Classifying the income groups based on 1994 US Census Data

Vineeth Bodduvaram
PES2201800046
PES University - EC Campus
bvineeth2000@gmail.com

Manish M
PES2201800428
PES University - EC Campus
manishmanikandan2000@gmail.com

Aryan Vijay
PES2201800029
PES University - EC Campus
aryanvijay2048@gmail.com

***Abstract*—A census is a procedure of systematically enumerating, and acquiring information of the members of a population. Our project handles the scenario of predicting which citizen is rich or poor. We have classified it based on the annual income, if the income is greater than $50k/year, the person is rich. We can predict based on different factors such as marital status, education, age, etc. We have used Barry Becker's 1994 dataset in our project.**

***Index Terms*—K-Nearest Neighbours , Gaussian Naive Bayes , Decision tree , Census**

I. Introduction

This problem is important as it analyses the dataset in a very structured manner as we need to look at each attribute and the effect it has on the income of a certain individual, although this dataset is more than two decades old, it's an important analysis as the same methods can be applied on modern datasets and the results from both the datasets could be compared.The dataset comprises of multiple attributes such as race, education, age, employee, marital status etc, which would directly affect the income of a person in 1994, as explained in our report which says that married people earned more than single people during the year 1994. We have used different methods such as logistic regression, K nearest neighbours and decision tree classifiers to compare the results.This dataset is extracted by Barry Becker in 1994 and it is found in the UCI Machine Learning Repository. The dataset contains different attributes such as age, marital status, education, gender, etc which are vital pieces of information for predicting the income of a citizen, the dataset also contains 41 additional countries apart from The United States of America for a broader range of users.The dataset contains multiple null values which will be filled in the methods as given in the report.

II. Dataset

The Census Income dataset has 32514 entries and 15 features

- **age**: Age of the individual.
  As shown in Fig.1 The average age group is between 20 and 45
- **Workclass**: Working in private sector or the public sector
- **fnlwgt**: The number of people the census believes the entry represents

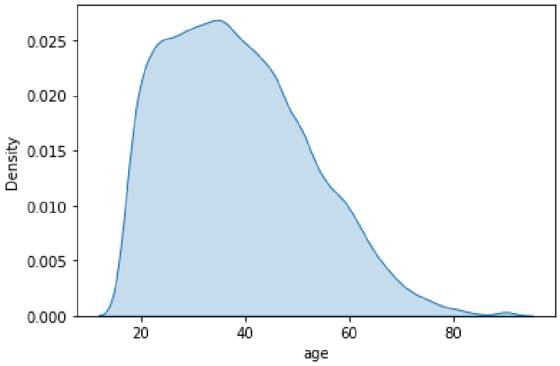


Fig. 1. Density distribution of age

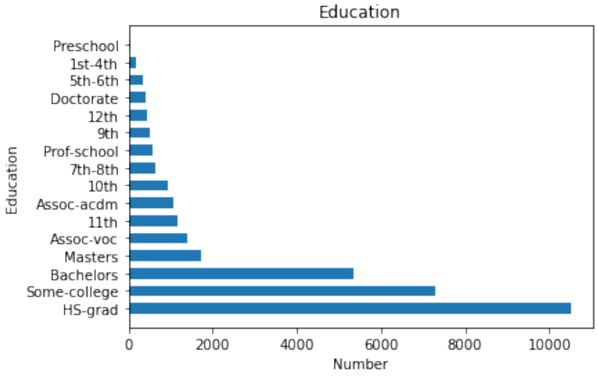


Fig. 2. Distribution of Highest education level

- **education**: The highest grade that has been finished or the top degree that has been achieved
- **education-num**: Numerical representation of the highest education accomplished. *The distribution is shown in Fig 2.*
- **relationship status**: Represents what this individual is relative to others.Ex Father of , Son of , Mother of etc
- **marital_status**: person is married/divorced/widowed or single
- **occupation**: The type of work done by the person. *The distribution is shown in Fig 3.*
- **workclass**: Working in private sector or the public sector

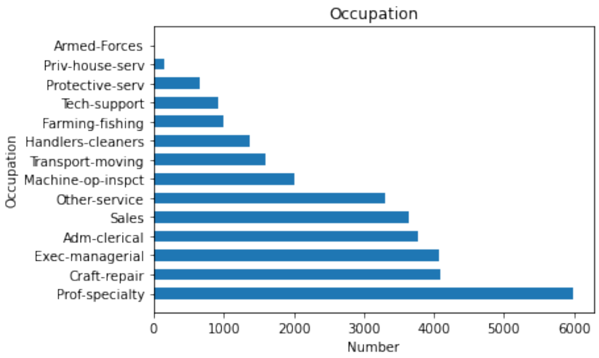


Fig. 3. Distribution of Occupation

- **Hours Worked per week**: The total number of hours work has been done by the individual in a week. *The density distribution is shown in Fig 4.*

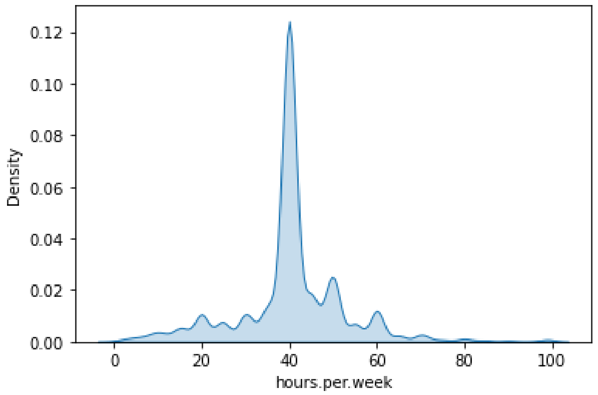


Fig. 4. Density Distribution of Hours worked per week

- **race**: the race to which a person belongs (ex. white, black , Hispanic etc.)
- **sex**: Biological gender of the person
- **capitalgain**: capitalgains for an individual
- **capitalloss**: capitalloss for an individual
- **label**: Which income category the individual belongs to.

III. Results From analysis

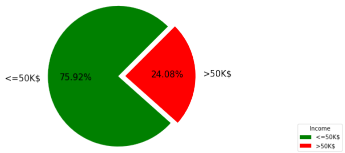


Fig. 5. Distribution of Income

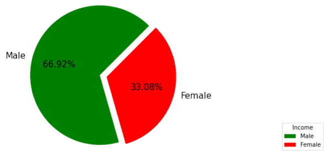


Fig. 6. Gender Distribution

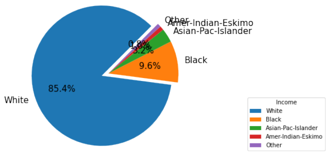


Fig. 7. Race Distribution

Fig 5 shows that 75.92% of entries have income below 50K$ therefore there is a possibility of high bias towards lower income during prediction
Fig 6 shows that 66.92% of entries are male.
Fig 7 shows that 85.4% of entries are white therefore the prediction model will have a very low accuracy for inputs where race is other than white as data is not equally distributed among all the races present in the dataset.
Fig 8 The above corr plot shows that education number , gender , age, and hours-per-week have the highest correlation with income

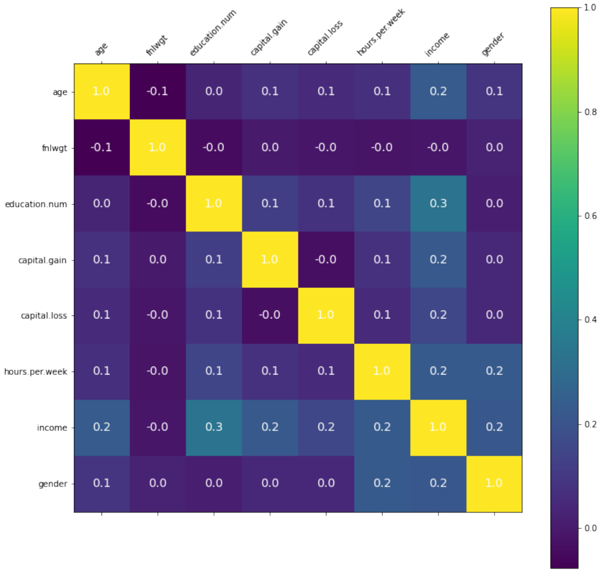


Fig. 8. Correlation plot

Fig 9 is the distribution of income with respect to age

Fig 10 is the distribution of income with respect to Capital diff

Fig 11 is the distribution of income with respect to Hours per week

Fig 12 is the distribution of income with respect to working class

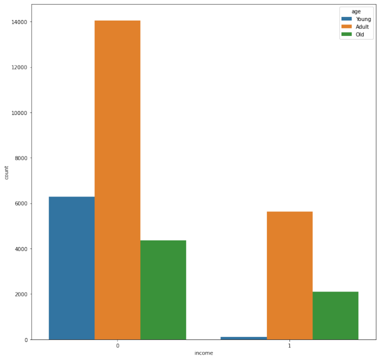


Fig. 9. Distribution ofAge with respect to Income

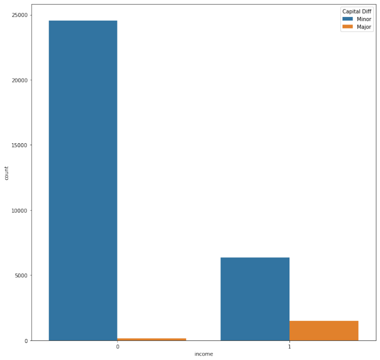


Fig. 10. Distribution ofCapital Diff with respect to Income

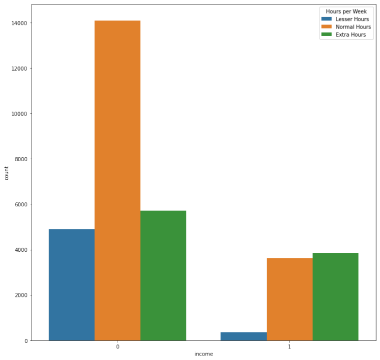


Fig. 11. Distribution of Hours per week with respect to Income

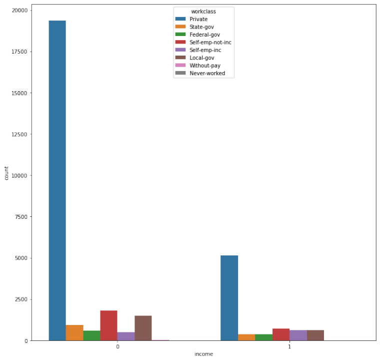


Fig. 12. Distribution of working class with respect to Income

Fig 13 shows that majority people who have their highest education level as high school don't earn have low income

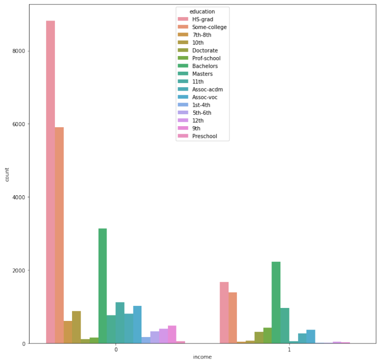


Fig. 13. Distribution of Education level with respect to Income

Fig 14 is the distribution of income with respect to occupation

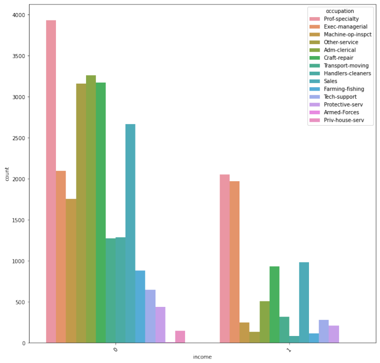


Fig. 14. Distribution of Occupation with respect to Income

Fig 15 is the distribution of income with respect to gender

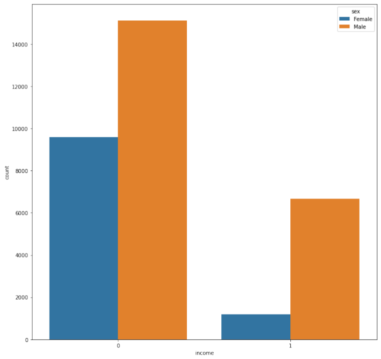


Fig. 15. Distribution of Native country with respect to Income

Fig 16 is the distribution of income with respect to native country being USA or not

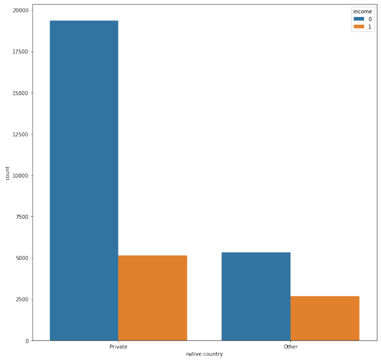


Fig. 16. Distribution of Gender with respect to Income

Fig 17 shows that white people have a higher chance of earning more than compared to other races

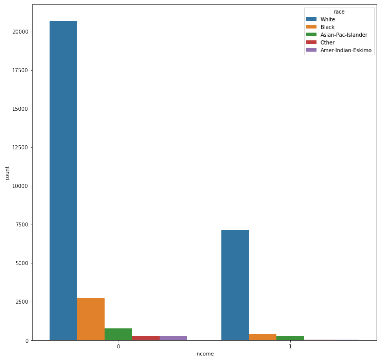


Fig. 17. Distribution of Race with respect to Income

IV. Pre-processing

We have done the pre processing for the dataset in the following steps

- There were dummy values like '?' in the dataset, we replaced those with Null values, which in turn were replaced by mode and median for categorical and numerical values respectively
- The income values were mapped to 0 and 1, if the value is 0 income is greater than $50000, else it is less than or equal to $50000
- We applied one hot encoding for categorical values with more than two features
- After encoding all features, we scale them.(we have use a simple min-max scaling method here)

V. Feature selection and ranking

We select the 10 most important features from the dataset using the Random Forest Classifier algorithm. The Random Forest Classifier will be included during the hyperparameter tuning phase to pick the top 10 features. This is used to identify which feature works best with which classifier.

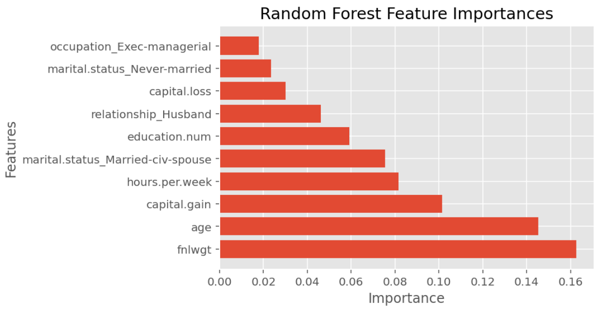


Fig. 18. Shows the most relevant features after RFI ranking

VI. Models Used

A. Splitting the data for training and testing

We split the 32000 odd rows into training and testing sample using the scikit-learn package. We follow the basic rule of splitting it into test size of 30% and training size of 70%. The only prerequisite would be to install the scikit-learn package beforehand.

B. K-Nearest Neighbours(KNN)

KNN is an instance of lazy learning where the algorithm predicts the best value for the output based on the following factors

- We predict the output based on the vote of it's neighbours
- We predict the output based on the property value of the object

We create a pipeline for the K-Nearest Neighbours with the following parameters:

- The Number of neighbours(N)
- The distance method we used for finding the centroid of the cluster(the cluster formed by the data point and it's nearest neighbours, this is where the it gets the name KNN from).

To find the best value of K we have plotted a K vs AUC score curve and from the results of that curve we found that the optimal value of K is 10.0. And the distance metric used for calculating the distance was Manhattan distance The mean AUC score for the most optimal value of K was found to be 0.870
To find the best performing we graphed *Fig19* the results of running the model with the number of vertices vs the ACU score and checked whether Manhattan or Euclidean distance gave higher accuracy.

We observe that the difference between the hyperparameter combinations is not really much when conditioned on the number of features selected. Let's visualize the results of the grid search corresponding to 10 selected features.
Using the K-Nearest Neighbours we achieve an AUC score of 0.870

C. Naive Bayes

Naive Bayes Classifier is a probability-based classifier algorithm which has it's roots in Bayes Theorem in

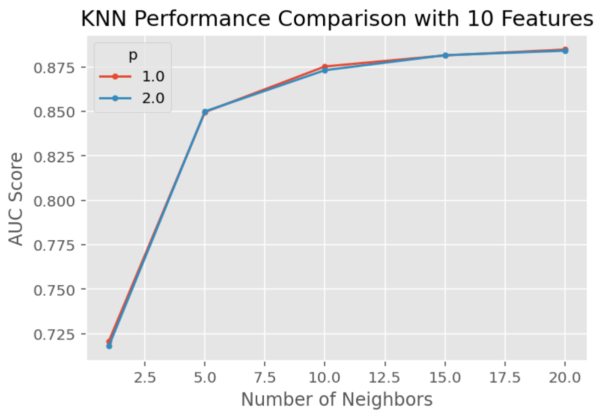


Fig. 19. KNN Performance Comparison

Probabilistic math We create an NB pipeline which is not coupled with Kernel density estimations as we get a decent enough result for our project without implementing the estimations Since we require a uniform Gaussian distribution to implement this algorithm, we need to perform a transformation to achieve that. We conduct the grid search over the powers of 10 so that we can achieve faster run times on our model, although this will provide us with lesser accurate results as we have prioritized time over accuracy We found the AUC score for the most optimal NB model to be 0.878(all of these assumptions are made by taking the top 10 features).
The best model is achieved by running the model with different smoothing as shown in*(Fig 20)*.

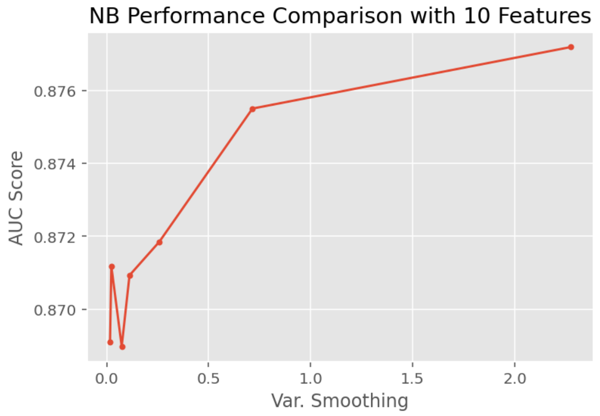


Fig. 20. Comparison with NB with Laplace Smoothing

The optimal Naive Bayes model yields an AUC score of 0.878 (with 10 features)

D. Decision Trees

We are using a decision tree model for this dataset because it uses a tree like model of decisions and it's possible consequences which will be very crucial for predicting the income based on the top 10 features that we have already selected We build a decision tree using the Gini index as it is a popular statistical measurement where we can determine the wealth variability or mutability in a specified group. Here the specified group implies to our dataset of 42 countries To find the best possible maximum depth for our decision tree, we set the values of max depth in multiples of 5 ex. 5,10,15, and so on. From the results found we found that the best possible value of the depth was found to be 5. The best Decision Tree model has a maximum depth of 5 with an AUC score of 0.881
.

The model's depth was decided by changing the max depth and running the model and verifying it.
Fig 21 shows that increasing the depth also increases the AUC score of the model.
The best Decision Tree model has a maximum depth of 5 with an AUC score of 0.881.

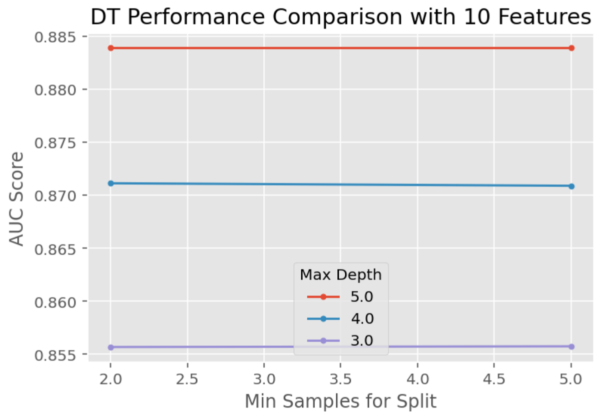


Fig. 21. Comparison of DT model with depths as 3,4,5

We notice that the optimal value of maximum depth hyperparameter is at the extreme end of its search space. Thus, we need to go beyond what we already tried to make sure that we are not missing out on even better values. For this reason, we try a new search as below.With max depth as 5,10,15.

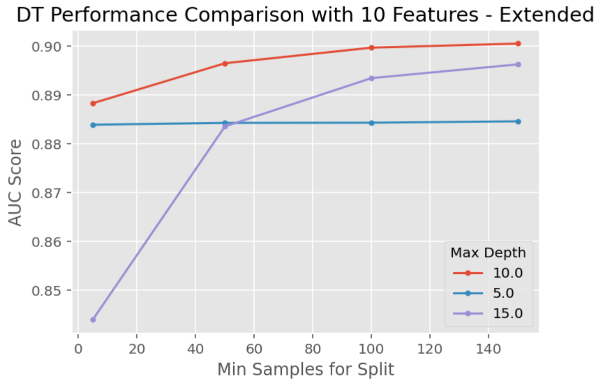


Fig. 22. Comparison of DT model with depths as 5,10,15

Fig 22 we can achieve slightly better results with the new search space.

VII. Performance Comparison

A. Cross-Validation

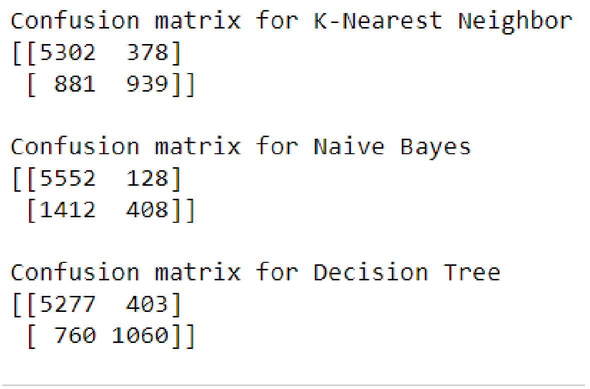
This is a very important phase in our evaluation as we need to select the most optimal value for the hyperparameter of each classification algorithm as mentioned above. The selection of hyperparameters is really important as it determines the architecture of our model for predicting the income based on the census dataset. During the hyperparameter tuning phase we used all of our training data. For instance, in the KNN algorithm we found that the optimal value of K=10.0, after cross-validating our training data with all the features in our dataset, as some of them may not be “seen” by the models. Since cross-validation in itself is a very random process, the best method would be to perform some basic statistical tests om our dataset. In our case, we have performed pairwised t-tests and the results are as follows.

The Results are as follows:

Model	AUC Score
K-Nearest Neighbours	0.8692513514869009
Naive Bayes	0.8820906340095377
Decision Trees	0.8921394169921564

The above table shows that Decision trees gives the best model for this dataset

B. F1 score and Confusion Metrics



```
Confusion matrix for K-Nearest Neighbor
[[5302  378]
 [ 881  939]]

Confusion matrix for Naive Bayes
[[5552  128]
 [1412  408]]

Confusion matrix for Decision Tree
[[5277  403]
 [ 760 1060]]
```

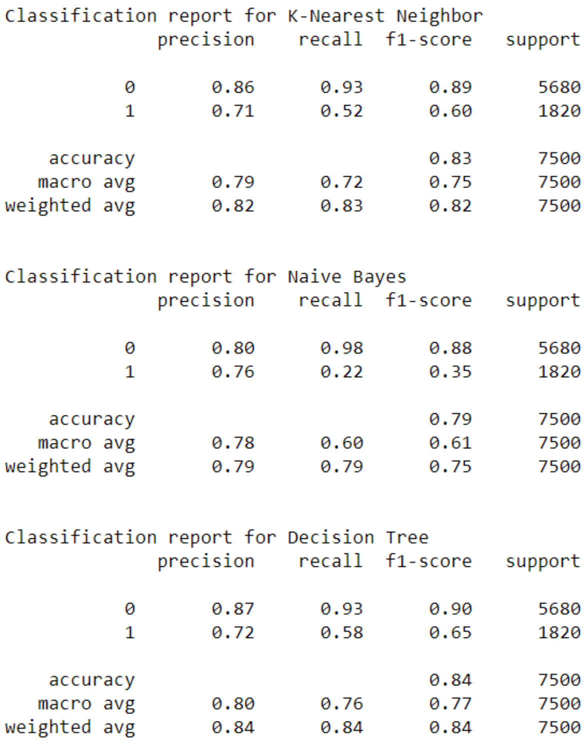
Fig 23 is the output of calculating the accuracy and F1 score.

Fig 24 is the confusion matrix of the models used.

VIII. Papers Used as Reference

A. Paper 1 [2]

The key points of this paper



```
Classification report for K-Nearest Neighbor
              precision    recall  f1-score   support

           0       0.86      0.93      0.89      5680
           1       0.71      0.52      0.60      1820

    accuracy                           0.83      7500
   macro avg       0.79      0.72      0.75      7500
weighted avg       0.82      0.83      0.82      7500


Classification report for Naive Bayes
              precision    recall  f1-score   support

           0       0.80      0.98      0.88      5680
           1       0.76      0.22      0.35      1820

    accuracy                           0.79      7500
   macro avg       0.78      0.60      0.61      7500
weighted avg       0.79      0.79      0.75      7500


Classification report for Decision Tree
              precision    recall  f1-score   support

           0       0.87      0.93      0.90      5680
           1       0.72      0.58      0.65      1820

    accuracy                           0.84      7500
   macro avg       0.80      0.76      0.77      7500
weighted avg       0.84      0.84      0.84      7500
```

- The method used in cleaning the samples was to remove all the entries with default or NA values
- Features such as “fnlwgt” ,”relationships”,”Capital gain/loss” were removed as they had a large amount of bad data or the data wasn’t useful
- Features such as race were also removed as most of the entries under that feature were ”White” therefore introducing bias to the models.
- The main Features selected to run the model on were:
  - Age
  - Education
  - Hours per week
  - Occupation
  - Gender
- The original distribution of the gender feature with respect to income would be maintained while sampling and in the training and testing phase of the model to get an accurate model to the population of the dataset
- The methods used were
  - Baseline
  - Naive Bayes
  - Naive Bayes (grouped) (ie. where the education qualification below 12th grade was grouped into one
  - Logistic Regression
  - K Nearest Neighbour
  - Decision tree
- Logistic Regression didn’t give any acceptable results as the error rate was extremely high
- Computing the model using the K nearest neighbours algorithm requires a lot of processing which in turn took

an extremely large amount of time

Limitations

- Ignoring the Feature ‘Race’ would give a different picture as there was a large amount of Racism in the US in 1994 and only now seen a moderate decline in it.
- Marital status was not taken into consideration as people with family usually tend to work harder
- As job availability ,average education , economic growth of the country and many factors change over a period of 26 years , the conclusion during that time period cannot be extended to our time.

B. Paper 2 [1]

The key points of this paper

- Three tasks were done to get the dataset ready for running the decision tree classifier
  - Using One hot encoding on categorical attributes
  - Dropping unnecessary columns and combining others
  - Checking for NaN values and preparing for separate features and target data frames
- The Random Forest classifier was used to for this paper for primarily two reasons
  - The target variable is a binary variable, therefore classification algorithms are better than regression algorithms
  - Random Forest classifier gave a higher accuracy score compared to other standard classifier algorithms such as Gaussian NB classifier.
- Education, age, capital and working more attributes were found to have a positive correlation with income
- The paper also concluded the following points
  - Self employed and government employees tend to have a higher income
  - Married people tend to have higher income
  - Whites earn more than other races
  - Males earn more than Females
- The top 5 features which were crucial to the income was attributed as follows
  - Age
  - Capital Gain
  - Hours per week
  - Education
  - Marital Status

Limitations

- Fitting 42 different countries is difficult as the environment and the value of currency and various other factors may not be the same giving rise to discrepancies.
- Cannot use the computed model in the present year as there is a large generation gap which leads to change in environment and the sources of income, various trends change across the years as well..

C. Paper 3 [3]

The Census dataset used in this paper is large with about 48,842 entries with 14 different features.

There were surprisingly less amount of data loss with only 3,620 entries having missing data. Therefore they removed the entries as it barely amounts to 10% of the dataset.With only a quarter of the entries earning an annual income of more than $50K The dataset from the UCI are already randomly arranged. The partition ratio used for testing and training were 33:66 Further the training dataset was split into training and Validation again split in the same ratio.

Models

The author of this paper decided on using the Naive Bayes model to predict the labels for the entries as there were no requirement to regularize the parameters.

The test errors had a very close representation , the high correlation between the education number and the age resulted in the overcounting in the case of Naive Bayes model. Discarding the age feature altogether reduced the overall test error.

Logistic Regression:

The logistic regression was chosen as it does not encounter the problem of overcounting like in the case of Naive Bayes . To avoid the problem of generating a model which ends trapped in a local minimum, the model was run a total of 10 times and the best performing model among them was chosen.Even though this does not assure that the model has achieved global minimum, this helps in avoiding in entering a very bad local minimum. In this model the parameters have to be regularized.

From the experiments with feature representation, in many cases it’s better to represent a field using categories than numerical values, especially when numerical encoding over complicates the data. For example, a 21 and 24 year-old would not be expected to make more than $50,000 a year in 1994. According to dollar signs.com, that would be the equivalent about $80,000 today. To reduce noisiness, it is preferable to group them up if they tend to give the same inference. This can also help speed up the model’s training as well. For SVMs for example, reducing the feature space will greatly improve its training and classification time. And as we also saw with different models, it can be better to discard an attribute that positively correlates with another existing attribute, like age and level of education.

In case of logistic regression the author decided to represent the categorical valued fields as it is instead of a numerical representation , and also mentions that encoding the data just introduces complicatedly large data to process.
Expecting a teenager to earn more than 50,000 a year in 1994 which is equivalent to 80,000 , which are outliers in case they are true which has a huge impact on the model, therefore the author decided to have categorical features grouped up instead. Also he discarded the attributes which have a huge correlation

with other independent features to decrease the bias(ex education number and age).

Limitations

- The Split of the data-set was uneven as the race was highly dominated as white in the dataset the same was not accounted to an accurate level while splitting the dataset.
- Not combining some of the noisy values together, which led to some out of bound predictions

IX. Conclusion

As seen in our project report, we've used three classification methods to predict the income based on the top 10 attributes of the dataset which are KNN Classifier , Naïve Bayes classifier and Decision Tree classifier, from the results found from all three methods we can infer that the decision tree is the perfect fit for this dataset as it gives us the highest accuracy and gives us the best prediction
However there are some drawbacks to this as we've made some crucial assumptions while doing our project

- We have assumed that the features will have the same impact on all the countries when it's highly unlikely that this will be true
- We have also used the following methods on a dataset which is 25 years old and these trends won't be similar to the trends in 2020 as times have changed.
- There would be a significant difference in the value of the American dollar from 1994 and the present day due to inflation which prevents the model from being of any use

References

[1] C. Zalazo Lemon and C. Mulakaluri. Predicting if income exceeds $50000 per year on 1994 census dataset with simple classification techniques. *Aging Clinical and Experimental Research.*

[2] Sisay M. Predicting income from the 1994 dataset using decision trees and random forest classifier. 2014.

[3] Nham. Classifying income from the 1994 census dataset.a0994191. 2016.