به نام او

شرح پروژه نهایی درس هوش مصنوعی مقدماتی

نام استاد:

دکتر شهاب‌الدین نبوی

نام استادیاران:

خشایار محمدی

امیرحسین قضاتی

علی سلیمانی

نیم‌سال دوم سال تحصیلی ۱۴۰۲ - ۱۴۰۳

◆ **عنوان پروژه:** پیش‌بینی دیابت با استفاده از الگوریتم‌های یادگیری ماشین

◆ **مقدمه:** دیابت یک بیماری مزمن و شایع است که می‌تواند عوارض جدی برای سلامتی داشته باشد. تشخیص زودهنگام دیابت می‌تواند به مدیریت بهتر و پیشگیری از عوارض آن کمک کند. در این پروژه، هدف ما ساخت یک مدل یادگیری ماشین برای پیش‌بینی دیابت با استفاده از داده‌های پزشکی بیماران است.

◆ لینک دسترسی به مجموعه داده:

لطفاً اینجا کلیک کنید.

◆ معرفی مجموعه داده:

ویژگی‌ها:

- (Age): سن فرد (عددی)
- (Gender): جنسیت فرد (دسته‌ای: مرد/زن)
- (Polyuria): ادرار زیاد (دسته‌ای: بله/خیر)
- (Polydipsia): تشنگی زیاد (دسته‌ای: بله/خیر)
- (loss weight Sudden): کاهش ناگهانی وزن (دسته‌ای: بله/خیر)
- (Weakness): احساس ضعف (دسته‌ای: بله/خیر)
- (Polyphagia): افزایش شدید اشتها (دسته‌ای: بله/خیر)
- (thrush Genital): حضور عفونت قارچی در عضو تناسلی (دسته‌ای: بله/خیر)
- (blurring Visual): مشکلات تاری دید یا بینایی (دسته‌ای: بله/خیر)
- (Itching): حضور خارش (دسته‌ای: بله/خیر)
- (Irritability): عصبانیت یا تغییرات مزاجی (دسته‌ای: بله/خیر)
- (healing Delayed): بهبودی تاخیری زخم‌ها (دسته‌ای: بله/خیر)
- (paresis Partial): نیمی از مشکلات حرکتی (دسته‌ای: بله/خیر)
- (stiffness Muscle): سفتی عضلات (دسته‌ای: بله/خیر)
- (Alopecia): ریزش مو (دسته‌ای: بله/خیر)
- (Obesity): وضعیت چاقی (دسته‌ای: بله/خیر)

◆ متغیر هدف:

- (Class): تشخیص دیابت (دسته‌ای: مثبت/منفی)

◆ وظایف پروژه:

۱. بررسی داده:

- جایگزینی مقادیر گمشده (NAN) با استفاده از تکنیک‌های Imputation مانند میانگین هر ویژگی، یا میانگین مربوط به هر کلاستر پس از انجام Clustering.
- بررسی توزیع هر ویژگی (عددی و دسته‌ای)
- بررسی ارتباطات بین ویژگی‌ها (Correlation)
- انجام Ranking Feature پس از طبقه‌بندی با Decision Tree یا Random Forest

راهنمایی: استفاده از تحلیل اکتشافی داده‌ها(EDA) و استفاده از هیستوگرام‌ها و نمودارهای جعبه‌ای با استفاده از کتابخانه‌های Pandas و Seaborn. همچنین استفاده از Pandas و Scikit-learn برای وارسی و اعمال imputation روی داده‌ها

۲. پیش‌پردازش داده:

- رفع مشکلات موجود در داده‌ها (در صورت وجود) مانند تبدیل متغیرهای دسته‌ای به شکل عددی

راهنمایی: استفاده از Label Encoding برای دسته‌های دودویی و One-Hot Encoding برای چند دسته‌ای. همچنین استفاده از **StandardScaler** برای مقیاس‌بندی ویژگی‌های عددی.

۳. Clustering:

- انجام PCA برای کاهش ابعاد.
- انجام Clustering با کاهش ابعاد (Reduction Dimensionality with Clustering) مانند الگوریتم K-Means
- نمایش داده‌ها در دو بعد و نمایش کلاسترها.

۴. ساخت مدل:

- تقسیم مجموعه داده به مجموعه‌های Test و Train
- آموزش مدل با چهار روش مختلف SVM, XGBOOST, Logistic regression, Decision tree برای پیش‌بینی تشخیص دیابت

راهنمایی: استفاده از **train_test_split** برای تقسیم داده‌ها به مجموعه‌های آموزشی و آزمایشی.

۵. تفسیر و گزارش‌گیری:

- ارزیابی عملکرد مدل با استفاده از معیارهای مناسب (Accuracy ،Precision ،Recall ،F1Score)
- مقایسه مدل‌ها با یکدیگر براساس معیارها و زمان و انتخاب بهترین مدل
- نمودار ROC و AUC
- آماده‌سازی **گزارش دقیق و جامع** که شامل روش‌های پروژه، یافته‌ها و پیشنهادات است

◆ ضمیمه:

در صورت اعمال هرگونه ویژگی و برنامه‌ای علاوه بر شرح پروژه، نمره اضافی به پروژه شما اختصاص خواهد یافت. برای مثال: بهینه‌سازی پارامترهای مدل(Hyperparameters)

● زمان تحویل پروژه: تا چهاردهم تیر ماه ۱۴۰۳

با آرزوی موفقیت برای شما