتم رگرسیون لاجستیک از الگوریتم SVM بهتر بوده است. همچنین الگوریتم درخت تصمیم از نظر دقت عملکرد بهتری نسبت به الگوریتم رگرسیون لاجستیک داشته است. از نظر معیار precision نیز الگوریتم رگرسیون لاجستیک عملکرد بهتری از الگوریتم SVM داشته است. با این حال الگوریتم جنگل تصادفی بهترین عملکرد را از این نظر داشته است. در معیار Recall نیز الگوریتم رگرسیون لاجستیک از الگوریتم SVM بهتر بوده است با این حال الگوریتم درخت تصمیم از سایر الگوریتم ها عملکرد بهتری داشته است.