1. TÊN ĐỀ TÀI: **Annual Salary Prediction**
2. Danh sách nhóm thành viên thực hiện:

Họ và tên	**Mã số sinh viên**
Phạm Thị Lộc(NT)	*19515111*
Vũ Lê Tự Lương	*19476841*
Trần Quang Huy	*16060381*
Nguyễn Hoàng Khang	*19515421*
Vũ Tuấn Linh	*19510891*

3. Mô tả tóm tắt đề bài:

Dự án sử dụng một số thuật toán phân tích dữ liệu cơ bản để dự đoán lương của một người dựa vào các dữ liệu nhân nhẩu học đã được lấy trước đó. Theo các dữ liệu đã được thu thập thì các dữ liệu đó sẽ được xử lý và phân tích để hỗ trợ thuật toán phân tích, tính toán và đưa ra kết quả đúng nhất dựa theo dữ liệu đã phân tích.

4. Mô tả dữ liệu

- Bao gồm 15 cột
- Mục tiêu: Thu nhập
  - <= 50K
  - > 50K
- Số thuộc tính: 14 (là nhân khẩu học và các đặc điểm khác để mô tả một người)

Variable	**Kiểu dữ liệu**	**Giá trị**	**Mô tả**
age	integer	Min: 17 Max: 90	Age – Tuổi
workclass	factor	Federal-gov, Localgov, Never-worked, Private, Self-emp-inc, Self-emp-not-inc, State-gov, Withoutpay	Class of work - Cấp công việc
fnlwgt	integer	Min: 12285 Max: 1490400	Final weight of how much of the population it represents

education	Factor	10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th,	Education level - Trình độ học vấn
		9th, Assoc-acdm, Assoc-voc, Bachelors, Doctorate, HS-grad, Masters, Preschool, Prof-school, Somecollege	
education_num	Integer	Min: 1 Max: 16	Numeric education level - Số trình độ học vấn
marital_status	Factor	Divorced, MarriedAF-spouse, Marriedciv-spouse, Marriedspouse-absent, Nevermarried, Separated, Widowed	Marital status of the person – Tình trạng hôn nhân
occupation	Factor	Adm-clerical, ArmedForces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Other-service, Privhouse-serv, Profspecialty, Protectiveserv, Sales, Techsupport, Transportmoving	Occupation of the person – Nghề nghiệp
relationship	Factor	Husband, Not-infamily, Other-relative, Own-child, Unmarried, Wife	Type of relationship - Loại mối quan hệ
race	Factor	Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, White	Race of the person - Chủng tộc

sex	Factor	Female, Male	Sex of the person – Giới tính
capital_gain	Integer	Min: 0 Max: 99999	Capital gains obtained - Lợi nhuận vốn thu được
capital_loss	Integer	Min: 0 Max: 4356	Capital loss – Lỗ vốn
hours_per_week	Integer	Min: 1 Max: 99	Average number of hour working per week – Số giờ làm việc trung bình mỗi tuần
native_country	Factor	Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, OutlyingUS(Guam-USVI-etc), Peru, Philippines, Poland, Portugal, Puerto-Rico, Scotland, South, Taiwan, Thailand, Trinadad&Tobago,	Country of origin – Quốc gia

		United-States, Vietnam, Yugoslavia	
salary	Factor	<=50K, >50K	Income level – Mức thu nhập

5. Mô tả thuật toán sử dụng

- Naive Bayes Classification (NBC) là một thuật toán dựa trên định lý Bayes về lý thuyết xác suất để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê.
    - Naive Bayes Classification là một trong những thuật toán được ứng dụng rất nhiều trong các lĩnh vực Machine learning dùng để đưa các dự đoán chính xác nhất dự trên một tập dữ liệu đã được thu thập, vì nó khá dễ hiểu và độ chính xác cao. Nó thuộc vào nhóm Supervised Machine Learning Algorithms (thuật toán học có hướng dẫn), tức là máy học từ các ví dụ từ các mẫu dữ liệu đã có.
    - Gaussian Naive Bayes là một biến thể của Naive Bayes tuân theo phân phối chuẩn Gaussian và hỗ trợ dữ liệu liên tục.
- Thuật toán Feature Engineering: Là quá trình chuyển đổi dữ liệu thô thành các tính năng hữu ích giúp chúng ta hiểu rõ hơn về mô hình của mình và

tăng khả năng dự đoán của nó. Tôi sẽ thực hiện kỹ thuật tính năng trên các loại biến khác nhau.

− Thuật toán Feature Scaling:

+ Các điểm dữ liệu đôi khi được đo đạc với những đơn vị khác nhau, m và feet chẳng hạn. Hoặc có hai thành phần (của vector dữ liệu) chênh lệch nhau quá lớn, một thành phần có khoảng giá trị từ 0 đến 1000, thành phần kia chỉ có khoảng giá trị từ 0 đến 1 chẳng hạn. Lúc này, chúng ta cần chuẩn hóa dữ liệu trước khi thực hiện các bước tiếp theo. − Thuật toán Accuracy Score:

+ Thuật toán giúp đánh giá độ chính xác của mô hình để xem độ tin cậy của mô hình có nằm trong khoảng đang tin cậy hay không. Nếu thấp hơn chỉ số độ tin cậy đáng có thì mô hình này không đáng tin cậy.

− Thuật toán Confusion Matrix:

+ Confusion Matrix ma trận nhầm lẫn hay ma trận lỗi là một bố cục bảng cụ thể cho phép hình dung hiệu suất của một thuật toán.

+ Ma trận nhầm lẫn là một trong những kỹ thuật đo lường hiệu suất phổ biến nhất và được sử dụng rộng rãi cho các mô hình phân loại. Nhìn thuật ngữ của nó thì trông có vẻ khó hiểu nhưng thực tế nó lại rất dễ hiểu. Do đó, bài viết này có thể giúp nó trở lên dễ hình dung, dễ hiểu hơn.

− Thuật toán ROC – AUC:

+ AUC - ROC là một phương pháp tính toán hiệu suất của một mô hình phân loại theo các ngưỡng phân loại khác nhau. Giả sử với bài toán phân loại nhị phân (2 lớp) sử dụng hồi quy logistic (logistic regression), việc chọn các ngưỡng phân loại [0..1] khác nhau sẽ ảnh hưởng đến khả năng phân loại của mô hình và ta cần tính toán được mức độ ảnh hưởng của các ngưỡng. AUC là từ viết tắt của Area Under The Curve còn ROC viết tắt của Receiver Operating Characteristics. ROC là một đường cong biểu diễn xác suất và AUC biểu diễn mức độ phân loại của mô hình. AUC-ROC còn được biết đến dưới cái tên AUROC (Area Under The Receiver Operating Characteristics). Ý nghĩa của AUROC có thể diễn giải như sau: Là xác suất rằng một mẫu dương tính được lấy ngẫu nhiên sẽ được xếp hạng cao hơn một mẫu âm tính được lấy ngẫu nhiên. Biểu diễn theo công thức, ta có AUC = P(score(x+) > score(x-)). Chỉ số AUC càng cao thì mô hình càng chính xác trong việc phân loại các lớp.

+ Đường cong ROC biểu diễn các cặp chỉ số (TPR, FPR) tại mỗi ngưỡng với TPR là trục tục và FPR là trục hoành.

- Thuật toán k-Fold Cross Validation:
  + K-Fold CV sẽ giúp chúng ta đánh giá một model đầy đủ và chính xác hơn khi chúng ta có một tập dữ liệu không lớn. Để sau đó chúng ta đưa ra quyết định model đó có phù hợp với dữ liệu, bài toán hiện tại hay không để mà đưa ra next action.