



فرم معرفی درس اصول و روش های داده کاوی _ نیم سال نخست ۱۴۰۶-۱۴۰۵

دانشکده بهداشت

رشته آمارزیستی، مقطع کارشناسی ارشد

نام و شماره درس: اصول و روش های داده کاوی _ ۴۱۳۵۰۱۰

روز و ساعت برگزاری: سه شنبه _ ۱۰-۱۲

محل برگزاری: دانشکده بهداشت

تعداد و نوع واحد (نظری/عملی): ۲ واحد نظری

دروس پیش نیاز: -

تلفن و روزهای تماس: ۳۷۹۲۳۱۷۴ _ سه شنبه ۱۰-۸

نام مسوول درس: دکتر کامران مهربانی

آدرس دفتر: دانشکده بهداشت، گروه آمارزیستی و اپیدمیولوژی آدرس Email: k.mehrabani@hlth.mui.ac.ir

هدف کلی درس:

پس از گذراندن این دوره دانشجویان باید تکنیک های کشف الگو در مجموعه داده را بشناسند و توانایی انجام گام های پیش پردازش (پاک سازی/آماده سازی) مربوطه به منظور دستیابی درست به هدف پژوهشی مد نظر را داشته باشند. قابلیت های تقابل مدل های آماری و مدل های مبتنی بر ماشین را فراگیرد. در نهایت باید بتوانند با استفاده از نرم افزار R تحلیل مناسبی از الگو ارایه دهد.

اهداف اختصاصی درس:

حیطه دانش

یادگیری مفهوم و تعریف های مربوط به داده کاوی، آشنایی با شیوه های پیش پردازش مجموعه داده و آماده سازی درست مجموعه داده برای کشف الگو.

حیطه نگرش

پی بردن به اهمیت و جایگاه داده کاوی در تحلیل داده های سلامت.

حیطه مهارت

توانایی تشخیص تکنیک های لازم به منظور دستیابی به هدف پژوهشی و انجام پیش پردازش های لازم.

منبع های اصلی درس:

1. Hastie T, Tibshirani R, and Friedman J. The Elements of Statistical Learning: data mining, inference and prediction. Springer, Last edition.
2. C. C. Aggarwal. Data Mining: The Textbook, Springer, Last edition.
3. Gareth James, Daniela Witten, Trevor Hastie, An Introduction to Statistical Learning with Applications in R, Springer, Last edition.

منبع های فرعی درس: اسلایدها و موردهای درس داده شده در کلاس.

روش تدریس: پرسش و پاسخ، سخنرانی، بحث گروهی، آزمون و مرور آموزه های جلسه های پیشین.

مسئولیت های فراگیرها: حضور فعال در بحث های کلاسی، پاسخ به تمرین های کلاس، انجام تکلیف ها و پروژه ها، مرور آموزه ها پس از هر جلسه، بحث و بررسی آن ها در جلسه بعد.

نحوه ارزشیابی دانشجوی و بارم مربوط به هر ارزشیابی:

الف) در طول دوره:

• پروژه ها | ۷ نمره

• ارائه کلاسی | ۴ نمره

ب) پایان دوره:

• پروژه نهایی | ۹ نمره

آمادگی لازم دانشجویها قبل از شروع کلاس: مرور آموزه‌های جلسه قبل و منبع درس.

تاریخ پروژه پایان ترم: طبق تقویم آموزش. تاریخ امتحان میان ترم: -

سایر تذکرات مهم برای دانشجویان:

• در صورت تاخیر در ارسال پروژه، زمان مربوطه تمدید نخواهد شد.

تاریخ	اهداف اختصاصی مدرس	اهداف ابلاغ شده در کوریکولوم
جلسه ۱	ارایه نمای کلی درس، تعریف های نخستین.	مفهوم داده کاوی؛ اهمیت، اهداف، و کاربرد داده کاوی در داده های سلامت؛ مفهوم داده با بعد بالا و مه داده کاوی؛ انواع منابع داده؛ رابطه و تفاوت داده کاوی با روش های کلاسیک آمار؛ آشنایی اجمالی با فرآیند داده کاوی CRISP-DM.
جلسه ۲	تعریف های تکمیلی و آشنایی عملی با روش های مدیریت مجموعه داده.	مفهوم داده کاوی؛ اهمیت، اهداف، و کاربرد داده کاوی در داده های سلامت؛ مفهوم داده با بعد بالا و مه داده کاوی؛ انواع منابع داده؛ رابطه و تفاوت داده کاوی با روش های کلاسیک آمار؛ آشنایی اجمالی با فرآیند داده کاوی CRISP-DM.
جلسه ۳	معرفی تفاوت داده کاوی با آمار و ارایه مثال واقعی.	آموزش مقدمات برنامه نویسی پایتون و معرفی بسته R برای داده کاوی.
جلسه ۴	رویکردهای پیش پردازش مجموعه داده.	روش های پیش پردازش داده ها و شناسایی داده های پرت؛ مفاهیم پاک سازی داده (مواجهه با داده های گمشده، حذف داده های تکراری، هموارسازی داده های نویز، انواع داده های پرت و تشخیص و تعدیل آن ها)؛ یکپارچه سازی داده؛ انتخاب داده؛ تبدیل داده (نرمال سازی و استاندارد سازی داده ها، مفهوم hierarchy generation)؛ (کاهش بعد، کاهش Numerosity و data compression)؛ کاوش داده؛ ارزیابی الگو؛ ارایه دانش و نمایش یافته ها.
جلسه ۵	ادامه رویکردهای پیش پردازش مجموعه داده.	روش های پیش پردازش داده ها و شناسایی داده های پرت؛ مفاهیم پاک سازی داده (مواجهه با داده های گمشده، حذف داده های تکراری، هموارسازی داده های نویز، انواع داده های پرت و تشخیص و تعدیل آن ها)؛ یکپارچه سازی داده؛ انتخاب داده؛ تبدیل داده (نرمال سازی و استاندارد سازی داده ها، مفهوم hierarchy generation)؛ (کاهش بعد، کاهش Numerosity و data compression)؛ کاوش داده؛ ارزیابی الگو؛ ارایه دانش و نمایش یافته ها.
جلسه ۶	آشنایی با مفهوم یادگیری و انواع آن.	آشنایی با مفهوم یادگیری آماری (Statistical Learning)؛ یادگیری با نظارت و بدون نظارت؛ آشنایی با رگرسیون خطی؛ رگرسیون لجستیک؛ روش های کلاسیک انتخاب متغیر همچون گام به گام.
جلسه ۷	ادامه آشنایی با مفهوم یادگیری و انواع آن.	برش داده ها (Data Splitting)؛ ساخت مدل و روش های ارزیابی مدل.
جلسه ۸	حل پروژه های عملی.	آموزش مقدمات برنامه نویسی پایتون و معرفی بسته R برای داده کاوی.

جلسه ۹	اعتبارسنجی.	آشنایی با الگوریتم بروتا و ارزیابی عملکرد الگوریتم‌ها (اعتباردرونی، اعتباربیرونی و اعتبارسنجی متقابل).
جلسه ۱۰	آشنایی با نشت اطلاعات.	آشنایی با الگوریتم بروتا و ارزیابی عملکرد الگوریتم‌ها (اعتباردرونی، اعتباربیرونی و اعتبارسنجی متقابل).
جلسه ۱۱	روش‌های رده‌بندی، رویکردهای احتمالاتی و غیراحتمالاتی.	آشنایی با روش‌های رده‌بندی بر مبنای احتمال و غیر احتمالاتی؛ برای داده‌های کیفی و کمی.
جلسه ۱۲	ادامه روش‌های رده‌بندی، رویکردهای احتمالاتی و غیراحتمالاتی.	آشنایی با روش‌های رده‌بندی بر مبنای احتمال و غیر احتمالاتی؛ برای داده‌های کیفی و کمی.
جلسه ۱۳	آشنایی با روش‌های خوشه‌بندی.	روش‌های خوشه‌بندی همچون سلسله مراتبی و غیر سلسله مراتبی، K-میانگین و KNN.
جلسه ۱۴	ادامه آشنایی با روش‌های خوشه‌بندی.	روش‌های خوشه‌بندی همچون سلسله مراتبی و غیر سلسله مراتبی، K-میانگین و KNN.
جلسه ۱۵	آشنایی به شیوه‌ها و تکنیک‌های بصری سازی حجم زیاد اطلاعات.	بصری سازی داده‌ها و مدل‌های گرافیکی.
جلسه ۱۶	مرور مبحث‌ها و رفع اشکال.	بصری سازی داده‌ها و مدل‌های گرافیکی.

با احترام، مهربانی