TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN - LINEAR REGRESSION

TOÁN ỨNG DỤNG VÀ THỐNG KÊ

HỌ VÀ TÊN: BÙI ĐỖ DUY QUÂN

MÃ SỐ SINH VIÊN: 21127141

LỚP: 21CLC02

Giảng viên hướng dẫn:
Phan Thị Phương Uyên

Ngày 19 tháng 8 năm 2023

Mục lục

1 YÊU CẦU CỦA ĐỒ ÁN

- Xây dựng mô hình dự đoán **mức lương** của kỹ sư sử dụng **mô hình hồi quy tuyến tính (linear regression)** với các yêu cầu sau:
  1. Sử dụng 11 đặc trưng gồm: **Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain**
  2. Sử dụng 5 đặc trưng tính cách: **conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience** và sử dụng phương pháp **k-fold cross validation** tìm ra đặc trưng tốt nhất trong các đặc trưng tính cách.
  3. Sử dụng 3 đặc trưng: **English, Logical, Quant** và sử dụng phương pháp **k-fold cross validation** tìm ra đặc trưng tốt nhất.
  4. Sinh viên tự xây dựng các mô hình (tối thiểu 3) và tìm mô hình cho kết quả tốt nhất qua phương pháp **k-fold cross validation**.
- Các thư viện được cho trước: **Numpy**, **pandas**.

2 BỘ DỮ LIỆU

- Bộ dữ liệu **Engineering Graduate Salary** gồm 2998 dòng và 34 cột. Sau quá trình tiền xử lý là loại bỏ các cột có **giá trị chuỗi** và **giá trị liên quan đến định danh và năm** thì còn lại 2998 dòng và 24 cột như sau:
  - Giá trị mục tiêu (y): **Salary**
  - 23 đặc trưng giải thích (X) gồm: Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain, ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg, conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience.
- Sinh viên đã được cung cấp 2 bộ dữ liệu: **train.csv** và **test.csv**. Bộ dữ liệu **train.csv** gồm 2248 mẫu để huấn luyện mô hình , và bộ dữ liệu **test.csv** gồm 750 mẫu để kiểm tra mô hình.

3 CÁC THƯ VIỆN SỬ DỤNG THÊM

Ngoài việc sử dụng 2 thư viện được cung cấp là **Numpy** và **pandas**, sinh viên còn sử dụng thêm thư viện **sklearn** và sử dụng module **feature_selection** của

thư viện này. Trong module này sẽ sử dụng 3 lớp:

- **VarianceThreshold**: Loại bỏ các đặc trưng có **phương sai** nhỏ hơn ngưỡng được đặt trước.
- **mutual_info_regression**: Tính độ tương quan giữa các đặc trưng và giá trị mục tiêu.
- **SelectPercentile**: Chọn ra nhóm các đặc trưng có độ tương quan cao nhất với giá trị mục tiêu.

4 KIẾN THỨC NGHIÊN CỨU

4.1 Mô hình hồi quy tuyến tính

5 TÀI LIỆU THAM KHẢO

- Cô Phan Thị Phương Uyên.
- Công thức thay đổi độ sáng và độ tương phản.
- Thay đổi độ tương phản.
- Công thức chuyển thành hình xám.
- Công thức chuyển thànhh hình màu sepia.
- Công thức và ma trận các kernel để làm mờ và rõ nét ảnh.
- Triển khai thuật toán cho làm mờ ảnh.
- Phương trình ellip.
- Công thức của ellip khi xoay 45 độ.
- Tài liệu các hàm trong thư viện numpy.
- Thư viện matplotlib hỗ trợ xuất ảnh và lưu ảnh.
- Thư viện Pillow hỗ trợ đọc ảnh.