TITLE : ADULT CENSUS INCOME DATASET USING MACHINE LEARNING ALGORITHMS

Project Title : Adult census income dataset
using machine learning algorithms
Mail-id : moezzema10@gmail.com
GitHub:https://github.com/sayyedamoezzema

ABSTRACT : This study aims to show the usage of machine learning and data mining techniques in providing a solution to the income equality problem. The Adult Dataset has been used for the purpose. Classification has been done to predict whether a person's yearly income in US falls in the income category of either greater than 50K or less equal to 50K category based on a certain set of attributes. We aim to predict whether an individual’s income will be greater than 50,000 per year based on several attributes from the census data.

In this dataset, we will conduct Data Balance Analysis which consists of measures on the Adult Census income dataset to determine how well features and feature values are represented in the dataset. In real time the adult information at the time of census given by adult person while filling out an online application form. It is expected that the development of machine learning models that can help the company to predict adult income ,education and age etc.

In this dataset set we will used multiple classification algorithms such as logistic regression, Decision tree classifier, SVC(support vector classifier), KNN(K-nearest neighbours) classifier and naive bayes classifier. This algorithms used to which one is best fitted and accuracy is good. Finally, we will concluded that the performance of our all model using metrics such as accuracy, precision, recall, and F1 score.

1.INTRODUCTION: The Adult Census dataset is our first section, we explore the data at face value in order to understand the trends and representations of certain demographics in the corpus. We then use this information in section two to form models to predict whether an individual made more or less than 50,000 . In the third section, we look into a couple papers written on the dataset to find out what methods they are using to gain insight on the same data. Finally, in the fourth section, we compare our models as well as that of others in order to find out what features are of significance, what methods are most effective, and gain an understanding of some of the intuition behind the numbers.

Humans have grown a lot of dependence on data and information in society and with this advent growth, technologies have evolved for their storage, analysis and processing on a huge scale. The fields of Data Mining and Machine Learning have not only exploited them for knowledge and discovery but also to explore certain hidden patterns and concepts which led to the prediction of future events, not easy to obtain. The problem of income inequality has been of great concern in the recent years. Making the poor better off does not seem to be the sole criteria to be in quest for eradicating this issue. People of the United States believe that the advent of economic inequality is unacceptable and demands a fair share of wealth in the society. This model actually aims to conduct a comprehensive analysis to highlight the key factors that are necessary in improving an individual's income. Such an analysis helps to set focus on the important areas which can significantly improve the income levels of individuals.

1.1 About the data : The adult Census Income dataset has 32561 entries. Each entry contains the following information about an individual:

- Age : the age of an individual Integer greater than 0

- Workclass : a general term to represent the employment status of an individual Private, Selfempnotinc, Selfempinc, Federalgov, Local-gov, Stategov, Withoutpay, Neverworked.

- Fnlwgt : final weight. In other words, this is the number of people the census believes the entry represents Integer greater than 0

- Education : the highest level of education achieved by an individual. Bachelors, Some college, 11th, Haggard, Preschool, Assocacdm, Assocvoc, 9th, 7th8th, 12th, Masters, 1st4th, 10th, Doctorate, 5th6th, Preschool.

- Education_num : the highest level of education achieved in numerical form. Integer greater than 0 .

- Marital _status : Marriedcivspouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

- Occupation : the general type of occupation of an individual.

- Relationship : represents what this individual is relative to others.

- Race : Descriptions of an individual's race black, White etc.

- Sex : the biological sex of the individual Male, Female.

- Capital-gain : capital gains for an individual Integer greater than or equal to 0 .

- Capital-loss : capital loss for an individual Integer greater than or equal to 0.

- Hours-per-week : the hours an individual has reported to work per week continuous.

- Native-country : country of origin for an individual.

- Income: income is less than 50k and more than 50k.

Preprocessing :

The collected data may conatin missing values that may lead to in consistency.To gain better result data need to be preprocessing and so it will better the effectiveness of the algorithms.

Train model on training data set :

Now we should train the model on the training dataset and make for soothsaying the test dataset.We can divide our train dataset into two tract train and testimony.We can train the model on this training part and using that make predict for the testimony part.In this way we can validate our soothsaying as we've the true soothsaying for the true soothsayings for.

Correlation attributes :

Grounded on the correlation among attributes it was observed more likely adult income. The attribute that are individual and significant can include education,occupation, age, sex, relationship, income, hourse per week which is since by insight it's considered as important. The correlation among attributes can associated using count in python platform.

Model evaluation :

Evaluate the performance of the model on separate test dataset to determine its accuracy, precision, recall ,support and F1-score.

2.Data exploration and analysis :

In this data set we done some exploratory data analysis .we have plotted some countplot and pairplot of education.

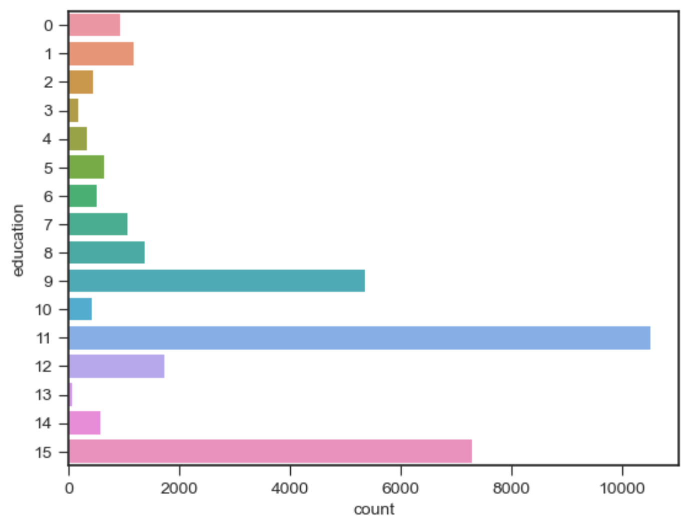

Fig 2.1 : Countplot of Education

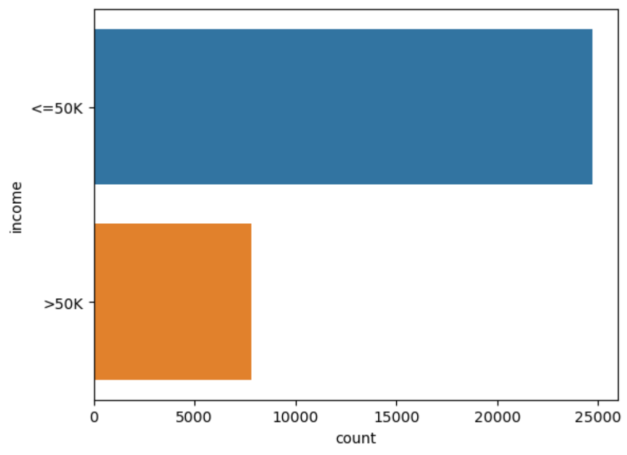

Fig 2.2 : Countplot of Income

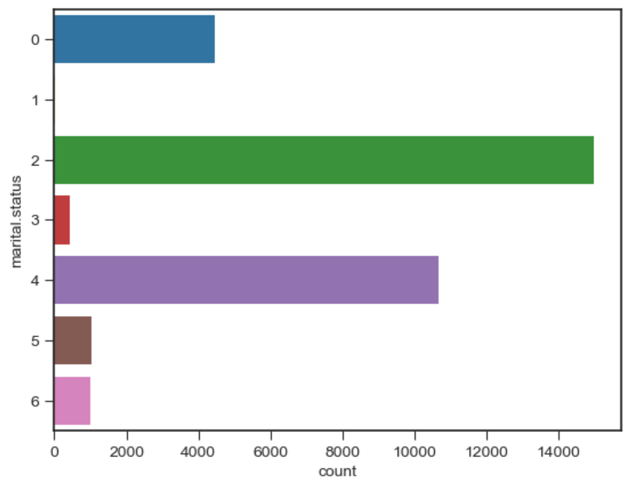

Fig 2.3 : Countplot of marital_status

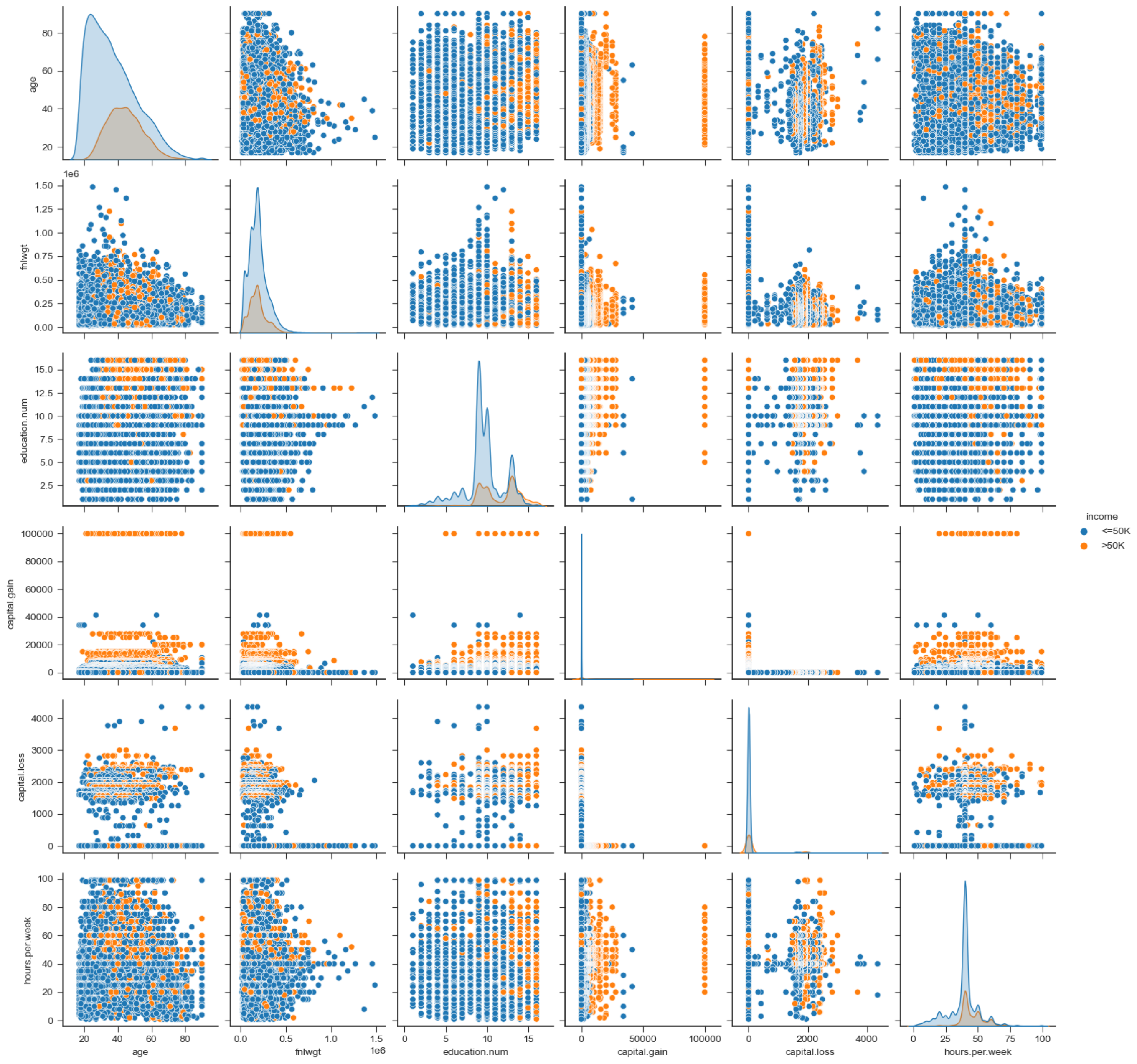

Fig 2.4 : Pairplot of income

Above plot is pair plot in this income related to hours_per_week, capital_gain, capital_loss, eduaction_num, fnlwght and age.

3 . Experiments :

The dataset contains the labels which we have to predict which is the dependent feature ‘Income level’. This feature is discrete consisting of two categories income less than 50k and more than 50k. So the problem we have is a Supervised Binary Classification type.

3.1 Evaluation Metrics :

Accuracy score is not usually the best form of evaluation in machine learning. Evaluation metrics where we build the model and apply that model to our given dataset that means how well a model is able to make prediction or classifications based on a set of input data. The following evaluation metrics have been choosen.

3.1.1 Precision :

Precision shows the ratio of true positives to the sum of true positive and false positives.

$$\text{Precision} = \frac{\text{True\_positives}}{\text{True\_posoitives} + \text{False\_posoitives}}$$

From the above equation,It is clear that for a good model.False_positives should be as small as possible.precision lies between 1(good) and 0(bad).

3.1.2 Recall :

This is the ration of true positives to the sum of true positives and false negatives.

$$\text{Recall} = \frac{\text{True\_positives}}{\text{True\_positives} + \text{False\_negatives}}$$

Here, for a good model,Flase_negatives should be as small as possible.Recall also lies between 1(good) and 0(bad).

3.1.3 Accuracy :

Accuracy summarieses the whole modle. It is the ratio of the correctly classified prediction to the entire prediction.

$$\text{Accuracy} = \frac{\text{Correct \_Predictions}}{\text{All\_Predictions}}$$

3.1.4 F1 Score :

The F1 score seems to be the most suitable metrics for evaluating imbalanced data problems.This is because it in corporate both the precision and recall scores.

$$\text{F1\_Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.2 procedure

3.2.1 any further processing?

Further data processing such as to use standardscaler is to improve the performance of the machine learning models. Many machine learning algorithms such as regression are sensitive to the scale of input feature. when some features have a larger scale than other so it can lead to biased results and reduced performance. By StanderScaler, we ensure that all feature have similar scale and thus avoid this issue. StandardScaler is typically used as a preprocessing step before training a machine learning model. It involves fitting the StandardScaler on the training data and then transforming both the training and test data using the same scaler.

This ensures that the scaling is consistent across the training and test data. The main use of StandardScaler is to standardize the features of a dataset to improve the performance of machine learning models, especially those that are sensitive to the scale of the input features.

3.2.2 Models used :

There are several models used in machine learning supervised binary classification, but for this use case we have label data i.e. target column contains adult_census income. Some of the popular one is Logistic Regression, SVC(support vector claasifier), KNN (K-Nearest Neighbour)claasifier and naive bayes, decision tree. and check which model gives best suitable accuracy and which model we can deploy further.

4.Result :

In the below table we compare of the multiple algorithms that I used.In our dataset we have used different algorithms like logistic regression, decision tree classifier, SVC(support vector classifier), KNN(k-nearest neighbour)classifier, and Navie bayes that accuracy as below.

classifier	class	labels	PREC	REC	F1- SCORE	SUPPORT	ACC

Logistic Regression	Adult_census Income using ml	0(<=50k)
		1(>50k)
			0.81
			0.62
				0.95
				0.28
					0.87
					0.38
						4963
						1550
							78.81%

Decision tree classifier	Adult_census Income using ml	0(<=50k)
		1(>50k)
			0.85
			0.75
				0.95
				0.47
					0.90
					0.58
						4963
						1550
							100%

Naive bayes classifier	Adult_census Income using ml	0(<=50k)
		1(>50k)
			0.85
			0.68
				0.93
				0.48
					0.89
					0.56
						4963
						1550
							82.28%

Support vector classifier	Adult_census Income using ml	0(<=50k)
		1(>50k)
			0.94
			0.54
				0.87
				0.75
					0.90
					0.62
						8115
						1654
							Train-85.42%
							Test-84.76%

K-Nearest neighbor classifier	Adult_census Income using ml	0(<=50k)
		1(>50k)
			0.88
			0.64
				0.90
				0.61
					0.89
					0.63
						7457
						2312
							82.83%

5. DISCUSSION

Justify five algorithms :

Supervised machine learning algorithms are a type of machine learning algorithm that learns to make predictions based on labeled training data. In supervised learning, the algorithm is trained using a set of inputoutput pairs, where the input is the data and the output is the corresponding label or target value that we want the algorithm to predict. The goal of the algorithm is to learn a mapping between the inputs and outputs, so that it can make accurate predictions on new, unseen data. There are many different types of supervised learning algorithms, including: Regression: Regression algorithms are used to predict a continuous value, such as a price or a temperature. Classification: Classification algorithms are used to predict a categorical value, such as whether an email is spam or not.

5.1 Logistic regression :

This model was the least successful model as the data had a fatal flaw causing incompatibility with the way the classifier works. The binary data points make for poor training in the secondary Y vector and also has issues in not being totally linearly separable.

5.2 Decision trees:

Decision trees are a type of algorithm that make predictions by partitioning the data into smaller subsets based on a set of rules or conditions. A decision tree is a type of supervised learning algorithm used in machine learning for both classification and regression tasks. It is a treelike model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each internal node of the tree represents a decision on a feature, and each leaf node represents a class label or a numerical value. The tree is built by recursively partitioning the feature space into smaller and smaller regions, based on the values of the features, until the regions are pure or meet some other stopping criterion.

The decision tree algorithm works by selecting the feature that provides the most information gain, or reduction in impurity, at each node. Information gain is typically measured by entropy or Gini impurity, which are measures of the degree of randomness or uncertainty in the class distribution at a given node. The algorithm stops when all the observations in a node belong to the same class, or when some other stopping criterion is met, such as a maximum depth or a minimum number of observations per node.

5.3Naive Bayes :

It is a type of probabilistic classification algorithm used in machine learning. It is based on Bayes' theorem, which describes the probability of an event occurring based on prior knowledge of conditions that might be related to the event. Naive Bayes is called "naive" because it assumes that the features used for classification are independent of each other, which is often not the case in practice.

Naive Bayes, the algorithm builds a probabilistic model of the training data, assuming that the distribution of each feature is independent of the other features. Then, for a new observation, the algorithm calculates the probability of each class given the observed values of the features, using Bayes' theorem.

The class with the highest probability is assigned as the predicted class for the new observation. here are three main types of Naive Bayes classifiers: Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes. Gaussian Naive Bayes is used for continuous numerical data, Multinomial Naive Bayes is used for discrete count data, such as text data, and Bernoulli Naive Bayes is used for binary data.

The Naive Bayes algorithm is computationally efficient and requires relatively small amounts of training data to achieve high accuracy. It has been used in a variety of applications, such as text classification, sentiment analysis, spam filtering, and medical diagnosis. However, its performance may suffer if the assumption of independence between features is strongly violated, or if the training data is imbalanced or noisy.

5.4 SVC(support vector classifier) :

The most applicable machine learning algorithm for our problem is Linear SVC. Before hopping into Linear SVC with our data, we're going to show a very simple example that should help solidify your understanding of working with Linear SVC.

The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses, though you can use this for many situations. SVC, or Support Vector Classifier, is a supervised machine learning algorithm typically used for classification tasks. SVC works by mapping data points to a high-dimensional space and then finding the optimal hyperplane that divides the data into two classes.

5.5 KNN(k-nearest neighbour) :

In KNN classification, the K nearest neighbours is found based on a distance metric, such as Euclidean or Manhattan distance, and the class of the new observation is assigned based on the majority class of the K neighbors. In KNN regression, the K nearest neighbours is used to calculate the average or median value, which is assigned as the predicted value for the new observation.

The choice of the value of K is an important parameter in KNN. A larger value of K makes the algorithm more robust to outliers and noise, but can lead to overfitting in some cases. A smaller value of K makes the algorithm more sensitive to noise and may result in overfitting. The optimal value of K can be determined through crossvalidation or other model selection techniques.

6 Conclusion :

It is clear that one of the main advantage of feature engineering lies with the modelbuilding process. The dataset used in this notebook only has 15 columns.Even then we can see that the time needed to train between the model with and without feature engineering are nearly 2 times apart. With a more complex model,such as the neural network and with more features, the time needed to train the model can go up signficantly.Additionally, with feature engineering, it significantly reduces the time needed to tune the parameters and the danger of overfitting on the given dataset.Although the model without feature engineering scores better on both the train and test set, I would have more confidently that the feature engineered dataset would perform better and more consistent because the score between train and test set is much closer as compared to the non-feature engineered dataset.

References :

1.Dataset from https://www.kaggle.com/datasets/uciml/adult-census-income

2.”what is machine learning”by https://monkeylearn.com/machine-learning

3.”logistic regression”by https://www.capitalone.com/tech/machinelearning/what-is-logistic-regression

4. "Introduction to Machine Learning with Python" by Andreas C. Müller and Sarah Guido

5. "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

6. "Machine Learning Yearning" by Andrew Ng "Bayesian Reasoning and Machine Learning" by David Barber

7.”SVC”byhttps://vitalflux.com/svm-classifier-scikit-learn-code.