The classification of data professionals’ salaries based on results from an online survey

Rushil Umaretiya and Utkarsh Goyal

Thoma Jefferson High School for Science and Technology

Table of Contents

[**Project Goal**](#_jv5crre5dv9j) **3**

[**Description of Dataset**](#_ymrfgmkot4bs) **3**

[**Preprocessing + WEKA**](#_cjyan6ed2eb1) **4**

[**skLearn Preprocessing**](#_ermuejai9lh3) **12**

[**Attribute Selection Algorithms & Model Classifiers Used**](#_f2o200fxqgkd) **13**

[Results and Analysis](#_fku771tx04si) **16**

[Decision Trees](#_orwx10tmiwdq) 16

[Random Forest Classifiers](#_41nl6jneidzn) 17

[Naive Bayes Gaussian Classifier](#_f3m0xipilb4h) 17

[KNeighborsClassifiers](#_fumkzbt20m66) 17

[Attribute Selection](#_uddyaogc34h8) **18**

[Variance Threshold](#_xaigh3vklsb1) 18

[Decision Trees](#_lfiy2kf6bky5) 18

[Random Forest Classifiers](#_3s6asuqh1sj7) 19

[Naive Bayesian Gaussian Classifiers](#_pshlrqygktes) 19

[KNeighborsClassifiers](#_nv84io6ilsq1) 20

[Chi Squared Test](#_rt90o7lznsxa) 21

[Decision Trees](#_dv49sxftibo4) 21

[Random Forest Classifier](#_qsu4azcxtgdb) 21

[Naive Bayesian Gaussian Classifier](#_k2dpkqhy0c8e) 22

[KNeighborsClassifier](#_yi2mxmf1whi2) 23

[Recursive Feature Elimination (Random Forest Classifier)](#_yqpw5wf05asr) 23

[Decision Trees](#_hf1ftkk9cx6a) 23

[Random Forest Classifiers](#_vo5ecl2tgj4z) 24

[Naive Bayesian Gaussian Classification](#_iw7lzz84b0ql) 24

[KNeighbors Classifiers](#_k2pny9gcqm1) 25

[Principal Component Analysis](#_34sfawrzudpq) 25

[Decision Trees](#_8hmyhvpck299) 25

[Random Forest Classifier](#_qfc8jk5xpj2p) 26

[Naive Bayesian Gaussian Classification](#_wmiyrzccog5f) 26

[KNeighborsClassifiers](#_1czcom24n2ur) 26

[**Conclusion/How to Reproduce Our Model**](#_dz4dqx4hvrsj) **27**

[**Team Members and Task Performed**](#_8mi88ydiaxd3) **28**

[**Appendix and Sources**](#_v88jm4diu7z5) **29**

# Project Goal

This project revolves around the processing of survey results from 2017 to 2021 of database administrators, analysts, architects, developers, and data scientists. This data includes a set of attributes along with salary (USD), which was binned, and with this, we were able to predict salary ranges of data professionals internationally with varying degrees of success. You can find all of our data in this [folder](https://drive.google.com/drive/folders/18qzVGtcs2nyDgdE915iTCctlhH4nVfTn?usp=sharing).

# Description of Dataset

Our dataset consisted of raw responses from a survey that [Brent Ozar](http://brentozar.com), SQL expert, conducted over 4 years. His article can be found along with the raw dataset can be found [here](https://www.brentozar.com/archive/2021/09/what-should-we-change-about-this-years-data-professional-salary-survey-2/) ([Excel file](https://downloads.brentozar.com/Data_Professional_Salary_Survey_Responses.xlsx)), which holds a total of 31 columns and 10,341 rows. In 2021, he asked the following questions:

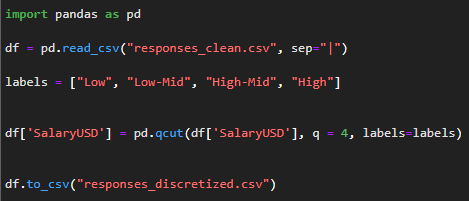
* What’s your total salary in US dollars, annual before taxes?
* Your country
* (Optional) Postal/zip code
* Primary database you work with
* Years that you’ve worked with this database
* Other databases you worked with in 2021
* Job type: (FTE, FTE of consulting/contracting company, independent, part time)
* Job title
* Do you manage other staff?
* Years of doing this job
* At how many companies have you held this job?
* How many other people on your team do the same job as you?
* How many database servers does your team work with?
* What is the population of the largest city within 20 miles of where you work?
* Employer sector (private, government, non-profit, etc.)
* What are your career plans for the year 2022?
* To which gender do you most identify?

# Preprocessing + WEKA

First, we faced an issue with importing our CSV file to WEKA, there was an issue with the file formatting, and so we had to remap some of our attributes’ values and remove others in order to allow ourselves to import our data into WEKA. We removed attributes that were flagrantly irrelevant, empty, or that would negatively affect our model. Here was our rationale for each removal before feature selection:

| **Attribute** | **Rationale** |
| --- | --- |
| Survey Year | Date is irrelevant |
| Timestamp | Date is irrelevant |
| Postal Code | International, Postal code is irrelevant |
| Other Databases | Mixed data with multiple select |
| Database Servers | Mixed data with multiple select |
| Counter | Useless |
| CompanyEmployeesOverall | Data is not fully numeric, 76% of data is missing |
| EducationIsComputerRelated | 46% of attribute is missing |
| NewestVersionInProduction | Mixed data with multiple select |
| PopulationOfLargestCityWithin20Miles | 58% of attribute is missing |

We then discretized our class into four quantile groups where the data was evenly distributed amongst all four bins. This was done using Panda’s qcut function, which will cut a data frame into n bins of equal size. These four bins were split and labeled as such:



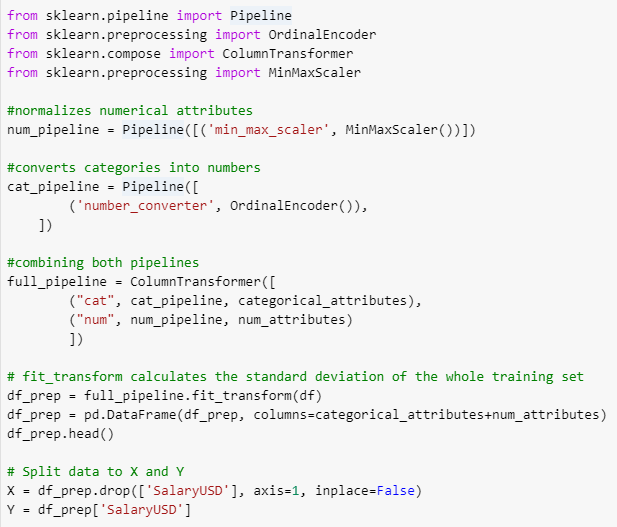
*(All salaries are in $USD)*

| Low | [13, 65,942.5) |
| --- | --- |
| Low-Mid | [65,942.5, 91,000) |
| High-Mid | [91,000, 115,650) |
| High | [115,650, 1,850,000] |

The rest of the features required minor edits/changes in order for them to be processable, accurate, and useful. All of these actions were done in WEKA/Excel.

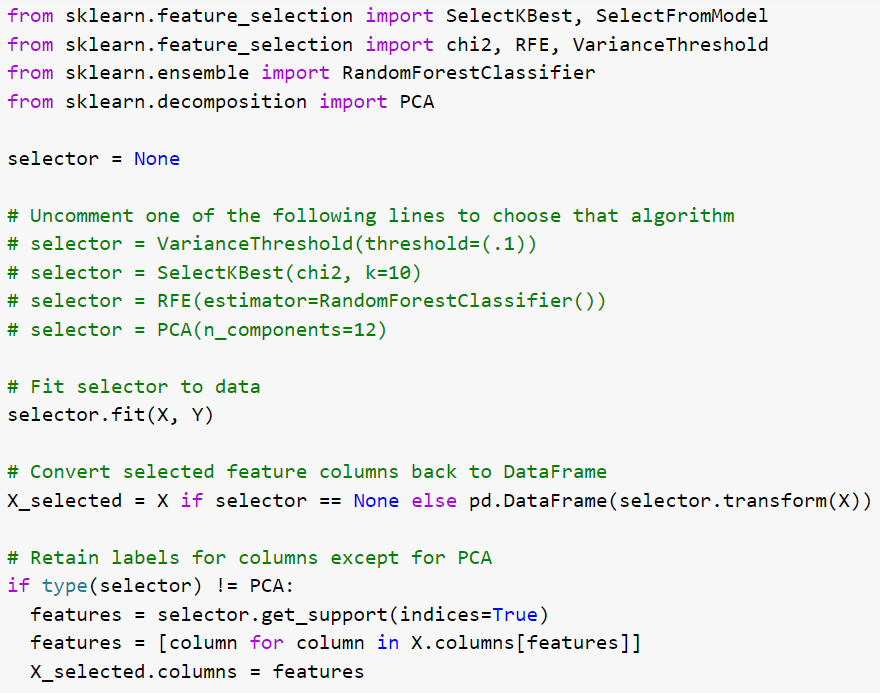
| Attribute | Preprocessing |
| --- | --- |
| Country | Hyphens added between country names |
| Primary Database | Values remapped:   | Microsoft SQL Server | M | | --- | --- | | PostgreSQL | P | | Other |  | | Azure SQL DB | A | | Microsoft Access | MA | | Oracle | O | | MySQL/MariaDB | MS | | Amazon RDS (any flavor) | RDS | | MongoDB | MDB | | DB2 | DB2 | | Cassandra | C | | SQLite | SQLI | | SAP | SAP | | Teradata | T | | Elasticsearch | E | |
| Years With this Database | Delete samples with unrealistic outliers:   * 53716 * 30331 * 1050   Fix years with real values:   * 2000 = 21 * 2003 = 18 * 2020 = 1 * 2017 = 4 * 1997 = 24 |
| Employment Status | | Full time employee | E | | --- | --- | | Full time employee of a consulting/contracting company | EC | | Independent consultant, contractor, freelancer, or company owner | IC | | Part time | P | | Independent or freelancer or company owner | I | |
| Job Title | | Developer: Business Intelligence (SSRS, PowerBI, etc) | D | | --- | --- | | DBA (Production Focus - build & troubleshoot servers, HA/DR) | DBAP | | DBA (General - splits time evenly between writing & tuning queries AND building & troubleshooting servers) | DBAG | | Manager | M | | Developer: App code (C#, JS, etc) | DA | | Developer: T-SQL | DT | | Architect | A | | DBA (Development Focus - tunes queries, indexes, does deployments) | DBAD | | Engineer | E | | Analyst | AN | | Other |  | | Data Scientist | DS | | DBA | DBA | | Principal database engineer | PDE | | DevOps, Sr Software Engineer DBA | DO | | Technician | T | | Database Specialist | DBS | | Consultant | C | | Systems Administrator | SA | | Sales | S | | DBA / BI Developer | DBABI | | Sr Consultant | SRC | | Analytics consultant | AC | |
| Manage Staff | Yes/No → Y/N |
| Years with this type of job | No change |
| How many Companies? | | 1 (this is the only company where I've had this kind of position) | 1 | | --- | --- | | 5 | 5 | | 4 | 4 | | 2 (I worked at another similar position elsewhere before this one) | 2 | | 3 | 3 | | 6 or more | 6 | | Not Asked |  | |
| How many other people are on your team? | | 5 or more | 6 | | --- | --- | | None | 0 | |
| How many employees does your company have overall? | Remove unrealistic outliers:   * 1e+8 |
| Employment | | Private business | PB | | --- | --- | | Education (K-12, college, university) | E | | State/province government | SG | | Local government | LG | | Non-profit | NP | | Federal government | FG | | Student | T | |
| Gender | | Male | M | | --- | --- | | Female | F | | NonBinary | NB | | Prefer not to say/None |  | |
| CareerPlan | | Stay with the same employer, same role | S | | --- | --- | | Prefer not to say |  | | Change both employers and roles | C | | Stay with the same role, but change employers | SR | | Stay with the same employer, but change roles | SE | |
| Years with this kind of job? | Delete samples with unrealistic outliers |
| Education | Unchanged |
| Certs? | No/YesValid/YesExpired |
| Hours | Unchanged |
| Telecommunication hours per week | Unchanged |

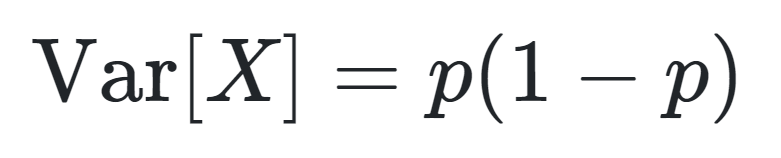
# skLearn Preprocessing

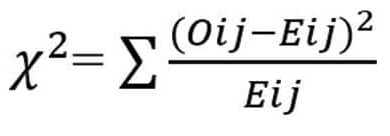
Since the classifiers in scikit-learn do not work with categorical data, we converted our categorical attributes into nominal attributes. Furthermore, the numerical attributes were normalized to be between 0 and 1, which can sometimes help with the accuracy of the classifiers. To do this, we used Ordinal Encoder pipelines as shown below.

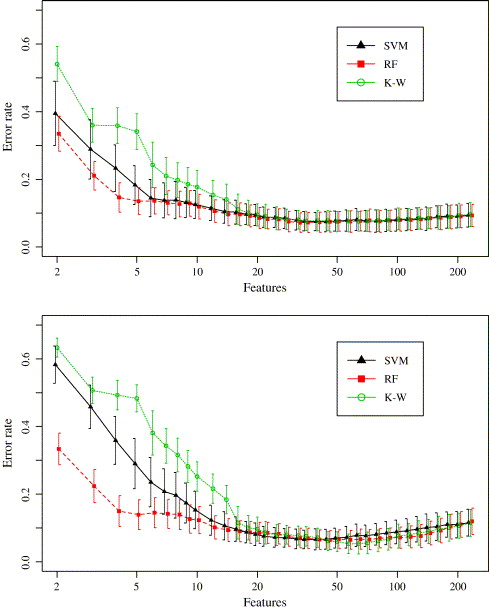
# Attribute Selection Algorithms & Model Classifiers Used

We used 4 different types of feature selection that skLearn offered: VarianceThreshold, Chi-squared, RFE based on RandomForest, and PCA. Different feature selection algorithms produced varied results for different classifiers, and some were boosted and others were weakened on a classifier-by-classifier basis.



The first, most simple, feature selection algorithm was VarianceThreshold. It will calculate the variance of each feature and remove all features whose variance don’t meet the threshold, and we selected a threshold of .25, or where 50% of the samples didn’t meet the threshold.

Second, we decided to use a Chi-square test, which is a statistical test of independence to determine the independence of every feature and the class, and if the class is found to be statistically independent of any given feature, it will be removed.

Third, we used Recursive Feature Elimination with RandomForests as our classifier. The classifier will assign each feature a weight, and RFE will select features by recursively evaluating smaller and smaller sets of features. First, the estimator is trained on the importance of each feature, and then the least important features would be pruned, and this would recur until the desired number of features is reached. We found an interesting article on RFE with RandomForests with a high success rate, and so this is what we used: [***Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products***](https://doi.org/10.1016/j.chemolab.2006.01.007)**.** Here is a graphic from the article illustrating RFE+RF performance rates as opposed to other classifiers:

Finally, we chose Principal Component analysis as our last selector. PCA will decompose features into smaller dimensions. skLearn’s implementation uses SVD, specifically the LAPACK implementation. Unfortunately, PCA compresses the feature space, so we are unable to tell which variables are being weighed and how, but our results are quite promising nevertheless.

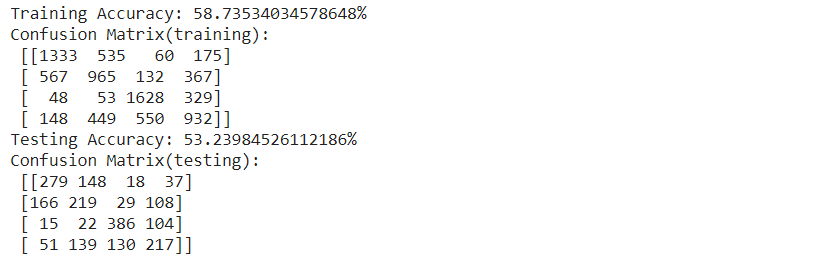
Attribute Selection Results

| **Variance Threshold** | **Chi-Squared Test** | **RFE + RandomForest** |
| --- | --- | --- |
| Country | Country | Country |
| Primary Database | Employment Status | Job Title |
| Employment Status | Job Title | YearsWithThisTypeOfDatabase |
| Job Title | ManageStaff | Years with this type of job |
| ManageStaff | Education | HowManyCompanies |
| EmploymentSector | Certifications | OtherPeopleonYourTeam |
| Career Plans this Year | Telecommute Days Per Week | DatabaseServers |
| Education | YearsWithThisDatabase | HoursWorkedPerWeek |
| Certifications | TelecommuteDaysPerWeek |  |
| TelecommuteDaysPerWeek | YearsWithThisTypeOfJob |  |
| OtherPeopleOnYourTeam | HowManyCompanies |  |

# Results and Analysis

The following screenshots show the accuracy of our model if we don’t use attribute selection algorithms. We can see that Random Forest gives us the best accuracy with an accuracy of 54.9%, while NaiveBayes gives us an accuracy of only 35%. It's important to note that NaiveBayes may be failing due to the assumption that all the attributes are considered independent of each other. We can only notice that Decision Trees and Random Forest Classifiers are doing really well comparatively for this dataset, meaning that entropy is a good measure of how to split the dataset. Furthermore, accuracy was also a good measure to determine how great our model was doing as the classes we discretized into had a relatively equal amount of instances.

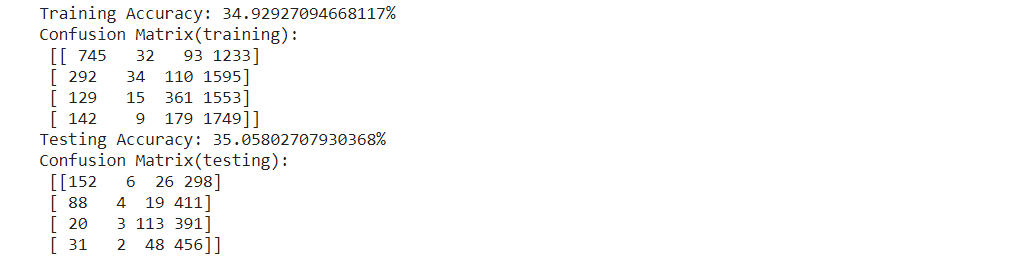
## Decision Trees



## Random Forest Classifiers



## Naive Bayes Gaussian Classifier



## KNeighborsClassifiers

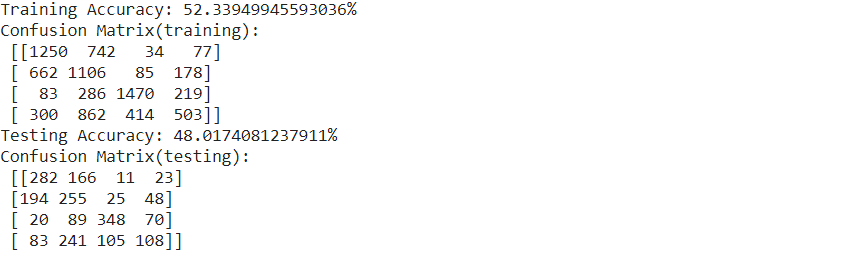


# Attribute Selection

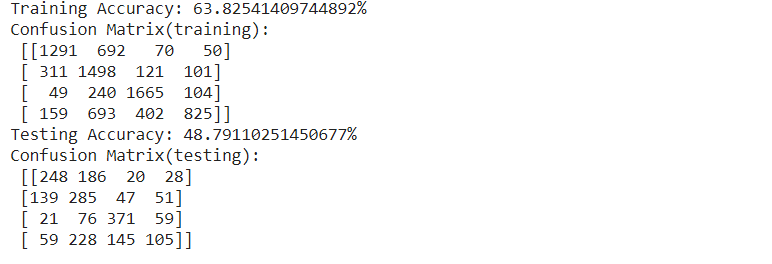
For attribute selection algorithms, there was significant improvement with the RFE classifier on Naive Bayes algorithm, which may be due to the fact it's an efficient algorithm in the sense that it will recursively eliminate features that would be considered “bad” or useless for the dataset, while checking to make sure that it helps. Compared to the other algorithms, they check for more correlation between the attributes and the classification, which isn’t necessarily helpful, if those “insignificant” attributes were distinguishing between some of the specific instances. Principal Component analysis only was able to slightly affect NaiveBayes, but this may be because of its expectation that all attributes are independent, which in this case it may not be the case that they are. Furthermore, we can see the comparison of our results with the following screenshots to the control of our dataset as shown above:

## Variance Threshold

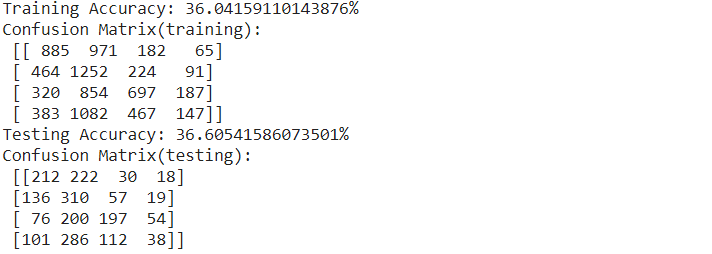
### Decision Trees

****

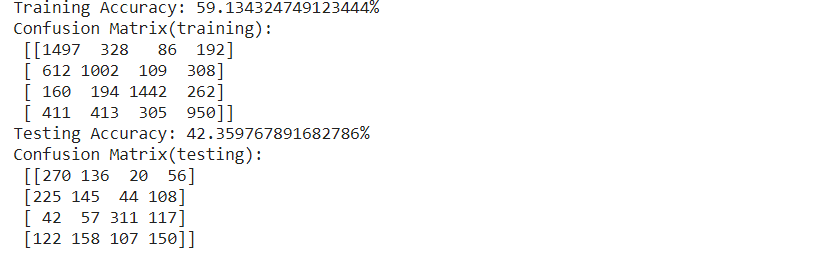
### Random Forest Classifiers



### Naive Bayesian Gaussian Classifiers

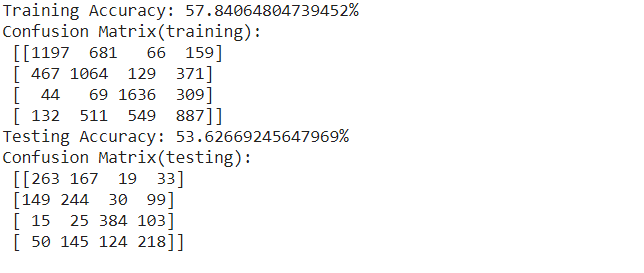


### KNeighborsClassifiers

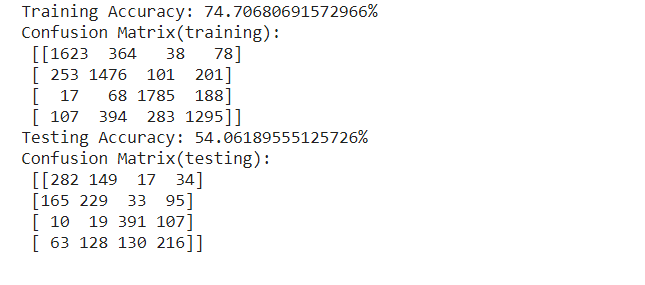


## Chi Squared Test

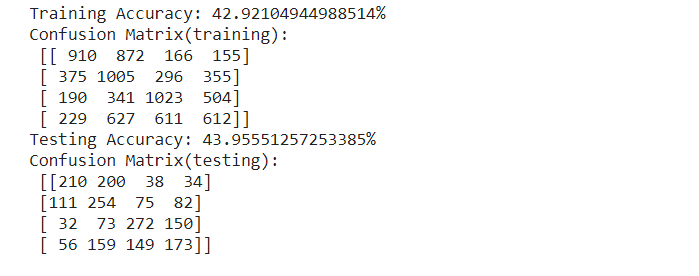
### Decision Trees

****

### Random Forest Classifier

****

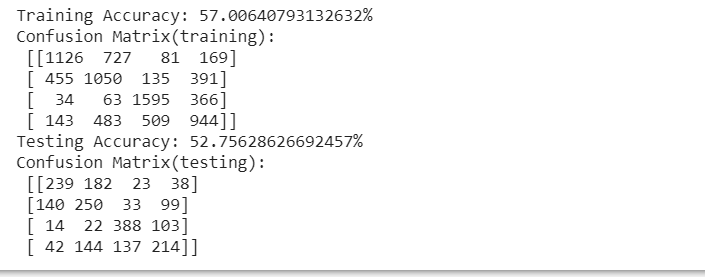
### Naive Bayesian Gaussian Classifier

****

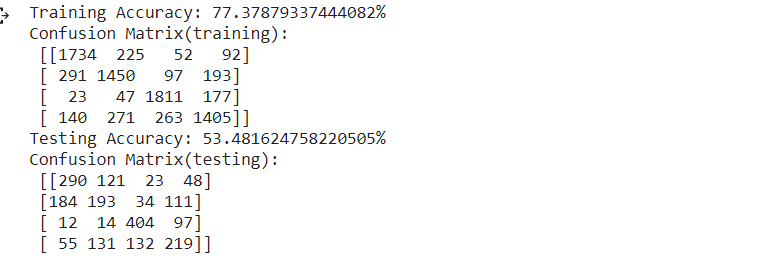
### KNeighborsClassifier

## Recursive Feature Elimination (Random Forest Classifier)

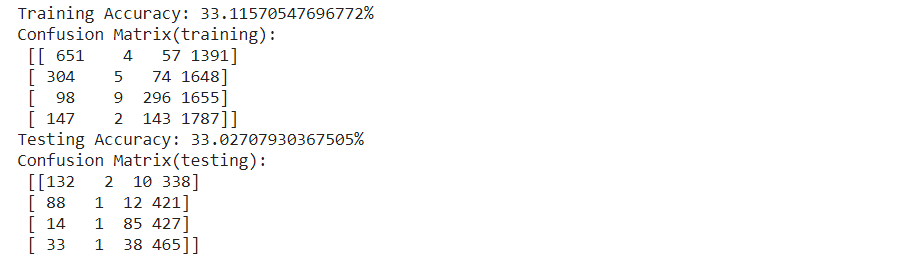
### Decision Trees

****

### Random Forest Classifiers

****

### Naive Bayesian Gaussian Classification

****

### KNeighbors Classifiers



## Principal Component Analysis

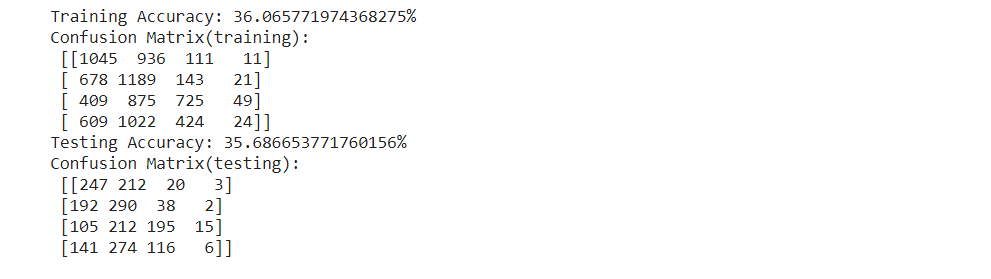
### Decision Trees

****

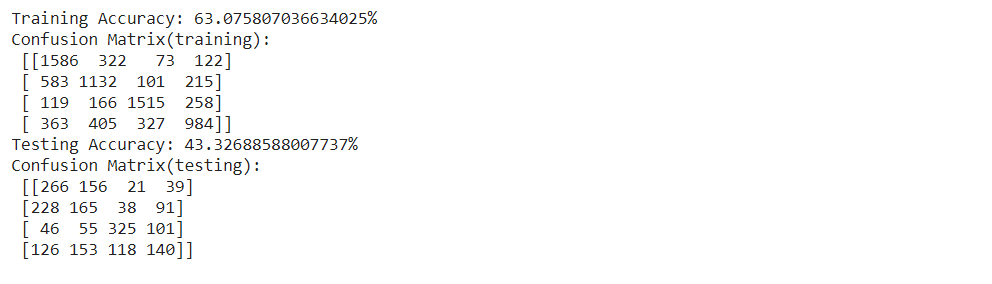
### Random Forest Classifier



### Naive Bayesian Gaussian Classification

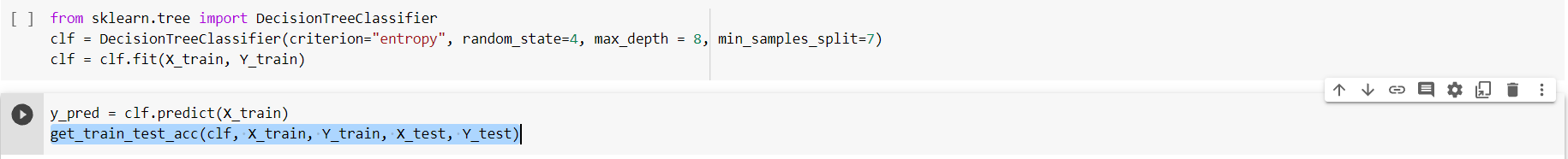


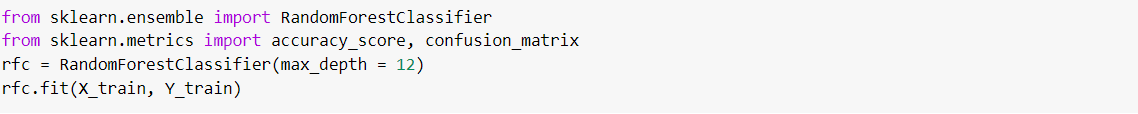
### KNeighborsClassifiers



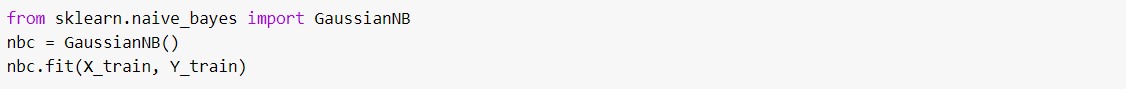
# Conclusion/How to Reproduce Our Model

**Decision Trees -** For the decision tree model, we used the entropy model, along with a max\_depth of 8 to prevent overfitting. Also, in our splits for the decision trees, we split it into a maximum of 7 different things, to make sure no section was thoroughly overfit. Other than this, the default parameters of Decision Tree Classifier were used for sklearn.



**Random Forest Classifier** - For this classifier, we used a similar parameter with decision trees, but instead set the max\_depth to be 12, because it was generalizing too much before that. The following code shows how to recreate the random forest classifier. 

**Naive Bayesian Gaussian Classifier -** This classifier wasn’t doing well no matter how we changed the parameters, so we ran the default parameters with the following code.



**KNeighborsClassifiers** - We tried this classifier as it seems interesting to group our data into clusters. For this one as well, we simply used the default classifier, and in terms of accuracy, it was in between Naive Bayesian and Decision Trees/Random Forest. Here’s how we did it:

# 

# Team Members and Task Performed

**Rushil Umaretiya -** I performed all of the preprocessing of the data, discretizing, splitting, and making it readable for WEKA and scikit-learn, along with the attribute selection explanations and the preprocessing section on the report.

**Utkarsh Goyal -** I did the majority of the classifier coding and worked on the analysis for the Final Project Report. I also worked on the tables for the attribute selection algorithms and inserted the finalized results into the report.

***We certify that the distribution of work was equal amongst both group members.***

# Appendix and Sources

*What should we change about this year's data professional salary survey?* Brent Ozar Unlimited®. (2021, September 28). Retrieved October 14, 2021, from https://www.brentozar.com/archive/2021/09/what-should-we-change-about-this-years-data-professional-salary-survey-2/.