**سوالات تمرين شماره دو BIG DATA**

تاريخ تحويل : 6/10/1400 نمره : 2

1. پايگاه داده BIG2015 که توضيحات آن در یکی از مقالات پيوست آورده شده است حاوی بدافزارهایی به دو صورت کد اسمبلی و کد باینری ميباشد. ميخواهيم با استفاده از استخراج ویژگی فرکانس تکرار کدهای اسمبلی برای هر بد افزار و به کمک RANDOM FOREST در SPARK دقت کلاسبندي و زمان اجرا را براي ۹ کلاس اصلي بد افزار در حالات زير بدست آوريم.

الف: دو نمودار دقت و زمان اجرا را براي تعداد درخت 10-20-30-40-50 و حداکثر عمق درخت 10 رسم کنيد.

ب: مقادير ماتریس سردرگمی و متريک هاي زير را براي بهترين جواب ارايه دهيد Precision, Recall, F-measure, True Positive Rate, False Positive Rate

خروجي گزارش شامل کد، نتايج، نمودار و تحليل آن مي باشد.

1. با استفاده از روشی که در مقاله دوم پیوست اورده شده است بردارهایی از تصاویر فایل باینری بدافزار براي آموزش و تست يک درخت تصميم گيري در SPARK استفاده کنید.

الف : نمودار دقت را با تغيير عمق درخت از k=3 to8 رسم نماييد.

ب: با در نظر گرفتن عمق بهينه درخت پارامتر MaxBins را بين مقادير 4 و 8 و 16 و 32 تغيير دهيد. نمودار دقت خروجي را بر حسب اين پارامتر رسم نماييد و تغييرات موجود در دقت نتايج را توضيح دهيد. گزارش شامل کد و نتايج و نمودارهاي هر مرحله و تحليل جواب ها مي باشد.