**پروژه درس داده کاوی**

**مجموعه داده ای:** برای انجام این پروژه از دیتاست Diabetes که از لینک <http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff> قابل دسترسی است استفاده کنید.

توجه:

1. برای انجام پروژه از هر ابزار داده کاوی موجود می­توانید استفاده کنید.
2. برنامه نویسی پروژه با پایتون نمره اضافی دارد.

**مراحل پروژه:** پروژه شامل چهار فاز است که در ادامه تشریح شده است.

فاز اول: شناخت داده ها

* توصیف ویژگی­های دیتاست: این توصیف شامل تعداد نمونه­ها، تعداد صفات، شرح هر صفت، نوع هر صفت (nominal, ordinal, …)، شرح و نوع ویژگی کلاس
* ترسیم نمودار boxplot و نمودار پراکندگی برای یک صفت انتخابی
* شرح روشی برای تعیین میزان همبستگی هر صفت با ویژگی کلاس

فاز دوم: پیش پردازش داده ها

* بکارگیری روش میانگین برای پر کردن مقادیر گمشده
* حذف داده­های خارج از محدوده

فاز سوم: کاوش داده­ها

* اعمال الگوریتم Apriori و استخراج 10 الگوی برتر و توصیف هر الگو
* دسته بندی داده­ها با استفاده از سه الگوریتم Bagging، SVM و درخت تصمیم
* خوشه­بندی داده­ها با استفاده از الگوریتم k-means به گونه ای که کمترین تضاد بین داده­های یک خوشه باشد (خوشه­ها را به صورت بصری نمایش داده و برای هر خوشه توصیفی ارائه دهید.)

فاز چهارم: ارزیابی مدل

در فاز قبل چندین مدل ساخته شد. در این فاز این مدلها ارزیابی و مقایسه می­شوند. برای ایجاد مجموعه تست و آموزش از 10-fold cross-validation استفاده کنید.

* دسته بندهای فاز قبل را بر اساس معیارهای Accuracy، Precision و Recall مقایسه کنید.
* بر اساس یک معیار انتخابی، کارایی الگوریتم خوشه بندی را ارزیابی کنید.