CIND820 - Big Data Analytics Project

Machine Learning Algorithms to Predict Income Level: A Comparative Study

Syeda Sunjida Aleya
500467980

Supervisor: Dr. Tamer Abdou

Ryerson University

Table of Contents

Abstract

In the current globalization, even though global inequality has reduced (between nations), it has increased inequality within nations. In the last 30 years, inequality within most nations has risen significantly, particularly among advanced countries. In this period, close to 90 percent of advanced economies have seen an increase in income inequality. Research suggests that economic growth and macroeconomic stability are greatly affected by inequality. Now a days inequality is at the center stage of economic policy debate across the globe, as government tax and spending policies have significant effects on income distribution. Governments of different countries have been trying their best to address this problem and provide an optimal solution. The aim of this study is to find a solution for the income inequality problem using machine learning and data mining techniques. The goal is to explore the U.S. Census-22 dataset and develop a classification model to predict whether an individual's income will be in “Low” or “High” category. This prediction will be based on several demographic and employment related variables that describe the personal income data.

In addition, I will apply three different feature selection methods (Filter, Wrapper, and Embedded) and will compare the performance of five the machine learning algorithms (Logistic regression, Naïve Bayes, K-Nearest Neighbours, Decision Tree, and Random Forest) on each selected group of features to see which method is best. This will help to investigate which variables are highly correlated with our target variable “Income”. Lastly, based on top attributes I will compare all five machine learning algorithms to see which one will outperform in terms of efficiency, effectiveness, and stability.

Dataset
https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.html

GitHub
SSAleya/CIND820_Census22 (github.com)

Introduction

Since 19th century, worldwide income inequality has been constantly increasing. Income inequality refers to the crucial imbalance of income distribution throughout a population. Less equal distribution results in higher income inequality. Dispersions of income inequality are an ongoing area of analysis for both local and global governing institutions. Researchers commonly analyze income distributions based on gender, ethnicity, geographic location, age, occupation, historical income etc. The aim of this capstone project is to re-examine the major features responsible for income inequality using machine learning and data mining techniques.

According to UNICEF, the United States is claimed to have a Gini index of 36 on an unweighted average basis in 2008, showing quite high inequality among the country's population. This project would inspect the dataset taken from U.S. Census Bureau website which consist of Current Population Survey Annual Social and Economic Supplement (ASEC) dataset, 2022. Following previous studies, 40 demographic and employment related variables are selected which describe the personal income data. The discreate target variable "Income" will be pre-processed, and optimum number of income levels will be selected for best prediction. Different feature selection methods such as Filter, Wrapper and Embedded methods will be applied and evaluated to select the best attributes.

The objective of this project is to identify the most efficient and best performing models that can be used to predict the income category based on the selected top attributes. The focus will be on evaluating and comparing machine learning classification models such as Decision Tree, Random Forest, Naive Bayes, Logistic Regression, and K-Nearest Neighbors (KNN) classifier. The dataset will be split into 80/20 for One-Hold (Train-Test Split) and for RepeatedKFold, k=10(n=5) will be considered for best results. The performance of each of these classification models (Efficiency, Effectiveness and Stability), would be evaluated based on several performance evaluation measures such as Confusion Matrix, Accuracy, F1-Score, Precision, Recall, and Matthews Correlation Coefficient (MCC) & Run Time.

Literature Review

Certain efforts using machine learning models have been made in the past by researchers for predicting income levels.

Navoneel Chakrabarty and Sanket Biswa [1] conducted a comprehensive analysis to highlight the key factors that are necessary in improving an individual's income. The had used 48,842 different records and 14 attributes for 42 nations. They had shown the Gradient Boosting Classifier Model achieved the highest accuracy of 88.16%, which broke the benchmark accuracy of existing works.

Using the same dataset, Alina Lazar [2] implemented Principal Component Analysis (PCA) and Support Vector Machine (SVM) methods to generate and evaluate income prediction. According to their study, “A detailed statistical study targeted for relevant feature selection was found to increase efficiency and even improve classification accuracy. A systematic study was performed on the influence of this statistical narrowing on the grid parameter search, training time, accuracy, and number of support vectors. Accuracy values as high as 84%, when compared against a test population, were obtained with a reduced set of parameters while the computational time was reduced by 60%”.

In his study, Sisay Menji Bekena [3] applied Random Forest Classifier machine learning algorithm to predict income levels of individuals based on the dataset acquired from UCI Machine Learning Repository. Based on 1994 census database, he has included 32,561 individual data on 13 attributes in his study. Using Random Forest classifier, he had shown that the predictive accuracy of the model on test data was 85%. According to his findings, important features prediction shows marital status, capital gain, education, age and hours per week are the top features which account for larger shares of the model accuracy. Also, using decision tree classifier, he had shown that these variables are the top 5 features in importance.

Ron Kohavi [4] showed that in some large databases, the accuracy of neither Naïve-Bayes nor Decision trees do scale up. In his study using Fisher’s Iris dataset he proposed a new algorithm, NBTree, which induces a hybrid of Decision-tree and Naïve-Bayes classifiers. In NBTress, the decision-tree nodes contain univariate splits as regular decision-trees, but the leaves contain Naïve-Bayesian classifiers. This is most suitable for scenarios where many attributes are significant in predicting the label, but they are not all necessarily conditionally independent. Chet Lemon, Chris Zelazo and Kesav Mulakaluri [5] also implemented this hybrid classifier using the US Adult Census dataset.

S. Deepajothi and Dr. S. Selvarajan [6] tried to present a comparative study of the classification accuracy provided by different classification algorithms (Naïve Bayesian, Random Forest, Zero R and K Star) on the census Adult Dataset. Their findings showed that the Naive Bayes classifier is simple and faster, and they also exhibit higher accuracy rate than the algorithms mentioned above.

Roshan Kumari and Saurabh Kr. Srivastava [7] in their research synthesizes binary classification in which various approaches for binary classification were discussed for the Sockpuppet detection.

Thus, after reviewing a variety of literatures on machine learning algorithms, I have decided to address the following research questions: What are the major factors effecting income levels?

Which feature selection technique gives the best set of features? And, out of five models, which one outperforms according to efficiency, effectiveness, and stability?

The data being used for this project will be the latest Current Population Survey Annual Social and Economic Supplement (ASEC) dataset,2022 collected from US Census Bureau website. For this study I have selected only “Person” dataset.

First, I have chosen three feature selection methods (Filter, Wrapper, and Embedded) to investigate the top features effecting income class. Next, I will apply five classification algorithms (Logistic regression, Naïve Bayes, K-Nearest Neighbours, Decision Tree, and Random Forest) to predict binary income class. Using a large income dataset containing 40 demographic and employment related variables and 152732 instances, a comparison of the classification algorithms will be also performed to choose the best model. The tools being used for this project are R Studio and Jupyter Notebook for the Python environment. Visuals will be provided using matplotlib, seaborn, and ggplots.

Methodology

Focusing on the comprehensive application of machine learning algorithms, I wanted to use a latest big dataset for my project. This U.S. Census dataset that I used, contains more than 100,000 instances and chosen 40 attributes gave me enough freedom to apply different methods for best possible outcomes. While investigating the best features to determine income levels, I have additionally implemented different feature selection techniques to compare. At the end, the comparison of the performances of five classifiers revealed the best model.

Literature Review	• Literatures based on U.S. Census Income. • Literaures based on diffrent machine learning algorithms.
Reasearch Questions	• What are the major factors effecting income levels? • Which feature selection technique gives the best set of features? • Which is the best model based on efficiency, effectiveness, and stability?
Data Collection	• Dataset Census Income 2022 collected from U.S. Census Bureau. • Selected 40 attributes based on previous study on U.S. Census Income -KDD (1994-1995) dataset • Initial instances: 152732 , attributes: 40
Initial Data Cleaning	• Removing all income <= 0 • Left 106534 instances and 40 attributes • Tool: RStudio
Exploratory Data Analysis	• Univariate Analysis : Histograms and Bar grapgs • Bivariate Analysis : Pearson and Spearman's Correlation
Preparing Target Attribute	• Converting discreate numerical attribute "Income" into categorical • Using quantiles for income categories. • 2QT, 3QT, 5QT, and 10QT used to make 2,3,5 and 10 categories of the income. • Selected 2QT (two categories: Low and High) evaluating performances of five algorithms.
Feature Engineering	• Feature Selection Methods : (1) Filter Method {MVR, MIG, LVF AND CC}, (2) Wrapper Method{Forward Selection}, and (3) Embedded Method {Sequential Feature Selection base on Random Forest}
Selecting best FS Method	• Comparing performance of five classification algorithms based on subset selected by each Feature Selection method.
Top Attributes Selection	• Selectiong 16 top attributes for prediction of income levels.
Modeling and Prediction	• Applying five machine learning algorithms: Decision Tree, Random Forest, KNN, Naive Bayes, and Logistic Regression.
Performance Analysis	• Evaluating five classifiers in terms of Effectiveness(Accuracy, F1-score, Precision, Recall and Matthews Correlation Cofficient), Efficiency (Run Time), and Stability (Test-Train Split and Repeated K Folds cross validation)
Conclusion	• Finding the answers of the research questions. • Suggesting scope for future improvements.

Data Description and Analysis

Data Source

This Census income dataset is collected from U.S. Census Bureau website which contains the Current Population Survey Annual Social and Economic Supplement (ASEC) dataset, 2022. This is an open-source data for publicly use. I have selected only "Person" dataset for this study. Out of 832 attributes I have selected 40 demographic and employment related attributes following previous study based on Census-Income (KDD) dataset. This data is also provided by the US Census Bureau and has been archived in the University of California, Irvine (UCI) repository which contains weighted census data extracted from the 1994 and 1995.

Data Cleaning

Out of 40, I have 8 numeric and 32 categorical variables. This large dataset initially contains 152732 instances. According to the "Public Use Benchmarks" [cpsmar22.pdf (census.gov) – page 391], I have excluded people without income; where PTOTVAL(Total persons income)=0. Also, after removing negative incomes the secondary dataset contains 106534 instances. This dataset has been encoded already; the detail data description is given in the Table1 [Appendix].

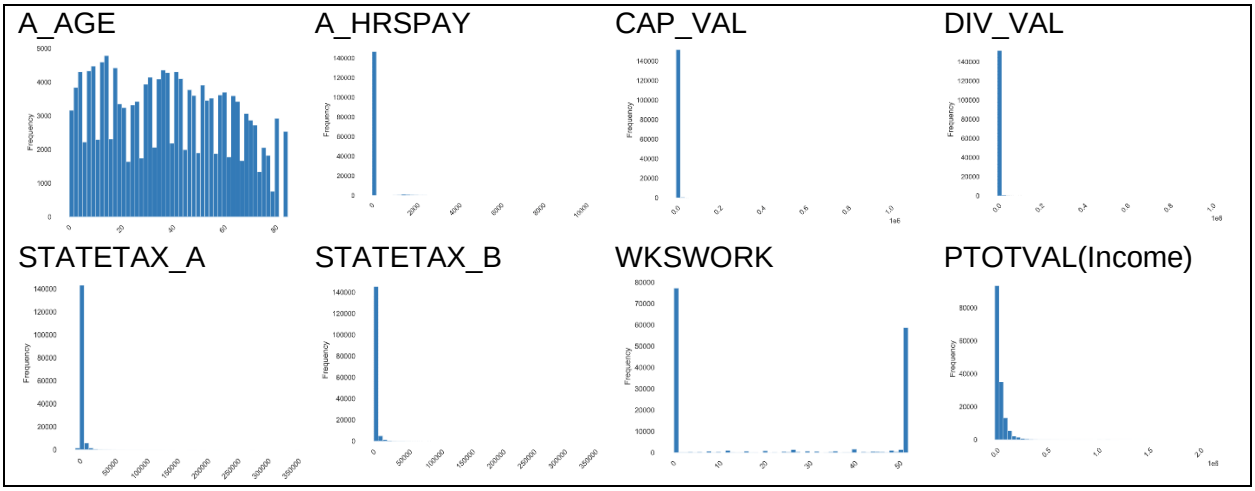
Exploratory Data Analysis (EDA)

I have performed an Exploratory Data Analysis (EDA), using Panda Profiling. The HTML report is available in my GitHub repository. This dataset has no missing values.

Univariate Analysis:

Numeric Attributes: Following are the eight numeric attributes, including our target attribute "Income".

Figure-1A: Distribution of numeric attributes



As we can see in the above Figure-1A, all the seven numeric attributes are left skewed except A_AGE. The Table-2 shows a statistical summary of these attributes.

Table-3: Data Summary of numeric attributes

Attributes	Mean	St. Dev	Min	Q1	Median	Q3	Max
A_AGE	38.28	23.22	0	17	38	57	85
A_HRSPAY	82.65	468.68	-1	-1	-1	0	9999
CAP_VAL	946.62	19628.68	0	0	0	0	999999
DIV_VAL	570.65	9545.69	0	0	0	0	999999
STATETAX_A	1227.15	5046.68	-8428	0	0	163	362021
STATETAX_B	1309.42	5075.38	0	0	0	425	362021
WKSWORK	23.26	24.87	0	0	0	52	52
PTOTVAL(Income)	39710.17	73070.69	-9999	0	19868.5	52005	2082782

Categorical Attributes: Following is the frequency plot of 24 categorical attributes which shows the percentage of each category along with the count.

Figure-1B: Frequency Plot of Categorical attributes (24)

Out of 32, 8 categorical attributes (PNETVTY, PEMNTVTY, PEFNTVTY, PRDTRACE, A_HGA, A_MJIND, A_MJOCC, and HHDFMX) has too many categories and that's why not presented in the above Figure-1B. Details of these eight attributes are presented in the following figure Figure-1C.

The three attributes "PENATVTY", "PEMNTVTY", and "PEFNTVTY" represents own, mother's and father's birth country respectively. The list of countries is presented in Table1-B[Appendix]. As we can see, out of 162 countries, U.S. (Code-57), is the birthplace of most of the people and the second majority is from Mexico (Code-303).

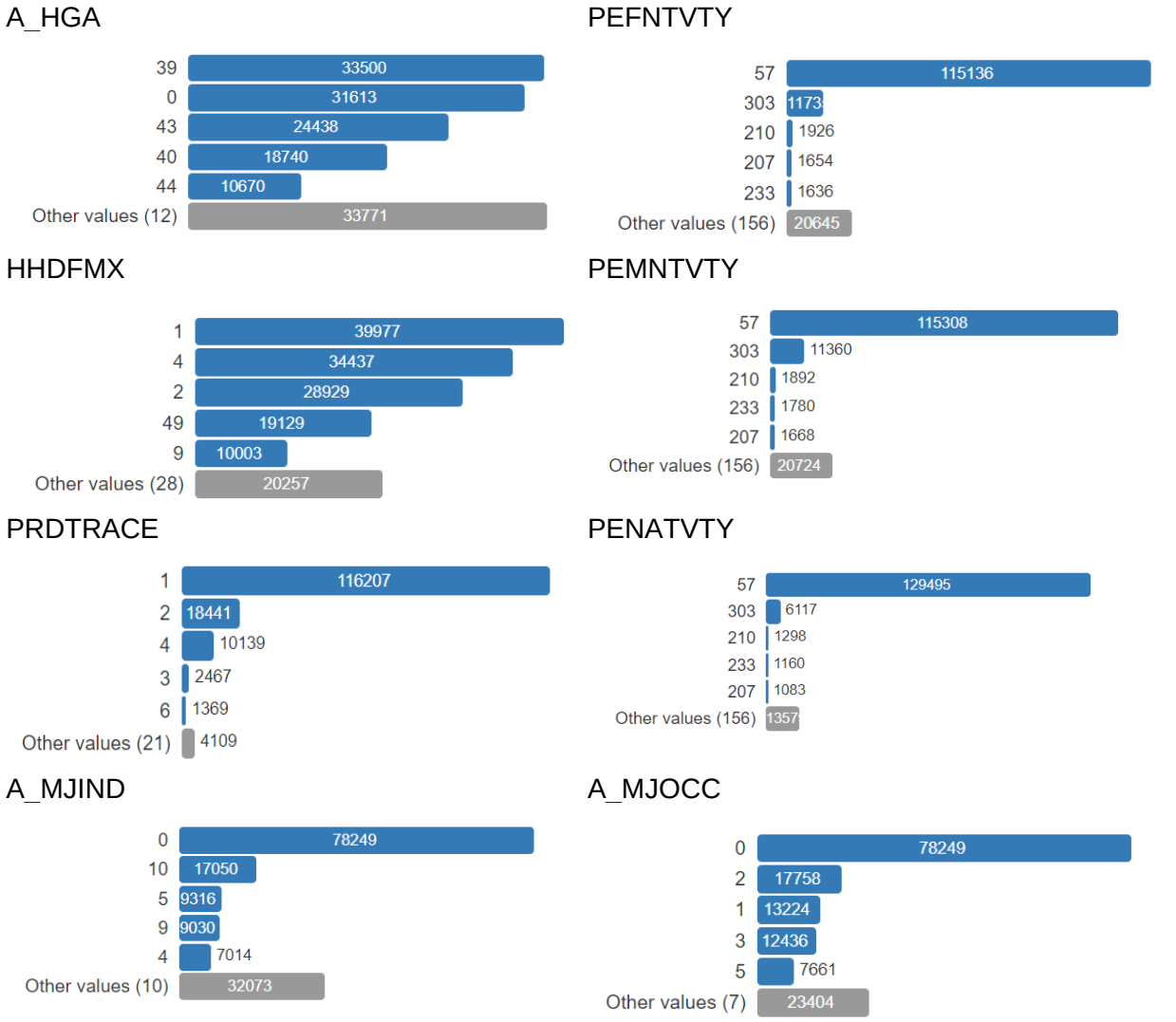
A_MJIND shows that majority of the population are either not in universe, or children (Code-0). The second major industry is educational services, health care, and social assistance (Code-10). The next two major industries are Wholesale and retail trade (Code-5) and Professional, scientific, management and administrative, and waste management services (Code-9).

Similarly, A_MJOCC represents that majority of the population are either not in universe, or children (Code-0). Second major occupation is Professional and related occupations (Code-2). Next comes Management, business, and financial occupations (Code-1) and Service occupations (Code-3) respectively.

PRDTRACE represents individual race, and we can see that majority are White only (Code-1). Second and third is Black only (Code-2) and Asian only (Code-4).

Detailed household and family status is presented in HHDFMX and A_HGA represents Educational attainment. The first major category is High school graduate - high school diploma or equivalent (Code-39), second is Children (Code-0). Next comes Bachelor's degree (for example: BA, AB, BS) (Code-43), Some college but no degree (Code-40) and Master's degree (for example: MA, MS, MENG, MED, MSW, MBA) (Code-44).

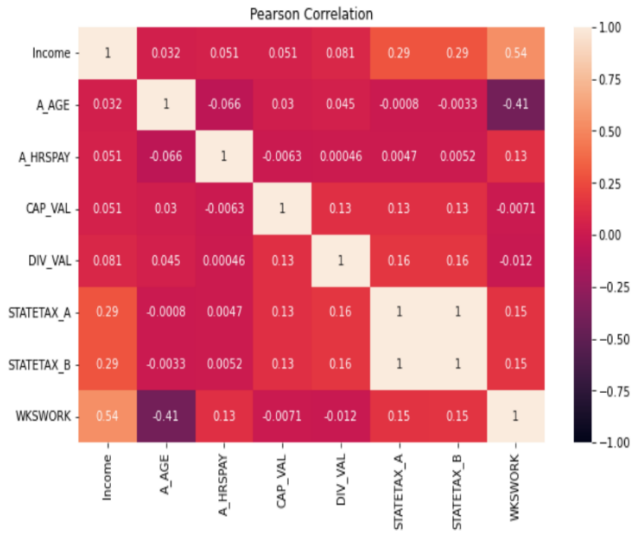
Figure - 1C: (PNETVTY, PEMNTVTY, PEFNTVTY, PRDTRACE, A_HGA, A_MJIND, A_MJOCC, and HHDFMX)



Bivariate Analysis:

Correlation Analysis: To visualize correlation, I have separated the numeric and nominal attributes. On numeric attributes Pearson's correlation and on categorical attributes Spearman's correlation test has been performed. Following are the heatmaps in Figure-2A and 2B.

Figure-2A: Pearson's correlation for numeric attributes



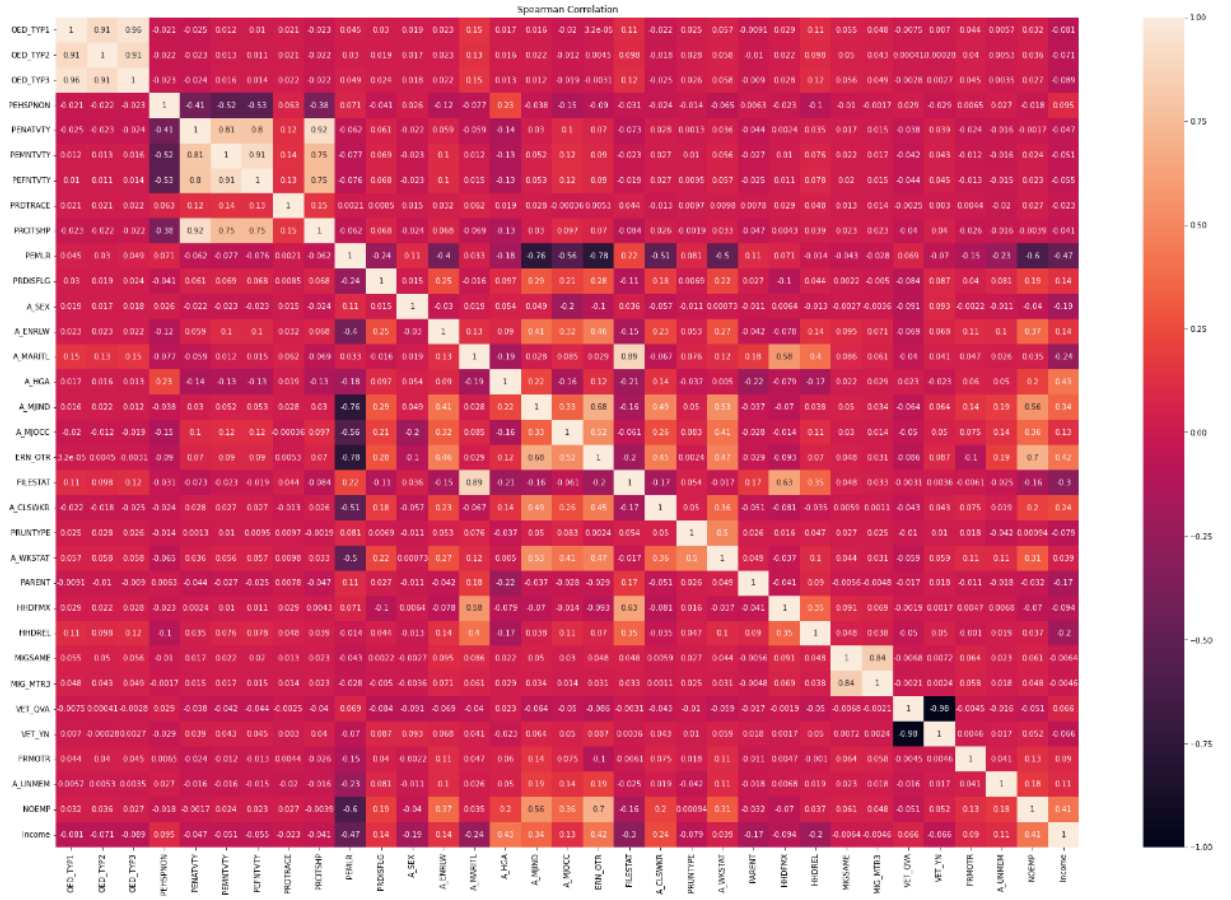
The following Table3 shows us very strong, strong, and moderate correlation between attributes, for both numeric and categorical.

Table 4: Correlation (Very strong, strong, and moderate)

Nominal:	
1)Very strong positive correlation (1.0 to 0.8) OED_TYP1 & OED_TYP2 OED_TYP1 & OED_TYP3 OED_TYP2 & OED_TYP3 PENATVTY & PEFNTVTY PENATVTY & PEMNTVTY PEFNTVTY & PEMNTVTY PENATVTY & PRCITSHP FILESTAT & A_MARITL MIGSAME & MIG_MTR3	4)Very strong negative correlation (-0.8 to -1.0) VET_YN & VET_QVA
2)Strong positive correlation (0.8 to 0.6) PEMNTVTY & PRCITSHP PEFNTVTY & PRCITSHP ERN_OTR & A_MJIND ERN_OTR & NOEMP HHDFMX & FILESTAT	5)Strong negative correlation (-0.6 to -0.8) PEMLR & A_MJIND PEMLR & ERN_OTR
3)Moderate positive correlation (0.6 to 0.4) HHDFMX & A_MARITL NOEMP & A_MJIND ERN_OTR & A_WKSTAT ERN_OTR & A_CLSWKR A_MJIND & A_WKSTST A_MJIND & A_CLSWKR	6)Moderate negative correlation (-0.4 to -0.6) PEMLR & Income PEMLR & NOEMP PEMLR & A_WKSTST PEMLR & A_CLSWKR PEMLR & A_MJOCC PEMLR & A_ENRLW

A_MJOCC & A_WKSTAT ERN_OTR & A_MJIND ERN_OTR & A_MJOCC HHDREL & A_MARITL	PENATVTY & PEHSPNON PEMNTVTY & PEHSPNON PEFNTVTY & PEHSPNON
Numeric:	
1)Very strong positive correlation (1.0 to 0.8) STATETAX_A & STATETAX_B	3)Moderate negative correlation (-0.4 to -0.6) A_AGE & WKSWORK
2)Moderate positive correlation (0.6 to 0.4) Income & WKSWORK	

Figure-2B: Spearman's correlation for categorical attributes



Preparing Target Variable “Income”

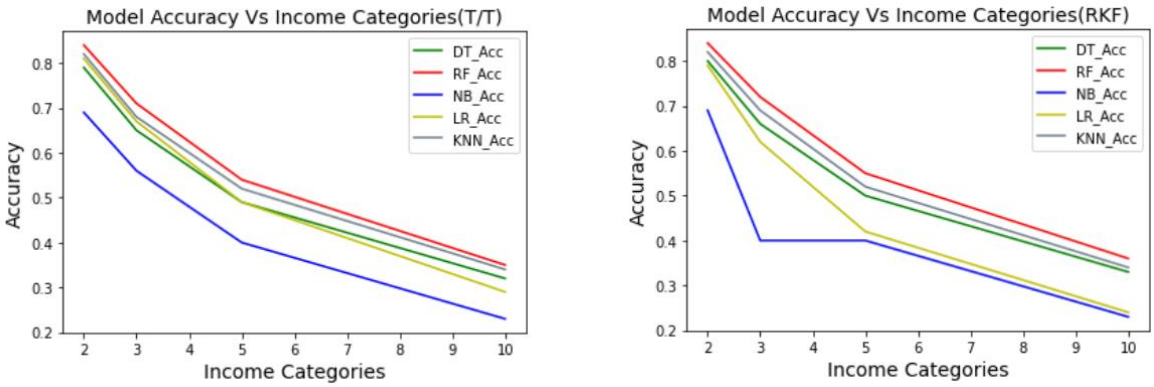
Our target variable “Income” is a discreate numerical variable which has been converted to categorical for classification problem. Initially our target variable “Income” was a numeric discrete variable. According to the aim of my study I have converted the dependent variable into categorical applying quantile method. To select the optimal number of categories I have created 2, 3, 5 and 10 categories of the target variable and applied five different classification algorithms one the whole dataset.

Table 5: Model accuracy based on the number of Income categories

	DT		RF		NB		LR		KNN	
Folds	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF
2QT	0.79	0.80	0.84	0.84	0.69	0.69	0.81	0.79	0.82	0.82
3QT	0.65	0.66	0.71	0.72	0.56	0.40	0.67	0.62	0.68	0.69
5QT	0.49	0.50	0.54	0.55	0.40	0.40	0.49	0.42	0.52	0.52
10QT	0.32	0.33	0.35	0.36	0.23	0.23	0.29	0.24	0.34	0.34

Both Train/Test Split and RepeatedKFold cross validation techniques used to get more accurate results. According to the performance of each algorithm it can be seen in the following Figure that, the more number of categories the less accurate the results are. We can see in the following figures that there is a clear decreasing trend of model accuracy. It is also true for other measures like Precession, Recall and F1 scores. According to this finding, as the optimal number of income classes, I have selected binary (“Low” & “High”) classification moving forward.

Figure-3: Model accuracy based on the number of Income categories



Feature Engineering

Implementing Feature Selection Methods

It is very common to have columns that are nothing but noise in real-world datasets. It is better to get rid off such variables to save memory space, time, and computational resources, especially in case of large datasets. Sometimes, it is difficult to be sure which features help in predicting the target variable.
This US Census dataset I have selected contains 39 independent attributes. To minimize number of features and to improve performance of the models, I have applied all three types of feature selection algorithms.

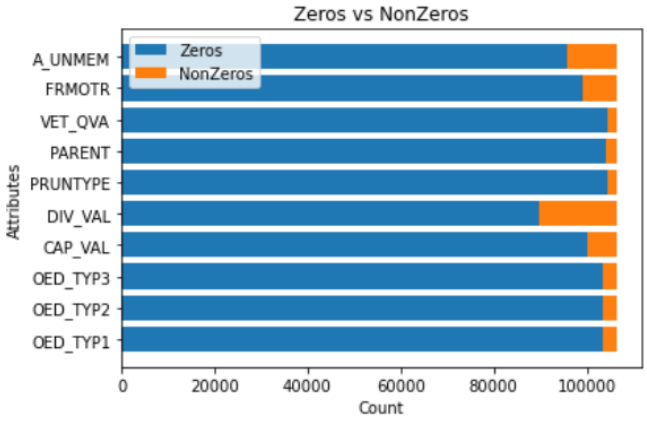
For Filter method, I have checked: (a)Missingness, (b) Mutual Information Gain, (c) Low Variance Filter, and (d) Correlation Coefficients. Next, for Wrapper method I have applied Forward Feature Selection method. Finally, I have used Sequential Forward Feature Selection technique as the Embedded method.

(1) Filter Method:

(a) Missing Value Ratio (MVR): The univariate analysis has shown us that ten attributes have very high numbers of zeros which represents None/Not in universe. These huge number of missingness can cause biasness. I would like to drop these attributes and compare the performance of the models.

Figure4: Missing Value Ratio

Attributes	% of Zeros
OED_TYP1	97.05
OED_TYP2	97.05
OED_TYP3	97.05
CAP_VAL	94.08
DIV_VAL	84.33
PRUNTYPE	98
PARENT	97.77
VET_QVA	97.97
FRMOTR	93.21
A_UNMEM	90.07

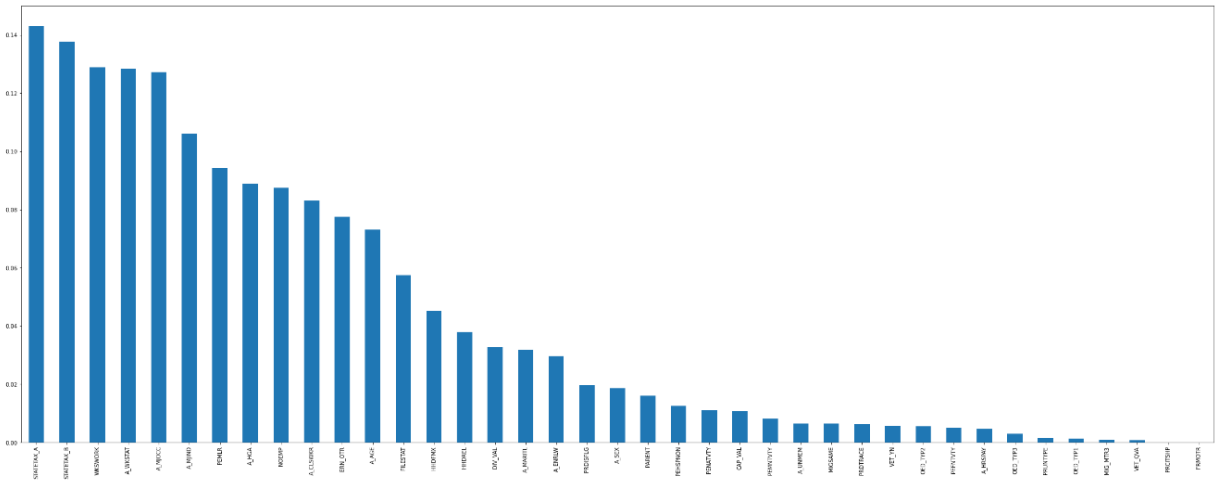


(b) Mutual Information Gain (MIG): Mutual information between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. Here this technique gave the following scores of the dataset, and I would like to select 24 attributes where MI >= 0.01. The reason behind this again, selecting the best performing attributes.

Table 6: Mutual Information Gain

#	Attribute	MI	#	Attribute	MI	#	Attribute	MI
1	STATETAX_A	0.14	14	HHDFMX	0.05	27	MIGSAME	0.006
2	STATETAX_B	0.14	15	HHDREL	0.05	28	PRDTRACE	0.006
3	WKSWORK	0.13	16	DIV_VAL	0.04	29	VET_YN	0.005
4	A_WKSTAT	0.13	17	A_MARITL	0.04	30	OED_TYP2	0.005
5	A_MJOCC	0.12	18	A_ENRLW	0.03	31	PEFNTVTY	0.005
6	A_MJIND	0.10	19	PRDISFLG	0.02	32	A_HRSPAY	0.004
7	PEMLR	0.09	20	A_SEX	0.02	33	OED_TYP3	0.003
8	A_HGA	0.09	21	PARENT	0.02	34	PRUNTYPE	0.001
9	NOEMP	0.09	22	PEHSPNON	0.01	35	OED_TYP1	0.001
10	A_CLSWKR	0.08	23	PEMNTVTY	0.01	36	MIG_MTR3	0.001
11	ERN_OTR	0.07	24	CAP_VAL	0.01	37	VET_QVA	0.001
12	A_AGE	0.06	25	PENATVTY	0.008	38	PRCITSHP	0.000
13	FILESTAT	0.06	26	A_UNMEM	0.006	39	FRMOTR	0.000

Figure 5: Mutual Information Gain from all attributes



(c) Low Variance Filter: As we know, variables with low variance have less impact on the target variable**.** We can set a threshold value of variance. And if the variance of a variable is less than that threshold, we can drop that variable. Since variance is range-dependent, therefore we need to do normalization before applying this technique. I have set my threshold >= 0.0001 on normalized data and after applying the technique I have got the following 19 attributes for further studies.

['PENATVTY', 'PEMNTVTY', 'PEFNTVTY', 'PEMLR', 'A_MARITL', 'A_HGA', 'A_AGE', 'A_MJIND', 'A_MJOCC', 'A_HRSPAY', 'CAP_VAL', 'DIV_VAL', 'FILESTAT', 'STATETAX_A', 'STATETAX_B', 'HHDFMX', 'HHDREL', 'WKSWORK', 'NOEMP']

(d) Correlation Coefficients: As we know, whenever two supposedly independent variables are highly correlated, it will be difficult to assess their relative importance in determining the dependent variable. The higher the correlation between independent variables the greater the sampling error of the partials**.** From bivariate analysis we have seen that some

of the attributes are highly corelated in our dataset. After excluding the following highly correlated 16 attributes (ideally >0.75), we got 23 attributes left.

Highly Correlated Attributes = [34, 10, 20, 17, 16, 23, 14, 6, 7, 5, 24, 3, 2, 36, 19, 33]
[“WKSWORK”, “PEMLR”, “ERN_OTR”, “A_MJIND”, “A_AGE”, “FILESTAT”, “A_MARITL”, “PEMNTVTY”, “PEFNTVTY”, “PENATVTY”, “STATETAX_A”, “OED_TYP3”, “OED_TYP2”, “VET_YN”, “A_HIRSPAY”, “MIG_MTR3”]

(2) Wrapper Method:

I have applied Forward Selection technique (in RStudio) and got 29 attributes but the coefficients of "PRDTRACE" and "PENATVTY" are insignificant, so have dropped them for further analysis.

(WKSWORK, A_HGA, A_AGE, A_SEX, A_WKSTAT, STATETAX_A, HHDREL, NOEMP, VET_YN, PRUNTYPE, PEMLR, A_MARITL, A_MJOCC, PRCITSHP, PRDISFLG, PARENT, HHDFMX, A_ENRLW, DIV_VAL, A_MJIND, ERN_OTR, PEHSPNON, CAP_VAL, A_CLSWKR, STATETAX_B, MIGSAME, MIG_MTR3)

(3) Embedded Method

I have applied Sequential Forward Selection technique (Using Random Forest Classifier, cv=10) and this embedded method has selected 29 independent attributes.

('OED_TYP1', 'OED_TYP2’, ‘OED_TYP3’, ‘PEHSPNON’, ‘PENATVTY’, ‘PRDTRACE', 'PEMLR', 'A_SEX', 'A_HGA', 'A_AGE', 'A_MJIND', 'A_MJOCC', 'A_HRSPAY', 'ERN_OTR', 'CAP_VAL’, ‘DIV_VAL', 'FILESTAT', 'STATETAX_A', 'STATETAX_B', 'A_CLSWKR', 'PRUNTYPE', 'A_WKSTAT', 'PARENT', 'HHDFMX', 'WKSWORK', 'VET_QVA', 'VET_YN', 'A_UNMEM', 'NOEMP')

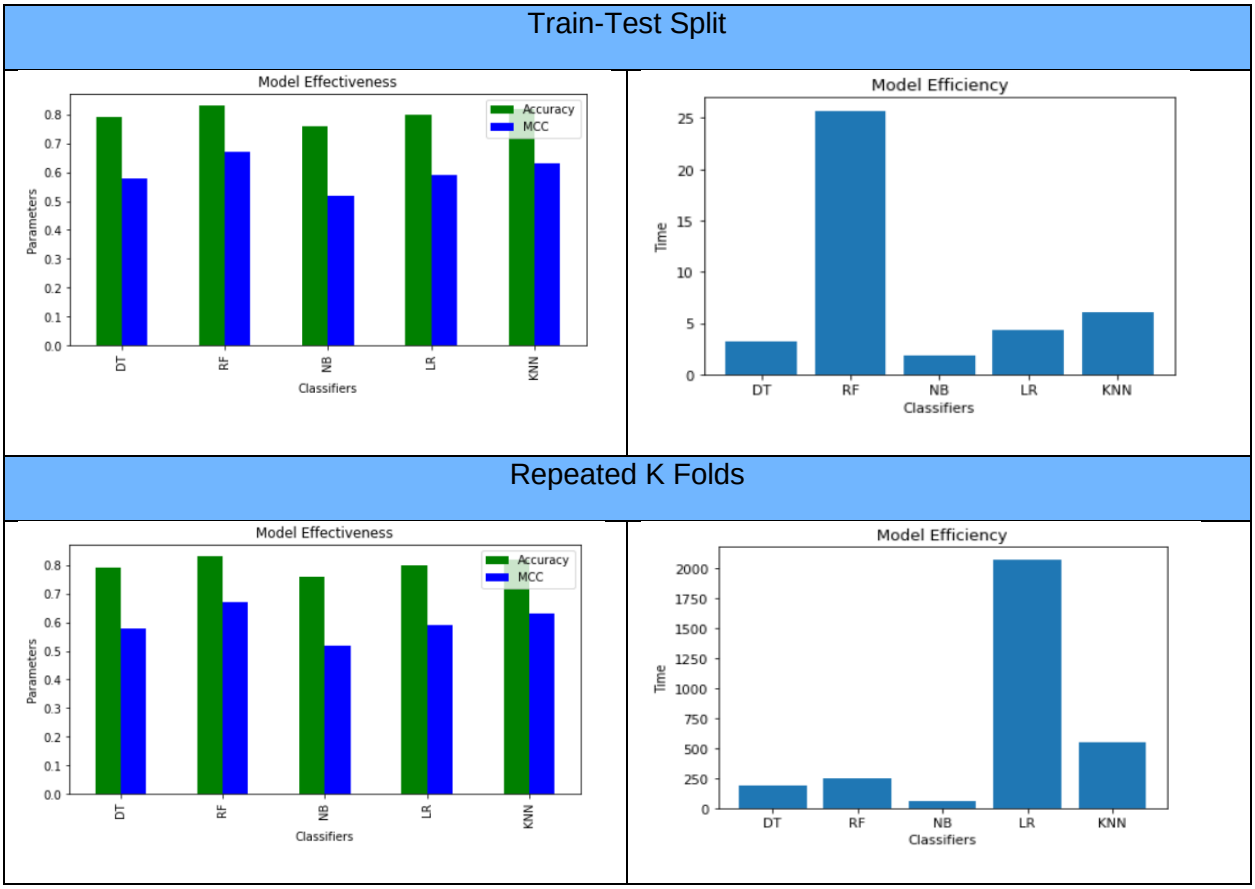
Comparing Feature Selection Methods

To compare the feature selection methods, I have applied five classification algorithms on each of the attribute’s group selected by each method. The focus is here to investigate models’ performance based on feature selection techniques. I have applied both Train/Test Split (80/20) and RepeatedKFold Cross Validation (k=10, n=5) techniques for more precise results.

(1) Filter method:

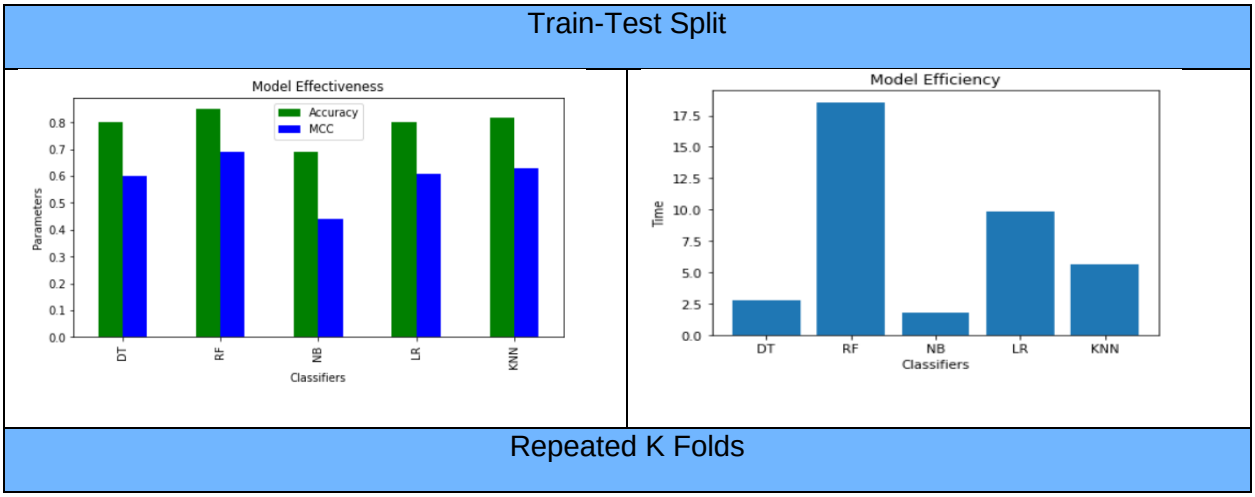
(a) Missing value ratio: After filtering the ten attributes with high numbers of zeros, the performance of the five classification algorithms is tabulated in Table 3 [Appendix].

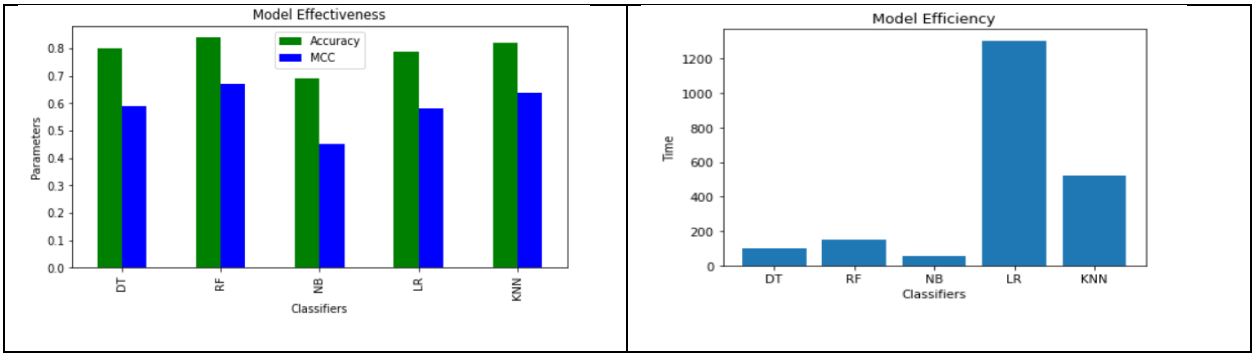
Figure 6: Models' performance based on Filter method (MVR)



(b) Mutual Information Gain: After applying this filter method we had left with 24 independent attributes. Using this subgroup, I have run all the five classifiers and tabulated the results below.

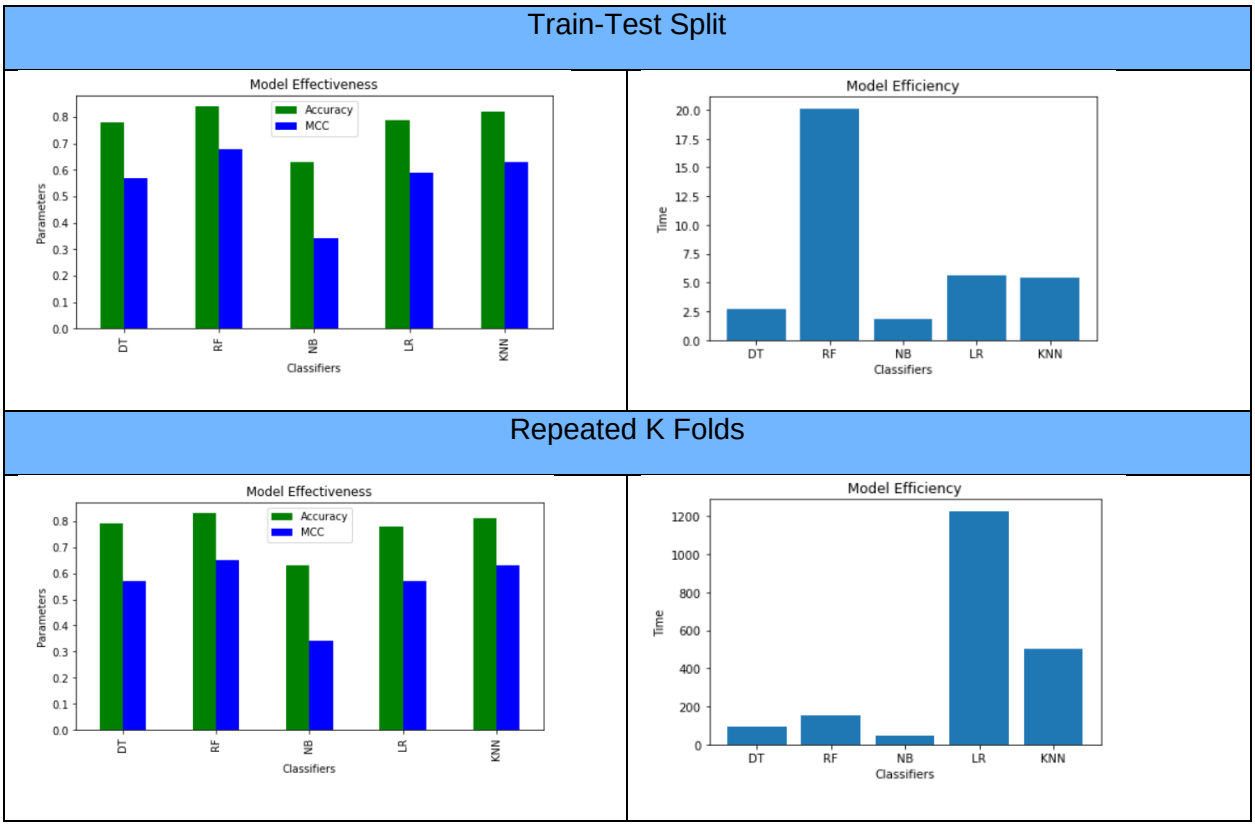
Figure 7: Models' performance based on Filter method (MIG)





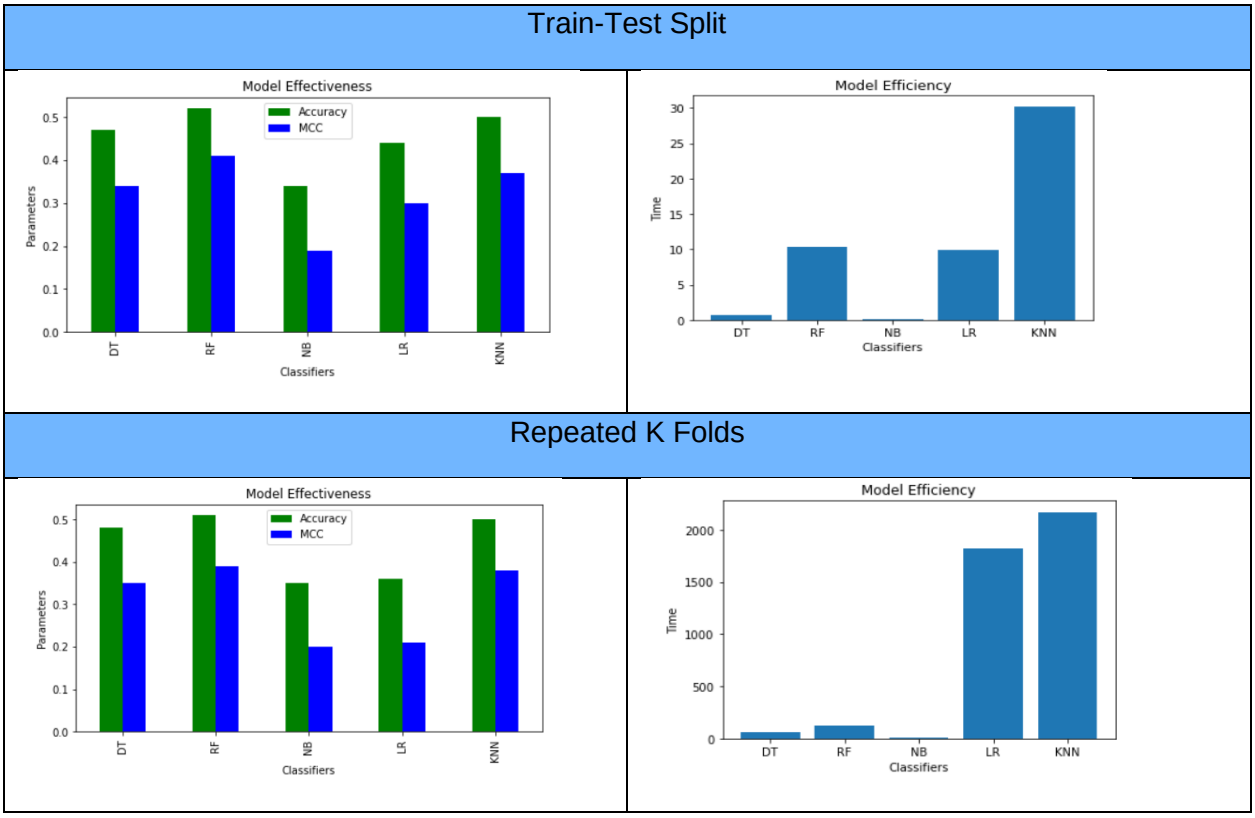
c) Low Variance Filter: After applying this technique (threshold=>=0.0001) on normalized dataset I had 19 independent attributes. The model performances on this subgroup are tabulated below.

Figure 8: Models' performance based on Filter method (LVF)



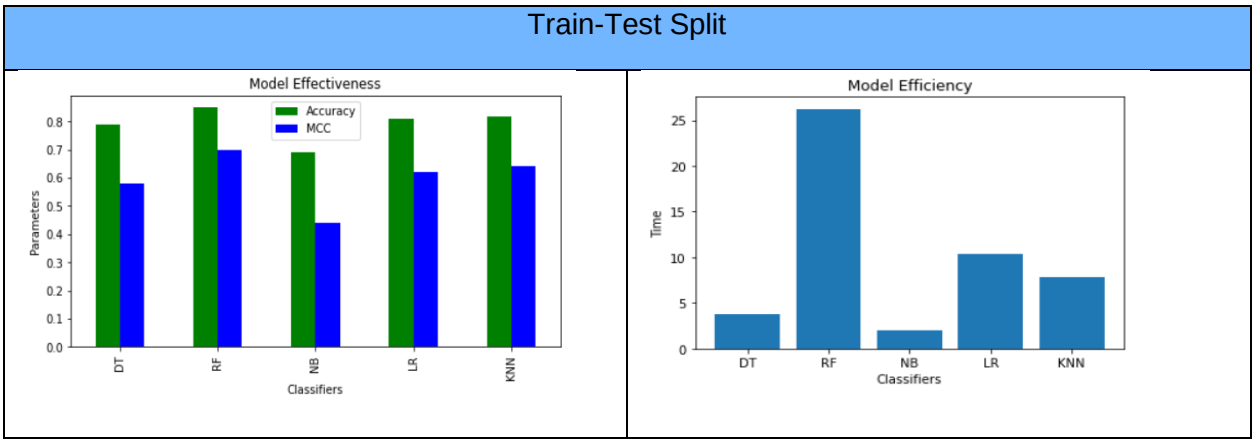
(d) Correlation Coefficients: After removing 16 highly correlated attributes (ideally >0.75), we got 23 independent attributes left. The models' performance based on this subgroup is in the following table.

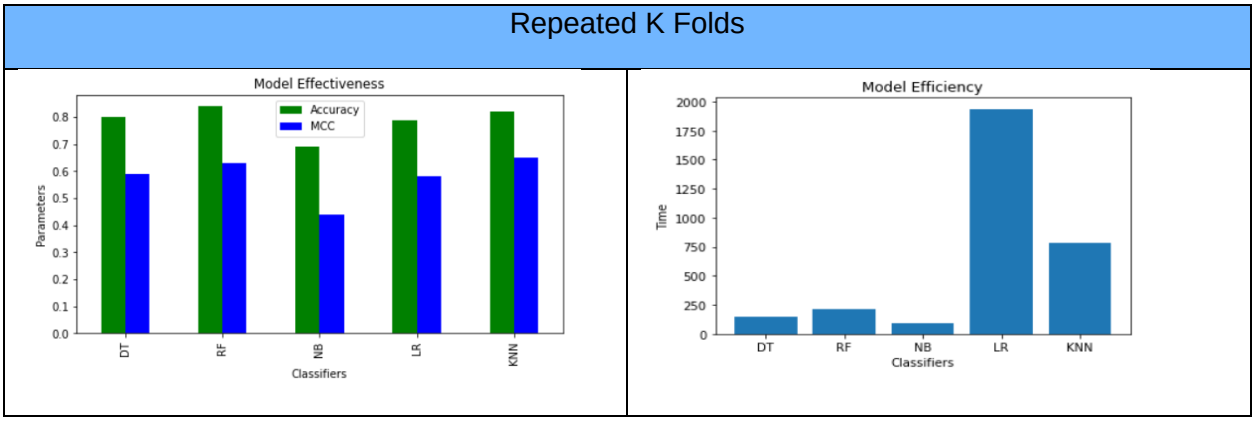
Figure 9: Models' performance based on Filter method (CC)



(2) Wrapper method: The Forward Selection method left us with 27 significant independent attributes. The models' performances on this subset are tabulated below.

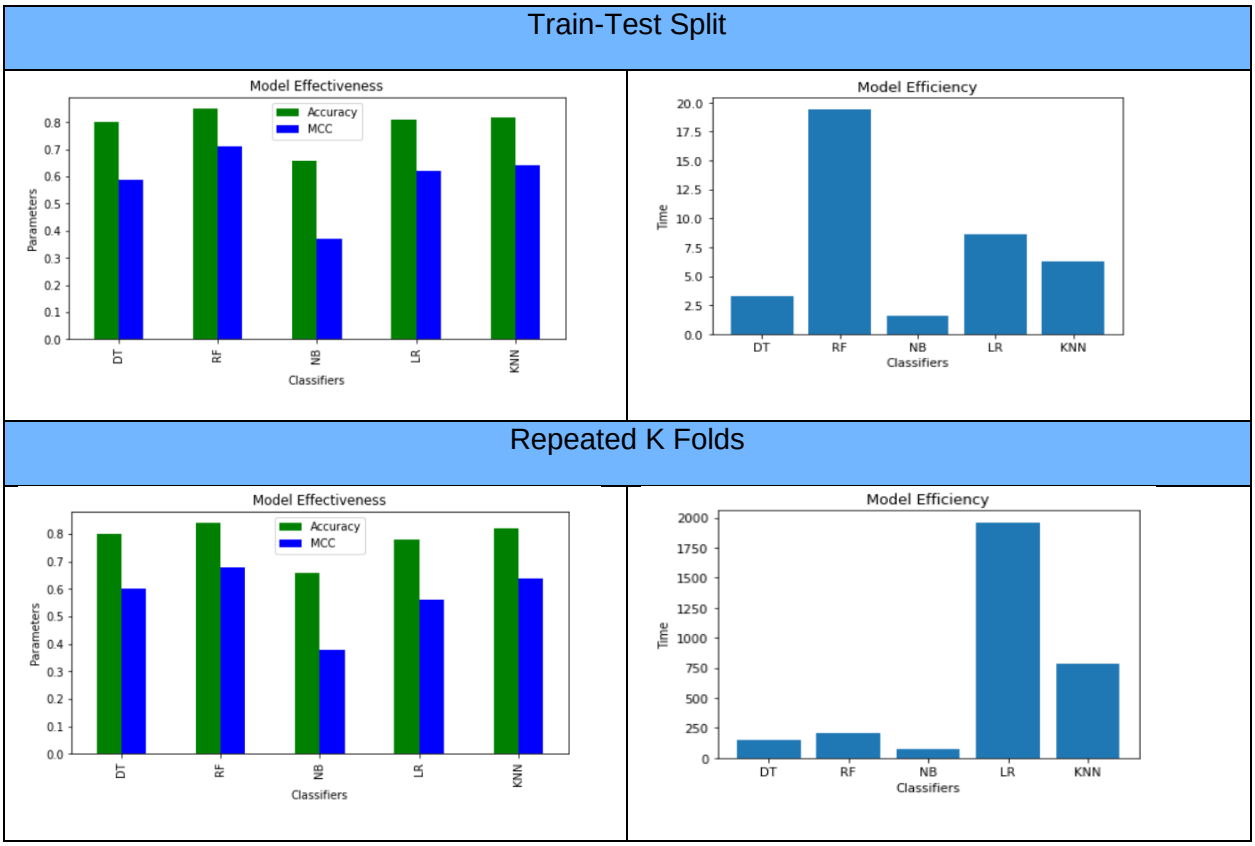
Figure 10: Models' performance based on Wrapper method (FS)





(2) Embedded Method: After applying Sequential Forward Selection (Using Random Forest Classifier, cv=10) this method gave us 29 independent attributes. The models' performances on this subset are tabulated below.

Figure 11: Models' performance based on Embedded method (SFS)



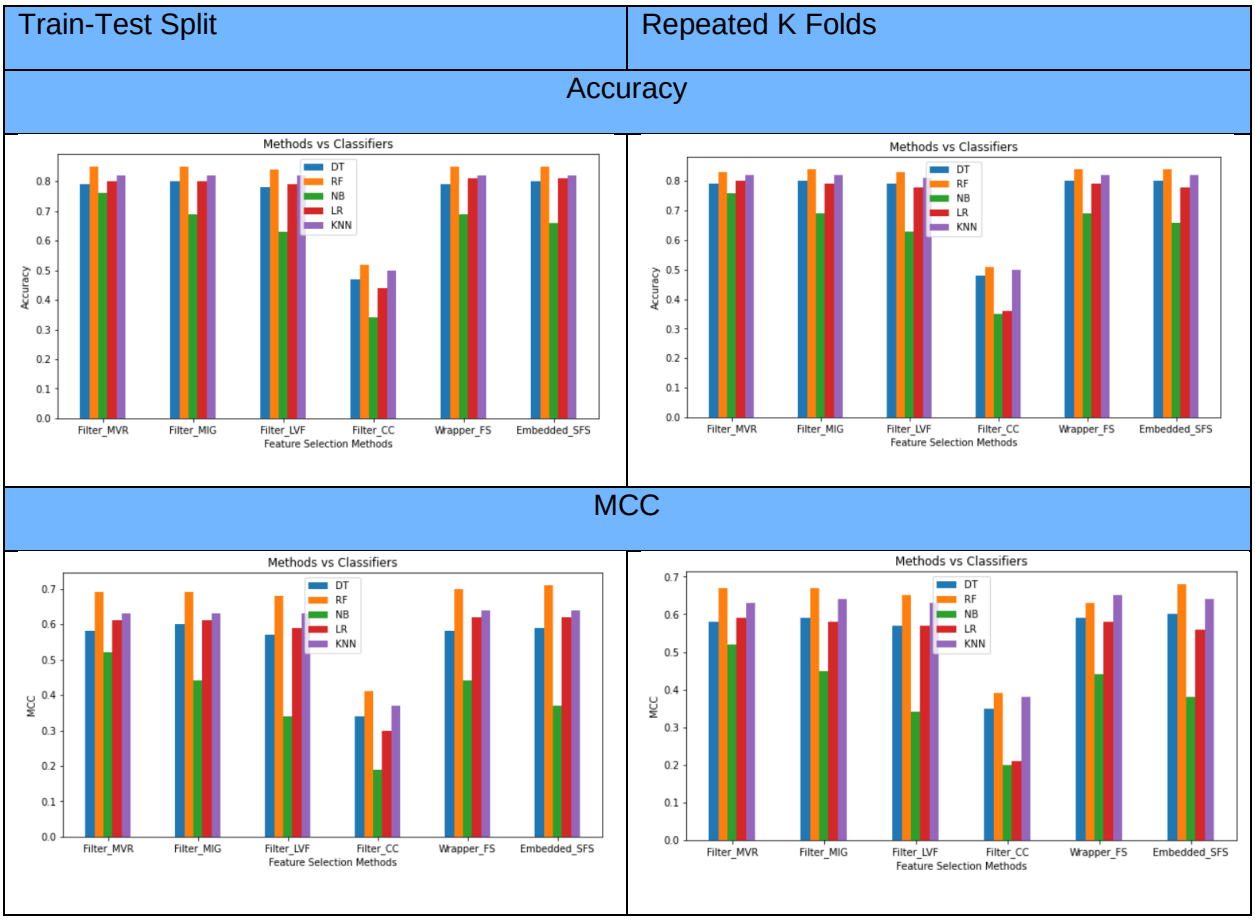
Above figures showing a comparison of Accuracy, MCC, and Run Time of each classifier based on each subset of dataset selected by three feature selection methods. As we can see, the effectiveness measures accuracy and MCC (for both T/T Split and RepeatedKFolds) are similar for all methods except filter method based on removing highly correlated attributes. All the models

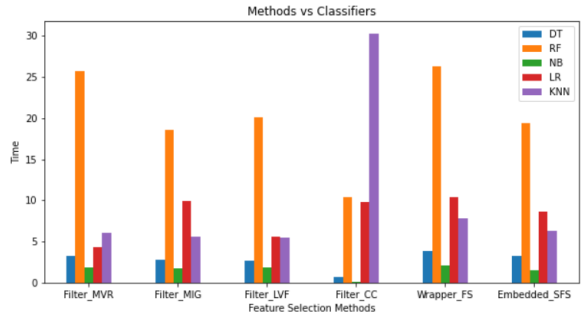
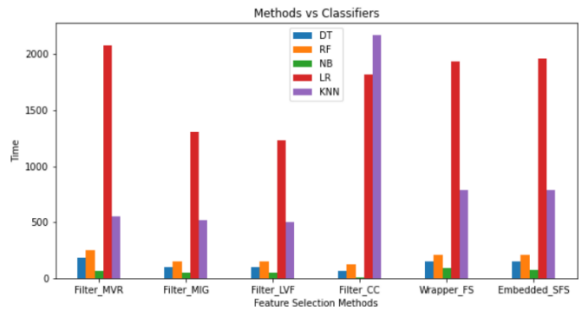
performing poor on this subset implies that, the highly correlated attributes most likely providing some independent information that our models can use for helping it train. In other words, these dropped attributes still useful for prediction accuracy and MCC. That means it's not surprising that we see a reduction in prediction accuracy by removing these.

The efficiency measure "Run Time" shows little different picture here. In both cases (T/T Split and RKF), Naïve Bayes is most efficient, and it worked best on filter method based on removal of highly correlated attributes, as well as Decision tree and Random Forest too. Where as KNN took longest time on this and worked fastest on the filter method based on Low Variance Filter. In case of Logistic Regression, fastest on filter method based on Missing Values Removal (for T/T Split) and on Low Variance Filter (for RKF).

Random Forest worked best on MIG, Wrapper and Embedded (85% and 84%), slightly better on Embedded, MCC (71% and 68%) on T/T split and RKF respectively. On the other hand, Naïve Bayes is best on Mutual Information Gain (MIG). DT, KNN, and LR are consistently sable on all the subsets except the one based on filtering highly correlated attributes.

Figure 12: Comparing Feature Selection methods based on Models' performance



Time	
	

Top Features Selection

Here I have selected top 16 attributes (in Green) to run the models. This selection is based on the Feature Selection methods. We can see in the above table that 5 attributes (A_HGA, A_MJOCC, STATETAX_B, HHDFMX, and NOEMP) are selected by all the methods. The rest 11 attributes (PEHSPNON, PEMLER, A_SEX, A_AGE, A_MJIND, CAP_VAL, DIV_VAL, STATETAX_A, A_CLSWKR, A_WKSTAT, and WKSWORK) are selected by 5 tests out of 6. Based on this subgroup I have applied all the five classifiers and have evaluated their performances based on effectiveness, efficiency, and stability.

Table 7: Features selected by three (Filter, Wrapper & Embedded) methods

	Filter Method					Wrapper Method	Embedded Method
Att #	Whole (39)	MVR (29)	MIG (24)	Variance (19)	Correlation (23)	FS (27)	Sequential FS (29)
1	OED_TYP1				OED_TYP1		OED_TYP1
2	OED_TYP2						OED_TYP2
3	OED_TYP3						OED_TYP3
4	PEHSPNON	PEHSPNON	PEHSPNON		PEHSPNON	PEHSPNON	PEHSPNON
5	PENATVTY	PENATVTY		PENATVTY			PENATVTY
6	PEMNTVTY	PEMNTVTY	PEMNTVTY	PEMNTVTY			
7	PEFNTVTY	PEFNTVTY		PEFNTVTY			
8	PRDTRACE	PRDTRACE			PRDTRACE		PRDTRACE
9	PRCITSHP	PRCITSHP			PRCITSHP	PRCITSHP	
10	PEMLR	PEMLR	PEMLR	PEMLR		PEMLR	PEMLR
11	PRDISFLG	PRDISFLG	PRDISFLG		PRDISFLG	PRDISFLG	
12	A_SEX	A_SEX	A_SEX		A_SEX	A_SEX	A_SEX
13	A_ENRLW	A_ENRLW	A_ENRLW		A_ENRLW	A_ENRLW	
14	A_MARITL	A_MARITL	A_MARITL	A_MARITL		A_MARITL	
15	A_HGA	A_HGA	A_HGA	A_HGA	A_HGA	A_HGA	A_HGA
16	A_AGE	A_AGE	A_AGE	A_AGE		A_AGE	A_AGE
17	A_MJIND	A_MJIND	A_MJIND	A_MJIND		A_MJIND	A_MJIND
18	A_MJOCC	A_MJOCC	A_MJOCC	A_MJOCC	A_MJOCC	A_MJOCC	A_MJOCC
19	A_HRSPAY	A_HRSPAY		A_HRSPAY			A_HRSPAY
20	ERN_OTR	ERN_OTR	ERN_OTR			ERN_OTR	ERN_OTR

21	CAP_VAL		CAP_VAL	CAP_VAL	CAP_VAL	CAP_VAL	CAP_VAL
22	DIV_VAL		DIV_VAL	DIV_VAL	DIV_VAL	DIV_VAL	DIV_VAL
23	FILESTAT	FILESTAT	FILESTAT	FILESTAT			FILESTAT
24	STATETAX_A	STATETAX_A	STATETAX_A	STATETAX_A		STATETAX_A	STATETAX_A
25	STATETAX_B	STATETAX_B	STATETAX_B	STATETAX_B	STATETAX_B	STATETAX_B	STATETAX_B
26	A_CLSWKR	A_CLSWKR	A_CLSWKR		A_CLSWKR	A_CLSWKR	A_CLSWKR
27	PRUNTYP				PRUNTYP	PRUNTYP	PRUNTYP
28	A_WKSTAT	A_WKSTAT	A_WKSTAT		A_WKSTAT	A_WKSTAT	A_WKSTAT
29	PARENT		PARENT		PARENT	PARENT	PARENT
30	HHDFMX	HHDFMX	HHDFMX	HHDFMX	HHDFMX	HHDFMX	HHDFMX
31	HHDREL	HHDREL	HHDREL	HHDREL	HHDREL	HHDREL	
32	MIGSAME	MIGSAME			MIGSAME	MIGSAME	
33	MIG_MTR3	MIG_MTR3				MIG_MTR3	
34	WKSWORK	WKSWORK	WKSWORK	WKSWORK		WKSWORK	WKSWORK
35	VET_QVA				VET_QVA		VET_QVA
36	VET_YN	VET_YN				VET_YN	VET_YN
37	FRMOTR				FRMOTR		
38	A_UNMEM				A_UNMEM		A_UNMEM
39	NOEMP	NOEMP	NOEMP	NOEMP	NOEMP	NOEMP	NOEMP

Comparison of Machine Learning Algorithms

Following are the confusion matrixes of each machine learning algorithm based on both T/T Split and RepeatedKFold cross validation. I will evaluate these models' performances in terms of efficiency, effectiveness, and stability.

1) Effectiveness

I will evaluate effectiveness based on Accuracy, Precision, F1-score, Recall, and MCC.

Accuracy = (TP + TN)/ (TP + FN + FP + TN)
Precision = TP/ (TP + FP)
Recall/Sensitivity/TP rate/hit rate= TP / (TP + FN)

F Score is the harmonic average of precision and recall.
F Score = 2 (Precision x Recall) / (Precision + Recall)

Matthews correlation coefficient (MCC):

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The MCC returns a value between -1 and 1:
+1 describes a perfect prediction.
0 unable to return any valid information (no better than random prediction).
-1 describes complete inconsistency between prediction and observation.

2) Efficiency

To evaluate the models' efficiency, it is important to look at the run time of each one. The shortest the run time the efficient that model is.

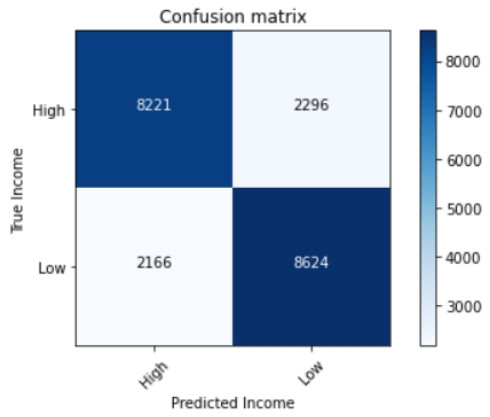
3) Stability

To verify the stability of a model one of the techniques is to compare results based on both Train-Test Split and Cross Validation. If the outcomes do not show significant changes than we can conclude that the model is stable.

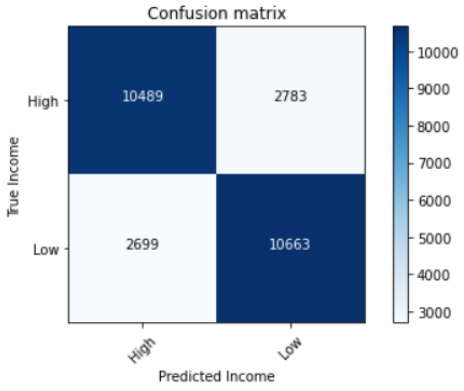
Following are the confusion matrixes for each of the classifiers.

Figure 13: Confusion Matrixes

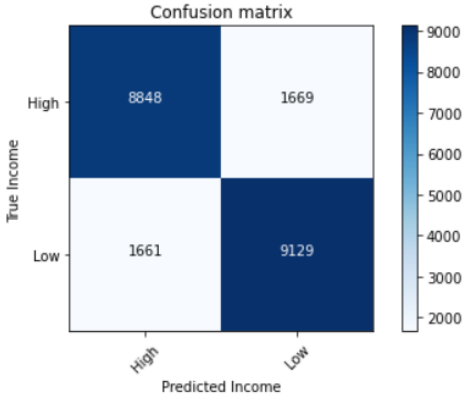
Decision Tree (T/T Split)



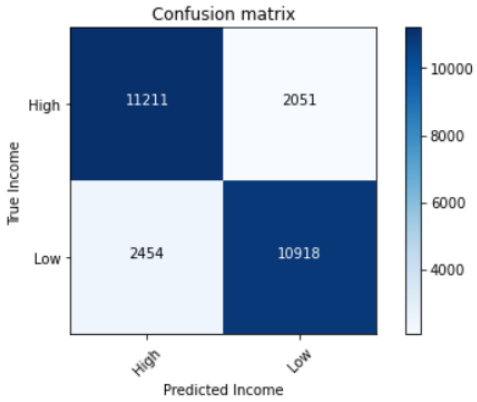
Decision Tree (RepeatedKFold)

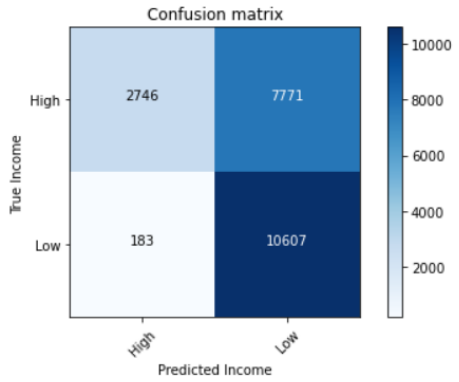
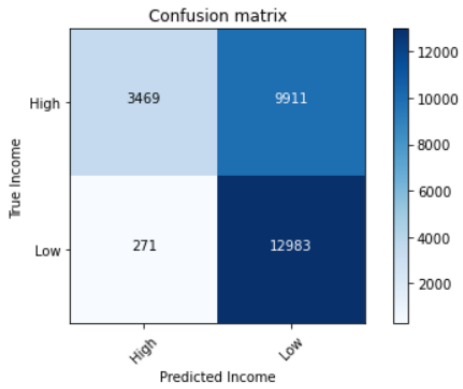
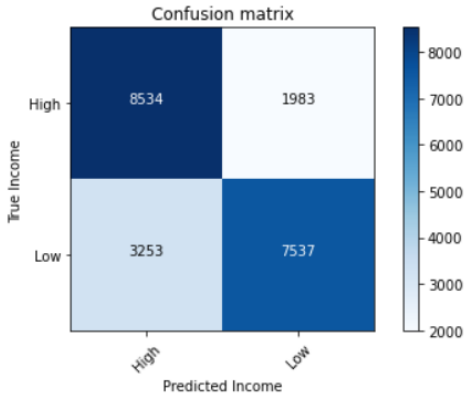
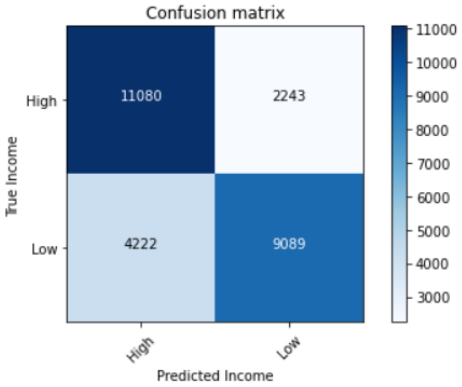
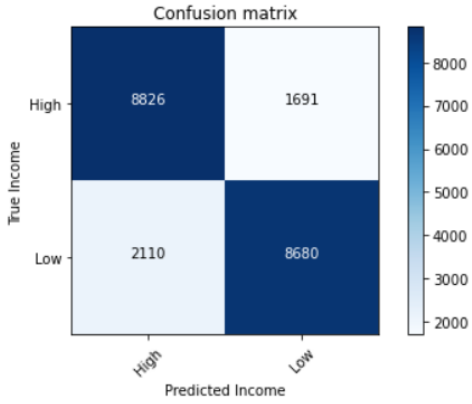
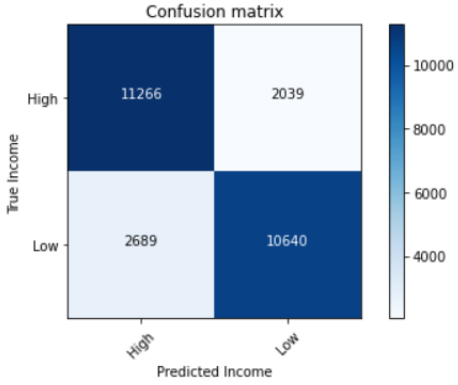


Random Forest (T/T Split)



Random Forest (RepeatedKFold)



Naïve Bayes (T/T Split) 	Naïve Bayes (RepeatedKFold) 
Logistic Regression (T/T Split) 	Logistic Regression (RepeatedKFold) 
KNN (T/T Split) 	KNN (RepeatedKFold) 

The Accuracy of Random Forest is the best among all the five classifiers (in both T/T Split and RKF). Same goes for Precision, Recall, F1-score, and Matthew's correlation coefficient (MCC).

The effectiveness of Decision Tree, KNN, and Logistic Regression are also very close to the Random Forest. Only, Naïve Bayes had performed consistently low in all cases.

Table 8: Model Performance on top 16 attributes

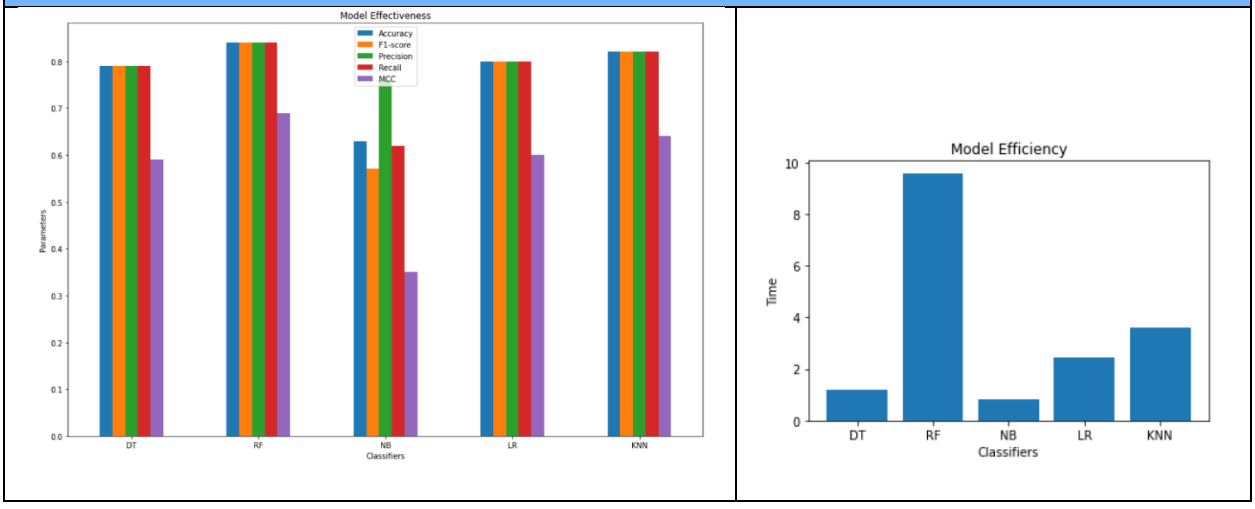
Parameter	DT		RF		NB		LR		KNN(k=11)	
	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF
Accuracy	0.79	0.79	0.84	0.83	0.63	0.62	0.80	0.78	0.82	0.82
F1-score	0.79	0.79	0.84	0.83	0.57	0.57	0.80	0.78	0.82	0.82
Precision	0.79	0.79	0.84	0.83	0.76	0.75	0.80	0.78	0.82	0.82
Recall	0.79	0.79	0.84	0.83	0.62	0.62	0.80	0.78	0.82	0.82
MCC	0.59	0.59	0.69	0.67	0.35	0.35	0.60	0.57	0.64	0.65
Time(sec)	1.21	66.59	9.59	118.12	0.83	37.38	2.43	1203.37	3.60	458.96

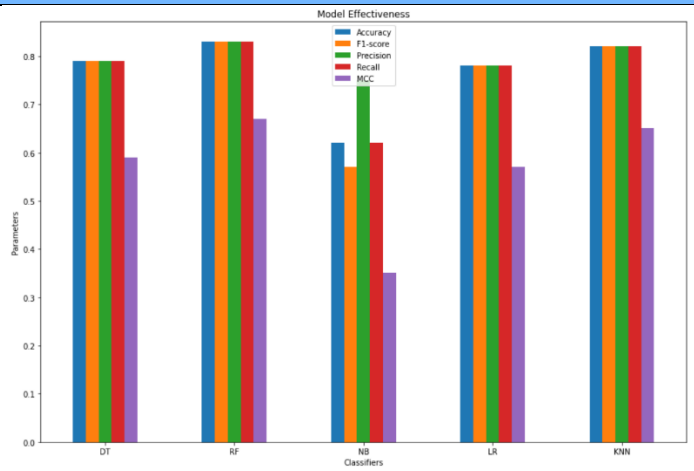
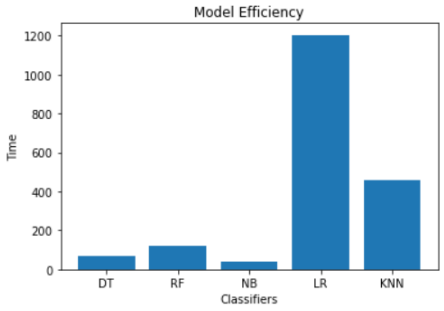
In terms of efficiency, the run time (measured in seconds) is fastest for Naïve Bayes (in both T/T Split and RKF). In case of Train-Test split, Random Forest took longest time; where as for RepeatedKFolds, Logistic Regression was the most inefficient.

In case of stability, we can see that all the models representing very similar performance in both Train-Test Split and RepeatedKFold cross validation techniques. This can be summarized that all these models are stable.

Figure 13: Model Performance on top 16 attributes

Train-Test Split



Repeated K Folds	
	

Conclusion

In this project, I have selected 40 attributes of the Current Population Survey Annual Social and Economic Supplement (ASEC) dataset, 2022 available on the U.S. Census Bureau. After initial cleaning of the dataset, I have created two categories ("Low" and "High") from the discrete target attribute Income. The focus of this project was to investigate the most important features that effects the income levels. Important features were finalized after completion of feature engineering. I have applied three Feature Selection techniques to select the best attributes that can predict effectively and efficiently if the person's income would be in "Low" or "High" group. I have compared all the five machine learning models' performances to see which Feature Selection method is more reliable. The findings show that removing highly correlated attributes in the Filter method is the worst performer here since all the models showed consistently low effectiveness. Most likely the reason is, these highly correlated attributes are providing some independent information that our models can use for helping it train. The models' performance on each subgroup selected by other Feature Selection methods are similar. Based on this, I have chosen 16 independent attributes, five of them (A_HGA, A_MJOCC, STATETAX_B, HHDFMX, and NOEMP) were selected by all methods which represent educational attainment, major occupation, state income tax liability before credits, detailed household and family status, and total number of persons who work for employers in all locations. The rest eleven (PEHSPNON, PEMLER, A_SEX, A_AGE, A_MJIND, CAP_VAL, DIV_VAL, STATETAX_A, A_CLSWKR, A_WKSTAT, and WKSWORK) were selected by five out of six methods representing Spanish/Hispanic/Latino, major labor force, sex, age, major industry, dividends income, capital gain, state income tax liability after credits, class of worker, and full/part-time status.

Next, five machine learning models were implemented and compared, to identify the most efficient and best performing model. Five different performance metrics were used to evaluate and validate the effectiveness of the models. Random Forest is the best effective model. Even though Naïve Bayes is most efficient but low in effectiveness. In terms of stability, all the models are stable.

Future Improvement

We can see even on the top 16 attributes, the best performer Random Forest gave accuracy, Precision, Recall, and F1-score only 84% and 83% and MCC 69% and 67% for T/T Split and ReapeatedKFold respectively. To improve the effectiveness, more income related attributes can be added in the dataset from the source. In addition, instead of taking percentiles for income categories, we can try setting a threshold for income levels to achieve more effective results.

References

[1] Navoneel Chakrabarty and Sanket Biswas: “A Statistical Approach to Adult Census Income Level prediction” Computer Science and Engineering, Jalpaiguri Government Engineering College, Jalpaiguri, West Bengal, India

[2] Alina Lazar: “Income Prediction via Support Vector Machine”, International Conference on Machine Learning and Applications - ICMLA 2004, 16-18 December 2004, Louisville, KY, USA.

[3] Sisay Menji Bekena: “Using decision tree classifier to predict income levels”, Munich Personal RePEc Archive 30th July, 2017

[4] Ron Kohavi: “Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision-Tree Hybrid”, Data Mining and Visualization, Silicon Graphics Inc., Mountain View, CA.

[5] Chet Lemon, Chris Zelazo and Kesav Mulakaluri: “Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques”, 048.pdf (ucsd.edu)

[6] S. Deepajothi and Dr. S. Selvarajan: “A Comparative Study of Classification Techniques on Adult Data Set”, International Journal of Engineering Research Technology (IJERT), ISSN: 2278-0181 Vol. 1 Issue 8, October2012

[7] Roshan Kumari and Saurabh Kr. Srivastava: “*Machine Learning: A Review on Binary Classification* “International Journal of Computer Applications (0975 – 8887) Volume 160 – No 7, February 2017

Appendix

Table1A: Detail Data Dictionary

(Note: NIU represents Not in Universe)

#	Data Name	Description	Type	Details
1	OED_TYP1	Other government assistance	Nominal	Values: 0 = NIU 1 = yes 2 = no
2	OED_TYP2	Scholarships, grants etc. from the school	Nominal	Values: 0 = NIU 1 = yes 2 = no
3	OED_TYP3	Other assistance (employers, friends, etc.)	Nominal	Values: 0 = NIU 1 = yes 2 = no
4	PEHSPNON	Spanish, Hispanic, or Latino?	Nominal	Values: 1 = Yes 2 = No
5	PENATVTY	Country of birth	Nominal	Number of Country:162
6	PEMNTVTY	Mother's country of birth	Nominal	Number of Country:162
7	PEFNTVTY	Father's country of birth	Nominal	Number of Country:162
8	PRDTRACE	Race	Nominal	Values: 01 = White only 02 = Black only 03 = American Indian, Alaskan Native only (AI) 04 = Asian only 05 =Hawaiian/Pacific Islander only (HP) 06 = White-Black 07 = White-AI 08 = White-Asian 09 = White-HP 10 = Black-AI 11 = Black-Asian 12 = Black-HP 13 = AI-Asian 14 = AI-HP 15 = Asian-HP 16 = White-Black-AI 17 = White-Black-Asian 18 = White-Black-HP 19 = White-AI-Asian 20 = White-AI-HP 21 = White-Asian-HP 22 = Black-AI-Asian 23 = White-Black-AI-Asian 24 = White-AI-Asian-HP 25 = Other 3 race comb. 26 = Other 4 or 5 race comb.
9	PRCITSHP	Citizenship Group	Nominal	Values: 1 = Native, born in US 12 = Native, born in PR or US outlying area 3 = Native, born abroad of US parent(s) 4 = Foreign born, US city by naturalization 5 = Foreign born, not a US citizen

10	NOEMP	Total number of persons who work for employers in all locations	Nominal	Values: 0 = NIU 1 = under 10 2 = 10 - 24 3 = 25 - 99 4 = 100 - 499 5 = 500 - 999 6 = 1000+
11	A_UNMEM	Member of a labor union or of an employee association	Nominal	Values: 0 = NIU or children and Armed Forces 1 = Yes 2 = No
12	FRMOTR	Receiving farm self-employment from secondary source	Nominal	Values: 0 = NIU 1 = yes 2 = no
13	PTOTVAL	Total persons income	Numeric	Values: 0 = none negative amt = income (loss) positive amt = income Universe: All Persons aged 15+
14	VET_YN	Receive veterans' payments?	Nominal	Values: 0 = NIU 1 = yes 2 = no
15	VET_QVA	Fill out an annual income questionnaire for the veteran's administration?	Nominal	Values: 0 = NIU 1 = yes 2 = no
16	WKSWORK	Number of weeks worked	Numeric	Values: 0 = NIU 1 = 1 week ... 52 = 52 weeks
17	MIG_MTR3	Within area moved	Nominal	Values: 1 = Nonmover 2 = Same County 3 = Different County, same state 4 = Different state, same division 5 = Different division, same region 6 = Different region 7 = Abroad 8 = Not in universe (children under 1 yr old)
18	MIGSAME	Living in same place last 1 year	Nominal	Values: 0 = NIU 1 = yes (nonmover) 2 = no, different house in U.S. (mover) 3 = no, outside the U.S.(mover)
19	HHDREL	Detailed household summary	Nominal	Values: <u>In household:</u> 1 = Householder 2 = Spouse of householder <u>Child of householder:</u> 3 = Under 18 years, single (never married)

				4 = Under 18 years, ever married 5 = 18 years and over Other household members: 6 = Other relative of householder 7 = Nonrelative of householder In group quarters: 8 = Secondary individual
20	HHDFMX	Detailed household and family status	Nominal	Values: In primary family: 01 = Householder 02 = Spouse of householder Child of householder: Under 18, single (never married): 03 = Reference person of subfamily 04 = Not in a subfamily Under 18, ever married: 05 = Reference person of subfamily 06 = Spouse of subfamily reference person 07 = Not in a subfamily 18 years and over, single (never married): 08 = Head of a subfamily 09 = Not in a subfamily 18 years and over, ever married: 10 = Reference person of subfamily 11 = Spouse of subfamily reference person 12 = Not in a subfamily Grandchild of householder: Under 18, single (never married): 23 = Reference person of subfamily 24 = Child of a subfamily 25 = Not in a subfamily Under 18, ever married: 26 = Reference person of subfamily 27 = Spouse of subfamily reference person 28 = Not used 29 = Not in a subfamily 18 years and over, single (never married): 30 = Reference person of a subfamily 31 = Not in a subfamily 18 years and over, ever married: 32 = Reference person of subfamily 33 = Spouse of subfamily reference person 34 = Not in a subfamily Other relative of householder: Under 18, single (never married): 35 = Reference person of subfamily 36 = Child of subfamily reference person 37 = Not in a subfamily

				Under 18, ever married: 38 = Reference person of subfamily 39 = Spouse of subfamily reference person 40 = Not in a subfamily 18 years and over, single (never married): 41 = Reference person of a subfamily 42 = Not in a subfamily 18 years and over, ever married: 43 = Reference person of subfamily 44 = Spouse of subfamily reference person 45 = Not in a subfamily In unrelated subfamily: 46 = Reference person of unrelated subfamily 47 = Spouse of unrelated subfamily reference person 48 = Child < 18, single (never married) of unrelated subfamily reference person Not in a family: 49 = Nonfamily householder 50 = Secondary individual 51 = In group quarters
21	PARENT	Presence of parents	Nominal	Values: 0 = Not in universe 1 = Both parents present 2 = Mother only present 3 = Father only present 4 = Neither parent present Universe: Family members under 18
22	A_WKSTAT	Full/part-time status	Nominal	Values: 0 = Children or Armed Forces 1 = Not in labor force 2 = Full-time schedules 3 = Part-time for economic reasons, usually FT 4 = Part-time for non-economic reasons, usually PT 5 = Part-time for economic reasons, usually PT 6 = Unemployed FT 7 = Unemployed PT
23	PRUNTYPE	Reason for unemployment	Nominal	Values: 0 = NIU 1 = Job loser/on layoff 2 = Other job loser 3 = Temporary job ended 4 = Job leaver 5 = Re-entrant 6 = New entrant
24	A_CLSWKR	Class of worker	Nominal	Values: 0 = Not in universe or children and Armed Forces 1 = Private

				2 = Federal government 3 = State government 4 = Local government 5 = Self-employed incorporated 6 = Self-employed-not incorporated 7 = Without pay 8 = Never worked
25	STATETAX_B	State income tax liability, before credits	Numeric	Values: 0 = none; dollar amount
26	STATETAX_A	State income tax liability, after all credits	Numeric	Values: 0 = none; dollar amount
27	FILESTAT	Tax filer status	Nominal	Values: 1 = joint, both 2 = joint, one ><65 & one 3 = joint, both 65+ 4 = head of household 5 = single 6 = non-filer
28	DIV_VAL	Dividends from stocks or mutual funds	Numeric	Values: 0 = none or NIU 1-999999 = dividends
29	CAP_VAL	Capital gains value	Numeric	Values: 0 = none or NIU 1-999999 = capital gains amount
30	ERN_OTR	Wage & salary from other work	Nominal	Values: 0 = NIU 1 = yes 2 = no
31	A_HRSPAY	Income per hour	Numeric	Values: 0000 = Not in universe or children and Armed Forces 0001-9999 = Entry (2 implied decimal places)
32	A_MJOCC	Major occupation recode	Nominal	Values: 0 = Not in universe or children 1 = Management, business, and financial occupations 2 = Professional and related occupations 3 = Service occupations 4 = Sales and related occupations 5 = Office and administrative support occupations 6 = Farming, fishing, and forestry occupations 7 = Construction and extraction occupations 8 = Installation, maintenance, and repair occupations 9 = Production occupations 10 = Transportation and material moving occupations 11 = Military specific occupations
33	A_MJIND	Major industry code	Nominal	Values: 0 = Not in universe, or children

				1 = Agriculture, forestry, fishing, and hunting 2 = Mining, quarrying, and oil and gas extraction 3 = Construction 4 = Manufacturing 5 = Wholesale and retail trade 6 = Transportation, warehousing, and utilities 7 = Information 8 = Finance and insurance, and real estate and rental and leasing 9 = Professional, scientific, management and administrative, and waste management services 10 = Educational services, and health care and social assistance 11 = Arts, entertainment, recreation and accommodation, and food services 12 = Other services, except public administration 13 = Public administration 14 = Military
34	A_HGA	Educational attainment	Nominal	Values: 0 = Children 31 = Less than 1st grade 32 = 1st,2nd,3rd, or 4th grade 33 = 5th or 6th grade 34 = 7th and 8th grade 35 = 9th grade 36 = 10th grade 37 = 11th grade 38 = 12th grade no diploma 39 = High school graduate - high school diploma or equivalent 40 = Some college but no degree 41 = Associate degree in college - occupation/vocation program 42 = Associate degree in college - academic program 43 = Bachelor's degree (for example: BA, AB, BS) 44 = Master's degree (for example: MA, MS, MENG, MED, MSW, MBA) 45 = Professional school degree (for example: MD, DDS, DVM, LLB, JD) 46 = Doctorate degree (for example: PHD, EDD)
35	A_MARITL	Marital status	Nominal	Values: 1 = Married - civilian spouse present 2 = Married - AF spouse present 3 = Married - spouse absent (separated) 4 = Widowed 5 = Divorced 6 = Separated

				7 = Never married
36	A_ENRLW	Last week was attending or enrolled in a high school, college, or university	Nominal	Values: 0 = Not in universe or children and Armed Forces 1 = Yes 2 = No
37	A_SEX	Sex	Nominal	Values: 1 = Male 2 = Female
38	PRDISFLG	Any disability conditions?	Nominal	Values: -1 = NIU 1 = Yes 2 = No
39	PEMLR	Major labor force recodes	Nominal	Values: 0 = NIU 1 = Employed - at work 2 = Employed - absent 3 = Unemployed - on layoff 4 = Unemployed - looking 5 = Not in labor force - retired 6 = Not in labor force - disabled 7 = Not in labor force - other
40	A_AGE	Age	Numeric	Values: 00-79 = 0-79 years of age 80 = 80-84 years of age 85 = 85+ years of age

Table1B: Country codes

Countries and Areas of the World

Numerical List of Countries and Areas of the World

Code	Name	Code	Name
057	United States	154	Serbia
060	American Samoa	155	Estonia
066	Guam	156	Latvia
069	Northern Marianas	157	Lithuania
073	Puerto Rico	158	Armenia
078	U.S. Virgin Islands	159	Azerbaijan
100	Albania	160	Belarus
102	Austria	161	Georgia
103	Belgium	162	Moldova
104	Bulgaria	163	Russia
105	Czechoslovakia	164	Ukraine
106	Denmark	165	USSR
108	Finland	166	Europe, not specified
109	France	168	Montenegro
110	Germany	200	Afghanistan
116	Greece	202	Bangladesh
117	Hungary	203	Bhutan
118	Iceland	205	Myanmar (Burma)
119	Ireland	206	Cambodia
120	Italy	207	China
126	Netherlands	209	Hong Kong
127	Norway	210	India
128	Poland	211	Indonesia
129	Portugal	212	Iran
130	Azores	213	Iraq
132	Romania	214	Israel
134	Spain	215	Japan
136	Sweden	216	Jordan
137	Switzerland	217	Korea
138	United Kingdom	218	Kazakhstan
139	England	220	South Korea
140	Scotland	222	Kuwait
142	Northern Ireland	223	Laos
147	Yugoslavia	224	Lebanon
148	Czech Republic	226	Malaysia
149	Slovakia	228	Mongolia
150	Bosnia & Herzegovina	229	Nepal
151	Croatia	231	Pakistan
152	Macedonia	233	Philippines

Code	Name
235	Saudi Arabia
236	Singapore
238	Sri Lanka
239	Syria
240	Taiwan
242	Thailand
243	Turkey
245	United Arab Emirates
246	Uzbekistan
247	Vietnam
248	Yemen
249	Asia, not specified
300	Bermuda
301	Canada
303	Mexico
310	Belize
311	Costa Rica
312	El Salvador
313	Guatemala
314	Honduras
315	Nicaragua
316	Panama
321	Antigua and Barbuda
323	Bahamas
324	Barbados
327	Cuba
328	Dominica
329	Dominican Republic
330	Grenada
332	Haiti
333	Jamaica
338	St. Kitts--Nevis
339	St. Lucia
340	St. Vincent and the Grenadines
341	Trinidad and Tobago
343	West Indies, not specified
360	Argentina
361	Bolivia
362	Brazil
363	Chile
364	Columbia
365	Ecuador
368	Guyana
369	Paraguay
370	Peru
372	Uruguay
373	Venezuela
374	South America, not specified
399	Americas, not specified
400	Algeria
407	Cameroon
408	Cape Verde
412	Congo
414	Egypt
416	Ethiopia
417	Eritrea
421	Ghana
423	Guinea
425	Ivory Coast
427	Kenya
429	Liberia
430	Libya
436	Morocco
440	Nigeria
444	Senegal
447	Sierra Leone
448	Somalia
449	South Africa
451	Sudan
453	Tanzania
454	Togo
457	Uganda
459	Zaire
460	Zambia
461	Zimbabwe
462	Africa, not specified
501	Australia
508	Fiji
511	Marshall Islands
512	Micronesia
515	New Zealand
523	Tonga
527	Samoa
555	Elsewhere

Table2: Models' performance based on Feature Selection Methods

Table 2A: Model Performance on 29 attributes (Filter method: MVR)

Parameter	DT		RF		NB		LR		KNN(k=11)	
	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF
Accuracy	0.79	0.79	0.85	0.83	0.76	0.76	0.80	0.80	0.82	0.82
F1-score	0.79	0.79	0.85	0.83	0.76	0.76	0.80	0.80	0.82	0.82
Precision	0.79	0.79	0.85	0.83	0.76	0.76	0.81	0.80	0.82	0.82
Recall	0.79	0.79	0.85	0.83	0.76	0.76	0.80	0.80	0.82	0.82
MCC	0.58	0.58	0.69	0.67	0.52	0.52	0.61	0.59	0.63	0.63
Time(sec)	3.26	184.43	25.71	251.81	1.85	61.97	4.31	2076.28	6.02	550.84

Table 2B: Model Performance on 24 attributes (Filter method: MIG)

Parameter	DT		RF		NB		LR		KNN(k=11)	
	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF
Accuracy	0.80	0.80	0.85	0.84	0.69	0.69	0.80	0.79	0.82	0.82
F1-score	0.80	0.80	0.85	0.84	0.67	0.67	0.80	0.79	0.82	0.82
Precision	0.80	0.80	0.85	0.84	0.76	0.76	0.80	0.79	0.82	0.82
Recall	0.80	0.80	0.85	0.84	0.69	0.69	0.80	0.79	0.82	0.82
MCC	0.60	0.59	0.69	0.67	0.44	0.45	0.61	0.58	0.63	0.64
Time(sec)	2.79	102.37	18.56	147.76	1.79	52.96	9.88	1304.28	5.60	519.73

Table 2C: Model Performance on 19 attributes (Filter method: LVF)

Parameter	DT		RF		NB		LR		KNN (K=11)	
	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF
Accuracy	0.78	0.79	0.84	0.83	0.63	0.63	0.79	0.78	0.82	0.81
F1-score	0.78	0.79	0.84	0.83	0.58	0.59	0.79	0.78	0.82	0.81
Precision	0.78	0.79	0.84	0.83	0.73	0.72	0.80	0.78	0.82	0.82
Recall	0.78	0.79	0.84	0.83	0.63	0.63	0.79	0.78	0.82	0.81
MCC	0.57	0.57	0.68	0.65	0.34	0.34	0.59	0.57	0.63	0.63
Time(sec)	2.72	96.60	20.12	152.42	1.83	45.79	5.64	1227.64	5.44	502.94

Table 2D: Model Performance on 23 attributes (Filter method: CC)

Parameter	DT		RF		NB		LR		KNN (K=11)	
	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF
Accuracy	0.47	0.48	0.52	0.51	0.34	0.35	0.44	0.36	0.50	0.50
F1-score	0.47	0.48	0.52	0.51	0.32	0.33	0.42	0.35	0.50	0.50
Precision	0.47	0.48	0.52	0.51	0.41	0.42	0.42	0.38	0.50	0.50
Recall	0.47	0.48	0.53	0.51	0.34	0.35	0.44	0.36	0.50	0.50
MCC	0.34	0.35	0.41	0.39	0.19	0.20	0.30	0.21	0.37	0.38
Time(sec)	0.66	65.92	10.39	124.26	0.16	10.28	9.86	1821.31	30.25	2169.55

Table 2E: Model Performance on 27 attributes (Wrapper method)

Parameter	DT		RF		NB		LR		KNN(k=11)	
	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF
Accuracy	0.79	0.80	0.85	0.84	0.69	0.69	0.81	0.79	0.82	0.82
F1-score	0.79	0.80	0.85	0.84	0.66	0.66	0.81	0.79	0.82	0.82
Precision	0.79	0.80	0.85	0.84	0.77	0.76	0.81	0.79	0.82	0.82
Recall	0.79	0.80	0.85	0.84	0.68	0.69	0.81	0.79	0.82	0.82
MCC	0.58	0.59	0.70	0.67	0.44	0.44	0.62	0.58	0.64	0.65
Time(sec)	3.79	148.72	26.24	211.85	2.06	87.63	10.38	1937.11	7.77	785.99

Table 2F: Model Performance on 29 attributes (Embedded method)

Parameter	DT		RF		NB		LR		KNN(k=11)	
	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF	T/T	RKF
Accuracy	0.80	0.80	0.85	0.84	0.66	0.66	0.81	0.78	0.82	0.82
F1-score	0.80	0.80	0.85	0.84	0.62	0.63	0.81	0.78	0.82	0.82
Precision	0.80	0.80	0.85	0.84	0.73	0.73	0.81	0.78	0.82	0.82
Recall	0.80	0.80	0.85	0.84	0.65	0.66	0.81	0.78	0.82	0.82
MCC	0.59	0.60	0.71	0.68	0.37	0.38	0.62	0.56	0.64	0.64
Time(sec)	3.28	146.74	19.44	210.68	1.56	76.11	8.59	1961.31	6.27	788.92