

گزارش تمرین سوم درس داده‌کاوی

**«آشنایی با نرم‌افزار Weka»**

گردآورنده: سعید دادخواه

استاد: دکتر ناظرفرد

بهمن ۱۳۹۵

# بخش اول: ساخت فایل arff

برای تبدیل فایل txt به فایل weka به فرمت arff از نرم‌افزار weka استفاده خواهد شد. بعد از اجرای نرم‌افزار weka در Weka GUI Chooser در شاخه Tools از گزینه Arff Viewer برای این کار استفاده می‌شود. البته برای اجرای همه مراحل به جای استفاده از GUI از یک اسکریپت برای انجام همه مراحل استفاده خواهد شد و برای انجام این مرحله از کلاس weka.core.converters.CSVLoader استفاده می‌شود.

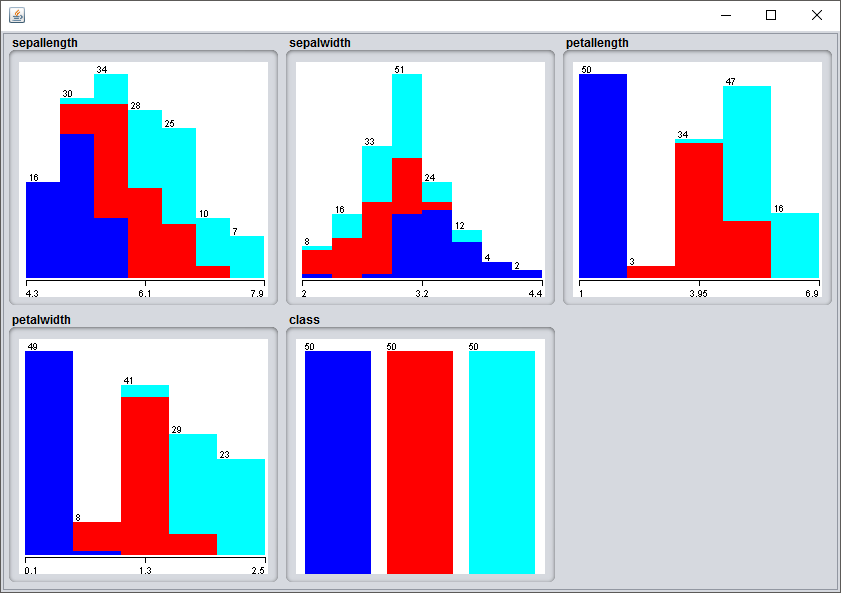
در صورتی که متد آغازین کلاس فوق اجرا شود برنامه به این شکل عمل خواهد کرد که آدرس یک فایل csv به عنوان ورودی می‌گیرد. ستون‌های این فایل csv باید دارای عنوان باشند. پس ابتدا یک برنامه به زبان پایتون اجرا می‌کنیم که یک فایل csv می‌سازد و عناوین ستون‌ها را نوشته و فایل sample.txt را به صورت csv می‌خواند و به ادامه فایل قبل اضافه می‌کند. پس از ساختن فایل csv کلاس بالا می‌تواند آن را خوانده و فایل arff را تولید کند.

برای انجام این مراحل باید پوشه‌های data و پوشه حاوی make\_weka.bat و add\_header.py در کنار یکدیگر قرار گیرند و فایل sample.txt نیز در پوشه data باشد. با اجرای make\_weka.bat ابتدا کد پایتون add\_header.py اجرا می‌شود و فایل sample.csv در پوشه data ایجاد می‌شود که همان فایل txt است با این تفاوت که عناوین ستون‌ها نیز به آن‌ها اضافه شده است. پس از آن کلاس بالا اجرا می‌شود و فایل sample.arff از روی فایل csv در همان پوشه data ساخته می‌شود.

# بخش دوم: Visualize All

با انتخاب گزینه Visualize All پنجره زیر باز می‌شود. در این نمودارها به هر کلاس یک رنگ اختصاص داده می‌شود و چگونگی توزیع هر یک از کلاس‌ها در هرکدام از ویژگی‌ها نمایش داده می‌شود. برای نمایش این موارد در ویژگی‌های عددی از نمودار هیستوگرام و برای مقادیر غیرعددی از نمودار میله‌ای استفاده می‌شود.

نمودار آخر که کلاس هرکدام از نمونه‌ها را نمایش می‌دهد کاملا مشخص می‌کند که از هر کلاس پنجاه نمونه داریم. از این پس کلاس‌های مربوط به رنگ آبی، قرمز و آسمانی به ترتیب کلاس اول تا سوم نامیده خواهند شد. نمودار اول نشان می‌دهد که به‌طوری کلی کلاس اول در این ویژگی کمتر از کلاس دوم و مخصوصا کلاس سوم است. نمودار دوم نشان می‌دهد که این ویژگی نمی‌تواند نقش موثری در جداسازی کلاس سوم از دیگر کلاس‌ها داشته باشد ولی می‌تواند به تشخیص کلاس اول و دوم کمک کند. نمودار سوم و چهارم نشان می‌دهند که این ویژگی‌ها در تشخیص کلاس اول می‌توانند بسیار خوب عمل کنند. در مورد دو کلاس دیگر نیز این ویژگی‌ها می‌توانند موثر واقع شوند.



# بخش سوم

با توجه به دستورات آورده شده در صورت پروژه مراحل زیر انجام شدند. برای نرمال‌سازی در تنظیمات فیلتر مقدار scale برابر دو و مقدار translation برابر منفی یک در نظر گرفته شدند. برای گسسته‌سازی نیز مقدار attribute indices برابر سه تا چهار و bins برابر پنج در نظر گرفته شدند. داده‌های به دست آمده در فایل iris-norm-disc.arff در پوشه data در دسترس هستند.