NOTTINGHAM TRENT UNIVERSITY

School of Science and Technology

COMP20121

MACHINE LEARNING FOR DATA ANALYSIS

by

Kypros Tsolakis

N0950765

Contents:

Part 1 - Introduction:

Data analysis is a technique of gleaning insights from data in order to improve business decisions. This data analysis technique usually moves through five iterative phases as follow:

- **Identify** the company request that you need to answer. What is the actual problem the company is facing? What do you need to in order to identify that and how you can target it?
- **Collect** the raw data sets in order to help you answer the identified question. Data collection can be obtained during feasibility study from internal users, like a company's managers, end users, CRM software, or other sources as portal website and social media programming activities.
- **Clean** the data to prepare it for analysis. This frequently includes removal duplicate and irregular data, integration of discrepancies, normalizing data structure and format, and dealing with syntax errors.
- **Analyse** the data. Manipulation of data can be done with a use of different data analysis techniques and tools. During this phase, you can apply data mining to discover patterns within databases or data visualization software to help transform data into graphical format that will be easier to understand and come up with conclusions.
- **Interpret** the outcomes of your analysis to see how well the data answered your original question. What are the possible recommendations you make based on this outcome? What are the limitations to your conclusions?

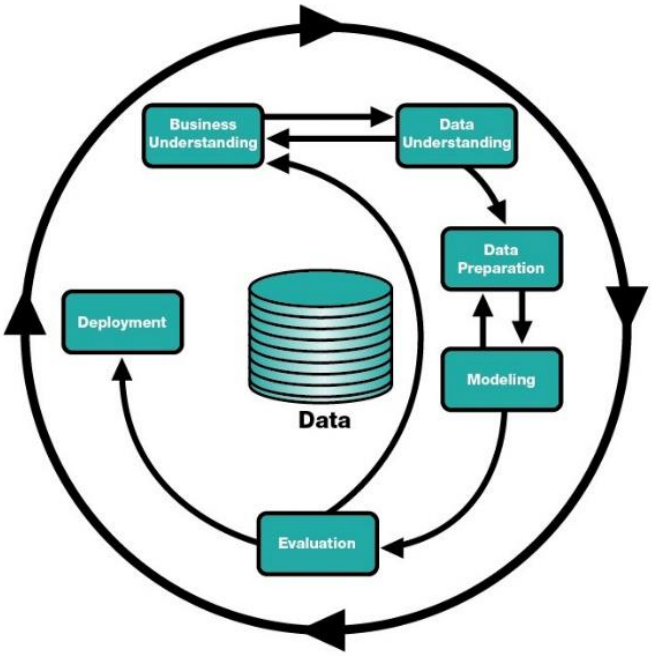
Data analyst tasks and responsibilities

A data analyst profession is to collects, cleans, and interprets data sets in order to answer an inquiry or solve a problem.

- **Gather data:** Analysts frequently gather data using different methods. This might consist of conducting surveys, trailing visitor characteristics on a company website.
- **Clean data:** Raw data might contain duplicates or errors. By cleaning the data, we can improve the quality of data in a spreadsheet or the output of a programming language.
- **Model data:** It involves creating and designing the structures of a database. You can choose what types of data to store and collect, establish how data categories are related to each other, and work through how the data actually appears.
- **Interpret data:** Interpreting data will include patterns discovery and trends in data that will help you answer the question.
- **Present:** Communicating the results of your analysis to the customer by using graphs, writing reports and diagrams it is important job of analysts.

What is CRISP-DM methodology:

CRISP-DM data mining methodology was developed during 90's and stands for Cross-industry standard process for data mining. It was a process model use during data mining in which data mining experts describe the common approaches they use to tackle problems. In this process diagram, iteration is explicitly stated as the rule rather than an exception. Going through the process once and not solving the problem generally isn't a failure. It is often the case that the whole process is centred around exploring the data, and after the first iteration the team knows much more. The method soon became successful as standard method for business and industrial applications.



Data mining is broken down into six phases:

Business Understanding: Emphasizes on understanding the objectives and requirements of the project. Business projects rarely come pre-packaged as clear and explicit problems for data mining. It is a need to define the data mining problem and to develop an initial plan to achieve the objectives.

Data Understanding: Focuses on identifying, collecting, describing, and analysing the data sets in order to achieve the appropriate results for the project. It is significant to detect important data and interesting subsets as also the relationship among them.

Data Preparation: The data preparation phase includes all operations that effect in the final dataset. This can be accomplished by selecting, cleaning, constructing, integrating, and formatting data.

Modelling: Is where data mining techniques are applied to the data. Determine which algorithms to try (e.g., decision trees, regression, neural net), generate test design, build model and assess model.

Evaluation: Before moving to the last phase, it's very critical to do a deep analysis and evaluation of the model and assess the processes used to build it to ensure that it meets the business objectives. One of the main goals is to see whether there is any critical business issue that has not been sufficiently addressed. A choice on how to use the data mining results should be made at the end of this step.

Deployment: Data mining findings — and increasingly, data mining techniques themselves — are put to real-world use in order to provide a return on investment. Implementing a predictive model in an information system or business process is one of the clearest cases of deployment.

Part 2 - Data Understanding, Data Pre-processing, Exploratory Data Analysis:

The survey was open live from 09/01/2021 to 10/04/2021, and the survey data was documented in December 2021. In this survey, more than 25000 people from 171 different countries and territories contributed, in which they were asked around 38 questions regarding their personal opinion about their personal carrier. As the purpose of the data set changes every year, the 2021 survey focused more on reflecting programmers' diversity. However, if someone sees the full data, he will understand how huge this survey was and almost amazing for someone to analyse the whole data. In the following table, you can see the attributes I chose to include in my data analysis. Those attributes in my opinion are the ones which should give us the most significant conclusions.

Selected Data Attributes:

No	Attributes Names	Explanation	Type
1	Age	The age of the developer.	Object
2	Country	The country that the developer lives.	Object
3	Education	The education level that the developer has.	Object
4	YearsOfCoding	How many years the developer codes.	Object
5	ProgrammingLanguage	What programming language the developer recommended to a new person in coding	Object
6	ComputingPlatform	What computing platform does the developer use most often for data science projects.	Object
7	Income	What is the salary that one developer makes in a year (approximate $USD)	Object

Figure 1 - Table describing all important data attributes.

Cleaning Data:

The first thing that we need to do is to clean the data that we were given so that we can focus on analysing only the attributes we have chosen for further analysis.

In order to clean the data, we must identify the respondents and any missing or incorrect values, as well as any outliers. Firstly, we will isolate, on a different table, the data attributes we have chosen. This will make our data cleaning and analysis tasks easier.

As you can see in the next table there some answers with the name ‘Nan’. Those answers are missing because Some people chose not to answer those particular questions.

	Age	Country	Education	YearsOfCoding	ProgrammingLanguage	ComputingPlatform	Income
1	50-54	India	Bachelor's degree	5-10 years	Python	A laptop	25,000-29,999
2	50-54	Indonesia	Master's degree	20+ years	Python	A cloud computing platform (AWS, Azure, GCP, h...	60,000-69,999
3	22-24	Pakistan	Master's degree	1-3 years	Python	A laptop	$0-999
4	45-49	Mexico	Doctoral degree	20+ years	Python	A cloud computing platform (AWS, Azure, GCP, h...	30,000-39,999
5	45-49	India	Doctoral degree	< 1 years	Python	A cloud computing platform (AWS, Azure, GCP, h...	30,000-39,999
...	...	...	...	...	...	...	...
25969	30-34	Egypt	Bachelor's degree	1-3 years	Python	A laptop	15,000-19,999
25970	22-24	China	Master's degree	1-3 years	Python	A personal computer / desktop	NaN
25971	50-54	Sweden	Doctoral degree	I have never written code	NaN	NaN	$0-999
25972	45-49	United States of America	Master's degree	5-10 years	Python	A laptop	NaN
25973	18-21	India	Bachelor's degree	I have never written code	NaN	NaN	$0-999

25973 rows × 7 columns

Figure 2 - Rows within the created table to display the important data attributes

In the following table, we list the number of the people who failed to answer the questions we chose to analyse. We can clearly see that all of the people answered the questions regarding their age, country, education, and years of coding, but many people fail to answer the questions regarding the programming language they use, their computing platform, as well as their salary. As it shows for the attributes Age, Country, Education and YearsOfCoding we got 100% percentage of answers but for ProgrammingLanguage, ComputingPlatform, and Income we have a lot of missing values.

```
Age                    0
Country                0
Education              0
YearsOfCoding          0
ProgrammingLanguage 1033
ComputingPlatform   1253
Income             10582
```

Figure 3 - Missing values 'NaN'

To handle those missing values, we have many options. For instance, we can delete the records or to choose to replace those values with the mean value calculated from the other records. Also, we could use various machine learning algorithms, like linear regression, to calculate the missing values or to remove them from our table with the result of missing a large loss of data. We can say for education or personal reasons they decided not to give a specific answer. One more option to handle those ‘NaN’ values we could put them in a separate category. Also, another action possible to do is to replace those values with the most given answer or to put our answer but taking this risk later when we are going to analyse our attributes will cause a lot of inaccurate results for our data. The final decision is to remove them completely from our table and as you can understand in the next table number of rows are down significantly.

	Age	Country	Education	YearsOfCoding	ProgrammingLanguage	ComputingPlatform	Income
1	50-54	India	Bachelor's degree	5-10 years	Python	A laptop	25,000-29,999
2	50-54	Indonesia	Master's degree	20+ years	Python	A cloud computing platform (AWS, Azure, GCP, h...	60,000-69,999
3	22-24	Pakistan	Master's degree	1-3 years	Python	A laptop	$0-999
4	45-49	Mexico	Doctoral degree	20+ years	Python	A cloud computing platform (AWS, Azure, GCP, h...	30,000-39,999
5	45-49	India	Doctoral degree	< 1 years	Python	A cloud computing platform (AWS, Azure, GCP, h...	30,000-39,999
...	...	...	...	...	...	...	...
25963	18-21	India	Bachelor's degree	1-3 years	Python	A laptop	$0-999
25964	60-69	United States of America	Bachelor's degree	20+ years	Python	A personal computer / desktop	300,000-499,999
25967	30-34	India	Bachelor's degree	1-3 years	Python	A cloud computing platform (AWS, Azure, GCP, h...	3,000-3,999
25968	35-39	South Korea	Bachelor's degree	5-10 years	Python	A personal computer / desktop	80,000-89,999
25969	30-34	Egypt	Bachelor's degree	1-3 years	Python	A laptop	15,000-19,999

14430 rows × 7 columns

Figure 4 - Important data attributes table after cleaning missing values.

Exploratory Data Analysis of Chosen Attributes:

Exploratory data analysis (EDA) is a method of data analysis that involves doing investigations on data in order to detect patterns or anomalies. Our next section focusses on examining the selected important attributes to gain a better understanding of the data.

We start by applying EDA on our first selected attribute, Age, and then we continue in a similar way for the rest of the attributes EDA to our first attribute, Age, and after this we will work in similar way to the next attributes.

Age:

To start analysing the Age attribute (first attribute) is to go through the data that is stored. People were asked to give their age in years, and for that reason, the attribute is a float number. People were asked to give their age in years, and for that reason, the attribute is a float number. An important thing to note is that every participant in the study answered that particular question and for that reason we do not have any missing data.

By using a bar chart to visualize the data, we can clearly see that most of the developers fall under the age groups between 25-29, and 18-21, for which the total number in both ranges is almost identical. Then as expected, there are less developers in the age group of over 70 years of age. The less developers are in the range 70+ years old.

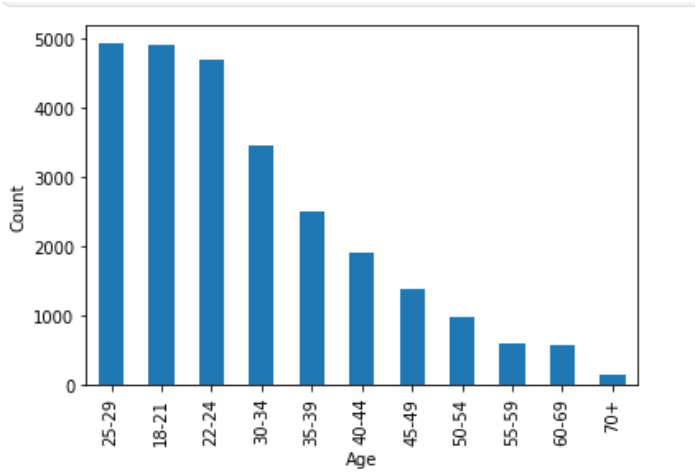


Figure 5 - Age bar chart

The same conclusion, but with a bit more numerical details, can be seen by using a pie chart instead of a bar chart for visualizing our data. Using this approach, we can immediately see that more than half of our data fall under the age group between 18 – 29.

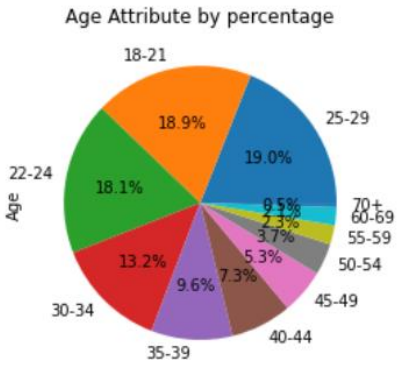


Figure 6 - Age pie chart with percentages

Country:

The next attribute that is going to be analysed is Country. For this one we had responses from 171 different countries and territories. Moreover, this question was answered from all the developers that were asked so we will have exact results when we show the data. To get a good first understanding of our data, we again visualize them on a world map using Choropleth. This helps us to better understand the countries from which the developers who answered the survey come. This helps us to better understand the countries from which the developers who answered the survey come. We can clearly see that most of the people who answered the survey come from India, followed up from the United States. Moreover, we can see that no developers from Central Africa, or Greenland answered the survey.

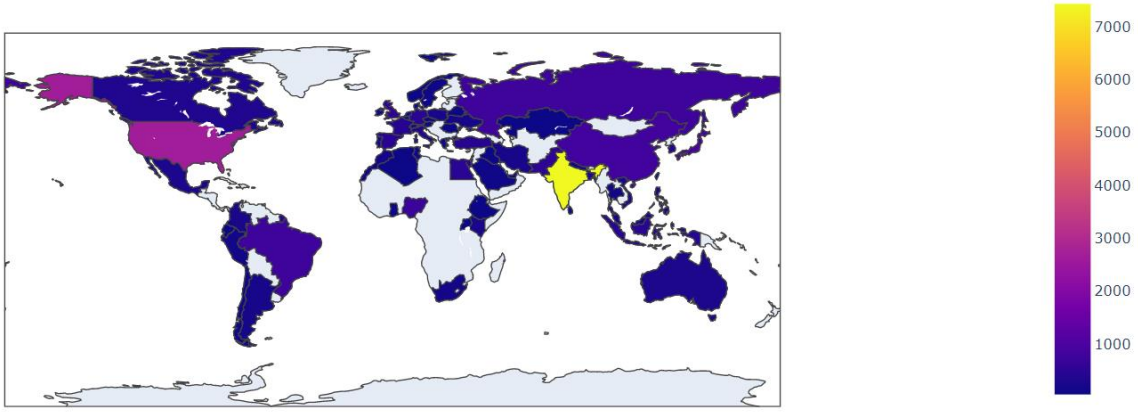


Figure 7 - Choropleth map of developers for each country

As previously, by using a pie chart we can get a numeric representation of our results for the countries. Due to the vast variation of answers of that particular question, during our analysis, we chose to look at the top six countries individually, and grouping all the other ones into one group in order to get a clearer picture of what is going on with our data. In the next pie charts, you can understand the difference between grouping the rest of the countries together. From the pie chart we can see our top country is India with a percentage 21.6% followed by the United States with 12.0% percentage. Those are the only two countries which have a percentage of more than 10%, while the other following four countries are others, Japan, Brazil, and Russia. The rest of the countries, which were grouped together, constitute into a percentage of 49.7%. This is almost half of the data sample.

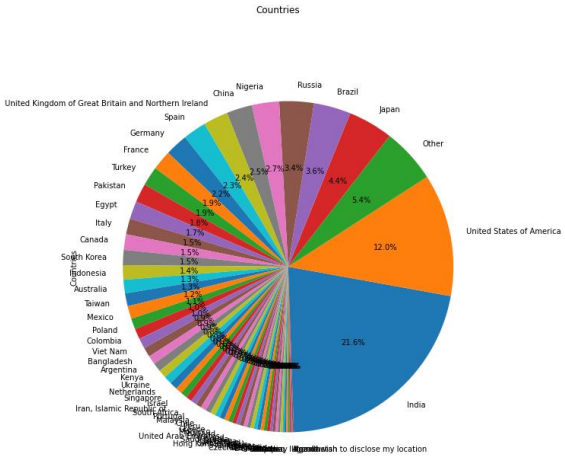


Figure 8 - Country pie chart before group countries percentages

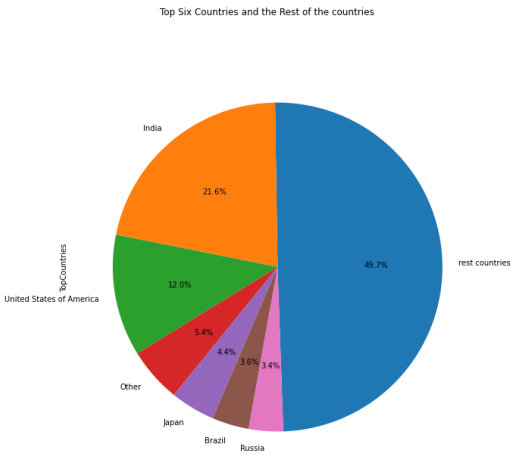


Figure 9 - Country pie chart after group countries percentages

Education (Education Level):

The next attribute we reviewed is the Education. For getting the answers, Stack Overflow asked the participants to state the highest formal education they have attained, or plan to attain within the next 2-

years. As the other two attributes analysed previously, this one was answers successfully from all the participants. For visualizing the answers of the participants regarding the Education attribute, we used a horizontal bar chart. From the results, we can see that the majority of the participants already have a Master's degree, or aim to get one, while the least of the participants, out of those who answered the question, either have no formal education, or have a PhD (or PhD candidates)

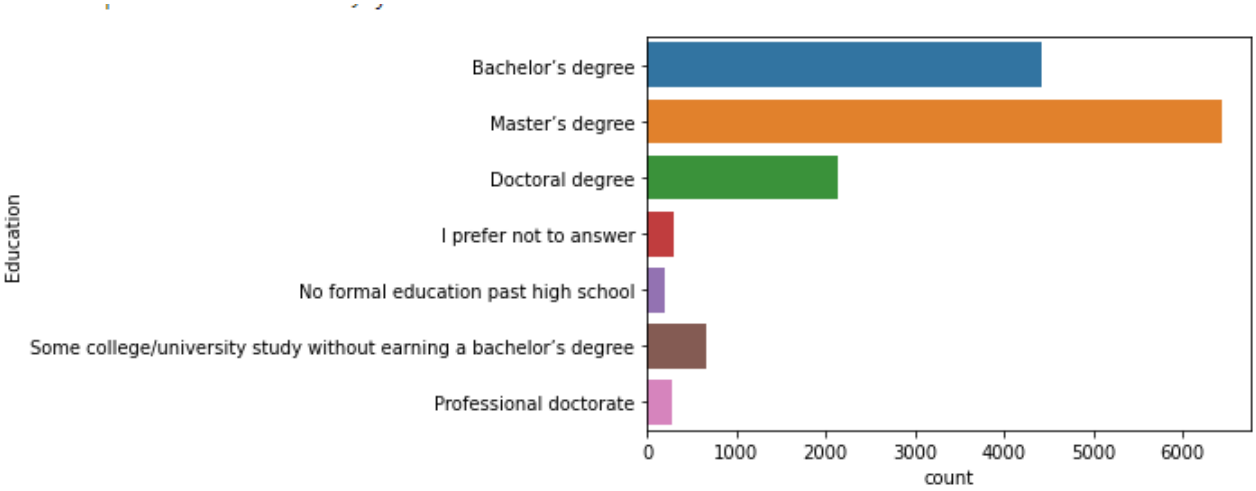


Figure 10 - Education Level horizontal bar chart

To better understand the Education question results, as previously, we graphically represent them by using a pie chart on the Education part, we will use one more technique so we will get the most accurate results from our data. The results clearly show that more than 50% of the participants either have a Master's or a Bachelor's degree, followed by the Doctoral degree with a percentage of 10.8%, while the rest of the participants constitute less than 10% of the total number.

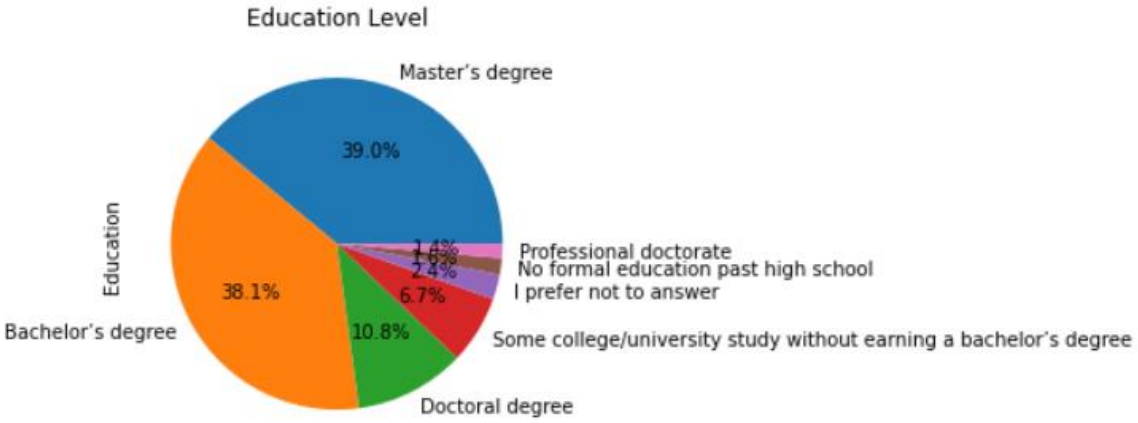


Figure 11 - Education Level pie chart percentages

YearsOfCoding:

Our next attribute is the YearsOfCoding as we can understand for this element the developers were asked how many years, they are writing code, or they are doing programming. The developers for this question

had to choose between seven answers. Again, this question from Stack Overflow was perfectly answer so for one more time our data does not need any clean to analyse it. The first methodology we are going to use is a bar chart. From this chart is observed that most humans that have been doing coding from 1 to 3 years. On the other hand, is obvious that less than 2000 thousand people have been coding for more than 20 years. The same observation can be concluded for people that have never written code. Additionally, the population of people that have been coding for more than 3 years is generally fewer that those that have been coding for below that 3 years.

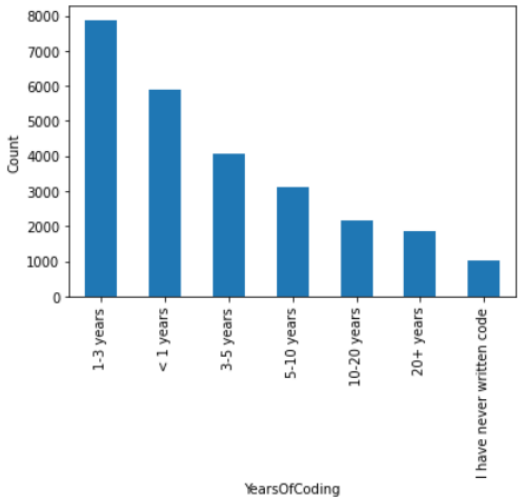


Figure 12 - YearsOfCoding bar chart

To further analyse the YearsOfCoding attribute we create one more bar chart but this time we are going to compare the YearsOfCoding data with the Education (Education Level) data. Most of the people that have already bachelor's or master's degree or planning to get one have been coding for more than 1 year but less than 5. Moreover, the people that graduate or planning to graduate with master's degree have more experience in coding then the doctoral and professional doctorate degree graduates (5 to 10 years of coding). Finally, people that choose to expand their knowledge are more familiar with coding than those that stopped the studies.

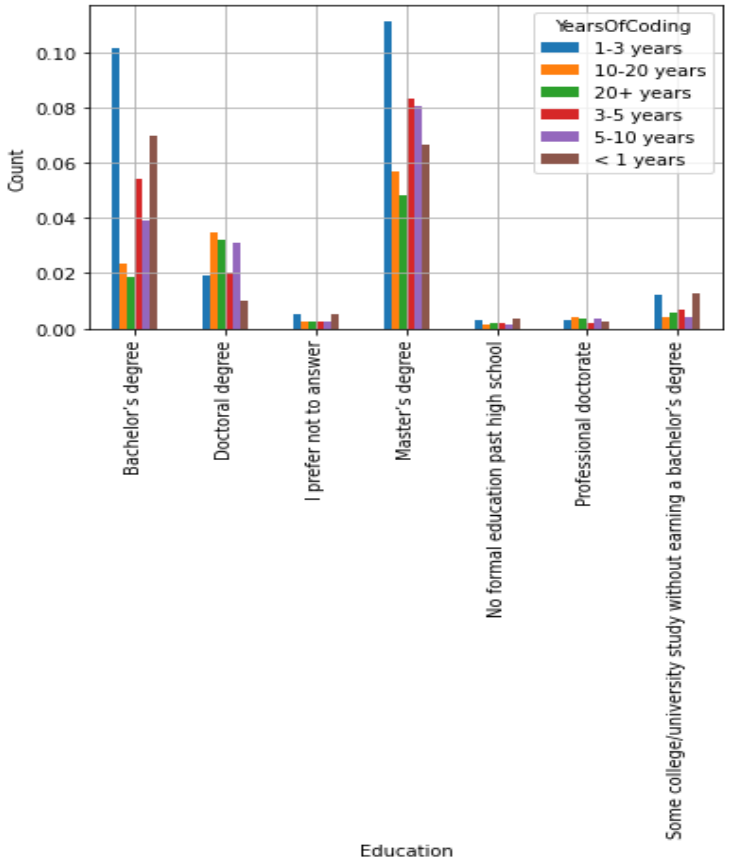


Figure 13 - YearsOfCoding compare with Education Level

ProgrammingLanguage:

ProgrammingLanguage attribute is about what programming language the developers recommend to a beginner that tries coding for the first time in their life to start with. In contrast with the previous element, not all people answered the question. To be exact 1033 people chose not to answer this question and gave the response 'NaN'. To have more accurate results and so our data are more comprehensive, we removed those wrong responses from our table.

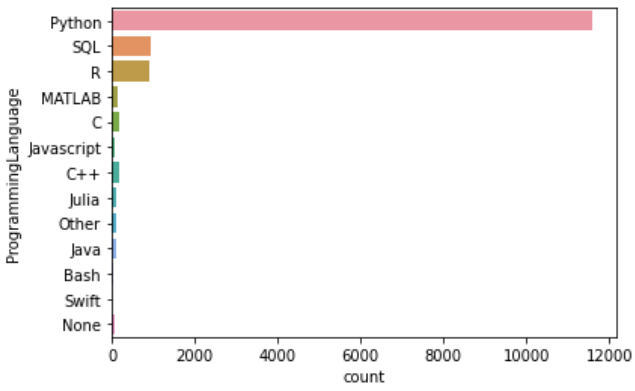


Figure 14 - ProgrammingLanguage horizontal bar chart

As you can see from the graph erasing the missing data helped us a lot with the results for ProgrammingLanguage attribute as it is clear now that more than half of the developers are suggesting starting with learning Python initially because in their opinion is the easiest way to start programming. Some of them think SQL or R language will be good to start with and a very small percentage vote for a different language.

ComputingPlatform:

We analyse ComputingPlatform attribute by asking the developers "what type of computing platform do they use for their data science projects?". Like the previous element, this one has some missing values as 1253 people answered 'NaN' and thus we removed those values again so when we analyse the attribute, our data will be understandable and more accurate. The methodology we will use for ComputingPlatform is a bar chart comparing the top six countries from our previous data, Country attribute. This way we can see what the favourite platform in our top six countries is. As it turns out from the chart most of the countries are using laptop and their second choice is a personal computers/desktop. Some of those countries are using a cloud computing platform (e.g., Aws, Azure) or a deep learning workstation (e.g., NVIDIA GTX, Lambdalabs). The difference between the numbers that use either laptop or computers/desktops and the numbers that use any other ComputingPlatform is gigantic in these countries.

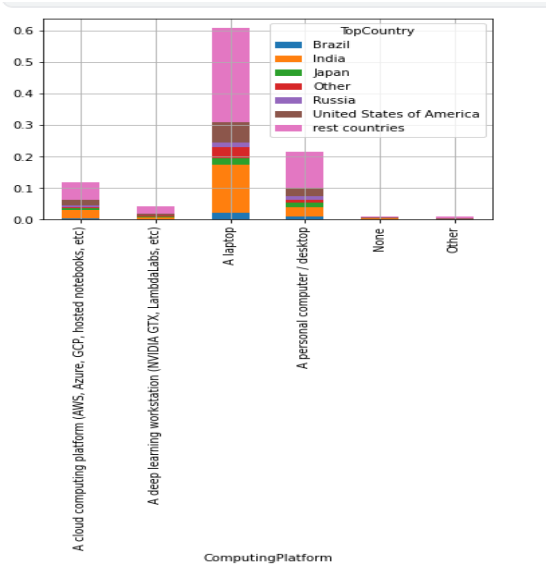


Figure 15 - ComputingPlatform compare with top 6 countries

Income (Income Salaries)

Our final attribute is ''Income''. This element is based on the salaries that every developer earns per year to program in Dollars ($). This last element has the most missing values. The number of those values is 10582 which is almost half the people that Stack Overflow asked about their year salary. Once again, we remove those missing values to analyse our data. Furthermore, this question has various answers as it can be seen. The bar chart below is not helpful, the only thing that can be concluded from the chart is that the answer $0-999 has be given much more frequently compared to the other, so we need to make some modifications.

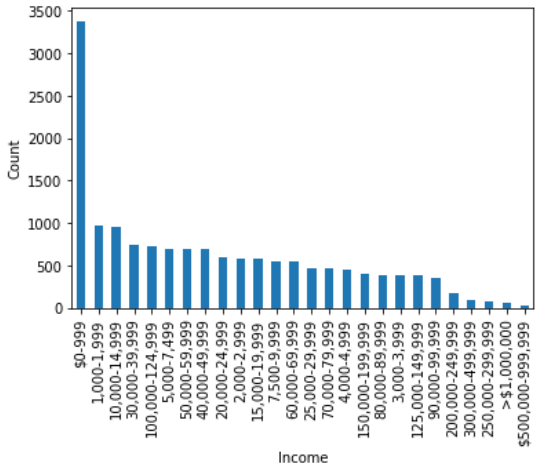


Figure 16 - Income bar chart

The first action is to select the first top nine salaries and put it in a pie chart so we can see their percentage and put the rest salaries together. As we observed before, the range of $0 to 999 is much more popular as an answer compared to the others by 21.3%. Taking this action now it can be easily viewed which ones are our top nine selections given from our developers. Also we see that some of the values have the same percentage with others values such as $1,000-1,999 and $10,000-14,999 with the percentage of 6.2% , $50,000-59,000 and $40,000-49,999 with the percentage of 4.6%. Additionally, the rest of the salaries are closed to 50% percent.

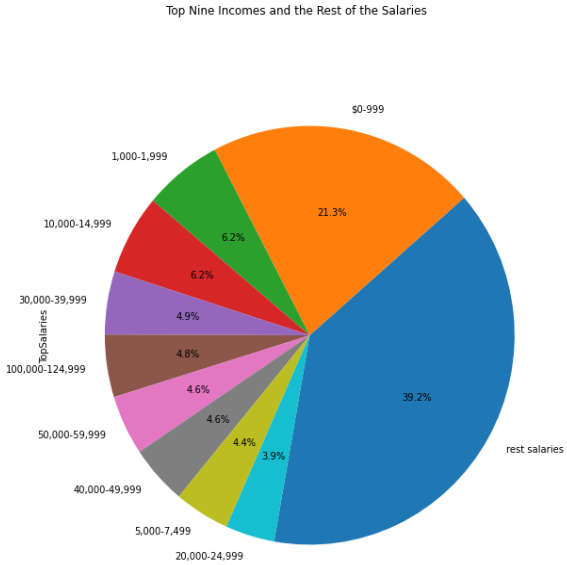


Figure 17 - Income pie chart after group the salaries percentages

The other action that we are taking is to compare the top nine salaries with the top six countries in order to have an idea about what is happening between those two attributes. As we can conclude for India the one third of their developers earn $0-999 and in United States a big number of the developers earn $100,000-124,999. For the remaining countries (Brazil, Japan, Russia and Other countries) is obvious that the salaries that people earn is roughly the same for all the categories. For the ''rest of the countries'' the main answers are ''the rest salaries'' and ''$0-999'' while the other answers are given in almost equal fequency.

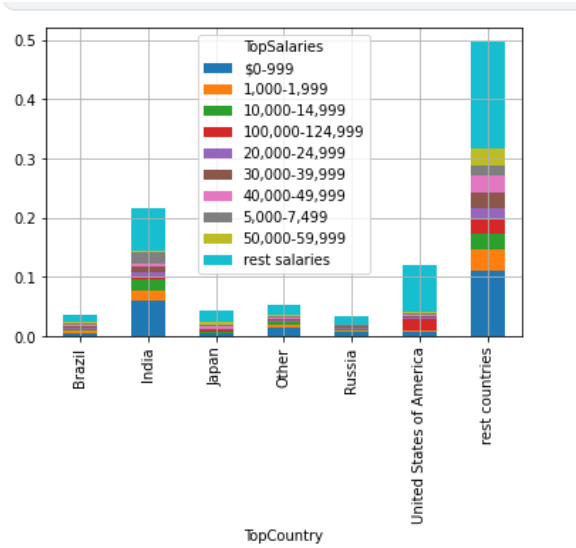


Figure 18 - Top salaries compare with the top six countries

Part 3 - Cluster Analysis:

Cluster analysis refers to combining records, data, or observations into categories with similar characteristics or features, and is associated with unsupervised learning. In this part we are splitting our data in two categories high-Income and the low-Income. The technique we are going to use to split the data is to find the median and if a value is higher than the median is high-Income and a value equal to or lower than the median is low-income. Furthermore, the salary earned, and income columns have been removed from the data frame since these are the target variables and disrupt clustering. We will then use cluster analysis to identify trends in the subsets based on their important features. Several different methods exist to uncover the optimal number of clusters, including k-means and 'the elbow method'.

The meaning of K:

The "elbow" approach of choosing the appropriate number of clusters for K-means clustering is implemented by the K-Elbow Visualizer. K-means is a straightforward unsupervised machine learning technique that divides data into a set of clusters (k). The procedure is relatively naïve in that it assigns all members to k clusters even if it is not the proper k for the dataset since the user must provide k in advance.

The elbow approach performs k-means clustering on the dataset for a range of k values (say, 1-10), and then computes an average score for all clusters for each value of k. The distortion score, which is the sum of square distances from each point to its assigned centre, is computed by default.

Other metrics, such as the silhouette score, the mean silhouette coefficient for all samples, or the calinski harabasz score, which computes the dispersion ratio between and within clusters, can also be utilized. When these overall measures for each model are shown, the optimal value for k may be visually determined. The "elbow" (the point of inflection on the curve) is the optimal value of k if the line chart resembles an arm. The "arm" can travel up or down, but if there is a sharp inflection point, it's a solid sign that the underlying model fits well there.

CH index:

We'd want our clustering assignments C to have a small W and a high B at the same time.

The CH index is based on this concept.

We keep track of CH score for clustering assignments that come from K clusters:

$CH(K) = B(K)/ (K - 1)\ W(K)/ (n - K)$

To choose K, just pick some maximum number of clusters to be considered Kmax (e.g., K = 20), and choose the value of K with the largest score CH(K), i.e.,

$K\hat{} = \text{argmax}\ CH(K)$

$K \in \{2,...Kmax\}$

We applied the elbow method for our first element the top six countries compare the high-Income and low-Income to find which value of K between the range 2 to 10 so we can do our clustering and as we can see the best option for our K is equal 5 so we are going to set K=5.

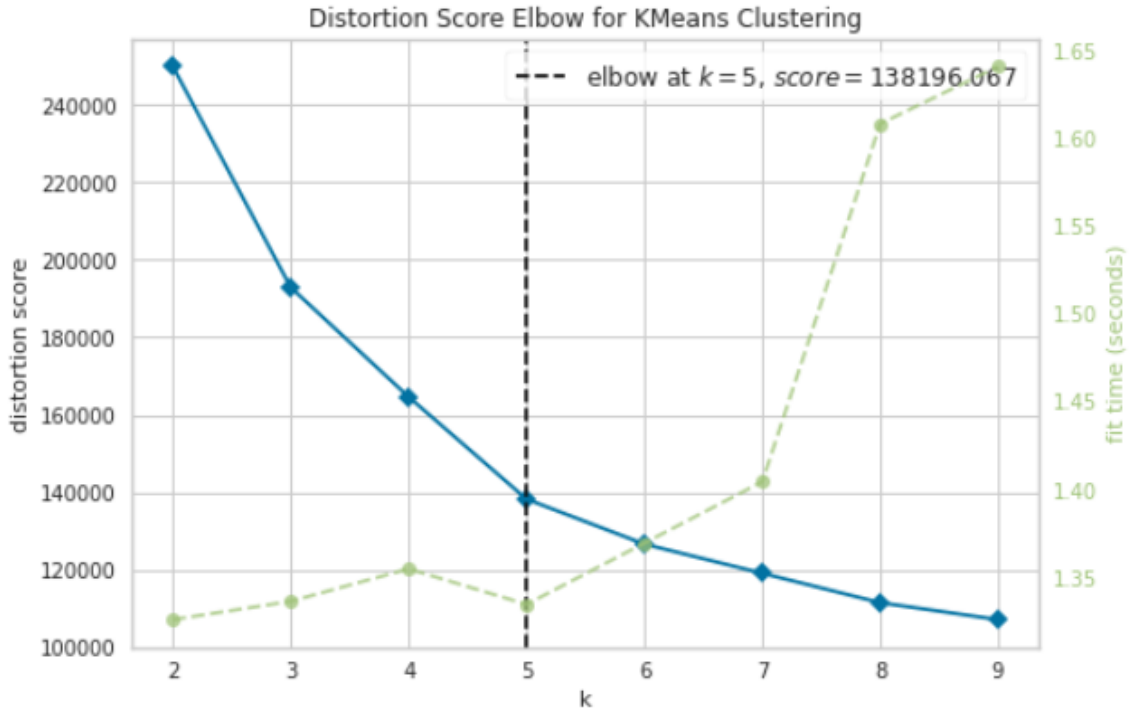


Figure 19 - Finding the best value for K

We are going to analyse the Income attribute with the top six countries. We set the K equals to 5 (K=5) as result our date divided in four categories, and then we apply the cluster method to make one separate scatter for each case. As you can understand the cluster can be clearly seen for the high-Income and low-

Income for our top six countries. What comes out from first scatter (high-Income) is that United States is first with a huge percentage compared to other top six countries, then we can see that India and Japan is almost in the same level for high-Income. For the second scatter what comes out is that for the category ‘rest countries’ is first for the low-Income followed up from India, and the all the other categories are in the same levels.

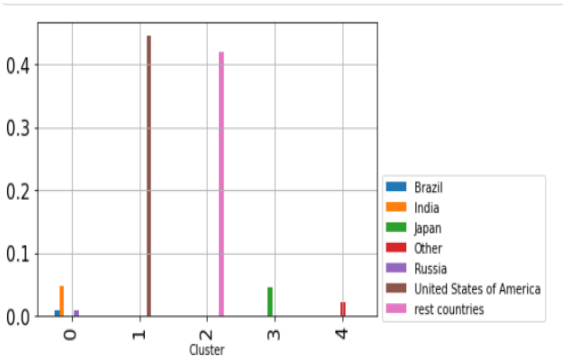


Figure 20 - Compare highIncome with top six countries

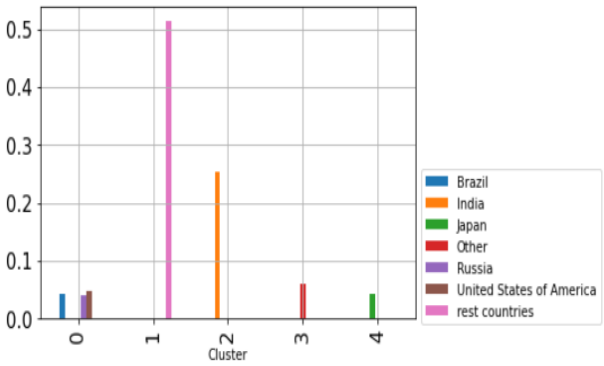


Figure 21 - Compare lowIncome with top six countries

Now we are going to do apply the same scatter method but this time we are going to replace the highIncome and lowIncome attribute and do for high-education and low-education for our top six countries. What we can see from the first scatter(high-education) is that for the top countries the category ‘India’ has the most values and ‘United States’ about the half of them. Also, the remain countries are about the same level. For the second scatter(low-education) is that all the countries are about the same level and ‘India’ and ‘Japan’ slighty above them. For both scatters we can clearly see that the categorie ‘rest countries’ have huge numbers compares the other categories and almost the both of them are the same.

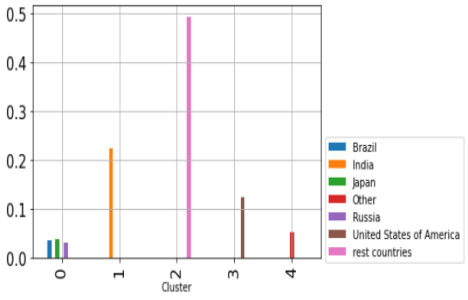


Figure 22 - Compare HighEducation with top six countries

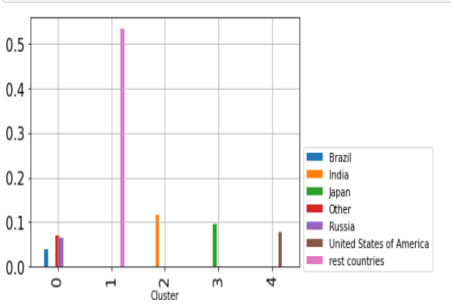


Figure 23 - Compare LowEducation with top six countries

The next step we are doing is now applying the scatter but change the top six countries with our top salaries. So for next two scatter with are going to have high-education and low-education with the top salaries for our developers. Is obviously that from the first scatter ‘$0-999 is most common answer and the other answers in our cluster for high-education are about the same. For the second scatter low-education with top salaries the answer ‘$0-999’ is again first but this time with higher answers.In cluster 3 and cluster 0 we

can see that '$10,000-14,999' and '1,000-1,999' is slighty more than the answers and the reamain answers in clustering 0 are about the same. For both scatter 'rest salaries' is the most common answer.

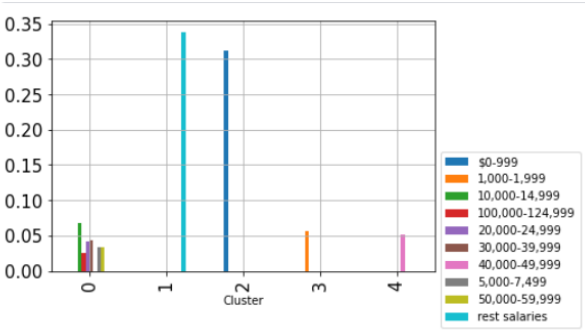


Figure 24 - Compare HighEducation with top nine salaries

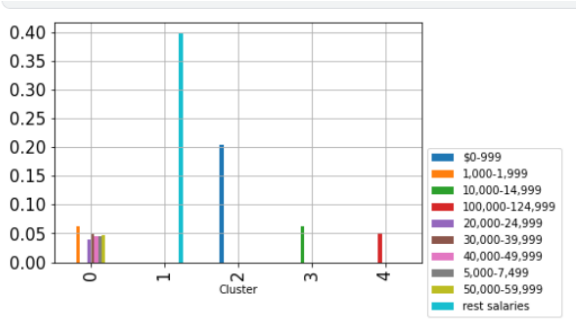


Figure 25 - Compare LowEducation with top nine salaries

Part – 4 Machine Learning Methods:

Workflow of machine learning for classification:

Machine Learning Workflow for classification it is a data flow diagram where data sets are mapped into classes depending on the business model. This diagram describes clearly the necessary phases we need to take in order to build our model and achieve our objective, which are:

- Collection and preparation of data, that include record and attribute selection, transformation and cleaning of data.
- Building datasets by splitting, training, validating and testing data. Machine learning algorithms implemented to it, which will in turn be made into a model.
- Model evaluation by predicting accuracy using the test data. Depends on accuracy of the models we will decide if a fine-tuned is needed or not and the validation data will be used to do so.
- Afterward a model has been constructed with the best accuracy, it may be used in a variety of ensemble models to improve accuracy even further.
- Finally, it will be ready for testing to assure high precision and accuracy before being deployed.

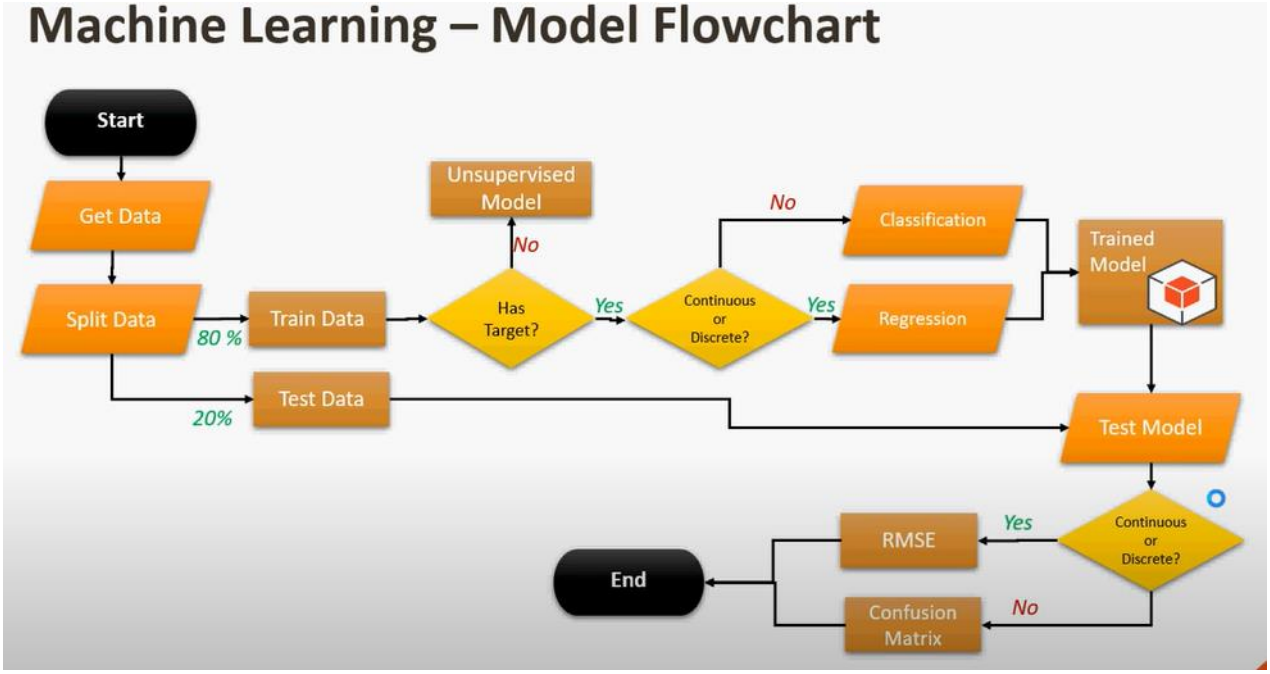


Figure 26 - Machine Learning-Model Flowchart

There are a number of elements to consider when deciding which machine learning model is superior than others. The question under investigation is a classification problem, since the outcome must be classified into one of two groups: high income or low income. As a result, classification models are preferable to regression models for classification problems. There are a variety of reasons why one model is chosen over another, including speed, accuracy, and data size.

Classification:

In machine learning, classification is a two-step process that includes both learning and prediction. The model is built based on the training data in the learning process. The model is used to forecast the response for provided data in the prediction stage.

Logistic Regression Model:

A statistical strategy for predicting binary classes is logistic regression. The result or goal variable is a binary variable. The term dichotomous refers to the fact that there are only two potential classifications. It can, for example, be utilized to solve cancer detection issues. It calculates the likelihood of an event occurring.

It's a kind of linear regression in which the target variable is categorical. The dependent variable is the log of odds. Using a logit function, logistic regression predicts the likelihood of a binary event occurring.

Linear Regression Equation:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta nXn$$

Where, y is dependent variable and x1, x2 ... and Xn are explanatory variables.

Sigmoid Function:

$$p = 1/1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 \ldots\ldots \beta nXn)}$$

Properties of Logistic Regression:

1. The dependent variable in logistic regression follows Bernoulli Distribution.
2. Estimation is done through maximum likelihood.
3. No R Square, Model fitness is calculated through Concordance, KS-Statistics.

Advantages of Logistic Regression Classification Algorithm:

1. Because it is a basic model, training takes relatively little time.
2. It has the ability to manage a huge number of features.

Disadvantages of Logistic Regression Classification Algorithm:

1. Despite the fact that it is called regression, we can only use it to solve classification issues because its range is always between 0 and 1.
2. It can only be used to solve binary classification issues and performs poorly in multi-class classification tasks.

Applications of Logistic Regression Classification Algorithm:

1. Credit Scoring: Predicting a person's creditworthiness (ability to repay a loan) based on data such as yearly income, account balance, and other characteristics.
2. Predicting User Behaviour: Many websites utilize logistic regression to predict user behaviour and lead them to links that they would find useful.
3. Discrete Choice Analysis: Logistic regression is a great way to forecast people's categorical preferences. This might include deciding which automobile to buy, which school or college to attend, and so on, based on people's characteristics and the many alternatives accessible to them.

We are applying the method of Logistic Regression for the features Age, TopCountry, Education, YearsOfCoding, ProgrammingLanguage and ComputingPlatform and with our target the feature HighIncome. We split the data up 30 % and 70% partition for training, validation, and test data, respectively and we are creating our model and prediction. The accuracy of the model that we made it came up 0.82 into decimal points.

Decision Tress:

The Decision Tree algorithm is part of the supervised learning algorithms family. The decision tree approach, unlike other supervised learning algorithms, may also be utilized to solve regression and classification issues.

By learning basic decision rules inferred from past data, the purpose of employing a Decision Tree is to develop a training model that can be used to predict the class or value of the target variable (training data).

We start at the root of the tree when using Decision Trees to forecast a class label for a record. The values of the root attribute and the record's attribute are compared. We follow the branch that corresponds to that value and go to the next node based on the comparison.

Advantages of Decision Tree Classification Algorithm:

1. This technique provides for a straightforward data representation. As a result, it's easier to decipher and explain to executives.
2. Decision trees are designed to imitate how people make decisions in real life.
3. They handle qualitative target variables with ease.
4. They can successfully handle non-linear data.

Disadvantages of Decision Tree Classification Algorithm:

1. They may produce complicated trees that are occasionally irrelevant.
2. When compared to other algorithms, they do not have the same level of prediction accuracy.

Applications of Decision Tree Classification Algorithm:

1. Sentiment Analysis: This is a text mining classification system for determining a customer's sentiment about a product.
2. Product Selection: Companies may use decision trees to figure out which products will generate the most profit when they first introduce them.

We are applying the method of Decision Trees as we did before with Logistic Regression for the features Age, TopCountry, Education, YearsOfCoding, ProgrammingLanguage and ComputingPlatform and with our target the feature HighIncome. Again, we are splitting the data up 30 % and 70% partition for training, validation, and test data, respectively and we are creating our model and prediction. Also, for this method we are building a huge decision, tree so the accuracy number will be as close as possible to 1.0. The accuracy of the model that we made it came up 0.83 with two decimal points.

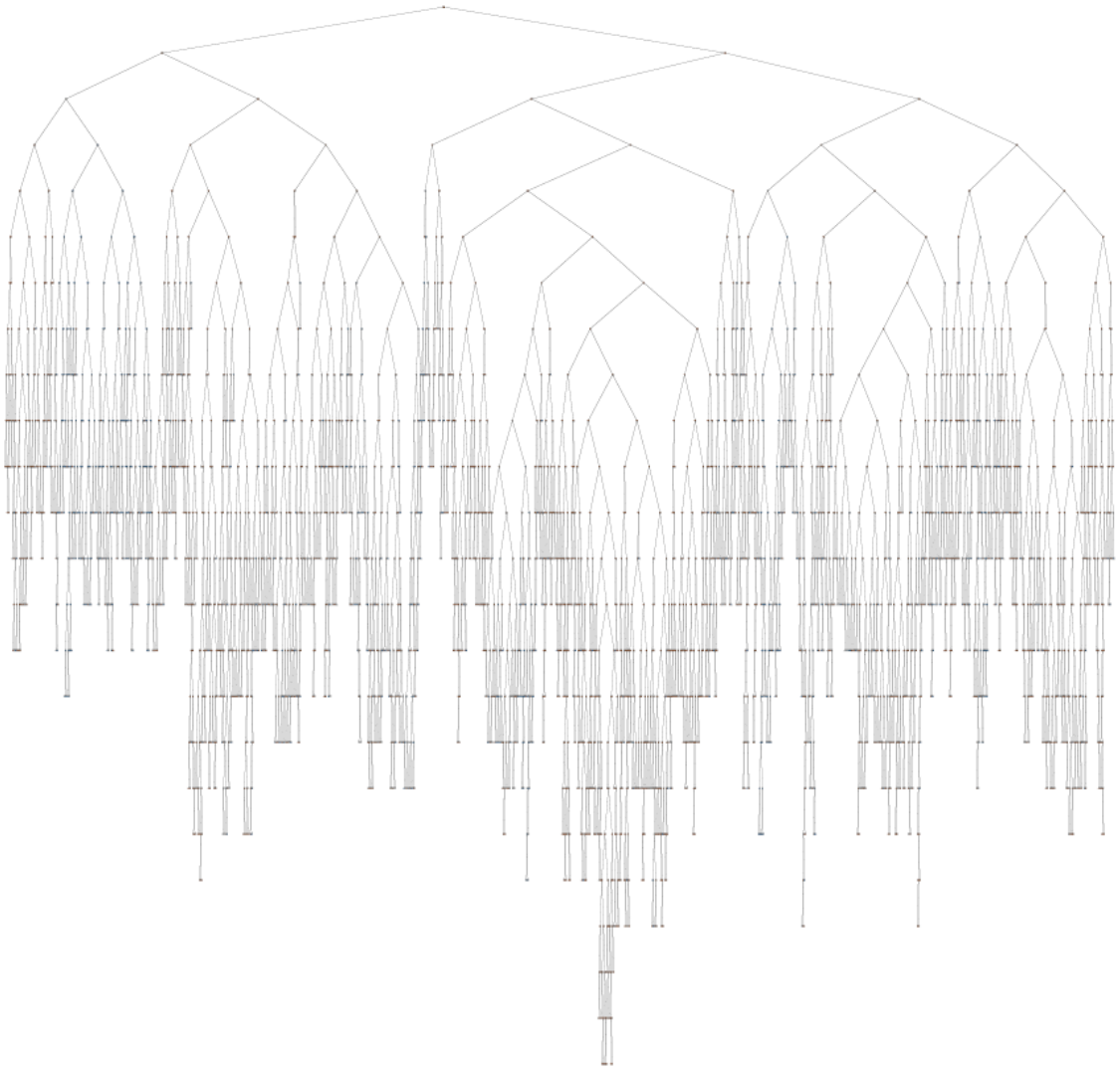


Figure 27 - Decision Tree

k-Nearest Neighbour Method:

The KNN algorithm identifies the K closest neighbours of a given observation point. It then uses the K points to calculate the proportions of each type of target variable and forecasts the target variable with the greatest ratio. Consider the following scenario, in which we must assign a target value to point X. The model will then predict that the point belongs to the pink-coloured class if we take four neighbours around it.

Advantages of K-Nearest Neighbour Classification Algorithm:

1. It may be used on datasets with any distribution.
2. It is simple to comprehend and intuitive.

Disadvantages of K-Nearest Neighbour Classification Algorithm:

1. Outliers have a big impact on it.

2. It favours a class with a larger number of occurrences in the dataset.
3. Finding the best value for K might be difficult at times.
4.

Applications of KNN Classification Algorithm:

1. Detecting Outliers: The algorithm can detect outliers since it is sensitive to outlier instances.
2. Recognizing Related Papers: To recognize documents that are semantically similar.

We are applying the k-Nearest Neighbour Method. Again, we are splitting the data up 30 % and 70% partition for training, validation, and test data, respectively. We are creating our model and prediction for a range of K equals 1 until it gets equal 24. For each time the K changes value we print the accuracy. After we print our accuracy, we also draw a graph so we can see that our results are correct. Is clear that the results match our plot. As the number of neighbours rose, accuracy improved considerably until there were two neighbours, at which time accuracy began to vary. The following were found to be the final hyper parameters for the KNN classifier: weights = “distance”, metric = “euclidean” and number of neighbours = k The highest accuracy achieved is 8.411 with 3 decimal points.

```
1 accuracy 0.7997227997227997
2 accuracy 0.8214368214368214
3 accuracy 0.823053823053823
4 accuracy 0.8313698313698313
5 accuracy 0.8304458304458304
6 accuracy 0.8318318318318318
7 accuracy 0.832986832986833
8 accuracy 0.8352968352968353
9 accuracy 0.8366828366828367
10 accuracy 0.8371448371448371
11 accuracy 0.8373758373758374
12 accuracy 0.8396858396858397
13 accuracy 0.8394548394548395
14 accuracy 0.8396858396858397
15 accuracy 0.8401478401478402
16 accuracy 0.8385308385308385
17 accuracy 0.8382998382998383
18 accuracy 0.83991683991684
19 accuracy 0.8410718410718411
20 accuracy 0.8413028413028413
21 accuracy 0.8403788403788404
22 accuracy 0.841995841995842
23 accuracy 0.8408408408408409
24 accuracy 0.8406098406098406
```

Figure 28 - Accuracy for KNN

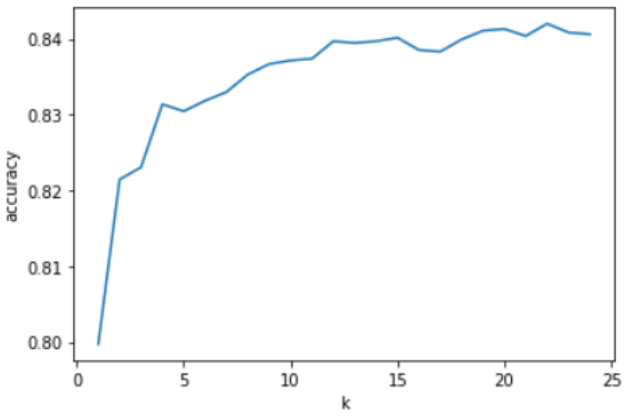


Figure 29 - Graph for KNN

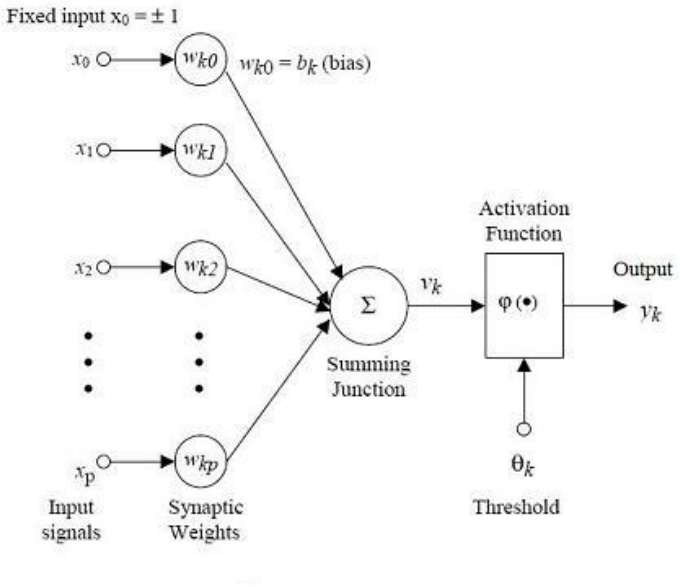
Artificial Neural networks:

Deep learning is a subset of ML that employs multilayer neural networks that automatically alter and adapt to new training data.
Through a learning process, artificial neural networks gain information. The ANNs improve their modelling performance with time.
The obtained knowledge is stored in the interconnections in the form of weights. These weights keep on changing as the network is trained and updated weights.

In artificial neural network model, all neuron layers must be interconnected and must be a process for updating the weights though learning from the model.

Always must be an activation function which essentially determines the output from neuron's weighted inputs. Perceptron are networks which consists of just one unit.

In mathematical terms process can be decided as below.



The output of yk neuron, will be the effect of some activation function on the value of vk.

$$v_k = \sum_{j=1}^{p} w_{kj} x_j$$

Multi-Layer Perceptron (MLP):

Multilayer perceptron is a network of multiple layers of neurons connect in feed-forward manner. Backpropagation algorithm considered to be most important for designing MLP. Each neuron in a layer is connected with every other neuron in the subsequent layer. An MLP has an input layer and an output layer with one or more hidden layers in between. The neurons in the hidden layers are not directly accessible, hence they are called hidden.

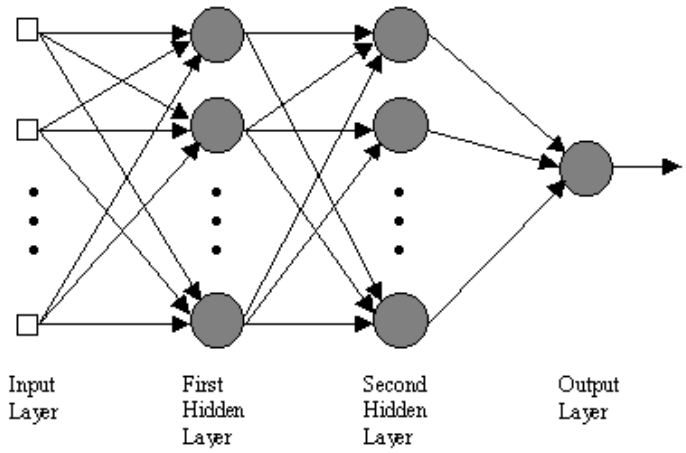


Image Reference: Feedforward Neural Networks: An Introduction, by Wiley.

Information passes in at the inputs and goes over the system, layer by layer until it reaches the output layer. Among typical operation, that is the point at which it goes about as a classifier, there is no feedback between layers. This is the reason they are called feedforward neural networks. Sigmoid function is one commonly used activation function in this case.

Sigmoid function

$$f(x) = \frac{1}{1 + e^{-x}}$$

There are 3-Types of Learning use in ANN.

1. Reinforcement learning algorithm uses a reinforcement signal instead of an output error of that neuron in order to tune the weights.
2. Supervised learning algorithm uses a supervisor that knows the desired outcomes and tunes the weights (W) consequently.
3. Unsupervised learning algorithm uses local data, instead of supervisor, according to emergent collective properties.

Advantages:

- A neural network can accomplish tasks that a linear program cannot.
- In case of an element failure the neural network can continue because of parallel nature.
- A neural network learns and does not need to be reprogrammed.
- It can be applied in any application.

Disadvantages:

- The neural network needs training in order to operate.

- The architecture of a neural network is different from the architecture of CPU'S so it is a necessity to be emulated.
- For large neural networks it is a need for use multiple processor machines.

Once again, we are applying the method this time for MLP for the features Age, TopCountry, Education, YearsOfCoding, ProgrammingLanguage and ComputingPlatform and with our target the feature HighIncome. We split the data up 30 % and 70% partition for training, validation, and test data, respectively and we are creating our model and prediction. The accuracy of the model that we made it came up 0.86 into two decimal points.

Ensemble methods:

Ensembles are predictive models that aggregate predictions from two or more different models to create a single forecast.
When the best performance on a predictive modelling assignment is the most essential outcome, ensemble learning approaches are popular and the go-to strategy.
However, they are not always the ideal methodology to apply, and many newcomers to the area of applied machine learning believe that ensembles or a certain ensemble method is always the best option.
On a predictive modelling project, ensembles provide two distinct benefits, and it's critical to understand what these benefits are and how to assess them to verify that using an ensemble is the best selection for your project

Lastly, we are going to apply the Ensemble method for our for-classification methods (Logic regression, Decision trees, k-Nearest Neighbour and MLP). In this stage we put all of them together and get the overall accuracy for them. Also, we are applying the voting method, so each method is voting for the range of number one to ten, we print out the results for each one and do again the accuracy of the results. For the Ensemble method we got for accuracy the number 0.86 in two d.p and for the voting method we got the number 0.85 in two d.p.

Voting\Stacking Classification:

When it comes to building a stacking/voting classifier, Scikit-Learn has several useful functions for us to employ. As parameters, the VotingClassifier accepts a list of distinct estimators as well as a voting method. The projected labels and a majority rules system are used in the hard voting technique, but the soft voting approach predicts a label based on the argmax/largest predicted value of the sum of the predicted probabilities.

Bootstrap Aggregation (Bagging):

Bootstrap Aggregation (also known as Bagging) is a basic yet effective ensemble approach.
An ensemble approach is a strategy for making more accurate predictions by combining forecasts from various machine learning algorithms.
Bootstrap Aggregation is a generic process that may be used to minimise variation in algorithms with a lot of it. Decision trees, such as classification and regression trees, are a high-variance approach (CART).
Decision trees are very dependent on the data used to train them. If the training data is modified (for example, a tree is trained on a part of the training data), the resultant decision tree and, as a result, the predictions may differ significantly. The Bootstrap approach is applied to a high-variance machine learning system, such as decision trees, in the process of bagging.

AdaBoost:

The AdaBoost algorithm, short for Adaptive Boosting, is a Boosting approach used in Machine Learning as an Ensemble Method. The weights are re-allocated to each instance, with larger weights applied to improperly identified instances. This is termed Adaptive Boosting. In supervised learning, boost is used to decrease bias and variation. It is based on the notion of successive learning. Each succeeding student, with the exception of the first, is produced from previously grown learners. In other words, weak students are transformed into strong students. With a little modification, the AdaBoost method operates on the same idea as boosting. Let's take a closer look at this distinction.

How AdaBoost works:

During the data training phase, it creates a certain number of decision trees. The improperly categorised record in the first model is given precedence when the first decision tree/model is constructed. Only these records are supplied to the second model as input. The procedure continues until we have decided on a number of base learners to develop. Remember that all boosting strategies allow for record repetition.

Random Forests Classification Algorithm:

A forest is made up of many different types of trees. A random forest, on the other hand, entails the processing of a large number of decision trees. Each tree predicts a probability value for the target variables. The probabilities are then averaged to obtain the final outcome.

We evaluate each tree as follows:

- The dataset's first samples are formed by picking data points and replacing them.
- To generate decision trees, we don't use all of the input variables. We simply use a small portion of the ones that are available.
- No trimming is done, and each tree is allowed to grow to the maximum length feasible.

Advantages of Random Forest Classification Algorithm:

1. It is effective for dealing with huge datasets.
2. It enables for the estimation of input variable relevance in categorization.
3. It outperforms decision trees in terms of accuracy.

Disadvantages of Random Forest Classification Algorithm:

1. It is more difficult to execute and, as a result, takes longer to analyse.
2.

Applications of Random Forest Classification Algorithm:

1. Credit Card Default: Credit card firms employ random forests to determine whether or not a cardholder will default on their loan.
2. Stock Market Prediction: It is used by stock investors to forecast a stock's developments and assess loss and profit.
3. Product Recommendation: It may be used to provide product recommendations based on a user's preferences.

Part 5 - Evaluation of Machine Learning Models:

A classification model is evaluated using three key metrics: accuracy, precision, and recall.

The percentage of correct predictions for the test data is identified as accuracy and it is calculate using the below equation.

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{all predictions}}$$

Precision is defined as the percentage of relevant instances (true positives) among all the examples expected to belong to a specific class.

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

The proportion of instances predicted to belong to a class compared to all of the examples that genuinely belong in the class is known as recall.

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Classification metrics:

There are four different sorts of outcomes that might occur though making categorization predictions.

1. True positives occur when you anticipate that an observation belongs to a particular class.
2. True negatives occur when you forecast that an observation does not belong to a class.
3. False positives happen when you assume an observation belongs to a class when it doesn't.
4. False negatives happen when you incorrectly forecast that an observation does not belong to a class.

On confusion matrix these four results are normally plotted. The confusion matrix below demonstrates the scenario of binary categorization. After making predictions on your test data and designating each forecast as one of the four probable outcomes indicated above, you'd create this matrix.

Confusion Matrices:

Decision Tree Classifier:

When the adjusted decision tree classifier is placed in a confusion matrix, as shown below, it is obvious that the decision tree model accurately predicted the majority of the time. The 'main diagonal' tiles with greater values demonstrate this. There were also more false-positives false-negatives than false-negatives in the model. This demonstrates that the model prefers to anticipate the condition as negative twice as often as it prefers to predict it as positive.

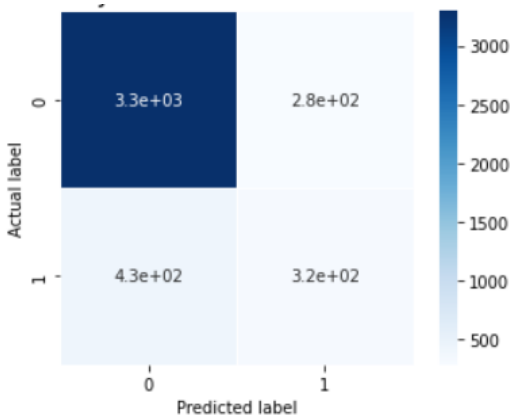


Figure 30 - Confusion Matrix for Decision Tree

Logistic Regression Classifier:

Putting the logistic regression model into a confusion matrix as shown below we can understand very clear that the true-negative (3547) compare with the true-positive (1) has a huge difference. Also, the same scenario is with false-negative (781) compare with the false Positive (0).

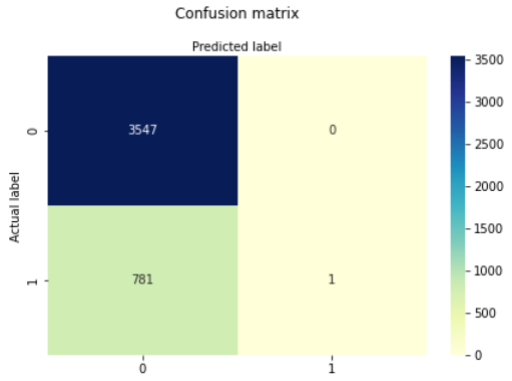


Figure 31 - Confusion Matrix for Logic Regression

k-Nearest Neighbour Classifier:

Below we can see the confusion matrix for the k-Nearest Neighbour Classifier and is clear that the true-negative are more than the true-positive 3.4+03 compare with the 2.5e+02. For the other two labels false-negative and false-positive the difference is huge as you can see the first one is 5.3e+02 and the second one is 1.6+02

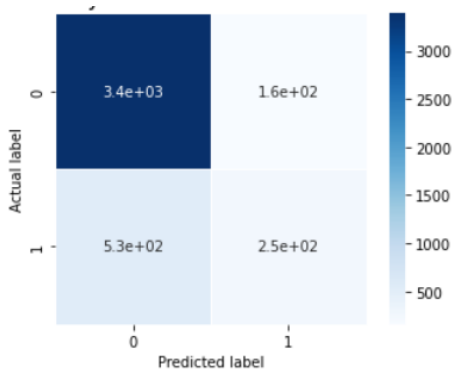


Figure 32 - Confusion Matrix for KNN

MLP Classifier:

Once again, we make a confusion matrix this time for MLP. As we can see the true-negative number is 3452 and the true-positive number is 307 so we can under a huge difference between two numbers. For false negative we got the number 460 and for false-positive 130 which are close compared to others.

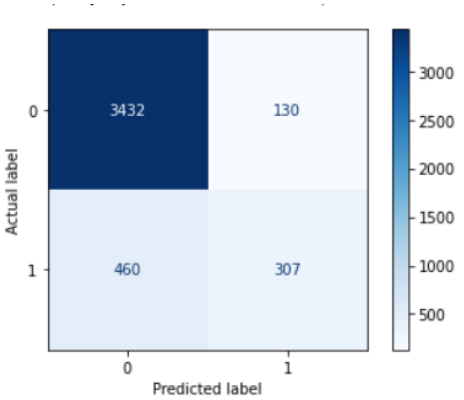


Figure 33 - Confusion Matrix for MLP

Ensemble method classifier:

Now we are applying the matrix model for all our four classifications meths (Logistic Regression, Decision Trees, KNN and MLP). As we can see from the confusion matrix the true-negative is more than the true-positive and if you have a look to the other classifiers this one it was also a common thing between them. Moreover, once again the false-negative are more than the false-positive which is also a common for all our matrices.

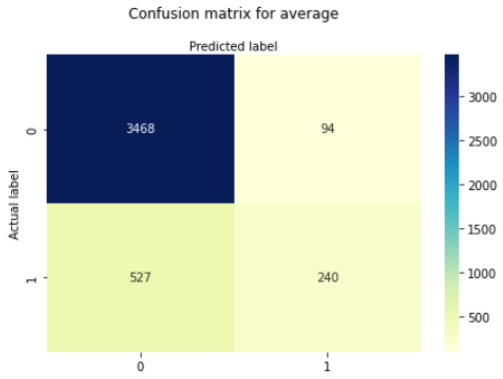


Figure 34 - Confusion Matrix for Ensemble Method

Prediction method classifier:

One more model that we are using is the voting for all our top classifications methods. Our classifications are voting, and we are taking the accuracy for each vote, and we put them all together. In the end we are building our confusion matrix. As it understandable the true-negative (3502) is more than the true-positive (199) and the false-negative (568) are more than the false-positive (60).

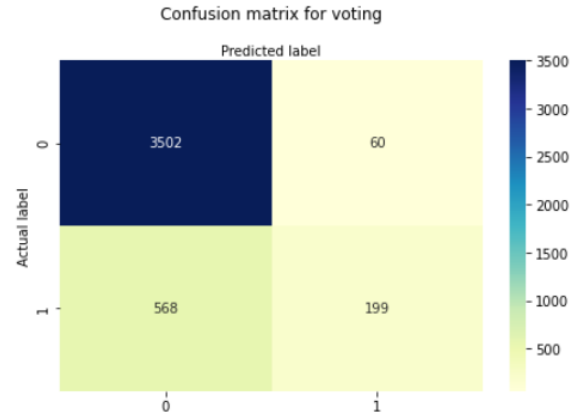


Figure 35 - Confusion Matrix for Prediction Method

Accuracy:

The figure below shows the average accuracy of each model, the average precision of each model and the average recall of each model after applying the ensemble method.

```
Accuracy for model 1: 0.8228228228228228
Precision for model 1: 0.0
Recall for model 1: 0.0

Accuracy for model 2: 0.8316008316008316
Precision for model 2: 0.5345454545454545
Recall for model 2: 0.3833116036505867

Accuracy for model 3: 0.8318318318318318
Precision for model 3: 0.5375722543352601
Recall for model 3: 0.363754889178618

Accuracy for model 4: 0.8657888657888658
Precision for model 4: 0.7336683417085427
Recall for model 4: 0.38070404172099087

Accuracy for average model: 0.8560868560868561
Precision for average model: 0.7155688622754491
Recall for average model: 0.3116036505867014

Accuracy for voting model: 0.8540078540078541
Precision for voting model: 0.7454545454545455
Recall for voting model: 0.26727509778357234
```

Figure 36 - Accuracy Results

Part 6 - Discussions and Conclusions:

Summarised Results:

To sum up all the work that was done during this project it can be determined that the first conclusion is that MLP classifier performed the best in terms of accuracy and precision after examining several performance indicators. The logistics regression was the model that performed the worst with the lowest accuracy compared to all the other methods used. This leave Decision trees and KNN in a middle position with fair to good accuracy numbers. At the end we applied the ensemble method for all four methods the number of their accuracy improve a bit, so we could get better results.

This coursework gave me the opportunity to experience new methods and technics in ML and I was able to analyse numerous elements of the data set and apply different models to forecast which salary bracket a developer is in based on. Also, I discovered significant patterns in the features of the developers based on their salary category and determined which models were the most accurate and precise.

Gained knowledge:

I'm much more comfortable with the CRISP-DM approach now than I was before this coursework. I've got a better grasp of how to handle data and conduct exploratory data analysis on it. I was also able to put my model-building skills to use by employing hyperparameter tuning and ensemble approaches. Before the coursework, I would have been hesitant to take on a categorization challenge and solve it but now that I have a better understanding of the process. I am confided that I would be able to take part on a more demanding project in the near future. I had to do a lot of background study on different areas of machine learning during this project, such as why certain models are better at tasks than others. As a result, I am able to recognise technical terminology that I would not have recognised otherwise as also to get a better idea about ML and the future of AI.

My next goal is to increase my capacity to apply my knowledge to other machine learning challenges. This might be accomplished by completing a variety of activities that need diverse approaches and processes. In terms of my talents and how to properly utilise them, I still have a lot to learn, and I believe I could do a

better job if I were to tackle this project again. Finally, I believe that this coursework has greatly increased my grasp of machine learning for data analytics, and that I still have a lot to learn in the future. It is obvious to me now that the future of technology belongs to machine learning and artificial intelligent.

References:

1. Linkedin.com. 2022. *CRISP-DM Methodology*. [online] Available at: <https://www.linkedin.com/pulse/crisp-dm-methodology-mani-ghaedi> [Accessed 22 April 2022].

2. Scikit-yb.org. 2022. *Elbow Method — Yellowbrick v1.4 documentation*. [online] Available at: <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html#:~:text=The%20elbow%20method%20runs%20k,point%20to%20its%20assigned%20center> [Accessed 22 April 2022].

3. Tibshirani, R., 2013. Clustering 3: Hierarchical clustering (continued); choosing the number of clusters. [online] pp.36-462/36-662. Available at: <https://www.stat.cmu.edu/~ryantibs/datamining/lectures/06-clus3.pdf> [Accessed 22 April 2022].

4. Navlani, A., 2019. Learn about Logistic Regression, its basic properties, and build a machine learning model on a real-world application in Python. *Understanding Logistic Regression in Python Tutorial*, [online] Available at: <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python> [Accessed 22 April 2022].

5. Chauhan, N., 2022. *Decision Tree Algorithm, Explained - KDnuggets*. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html#> [Accessed 22 April 2022].

6. Brownlee, J., 2022. *Why Use Ensemble Learning?*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/why-use-ensemble-learning/> [Accessed 22 April 2022].

7. *The Ultimate Guide to AdaBoost Algorithm | What is AdaBoost Algorithm?*, 2022. [online] Available at: <https://www.mygreatlearning.com/blog/adaboost-algorithm/#:~:text=AdaBoost%20algorithm%2C%20short%20for%20Adaptive,assigned%20to%20incorrectly%20classified%20instances> [Accessed 22 April 2022].

8. Brownlee, J., 2022. *Bagging and Random Forest Ensemble Algorithms for Machine Learning*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/> [Accessed 22 April 2022].

9. Nelson, D., 2022. *What are Ensemble Models in Machine Learning?*, [online] Available at: <https://stackabuse.com/ensemble-voting-classification-in-python-with-scikit-learn/> [Accessed 22 April 2022].

10. DeZyre. 2022. *Neural Network Tutorial*. [online] Available at: <https://www.projectpro.io/data-science-in-python-tutorial/neural-network-tutorial> [Accessed 22 April 2022].

11. ProjectPro. 2022. *Machine Learning and Data Science Code Examples | ProjectPro*. [online] Available at: <https://www.projectpro.io/recipes?utm_source=Blg435&utm_medium=RcpLink&utm_campaign=TXTCTA2> [Accessed 22 April 2022].

12. DeZyre. 2022. *K-Means Clustering Tutorial*. [online] Available at: <https://www.projectpro.io/data-science-in-r-programming-tutorial/k-means-clustering-techniques-tutorial> [Accessed 22 April 2022].

13. GeeksforGeeks. 2022. *Ensemble Methods in Python - GeeksforGeeks*. [online] Available at: <https://www.geeksforgeeks.org/ensemble-methods-in-python/?fbclid=IwAR0veK0_jA2BNLYmNYMcyXAk0nO9dOSZS5fIlixFf5oDnHmBQ2PGJC76bl4#:~:text=Ensemble%20means%20a%20group%20of,robustness%2Fgeneralizability%20of%20the%20model> [Accessed 22 April 2022].

14. Jeremy Jordan. 2022. *Evaluating a machine learning model.*. [online] Available at: <https://www.jeremyjordan.me/evaluating-a-machine-learning-model/> [Accessed 22 April 2022].