# **ENGINEERING GRADUATES EMPLOYMENT OUTCOMES**

**FINAL REPORT**

**Submitted By:**

**Group No. 1 | PGP-DSE MAR 2020 | GURGAON**

**Group Members:**

* Harish Gola (94PW3Q8QYI)
* Ketan Sharan (ILQ09BTYRX)
* Sanchit Mishra (RZNYC8URXN)
* Sandeep Kumar (69IVXINW9S)
* Shivoy Thakral (682O9T1VA8)

**Research Supervisor**

Dr. Dipanjan Goswami

**Great Lakes Institute of Management**



**CONTENTS**

1. Abstract 1
2. Scope-industry Review 1
3. Literature Survey 2
4. Introduction 4
5. Problem Statement 6
6. Dataset & Domain 6
7. Exploratory Data Analysis 9
8. Feature Engineering 15
9. Model Building 17
10. Business Recommendations 26
11. References and Bibliography 29
12. Appendix 30
13. **ABSTRACT**

In excess of 10,00,000 specialists enter the worldwide labour force each year. An applicable inquiry is the thing that decides the positions and pay rates these specialists are offered just after graduation. Past investigations have indicated the impact of different factors, for example, school notoriety, reviews, the field one spends significant time in and economic situations for explicit ventures.

Human capital is a company’s most important asset. It is also typically the largest operating expense, and thus can have a major impact on profitability and overall success of a company. To keep employees engaged and productive, companies must compensate them appropriately, without overpaying them. The cost of overpaying or underpaying can make a significant impact on a company’s business. Accurate recruitment of employees is a key element in the business strategy of every company due to its impact on companies’ productivity and competitiveness.

In this project we depict the subtleties of the dataset and talk about a range of inquiries around meritocracy in labour markets, predispositions in labour determination and other common market influences it can help reveal and answer.

1. **SCOPE-INDUSTRY REVIEW**

The Central Statistics Organization and International Monetary Fund has reported India as one of the world’s fastest growing major economy. Analyzing its complex set-up and making an assessment of its 8 main sectors excluding agriculture, the Wage-Index Report 2018 contributes to better understanding the interactions between the different structural issues and helps to identify strengths and pitfalls of the Indian labour market. Findings can be used to identify and tackle existing problems. Building on the existing research in the field, special attention is also given to issues such as gender, education or tenure groups. Some of the findings are given below-

The total median wages of all sectors in 2018 improved in absolute terms of INR 46.2 from 2017 (INR 265.6 in 2018 vs. INR 219.4 in 2017). Despite the recovery in the last year, wages in 2018 are still lower in all sectors when compared to 2016.

In 2018, bachelor graduates earned INR 162.54 on average and received 82.6% more than workers with secondary education (INR 89.01). Workers with a master’s degree earned 103.7% more than holders of bachelor degrees and 272% more than workers with secondary education. Compared to 2017, the wages are higher, however, showing the importance of higher education, the difference in wages between a worker with a master’s and bachelor’s degree is constantly growing: from 47% in 2016 it rose to 50.9% in 2017 and to 103.7% in 2018.

India’s overall gender pay gap was 22.5% in 2018. In 2016 men earned 24.9% (INR 86) more than women. The trend slowed down in 2017 to 20.0% (INR 46.2) but it rose again to 22,5% in 2018 (INR 64.95). Further analysis, reflecting the distribution of the sample, shows that gender pay gap increases with more years of tenure and higher education. Moreover, it should be noted that women are underrepresented in the Indian working force (approximately 1/4th) and they are underrepresented also in our sample where men represent more than

80%. [Resource: <https://my.monsterindia.com/salary-check.html>]

How company size is going to affect the wages (wages per hour) of employees, we can see that from below observations-

Lowest wages are paid in companies with less than 10 employees (INR 116) while biggest wages are paid in companies from 1000 to 5000+ employees (INR 461.9), meaning that the latter are paid around 4 times the former. Wages grow progressively in every companies’ size group drawing an ascending curve. Unlike other sectors in which the wages’ peak is reached in the companies with more than 5000 employees, here the peak is reached already in companies from 1000 to 5000 (INR 461.9) which is paid exactly the same hourly wage of the biggest companies with 5000+ employees.

According to the Manpower Employment Outlook Survey, the prime sectors that would propel job growth in the country are services, wholesale and retail, finance, education and manufacturing. Hiring in IT would be focused on niche skills such as artificial intelligence, robotics, internet of things, augmented and virtual reality, block chain, machine learning, data science/big data, python, cloud and so on. In the coming days, technological advancement and disruptive technology is set to impact employment structure in sectors other than IT as well including the manufacturing sector.

It is evident that employment structure and the skill requirements in the present-day scenario are rapidly changing throughout the globe, India being no exception. The Future of Jobs report by the World Economic Forum estimated that 65 percent of children entering primary school today will ultimately end up working in completely new job types that do not yet exist. 5 million jobs will be lost to automation by 2020 and that the number will keep growing in the following years. The situation is of particular concern for a country like India with its ever-increasing workforce. In view of the changing needs and demand, bridging skill gaps by targeted re-skilling and upskilling of the workforce has become imperative. Besides, revolutionizing the school curriculum itself, in order to make children future-ready, is the need of the hour.

In this prelude, the analytics sector has been conferred with immense data to understand the behavior of salary and job titles of employees on the basis of their skill, education and that will further strengthen their prediction models. Websites like Glassdoor, naukri.com, Linkedin, monster-india take help of these analyses and show the employment outcomes. This project is an attempt to analyze such data and predict the class of a salary.

1. **LITERATURE SURVEY**

There are literature studies which aim to understand different basic machine learning classification models and predict the class or range of salary.

In a case study by Ignacio Martín (Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study), analyses 4,000 job offers from a Spanish IT recruitment portal. In descriptive analytics it was found that more than 80% offers were from the Spain and Barcelona. The most demanded position was Programmer, followed by Administrator and Technical Support. Programmer positions are required in more than half of the available posts, thus suggesting a clear demand of highly technical profiles. The mean gross annual salary offered in this dataset was 27,340 € with a median of 27,000 € and the maximum observed salary at 76,000 €. K means clustering technique was implemented on job titles and skills. There were 9 clusters in job titles and it was found out that the amount of permanent positions increases significantly from early career jobs (low wage, less than two years of experience in average requirement) to established professionals (more than 3 years of experience, higher salaries), suggesting that movement is promoted by companies at early career stages, since experience appears to shape both salary and permanent contract probabilities. There were 5 clusters formed for skills and it was founded that IT and technology are the most demanded skills by recruiters. Some of the features like experience and job roles were contributing significantly in salary predictions. Several classification models were compared and random forest was performing better with accuracy of 84%.

In a study by Navyashree M, (Department of Computer Science, EWIT, Bangalore, India) 10 clusters were formed on the basis of salary. The study includes an investigation on which fields in a job post have greater influence on salary and how they are interrelated, The discovery of the main data-driven profiles obtained through skill-set based aggregation, The formulation of the salary prediction problem as a classification task in order to have a better accuracy by focusing on discrete ranges instead of continuous salary values, The comparison of several classifiers, including SVM, Decision tree, random forests, in order to find the model with the best accuracy in predicting the salary range. The performance of random forest was better as compare to other models with accuracy of 97% on test data.

In another study by Hamermesh and Donald (2008), the authors determined that more than half of the variation in income is explained by factors such as ability, high school performance, parent’s economic status and student’s demographic characteristic. The salary difference also significantly differs for different study majors.

Rumberger and Thomas (1993) stated that the three different types of qualitative factors labelled as individual and institutional factors namely college major, school quality and educational performance, have an impact on starting salaries. Furthermore, the impact of school quality and educational performance is not uniform across different majors.

Jones and Jackson (1990) investigated the role of college GPA on the salary after the five years of graduation into the job and observed that there is an 8.9% increase in salary per unit change in GPA. However, these findings are within the limited scope of experimental design scenario.

In another study conducted by Godofsky et al. (2011), authors examined the impact of internships and industrial training on the transition between the education and employment. The study showed that there was a significant difference between the starting salaries of the graduates who took up internships or industrial training against the students who did not participate in any internships during the course of their graduate program. Additionally, the graduates with previous work experience related to their choice of college major have a higher starting salary compared to the ones who don’t.

In a study by Rajveer Singh (Technological University Dublin), The project is aimed at understanding the primary salary determinants of entry level engineering graduates in Indian Labour Markets. The primary factors under examination against the salary were: academic features, cognitive skills, standardized test scores, and personality traits. In addition, another objective of the study was to select a best performing salary prediction model. It was found out that they used supervised linear regression to predict the salary because of which root mean square error was too high and R2 of the model was 0.26, which is very less.

1. **INTRODUCTION**

The new economic policies introduced in the early 1990s enhanced foreign investment portfolios into the domestic Indian Market. Since then, the increasing globalization has integrated Indian labour markets to global markets. Indian labour markets have seen a tremendous growth in the last two decades in the private sector. India, with one of the fastest growing economies in the world, also has one of the largest Information Technology (IT) industries in the world, hence generating a huge demand for skilled labour.

Every year a massive number of engineers are entering into Indian Markets. In recent times, the rapidly emerging economy and increasing technology sector have a significant impact on demand and supply for specific skills, practices, and employability of engineers. Even though this demand is diversified across industrial sectors, a large number of these engineers are employed within the IT sector. In recent years the supply of engineers has surpassed the demand in the Indian IT sectors. Recent survey studies have indicated that this quantity surge has also degraded the quality and employability of engineering graduates. Salary is considered one of the major reasons for pursuing engineering as college studies. The aim of this study is to determine the various factors that determine the salaries of engineering graduates in Indian Labour Markets.

**Undergraduate Students** make numerous decisions during their academic years which influence the course of their career. There are a number of factors which are in play such as choice of university, study majors, internships etc. and which may have a significant impact on the salaries and career options. This research study will focus on determining the impact of academic performance, cognitive skills, personality traits, standardized test scores and demographics on the starting salaries of undergraduate students specific to Indian Labour Markets.

**For Engineering Studies -** the student generally joins an engineering college through an entrance exam or merit-based criteria depending on the type of college, which then in turn has an affiliation to one of the Central, State or Deemed Universities. The curriculum and structure of study again differ depending on the universities. These diversities in education ecosystem make it difficult for the employer to evaluate students based on standard merit. So, in order to standardize the evaluation criteria more than 3500 organizations refers to AMCAT (Aspiring Minds) scores – A standardized test students take after the undergraduate course. The dataset used for this study is released by Aspiring Minds, the organization which facilitates this test.

**Research Project:** The main purpose of the research study is to examine the academic factors, cognitive skills and personality factors, which best predicts the salary of a recently graduated engineer in Indian Markets. The sample of study focuses on engineering graduates in India. Within this study, independent variables will include: Personal Information, Pre-University Information, Standardized Test Scores and Demographics Information of candidates. The dependent variables in the dataset are Starting Salary, Job location, and Job Title. Regression analysis will be performed to study the relationship between these variables. The study also performs a comparative analysis of various salary prediction models based on prediction accuracy to find an optimal salary predictor.

**The study aims to answer the following** **Research Question:** What are the primary factors, in determining the starting salary of a recently graduated engineer in Indian Labour Markets? The results of the study would allow an engineering graduate to best navigate through various choices to achieve higher salaries. The results will also help the leaders in the education system to develop programs and resources to align with the requirements of higher wages into the Indian Labour Markets. The outcomes of the study can significantly inform the students and education administrators in terms of choices and focus on achieving a higher return for both parties.

**Research Objectives and Hypotheses:** In order to answer the research question, a quantitative study will be conducted using the AMEO dataset. The research objectives of the research study are:

* To explore the existing knowledge base by measuring or evaluating the employability and salary dynamics of undergraduate and graduate students.
* To understand the impact of different factors such as cognitive skills, academic choices, academic performance, demographics, and personality traits on the salary of fresh engineering graduates.
* To examine which of the cognitive skills is contributing the most to Salary.
* To build and select the best classification model evaluated on the basis of Accuracy and ROC\_AUC as the performance measuring evaluation, in classifying the salary of recent engineering graduates.

**Research Methodology:** An exploratory research method is used to address the research question using secondary data. A quantitative research approach is used to conduct an investigation into the data to understand the quantitative properties and underlying relationships within the data. The research objectives defined earlier will be achieved using the course of action outlined below:

* An extensive literature review is conducted, which is used to summarize the existing research studies in the context to the research question. Additionally, the literature review will allow to objectively shape the course of the research project.
* The relationships between the various factors on determining the starting salary are explored using various visualizations and feature engineering combined with data preprocessing.
* Statistical tests are used for hypothesis testing
* Classification modelling with various techniques is used to build an accurate class predictor and the best performing model is selected using model accuracy based on Accuracy and ROC\_AUC scores.
* Unsupervised Learning is applied on the dataset for exploration purposes to check classification model accuracies built on labels generated by unsupervised models.

**Scope and Limitations:** The scope of the research study is targeted on recently graduated engineers within Indian Labour Markets. There is a significant amount of research literature available that is focused on examining the various factors that predict job seeking behavior and re-employment of experienced professionals. This study will not focus on experienced professionals but rather on newly graduated engineers. The study is concentrated around the first job placement of engineering students after graduation and not the successive job offers.

1. **PROBLEM STATEMENT**

More than a million engineers enter the global workforce every year. A relevant question is what determines the jobs and salaries these engineers are offered right after graduation. Various factors such as college grades, candidate skills, proximity of the college to industrial hubs, the specialization one has, market conditions for specific industries determine this. Based on this information, our group has derived two problem statements, which we aim to address in this project:

1.The factors that influence the salary and job titles of engineering graduates in the labour market.

2.Machine learning models to predict the class of salary of engineering graduates based on the relevant factors.

1. **DATASET AND DOMAIN**
   1. **DATA DICTIONARY:**

The initial dataset obtained for analysis has been obtained (Refer appendix 12.1) and the classification of variables of dataset has been specified in (Refer appendix 12.2)

* 1. **MISSING VALUES TREATMENT:**

The missing values in the dataset obtained is filled with -1 instead, so in order to check for missing values, -1 or any insignificant value mentioned in the entry is replaced with np.nan after which the results were obtained (Refer appendix 12.3)

* 1. **MISSING VALUES IMPUTATION**:

**Domain**: We chose mean and median imputation for this, comparing both the results it is observed that mean imputation has slightly better response in reducing the skewness of this feature.

**Technical Optional Subjects**: For the optional subjects it can be inferred that it’s up to the candidate whether or not he is choosing that subject or not. So, for the subject that was not chosen by the candidate value 0 has been imputed.

* 1. **REDUNDANT FEATURES:**

In the following dataset it can be inferred that there are some features which have no significant application or usage in our model dataset, such features are removed. Features removed are:

* **ID** - As it has too unique values hence this categorical variable is should be removed.
* **DOL** - As this dataset is very old (2015) we can’t determine whether the person who was offered a job has left that job or not hence it can be removed.
* **DOJ** -With DOL removed DOJ has no significance hence this can be also be removed.
* **Jobcity** - It has too many unique values and also missing values hence this feature cannot be used model building although it can be used for visualization.
* **CollegeCityID** - This feature comprises of lot of unique values but it can be used statistical tests and visualization.
* **CollegeCity** - This feature comprises of lot of unique values but it can be used statistical tests and visualization.
* **Designation** - It has too many unique values of categories which will make encoding complicated hence we can drop this feature but can be used for visualization.
  1. **COMPLEXITY INVOLVED:**
* Categorical features have too many unique values.
* Candidates have chosen different optional subjects mostly according to their domain so there are many 0s in the features that were not chosen by any candidate.
* Dataset is dated from 2015 which makes DOL feature unreliable.
* As some categorical features have too many unique values hence encoding in such cases can be huge limitation because of which we may to drop that feature while creating model.
* Null values in the initial dataset were represented by the value -1.
* Some Feature like 10board and 12board were filled with incorrect entries so those records had to be cleaned.
* Too many unique values in the Feature Specializations some of which are subcategories of other major specializations so a common platform of specialization needs to be introduced for the candidates belonging to nearly same domain.
* Scaling of some features required exploration.
* The relationship between most of the variables does not show any linear relationship in the initial plot.
* After encoding, correlation study becomes complicated with too many features.
  1. **PROJECT OUTCOME (COMMERCIAL, SOCIAL AND ACADEMIC):**
* **Commercial:**

From commercial perspective, the model reflects on the compensation class person belongs in, so it gives company idea whether, based on the skills, the candidate is overpaid or underpaid. This can have major impact on the profitability and overall success of the company. To keep employees engaged and productive, companies must compensate them appropriately, without overpaying them. This project can also give an idea to the recruiting companies about talents the applying candidates possess, where the various technical skills of the candidate are quantified in the form of scores and helps them to include best talents in their workforce.

* **Social:**

From Social aspects of the project, this project also lays emphasis on the social skills candidates must possess in order to get job. This aspires the future candidate not only to develop technical or language skills but also to develop and improve on their social and soft skills such as conscientiousness, agreeableness, extraversion, neuroticism and openness to experience. In the present era, apart from the technical skills companies also need candidates to have the abovementioned skills for their business and representation of company in the society so as to add social values.

* **Academic:**

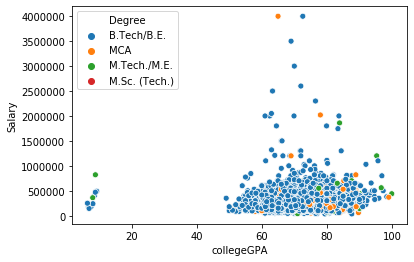
The projects show the factors that influence the salary class of the candidate the aggregate scored by candidate in 10th or 12th, the college CGPA, the specialization subject the candidate has opted for pursuing a Bachelor’s or Master’s degree. All these factors which helps the person to get job with particular compensation can help the future candidates to focus and opt for the subjects that is having more scope in career development. This will also give them an understanding of significance of marks scored in matriculation and HSC examinations.

* 1. **MODIFICATION OF FEATURES:**
* **Specialization**: initially containing around 46 categories the categories of this are clubbed together to mainly 6 categories so that encoding becomes easy.
* **10board**: This feature contains 275 categories, so to make the classification and encoding simple and uniform, different categories are clubbed together and now comprises of mainly 3 categories.
* **12board**: This feature contains 340 categories, so to make the classification and encoding simple and uniform, different categories are clubbed together and now comprises of mainly 3 categories.
  1. **ADDITION OF NEW FEATURES BASED ON EXISTING FEATURES:**

**Classification of salary**: The salary column is classified is into two categories namely 0 for salary<300000 and 1 for salary>300000. It is renamed as Salary\_class.

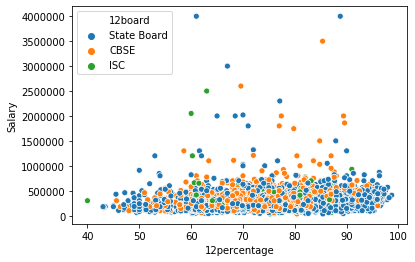
**Age:** From DOB feature the age of the candidate is determined and a new feature naming AGE is created.

1. **EXPLORATORY DATA ANALYSIS**
   1. **RELATIONSHIP BETWEEN VARIABLES:**
      1. **MULTIVARIATE ANALYSIS:**
   2. **Salary and College GPA:**



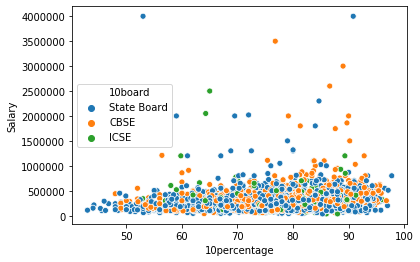
Let’s compare salaries and college GPA of candidates. If we observe the above scatter-plot we can see that almost all the candidates have an annual CTC whose college GPA was higher than 50. Some candidates whose college GPA is between 60 and 80, have got higher salaries. There are two candidates who have got the highest salaries, which is approximately around 40 lakhs. They are having B. Tech /B.E. and MCA degree.

* 1. **Salary and 12th percentage:**



In the above graph we are comparing salaries with 12th percentage, which is equally distributed. There is no significant relationship of 12 percentage with salary. Some candidates whose 12 percentage is between 60 and 90, have got higher salaries. Candidates who got the highest salaries are from State Board. Again, there is no significant relationship of 12th board and salaries.

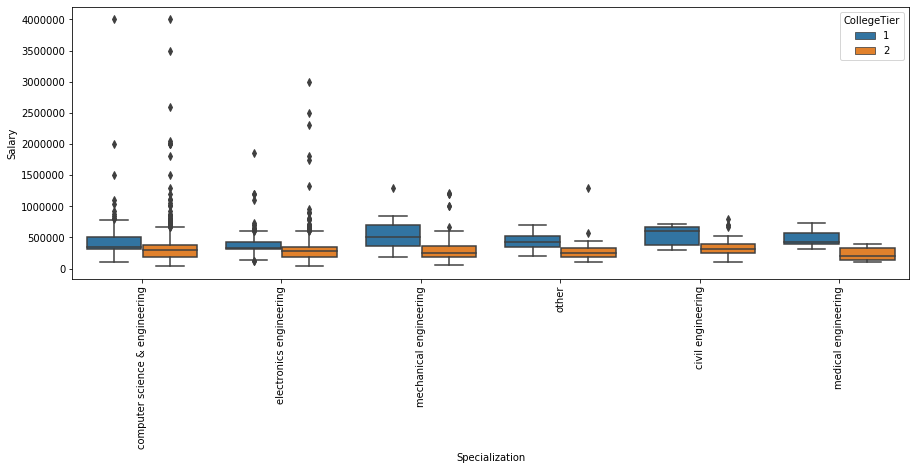
* 1. **Salary and 10th percentage:**



In the above graph we are comparing salaries with 10th percentage, which is equally distributed. There is no significant relationship of 10th percentage with salary. Some candidates whose 10th percentage is between 60 and 90, have got higher salaries. Candidates who got the highest salaries are from state board. Again, there is no relationship of 10th board and salaries. There is one interesting factor we can see, one of the candidates who got the highest salaries, has only 55% in 10th board. Let’s find out the other detail about that candidate which are mentioned below-

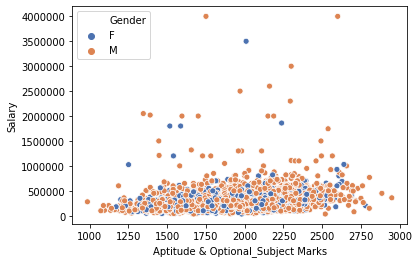


* 1. **Salary with Specialization:**



In the above graph we are comparing the salary with specialization and we can see that candidates from computer science and electronics engineering have got the higher salaries. Now if we want to see the effect of college tier on the salaries then we can conclude from the above graph that the mean salary of college tier1 is greater than the mean salary of college tier2 for all the specialization. Now question arises that does salary have any relation with the specialization? We can observe that in the graph by comparing the average salaries of specialization (Refer appendix 12.4). As per the below graph average salary of civil engineering is highest, followed by mechanical and computer science engineering. Though we cannot conclude any significant relationship between salary and specialization.

* 1. **Salary with AMCAT Aptitude and optional subject marks:**



In the above graph we are comparing salary with sum of Aptitude (Quant, Logical, English) and optional subject marks (optional subjects chosen by candidates with their respective domain) with gender as hue. Salary is equally distributed with the total marks and does not show any significant relationship. Candidates who got the highest salaries are both males. Average salaries of both male and female are almost equal, which are mentioned below-



* 1. **DISTRIBUTION OF DATA:**



The detailed analysis of variables has been specified (Refer appendix 12.5)

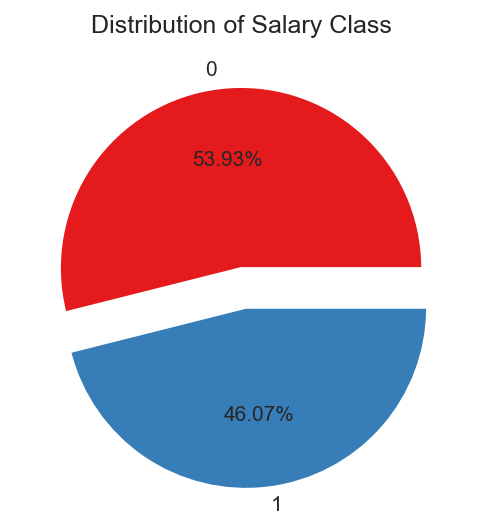
* 1. **MULTICOLLINEARITY:**

After checking for multicollinearity using variance inflation factor analysis we obtained the VIF values (Refer appendix 12.6), with respect to respective features, within the tolerance zone. Hence, we can these features for linear model such as logistic regression.

* 1. **OUTLIERS AND ITS TREATMENTS:**

After observing the outlier plot (Refer appendix 12.7), one can see that the 3 variables – agreeableness, Salary and openess\_to\_experience have maximum number of outliers in data. Moreover, the outliers in these variables can be referred to extreme values which are good for the learning phase for the model, so that if the model comes across such extreme values while testing phase it can produce good results and classification model are robust to outliers. The other variable outliers can be negotiated as the number of outliers is negligible compared to the whole data. Therefore, as part of this study we will not treat the outliers by capping them or any other technique.

**7.5 CLASS IMBALANCE:**



As shown in the above Figure, 53.93% of the total values of the target class fit in to the '0' class, connoting that the applicants classified compensation will be less than or equivalent 3 LPA. The remaining 46.07% in the '1' class, means that the applicants classified compensation will be more than 3 LPA. Since the end goal of our investigation is that we are fundamentally focusing on the compensation class which will be given to the applicant based on the range of abilities he/she presents and the imprints they have acquired in their past instructive vocations. It tends to be seen that there is no class irregularity between the given 2 compensation classes consequently we can utilize this information with important strategies like resampling, smote sampling, and so on to fabricate a decent machine learning model.

**7.6 STATISTICAL SIGNIFICANCE OF FEATURES:**

After performing statistical tests such as ttest and manwhitneyu test we have obtained pvalues (Refer appendix 12.8). Amongst all the categorical and numerical variables, the features ‘Extraversion’, openness\_to\_experience, all Specializations except Civil Engineering, Degree\_M.Sc. (Tech.) and Degree\_M.Tech./M.E. are statistically insignificant to the target variable. Hence these features will not be considered for building Parametric Models as they do not affect the target variables.

1. **FEATURE ENGINEERING**
   1. **TRANSFORMATION OF NUMERICAL FEATURES**

The numerical features of the dataset are assembled and skewness is checked.

As the skewness is within the acceptable range, hence it can be inferred that that transformation is not required for the variables.

* 1. **TRANSFORMATION OF CATEGORICAL FEATURES**

For categorical features the one hot encoding has been chosen to represent different categories in respective features mentioned below:

* **Gender**: This column having two categories (M and F).
* **10board**: This column has been simplified into mainly 4categories (CBSE, ICSE, State Board, N/A).
* **12board**: This column has been simplified into mainly 4categories (CBSE, ISC, State Board, N/A).
* **Degree**: This column comprises of 4 categories (B. Tech/B.E., MCA, M. Tech/M.E., M.Sc. (Tech.)).
* **Specialization**: This column consists of 6 categories (computer science & engineering, electronics engineering, mechanical engineering, other, civil engineering, medical engineering).

There are some columns already label encoded in the dataset. They are as following:

* **CollegeTier**: In this column, college has been label-encoded based on Tier of the college 1 for the best and 2 for good.
* **CollegeCityTier**: In this column, the city where the college is located has been label encoded as 0 and 1.
  1. **SCALING OF THE DATA**

For our dataset we are using the Standard Scaler for scaling our input variables. Normalization of data is also done in this case.

There are some features which are already scaled in the initially extracted dataset these were:

* **Domain**
* **conscientiousness**
* **agreeableness**
* **extraversion**
* **neuroticism**
* **Openess\_to\_experience**
  1. **FEATURE SELECTION:**

For our Logistic Regression Model/Parametric Models we will consider the features based on statistical tests and exploratory data analysis. (Refer appendix 12.9)

For our Non – Parametric Models, we will use the entire preprocessed data with all the one hot encoded categorical columns and numerical columns.

1. **MODEL BUILDING**

All the model assumptions required for model building has been specified in appendix (12.10).

**9.1 Base Model Results**

After EDA and feature engineering we split the data into train, test (70:30 ratio) for model building. In the beginning of our model building approach, we decided to proceed with all the basic models – Logistic Regression, Naïve-Bayes, Decision Tree, Random Forest and K-Nearest Neighbors. We ran all models with and without hyper parameter tuning, to get an idea of the base performances of these models. We ran the above-mentioned algorithms and yielded results using K-Fold Cross Validation of each model.

The 5-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way. We have also performed 10-fold cross validation procedure in the later parts of our exploration. The 5-fold cross validation results have been obtained (Refer appendix 12.11).

The figure represents the box and whisker plots showing the spread of the accuracy scores across each cross-validation fold for each algorithm.

If we observe the above algorithm comparison chart, we can see that Logistic, KNN and Random Forest models are giving better results as compare to other models or we can say that

focus on these 3 models for future model evaluation and performance. Now if we observe the accuracy score table, we found that none of the models have a good accuracy score so we can say that there is still some scope of improving the performance of the models.

In order to improve the performance of the model we applied Unsupervised Learning algorithms on our data and explored it further to uncover hidden patterns.

**9.2 Unsupervised Learning**

**9.2.1 K Means Clustering**

To find the number of optimal-cluster using K means we can use elbow plot which is the graphical representation between the number of clusters and its moment of inertia. (Refer appendix 12.12)

If we observe the elbow plot (Refer appendix 12.12) we can see that there are a lot of optimal number we can choose to form the cluster. After trial and error, we compared different number of clusters with its silhouette Score and we chose n = 2.

**9.2.2 Hierarchical clustering**

To understand the number of clusters through hierarchical clustering we can use dendrogram.

The dendrogram as mentioned (Refer appendix 12.13) has been drawn using the linkage method as ‘ward’. When we tried to compare the other linkage method it was found that others methods were forming the Highly imbalance graph. So, we chose ward as the linkage method as it forms the cluster on the basis of minimum moment of inertia and it was giving the better cophenetic distance as compared to the other models. Here also we are getting optimal number of clusters as 2 at the threshold of 100.

Now when we compared both methods of clustering, we chose K - Means because it has a better silhouette score and less moment of inertia as compare to hierarchical clustering.

**9.2.3 Visualization using K Means**

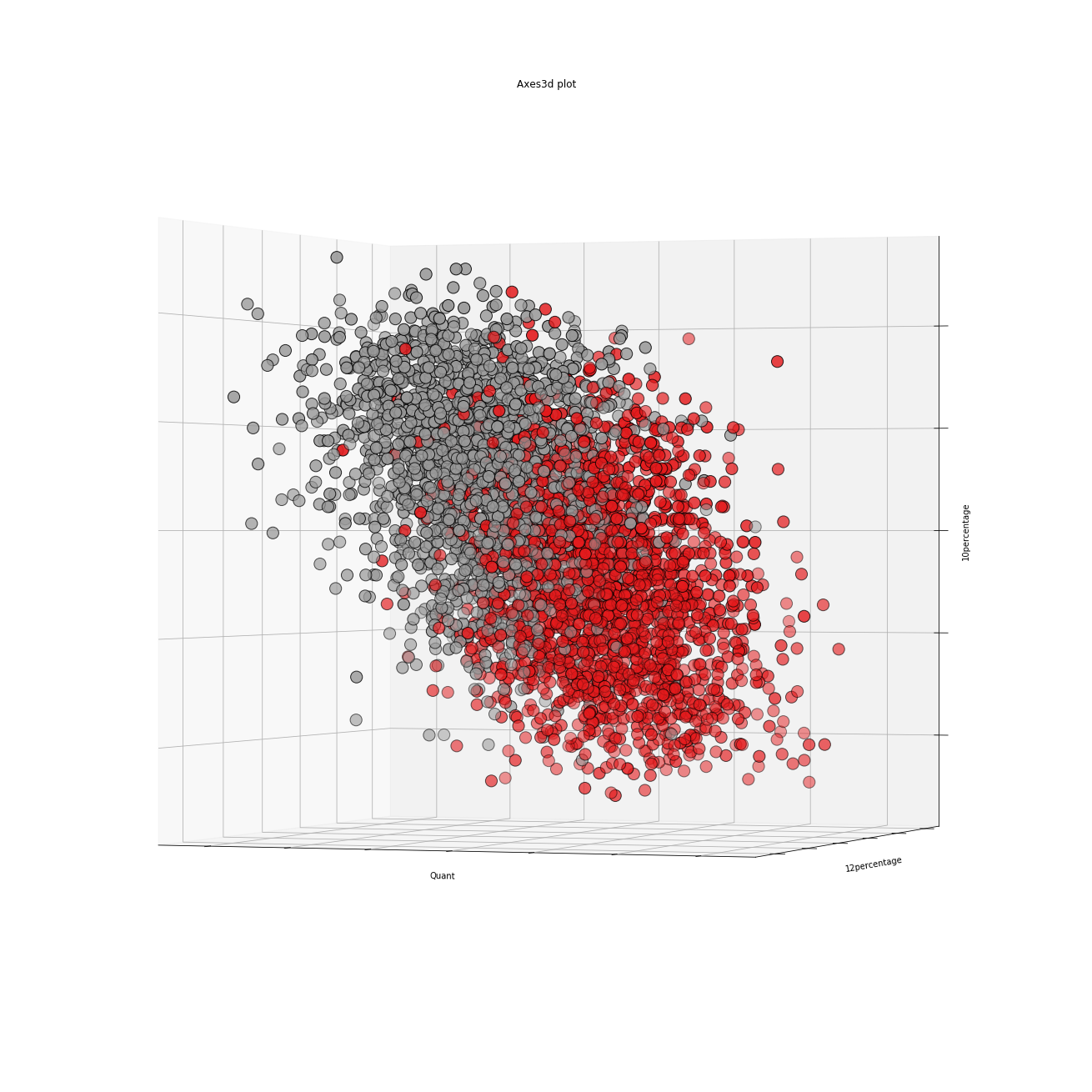


Fig. (i)

*Fig. (i) is 3D graphical representation of the column ‘Quant’, ‘12percentage’ and ‘10percentage.*

Now if we observe the above plots, we can see that clusters are not homogeneous. The ideal case should be that clusters are highly homogeneous within and highly heterogeneous between, which is not the case here. So, we will apply Principal Component Analysis technique which is going to reduce the dimensions of the data and will segregate the clusters in a better way through Visualization.

**9.2.4 Principal Component Analysis**

In statistics, Principal Components Analysis (PCA) is a practice that can be used to simplify a dataset., more formally it is a linear transformation that selects a new coordinate system for the dataset such that the highest variance by any projection of the dataset comes to lie on the first axis (then called the first principal component), the second highest variance on the second axis, and so on. PCA can be used for reducing dimensionality in a dataset while retaining those characteristics of the dataset that contribute most to its variance by removing the later principal components (by a more or less heuristic decision). These characteristics may be the "most important", but this is not necessarily the case, depending on the application.

PCA has the specialty of being the optimal linear transformation subspace that has largest variance. However, this comes at the price of greater computational requirement. Unlike other linear transforms, the PCA does not have a fixed set of basis vectors, its basis vectors depend on the data set.

Assuming zero empirical mean (the empirical mean of the distribution has been subtracted from the data set), the principal component *w*i of a dataset *x* can be calculate by finding the eigenvalues and eigenvectors of the covariance matrix of *x,* we find that the eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the dataset. The original measurements are finally projected onto the reduced vector space.

1. **Proportion of variance plot**

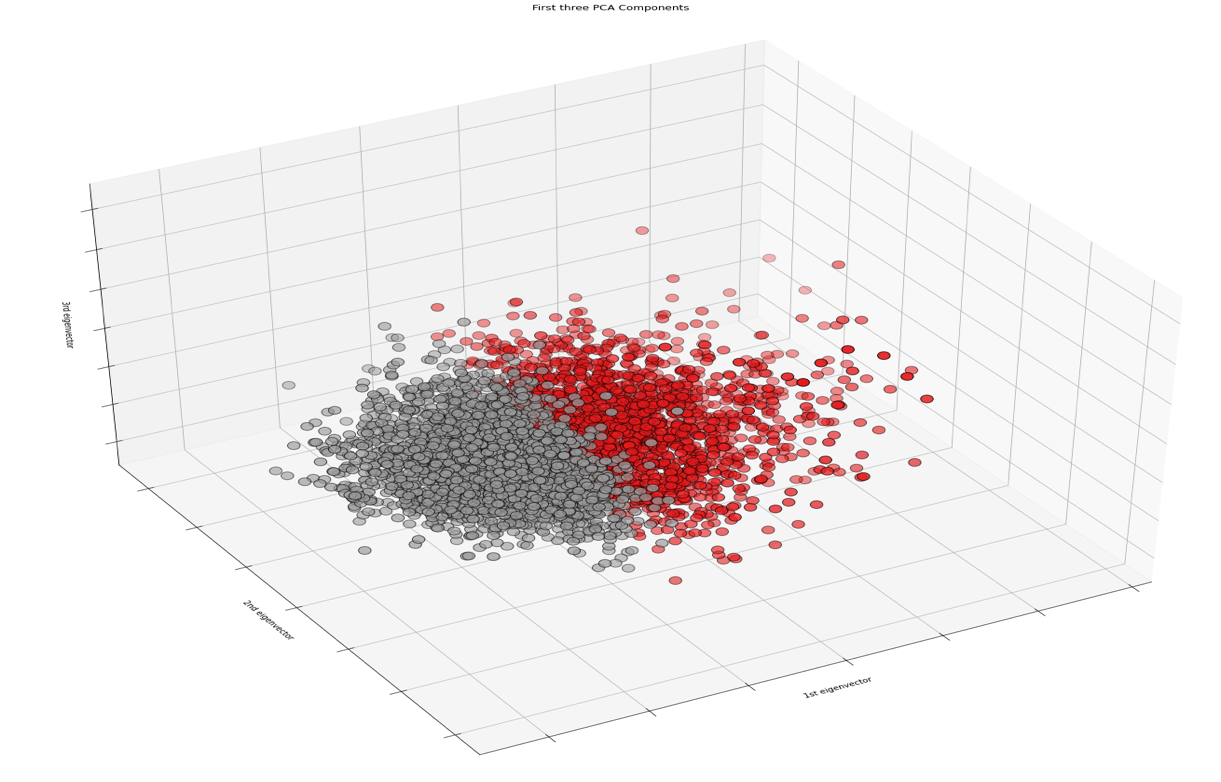
**Explained variance** is the amount of variance explained by each of the selected components.

**Explained variance ratio** is the percentage of variance explained by each of the selected components.

The proportion of variance plot is plotted (Refer appendix 12.14.1), where X-axis is the number of Principle Components and y axis is the percentage of explained variance.

We can see that the first 20 Principal Components explain over 95% variance of the data; hence we can choose the appropriate number of components accordingly. For our project we are choosing 25 Components which is explaining 99% variance of the data.

1. **Scatterplot of PCA1, PCA2 and PCA3**



From the above 3D plot, we can say that, though there is a certain degree of overlap, the points belonging to the same category are distinctly clustered and region bound. This proves that the data captured in the first three Principal Components is informative enough to discriminate the categories from each other.

In other words, we now have evidence that the data is not completely random, but rather can be used to discriminate or explain the output Y.

1. **Effect of variables on each component**

The components attribute in PCA provides principal axes in feature space, representing the directions of maximum variance in the data. This means, we can see influence on each of the components by features.

## The heatmap and the color bar (Refer appendix 12.14.2) basically represent the correlation between the various feature and the principal component itself. The PCA biplot is an interesting plot and contains lot of useful information.

It contains two plots:

* 1. *PCA scatter plot* which shows first 3 components (We already plotted this above)

ii) *PCA loading plot* which shows how strongly each characteristic influences a principal component.

1. **Loading Plot**

All vectors start at origin and their projected values on components explains how much weight they have on that component. Also, angles between individual vectors tells about correlation between them.

A biplot simultaneously plots information on the observations and the variables in a multidimensional dataset.

A biplot can optimally represent any two of the following characteristics:

* distances between observations
* relationships between variables
* inner products between observations and variables

A 2-dimensional biplot represents the information contained in two of the principal components. It is an approximation of the original multidimensional space.

The PCA biplot (Refer appendix 12.14.3) represents the variables with calibrated axes and observations as points allowing you to project the observations onto the axes to make an approximation of the original values of the variables.

The angles between the vectors tell us how characteristics correlate with one another

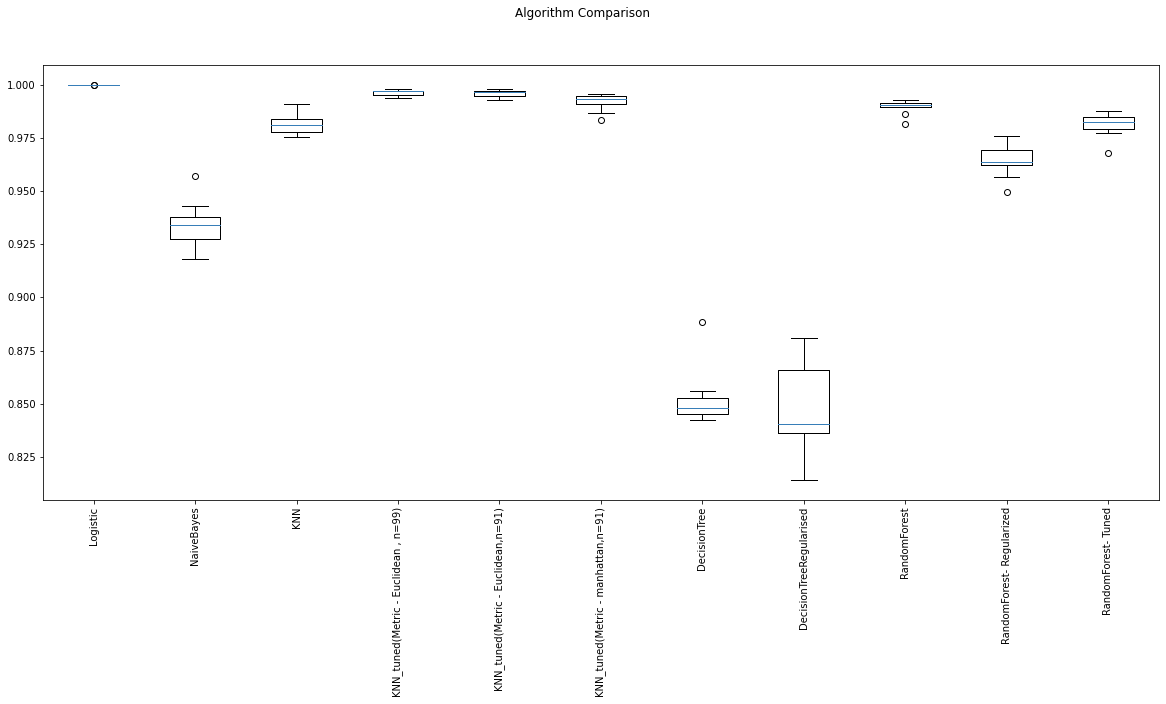
* When two vectors are close, forming a small angle, the two variables they represent are positively correlated. [10 Percentage and 12 Percentage]
* If they meet each other at 90°, they are not likely to be correlated. [Age and Open to experience]
* When they diverge and form a large angle (close to 180°), they are negative correlated. [Logical and Neuroticism]

**9.2.5 Classification Models Evaluation Using PCA Components**

We applied various classification models on PCA components that explain 99% variance of the data and evaluate the results. The result metrics for the base and tuned models obtained are shown below:



The algorithm comparison of the models is shown below-



Based on these models, we derived the following inferences:

1. As we saw earlier in base model that logistic, KNN and Random forest were giving us the better results. Now after applying PCA we can see that logistic has become highly overfit

Giving us the accuracy score as high as 99%.

1. Now we can see that KNN tuned and Random Forest Tuned both models are giving better results with accuracy around 93% but if we observe the gap between bias and variance

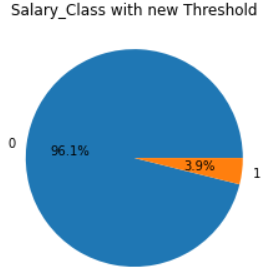
Error of both of the models then we can see that KNN tuned has the least gap.

**9.3 Classification Based On New Threshold Value Of Salary**

The mentioned dataset dates back to year 2015, however after that period with the advancement in technologies and rapid industrialization, the standard of living along with its respective cost has increased. As a result of which engineers who are getting job need higher annual compensation in order to meet their expenses.

So, keeping in mind the abovementioned conditions, it is essential to increase the threshold value for the classification of Salary. So, the new threshold value considered here is 600000.

After setting up these conditions, the distribution of records with respective class is mentioned below in terms of percentage:



**9.3.1 Base and tuned models with new classification class:**

Applying Base and tuned models to this dataset gives us the performance metrics (Refer appendix 12.15).

As it can be seen in plot (Refer appendix 12.15) the recall and precision score values for most of the above models are very low. This is due to highly imbalanced Salary class due to higher threshold that generates a minority class of 1.

This implicates that the above models are only accurate in predicting the majority class making the model weak for prediction of the significant class (which in this case is the minority class).

The comparative study shows that the recall values of all the models other than the naïve bayes model is very low which makes these models insignificant for prediction. Tuned models of KNN and Random forest algorithms have the lowest recall values (0 in this case). Naïve bayes model has the highest recall value which makes it the only model with moderate performance due to low accuracy.

From the above result metrics, it can be concluded that there is necessity to remove the imbalance of salary class without changing the threshold value. So different imbalance removing algorithms can be applied here to serve the purpose. As the data is important to us so we can’t lose the data hence we can rule out under sampling.

The sampling algorithms which will be used here are the following:

* Oversampling
* SMOTE

**9.3.2 Oversampling of Data**

After applying the algorithm to the newly classified dataset the distribution of the dataset in accordance with their classes results to the perfectly balanced class dataset in the ratio 1:1.

Now base model algorithms and tuning is applied on the newly generated dataset obtained by oversampling which gives us the following tabulations and plots (Refer appendix 12.16):



From the metrics tabulation and plot it can be observed that the recall and precision values of non-parametric models have significantly increased compared to previously imbalanced data hence improving the performance for the prediction of significant class.

However, it is observed that the accuracy and recall value are close to 100% in some cases which brings out the possibility that the model can be overfit. Random Forest model has extreme high biased error and extreme very low variance error giving a high gap between them, with highest accuracy of 99.99% where coefficient of variation is zero which doesn’t make the model good for prediction. Similarly, KNN and tuned random forest also have high gap between bias and variance error.

As oversampling only duplicates the records, hence its faster than SMOTE.

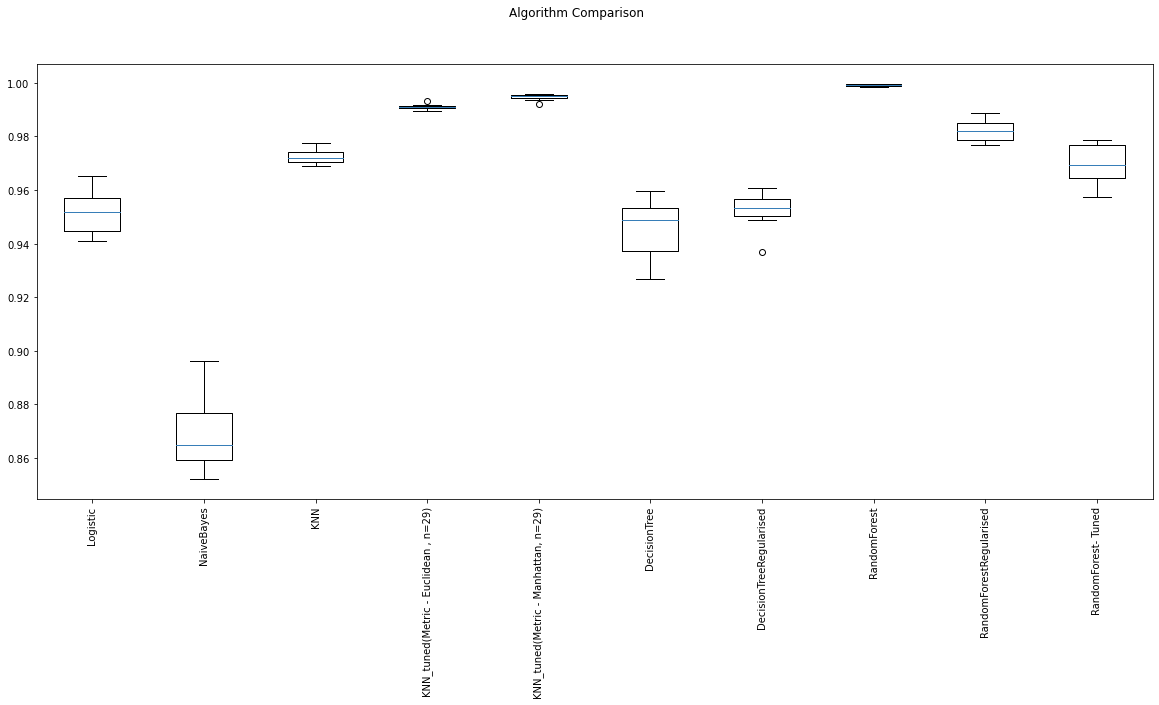
The result metrics and the comparison plot for the base and tuned models has been obtained (refer appendix)

**9.3.3 SMOTE Treatment**

After SMOTE treatment the salary class has become perfectly balanced.



The result metrics and comparison plot for different base and tuned models for SMOTE treated dataset can be seen below:

****

As we can see the tabulation above, it can be inferred that the accuracy of Logistic regression model has significantly increased when compared to Oversampling tabulation. As our main focus is recall and precision in this case, we can see the recall and precision of Logistic regression has also significantly increased in SMOTE treated data when compared to Oversampling treated data.

While the accuracy and other parameters for naïve bayes and other non-parametric models are more or less in +/-10% tolerance limit in both cases. We can say that Logistic Regression model has significantly improved in SMOTE when compared with Oversampling.

****



1. **BUSINESS RECOMMENDATIONS**

**10.1 Methodological improvements:**

1. **Data Collection:**

One important thing to take away from this dataset is that it had too many features. Certain features were found out to be insignificant during the statistical significance tests. It would be advisable to drop those features during the upcoming data collection process. There were lot of columns which were related to the optional subject of the candidates, which can be converted into a single

Column by taking the average so in future data collection process it can be transformed at the initial stage itself.

1. **Preprocessing:**

One hot encoding for the categorical features proved to be very exhaustive as the features themselves had too many numbers of sub-categories. This resulted in the increase of dimensionality. Other encoding methods can be tried out if there is a time availability constraint. The numerical features were found out to have a skewed distribution and primary transformations were applied to bring it into Gaussian distribution. Advanced transformation techniques can be experimented to treat the extreme values.

1. **Modeling:**

Many classification models were used for model building. Most of them resulted in models with a commendable ROC\_AUC score. Hyper-parameters can be tuned if the improvement of f1-score is the intention. Different threshold can be used to compare the models.

**10.2 Research/Commercial Value:**

1. **Research Value:**

There is further scope for research with respect to the clustering techniques and threshold value of salary. As this dataset contained high number of features, k-means was applied. Other clustering techniques can be applied to finding out hidden patterns. This way, it could help in clusters with better similarity and certain candidates can be offered the salary with respect to their cluster.

1. **Commercial Value:**

This study can be used to predict the salary class of the engineering graduates. Our best model can be used to fix the CTC of graduates. As we know that human capital is a company’s most important asset. It is also typically the largest operating expense, and thus can have a major impact on profitability and overall success of a company. To keep employees engaged and productive, companies must compensate them appropriately, without overpaying them. The cost of overpaying or underpaying can make a significant impact on a company’s business. As per the EDA of our project it was found that computer science graduates have got the higher salary. Most of the candidates who have higher packages they had college GPA between 60 to 80. There was no linear relationship of any feature with the salary column. So, an HR can keep these things in mind before going for the recruitment.

**10.3 Recommendations Based on Insights:**

Taking into consideration of our analysis and understanding of the dataset, EDA and model performances, we tried to build the model in two different ways, which are:

1. Using clustering techniques
2. By changing threshold value.

If we talk about the clustering technique, after applying PCA, KNN tuned is giving the **accuracy score of 95.94%, precision 96%, recall 95% and F1\_score 95%.** KNN tuned is giving the least bias, variance error gap. Now we tried to build the model with our 2nd technique which is by changing the threshold value of the salary. We kept it at 6 lakhs.

After implementing the oversampling and smote technique, it was found that smote technique was giving better results. After applying the smote technique Random Forest is giving the accuracy **score of 98.63%,** **precision 98%, recall 98% and F1\_score 98%.**

Regarding the studies related to Data Science studies, it is to be made clear that this study can be used only for prediction purposes and decision making despite good mathematical results. Despite the impressive results and also the trend of the market may be subjected to change in the upcoming days. That being said, the two objectives of our study can be used in different ways to be incorporated into a salary prediction.

1. Firstly, if we want to predict the salary and implement the product in the market then we can go ahead with KNN tuned model which is giving the good results.
2. Secondly it depends on the market trend if the economy is booming and candidates are getting higher salaries then we can change the threshold of the salary and build the model accordingly.
3. **REFERENCES AND BIBLIOGRAPHY**

# [1] Sidart asiri – “Machine Learning Classifiers” - *Towards Data Science [Jan, 2019].*

# [2] Varun Aggarwal, Shashank Srikant, Harsh Nisar ‘AMEO 2015 - A dataset comprising AMCAT test scores, biodata details and employment outcomes of job seekers.’ *3rd IKDD conference [March,2016].*

[3] Bekir Karlik, Baris Kocer, Abdulkerim Mohammed Yibre, ‘Comprising Feature Selection and Classifier Methods with SMOTE for Prediction of Male Infertility’ , *International journal of Fuzzy Systems* *[February 2016]*

[4] Sara Zahi and Boujemâa Achchab. Modeling car loan prepayment using supervised machine learning. *Procedia Computer Science 170 (2020) 1128–1133* [April 2020]

## [5] Deepti Sisodia and Dilip Singh Sisodia. Prediction of Diabetes using Classification Algorithms. [*Procedia Computer Science*](https://www.sciencedirect.com/science/journal/18770509)[*Volume 132*](https://www.sciencedirect.com/science/journal/18770509/132/supp/C)*, 2018, Pages 1578-1585 [2018]*

# [6] S.B Ostinatos. Supervised Machine Learning: A Review of Classification Techniques. *Emerging Artificial Intelligence Applications in Computer Engineering [2007].*

[7] Qing Li, PhD, Xueqin Yao, BSc, Damien Échevin. How Good Is Machine Learning in Predicting All-Cause 30-Day Hospital Readmission? Evidence from Administrative Data. *Value in Health.* [*Volume 23, Issue 10*](https://www.sciencedirect.com/science/journal/10983015/23/10)*, [October 2020]*

[8] Will Badr. 6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples) Popular strategies to statistically impute missing values in a dataset *Towards Data Science. [Jan,2019]*

[9] Mohammad Reza Mahmoudi, Mohammad Hossein Heydari, Sultan Noman Qasem, Amirhosein Mosavi, Shahab S. Band. Principal component analysis to study the relations

between the spread rates of COVID-19 in high risks countries. *Alexandria Engineering Journal [Sept,2020]*

[10] Statistics and Data Analysis for Financial Engineering by David Ruppert, David S. Matteson.

[11] Introduction to Statistics and Data Analysis by Christian Heumann, Micahel Schomaker Shalabh

[12] The Data Design Manual by Steven S. Skiena.

[13] Introductory Statistics by Prem S. Mann**.**

[14] Applied Statistics and Probability for Engineers by Douglas C. Montgomery, Douglas

Montgomery, and George C. Runger.

[15] https://www.payscale.com/compensation-today/2016/03/how-compensation-impacts-your-business

[16] https://www.entrepreneur.com/article/159438

[17] <https://www.atlantis-press.com/journals/ijcis/25899235/view>

[18] <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>

[19]<https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/>

[20]<https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/>

[21] <https://medium.com/@b.terryjack/unsupervised-learning-to-aid-labelling-for-supervised-learning-253fe2d8e06b>

1. **APPENDIX**

**12.1 DATA DICTIONARY:**

The initial dataset obtained for analysis has following data dictionary:



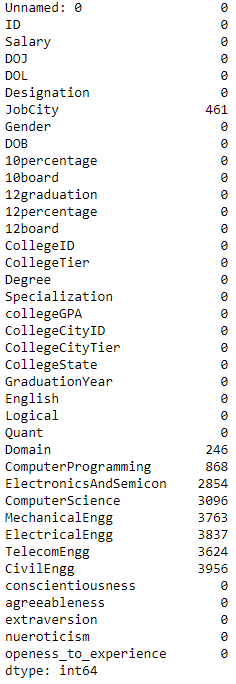
* 1. **CLASSIFICATION OF VARIABLES:**

**Numerical:** Salary, DOJ, DOL, DOB, 10pecentage ,12percentage, CollegeGPA, English, Logical, Quant, Domain, ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecommEngg, CivilEngg, conscientiousness, agreeableness, extraversion, neuroticism, openness\_to\_experience.

**Categorical:**Designation, JobCity, Gender, 10board, 12board, CollegeTier, Degree, Specialization, CollegeCityTier, CollegeState, ID, CollegeID, CollegeCityID, Graduationyear, 12graduation

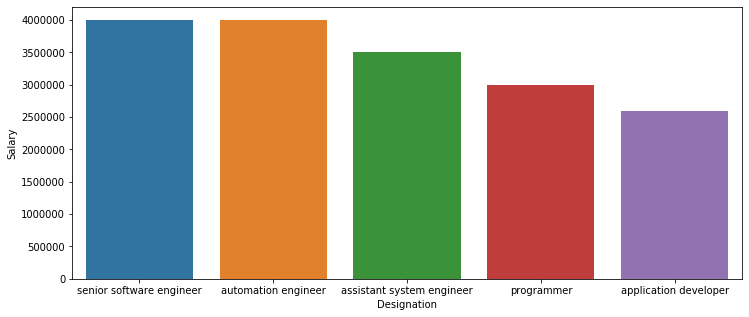
* 1. **MISSING VALUES TREATMENT:**

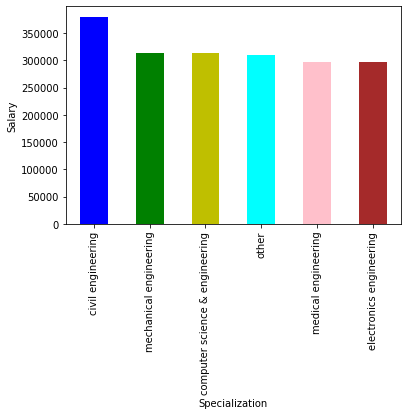
The missing values in the dataset obtained is filled with -1 instead, so in order to check for missing values, -1 or any insignificant value mentioned in the entry is replaced with np.nan after which the following results were obtained:



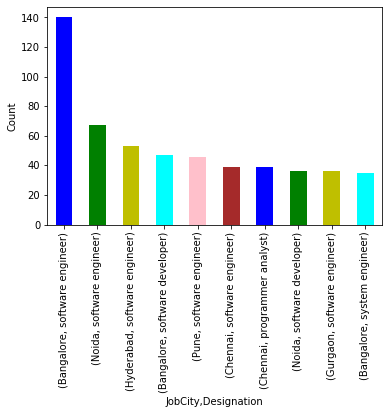
* 1. **EXPLORATORY DATA ANALYSIS:**

If we observe the below graph, we can see that candidates who got the highest salaries their designations are senior software engineer and automation engineer. Other designations are assistant system engineer, programmer and application developer.





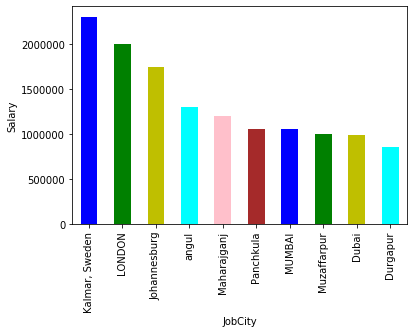
**Job City and Designation:**



In the above graph we can observe the details about the job city and the highest number of designations provided in the city. We can see that most of the roles were offered to the candidates in cities like Bangalore, Noida and Hyderabad. Maximum number of roles offered to the candidates were software engineering roles. We can see the top 5 job city and designation offered to the candidates in the below table:

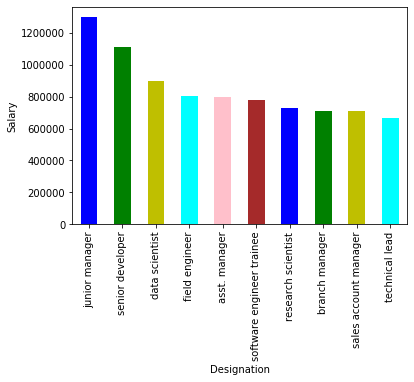


**Job City and Salary:**

****

The above graph is representing the job city and their average salary. From above observation we are getting to know that there are some candidates who got the international package. The highest average salary offered in a job city is Kalmar, London and Johannesburg.

**Designation and Salary:**

****

The above graph represents the average salary and designation. Best average salary offered to candidates are with designation junior manager, senior developer and data scientist.

**12.5 DISTRIBUTION OF DATA:**

* **SALARY**

From the above table, this column is observed to have a skewness of 6.4510 which means the data is right skewed. It has a kurtosis of 80.92 referring to the significant weight in the tails and is called a Leptokurtic distribution.

* **10 PERCENTAGE**

From the above table, this column is observed to have a skewness of -0.591019 which means the data is near to normal distribution. It has a kurtosis of -0.110284 referring to the significant weight in the near the mean and is called Platykurtic distribution.

* **12 PERCENTAGE**

From the above table, this column is observed to have a skewness of -0.032607 which means the data is near to normal distribution. It has a kurtosis of -0.630737 referring to the significant weight in the near the mean and is called Platykurtic distribution.

* **COLLEGE GPA**

From the above table, this column is observed to have a skewness of -1.2491241 which means the data left skewed. It has a kurtosis of 0.232088 referring to the significant weight in the tail and is called Leptokurtic distribution.

* **ENGLISH**

From the above table, this column is observed to have a skewness of 0.191997 which means the data is near to normal distribution. It has a kurtosis of -0.254133 referring to the significant weight near the mean and is called Platykurtic distribution.

* **LOGICAL**

From the above table, this column is observed to have a skewness of -0.216602 which means the data is near to normal distribution. It has a kurtosis of -0.224761 referring to the significant weight near the mean and is called Platykurtic distribution.

* **QUANTS**

From the above table, this column is observed to have a skewness of -0.019399 which means the data is normally distributed. It has a kurtosis of -0.102472 referring to the significant weight near the mean and is called Platykurtic distribution.

* **DOMAIN**

From the above table, this column is observed to have a skewness of -0.398797 which means the data is near to normal distribution. It has a kurtosis of -0.826813 referring to the significant weight near the mean value and is called Platykurtic distribution.

* **CONSCIENTIOUSNESS**

From the above table, this column is observed to have a skewness of -0.527003 which means the data is near to normal distribution. It has a kurtosis of 0.122596 referring to the significant weight in the tails and is called Leptokurtic distribution.

* **AGREEABLENESS**

From the above table, this column is observed to have a skewness of -1.204915 which means the data left skewed. It has a kurtosis of 3.391242 referring to the significant weight near in the tails and is called Leptokurtic distribution.

* **EXTRAVERSION**

From the above table, this column is observed to have a skewness of -0.523267 which means the data is near to normal distribution. It has a kurtosis of 0.643969 referring to the significant weight near the tails and is called Leptokurtic distribution.

* **NUEROTICISM**

From the above table, this column is observed to have a skewness of 0.165710 which means the data is near to normal distribution. It has a kurtosis of -0.191539 referring to the significant weight near the mean value and is called Platykurtic distribution.

* **OPENESS TO EXPERIENCE**

From the above table, this column is observed to have a skewness of -1.506962 which means the data is left skewed. It has a kurtosis of 5.788327 referring to the significant weight near the tails and is called Leptokurtic distribution.

* **AGE**

From the above table, this column is observed to have a skewness of 0.887271 which means the data is right skewed. It has a kurtosis of 1.826944 referring to the significant weight near the tails and is called Leptokurtic distribution.

* **NUMBER OF OPTIONAL SUBJECTS**

From the above table, this column is observed to have a skewness of 0.425634 which means the data is near to normal distribution. It has a kurtosis of 0.365757 referring to the significant weight near the tails and is called Leptokurtic distribution.

* **OPTIONAL MARKS**

From the above table, this column is observed to have a skewness of -0.96705 which means the data is left skewed. It has a kurtosis of 2.691225 referring to the significant weight near the tails and is called Leptokurtic distribution.

* **SALARY CLASS**

From the above table, this column is observed to have a skewness of 0.157624 which means the data is near to normal distribution. It has a kurtosis of -1.976143 referring to the significant weight near the mean value and is called Platykurtic distribution.

* 1. **MULTICOLLINEARITY**



**12.7 OUTLIERS AND ITS TREATMENTS:**

Chart, funnel chart

Description automatically generated

**12.8 STATISTICAL SIGNIFICANCE OF FEATURES:**



\*- Significant

\*\*- Insignificant

**12.9 FEATURE SELECTION:**

For our Logistic Regression Model/Parametric Models we will consider the following features based on statistical tests and exploratory data analysis.



**12.10 MODEL ASSUMPTIONS**

**Assumptions (Unsupervised Learning)**

**K-Means Clustering**

* This method considers two assumptions regarding the clusters:
  + The Clusters are spherical.
  + The Clusters are of similar size.
* Spherical Assumption helps in separating the clusters when the algorithm works on the data and forms clusters. If the assumption is violated, the clusters formed may not be what one expects. On the other hand, assumptions over the size of clusters helps in deciding the boundaries of the cluster.
* Certain measures (Example: Euclidean Distance) assume that the variables are uncorrelated within clusters.

**Hierarchical Clustering (Agglomerative Clustering)**

* This type of clustering has got no separate assumption of its own.

**Principal Component Analysis (PCA)**

* PCA is a non – parametric method.
* PCA should mainly be used for variables which are strongly correlated. If the relationship is weak between variables, PCA does not work well to reduce data.
* PCA makes the assumption that there is no unique variance, the total variance is equal to common variance.
* It is not recommended to use PCA with categorical data.

**Assumptions (Supervised Learning):**

**Logistic Regression**

* Logistic Regression does not make many of the key assumptions of Linear Regression and models based on Ordinary Least Squares Algorithms – regarding Linearity, Normality, Homoscedasticity and measurement levels.
* Linear Regression is used for predicting a continuous dependent variable and Logistic Regression is used to predict a categorical dependent variable

However, some assumptions still apply:

* The logistic regression assumes that there is minimal or no multicollinearity among the independent variables.
* The Logistic regression assumes that the independent variables are linearly related to the log of odds. (logit(p) = log(p/(1-p)), where p is the probabilities of the outcome. This can be checked by visually inspecting the scatter plot between each predictor and the logit values).
* The logistic regression usually requires a large sample size to predict properly.
* The Logistic regression which has two classes assumes that the dependent variable is binary and ordered logistic regression requires the dependent variable to be ordered.
* The Logistic regression assumes the observations to be independent of each other.

**Naïve Bayes**

* In case of Naïve Bayes Classifier an assumption is that the predictors/features are independent.
* Another assumption made here is that all the features have an equal effect on the outcome.

**K-Nearest Neighbours**

* KNN is a non-parametric, lazy learning algorithm. Non – Parametric model means it does not make any assumptions on the underlying data distribution.

**Decision Tree**

* In Decision Tree there is no probabilistic model, just the binary split which leads to maximum information gain. It is a non – parametric model so it does not make any assumptions on the underlying data distribution.

**Random Forest**

* Random Forest is a collection of Decision Trees and it is also a non-parametric model. The difference is that for Random Forest we use Bootstrap Aggregation. The only assumption that it relies on is that sampling is representative.

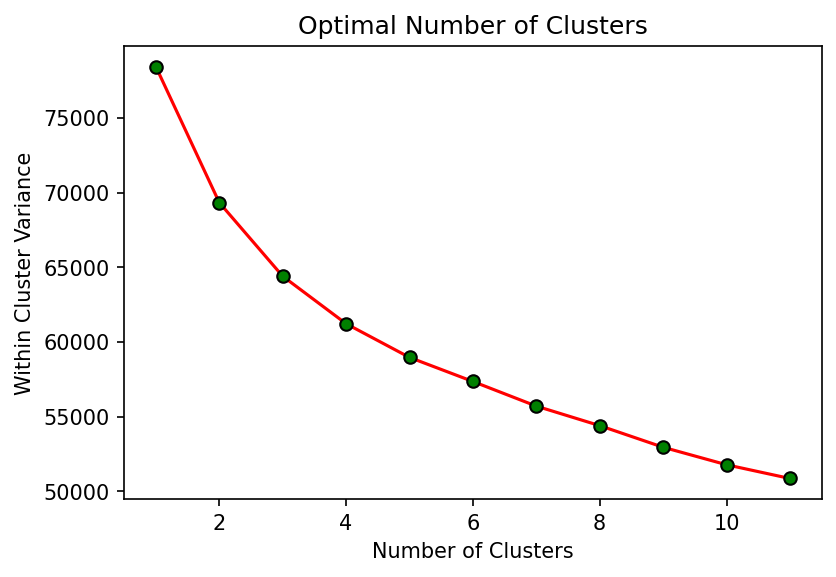
**12.11 Base Model Results**



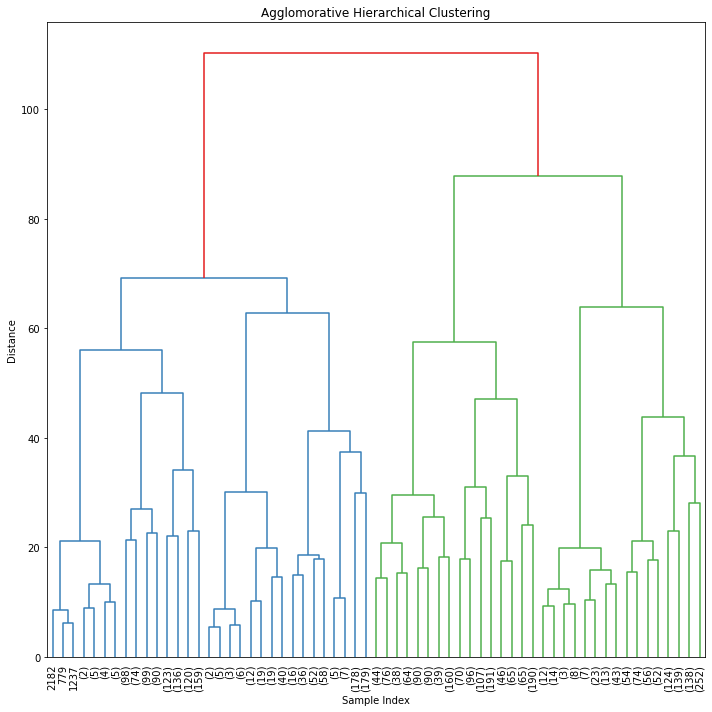
The below figure represents the box and whisker plots showing the spread of the accuracy scores across each cross-validation fold for each algorithm.



**12.12 K Means Clustering**

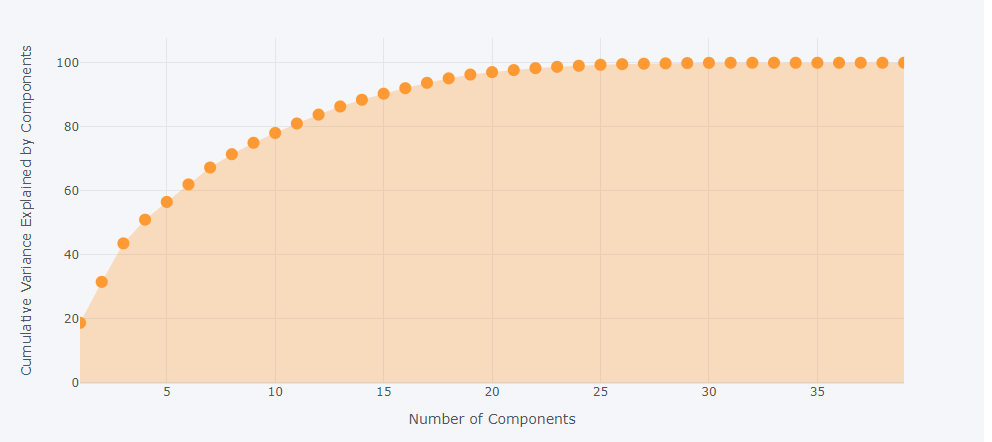


* 1. **Hierarchical clustering**



* 1. **Principal Component Analysis**

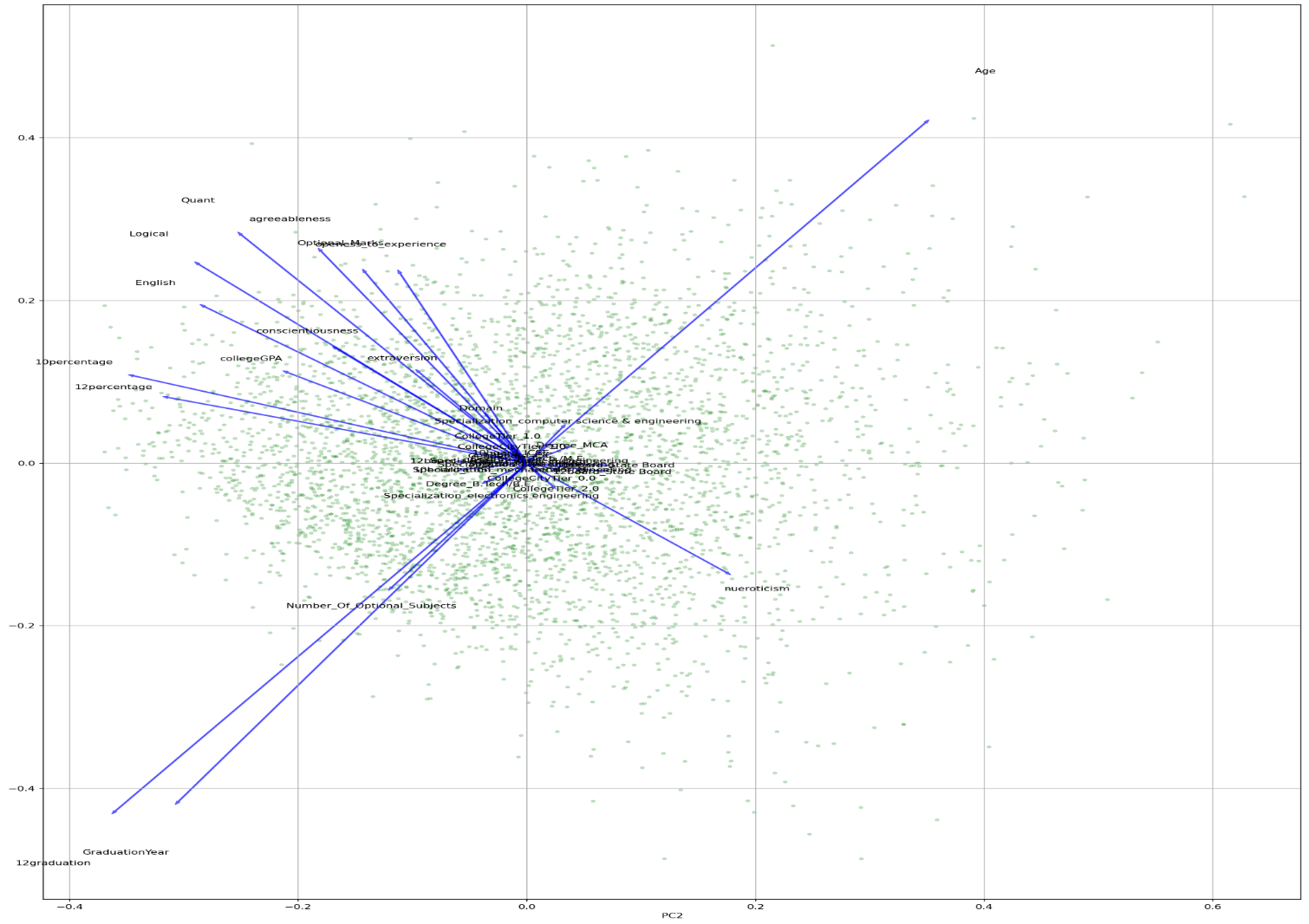
**12.14.1 Proportion of variance plot**



**12.14.2Effect of variables on each component**

****

**12.14.3 Loading Plot**

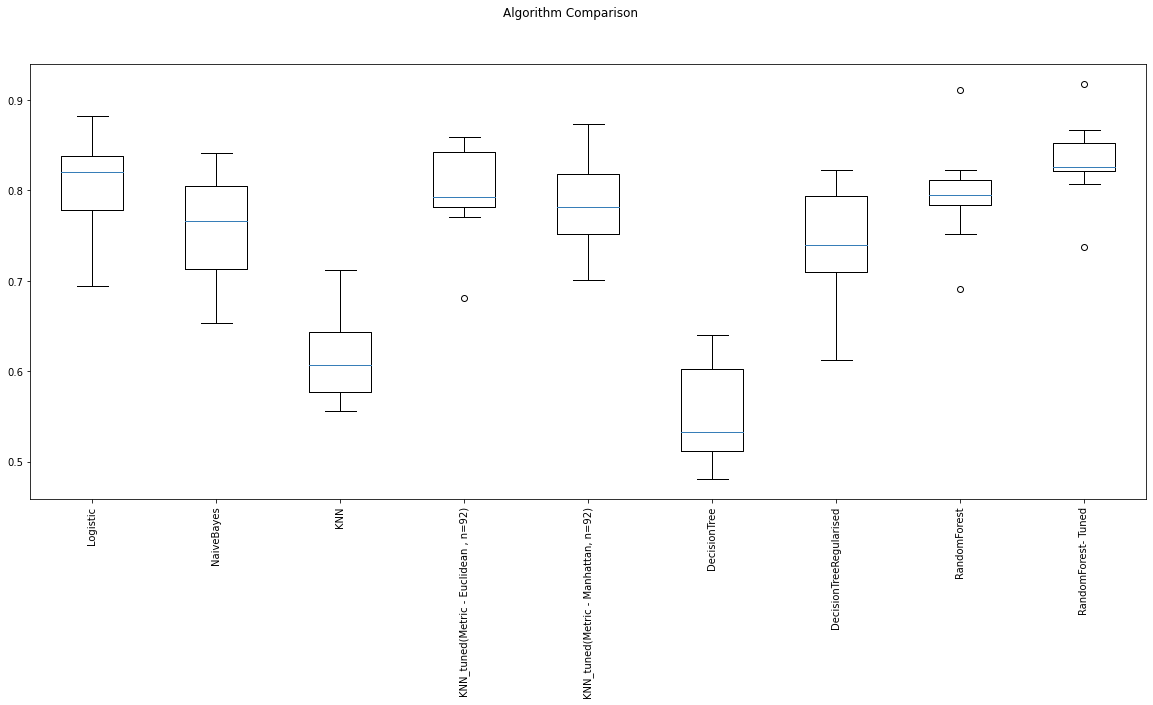


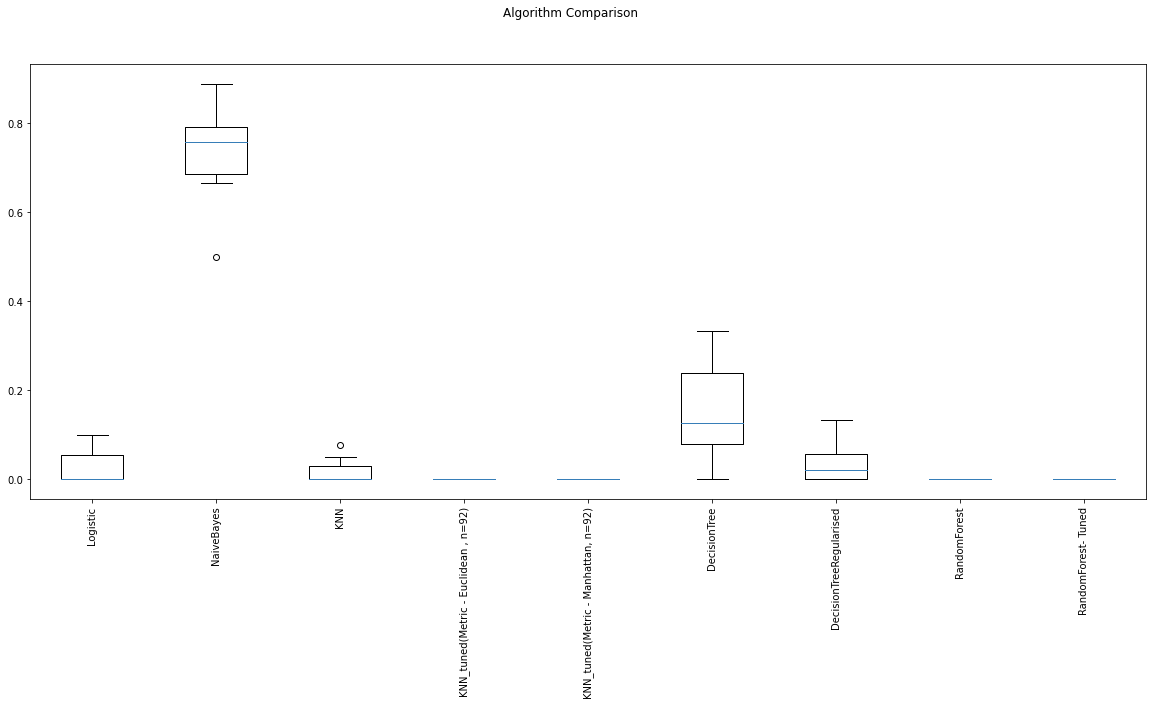
**12.15 Base and tuned models with new classification class:**

Applying Base and tuned models to this dataset gives us the following performance matrix.

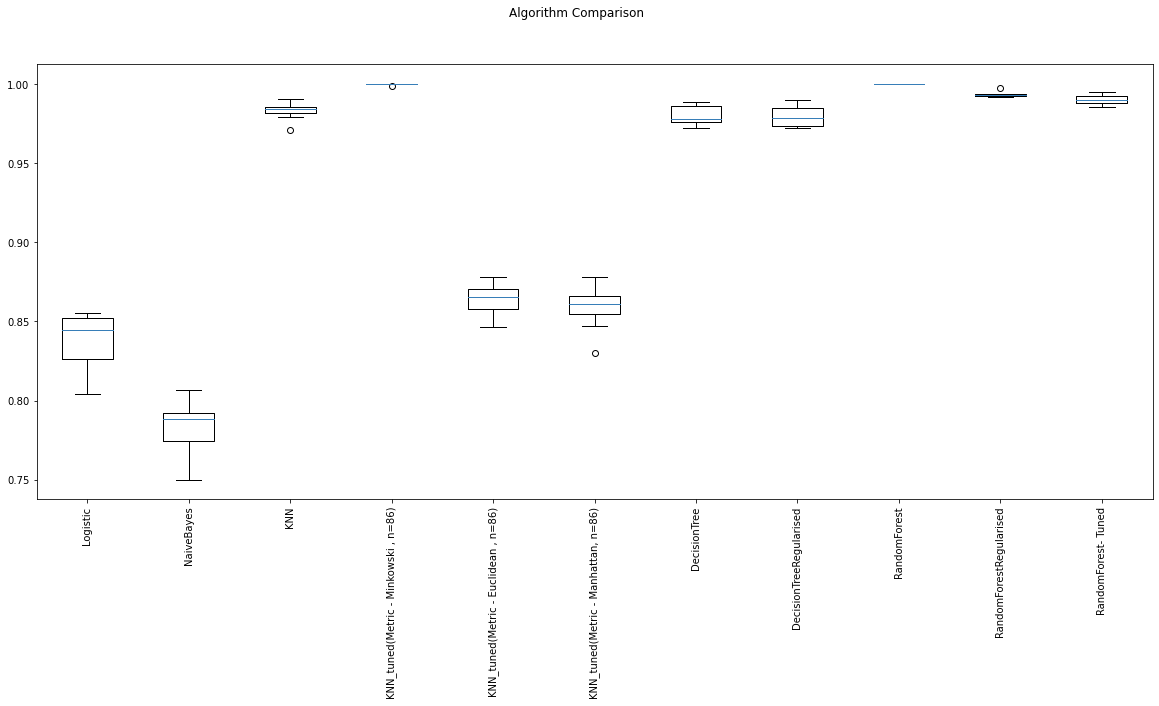
The latter plot shows results based on recall values.







**12.16 Oversampling of data**



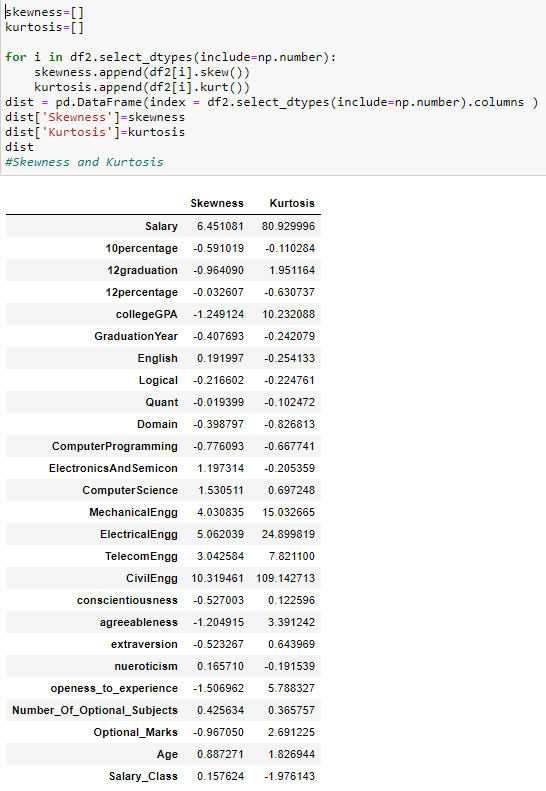
**12.17 Source code for plots and algorithms**

Here we have included the source code for each of the figures included in our report. They have been referenced by the figure and page number.

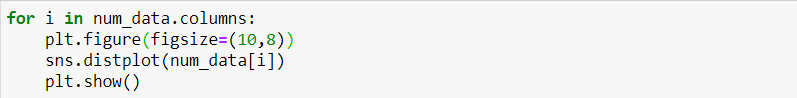
**EXPLORATORY DATA ANALYSIS**

**DISTRIBUTION OF DATA**

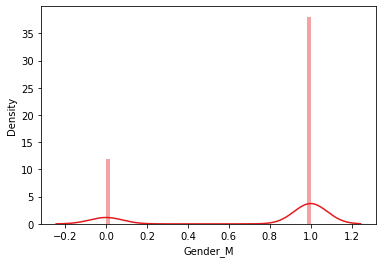
**SKEWNESS & KURTOSIS**

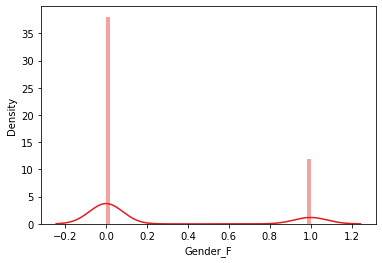


**DISTRIBUTION OF VARIABLES**

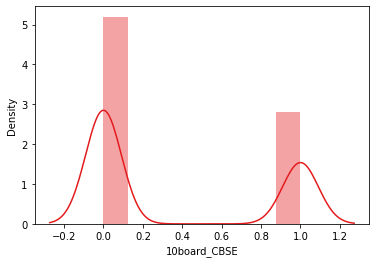


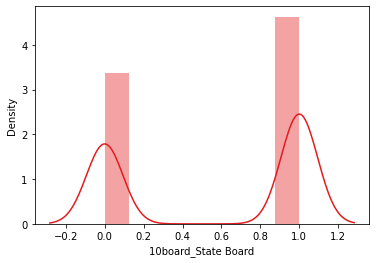
**GENDER**



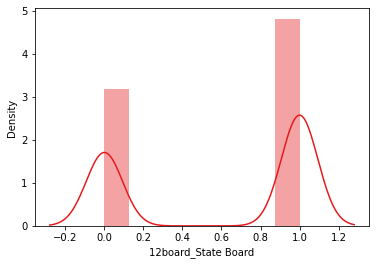
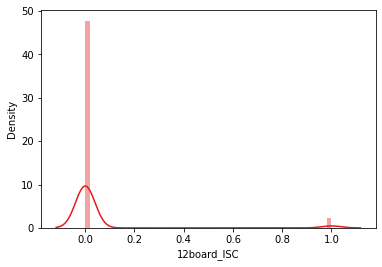
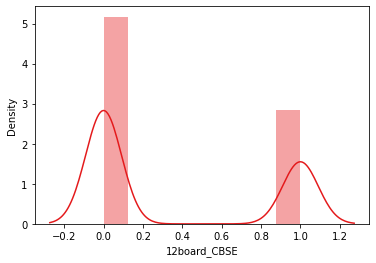


**10TH BOARD**

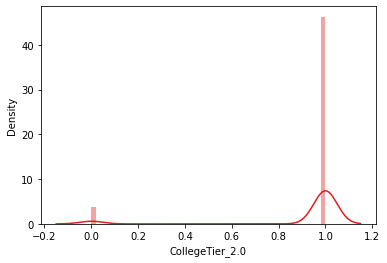
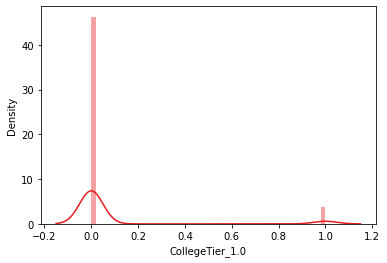




**12TH BOARD**

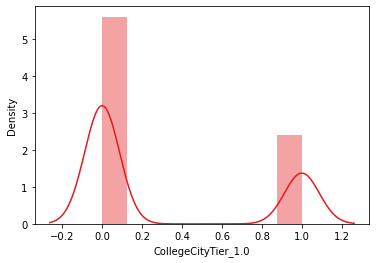


**COLLEGE TIER**

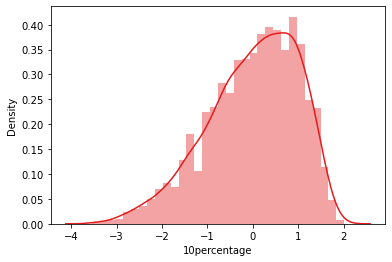
****

**COLLEGE TIER CITY**

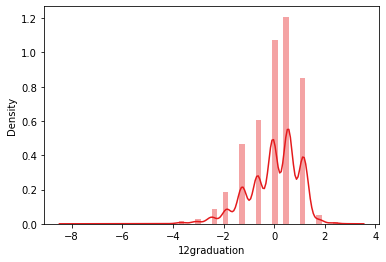
****

****

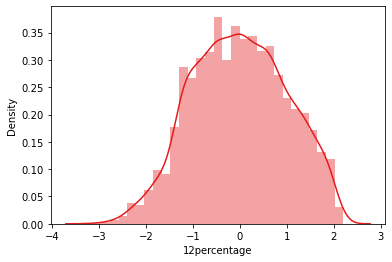
**10 PERCENTAGE**

****

**12graduation**

****

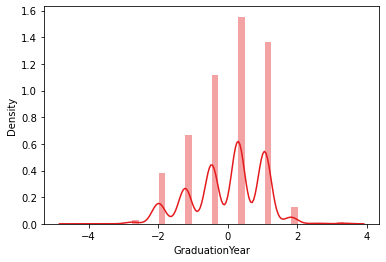
**12percentage**

****

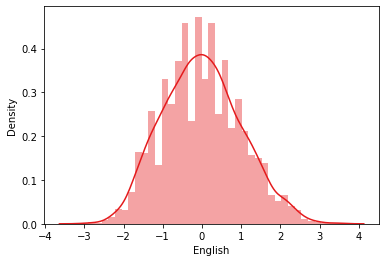
**collegeGPA**

****

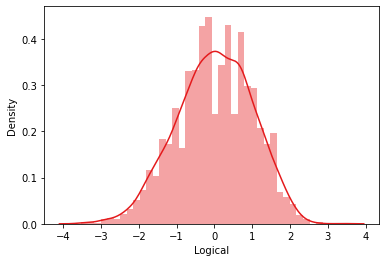
**GraduationYear**

****

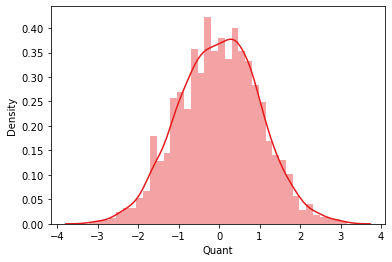
**English**

****

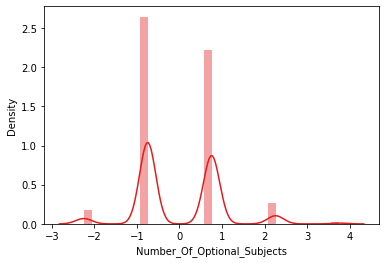
**Logical**

****

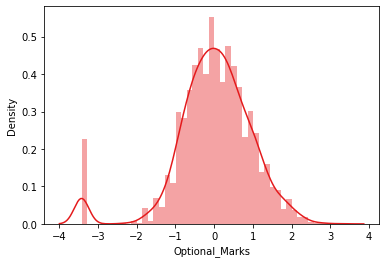
**Quant**

****

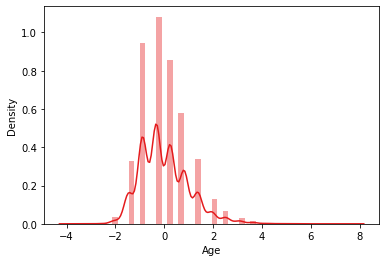
**Number\_of\_Optional\_Subjects**

****

**Optional\_Marks**

****

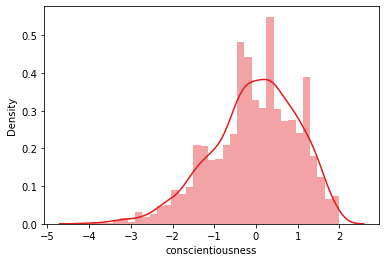
**Age**

****

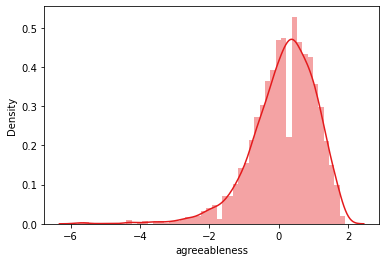
**Domain**

****

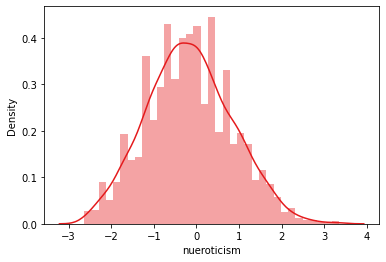
**Conscientiousness**

****

**Agreeableness**

****

**Neuroticism**

****

**MULTICOLLINEARITY**

**Graphical user interface, text, application, email

Description automatically generated**

**Table

Description automatically generated**

**FEATURE ENGINEERING**

**Graphical user interface, text, application

Description automatically generated**

**Table

Description automatically generated**

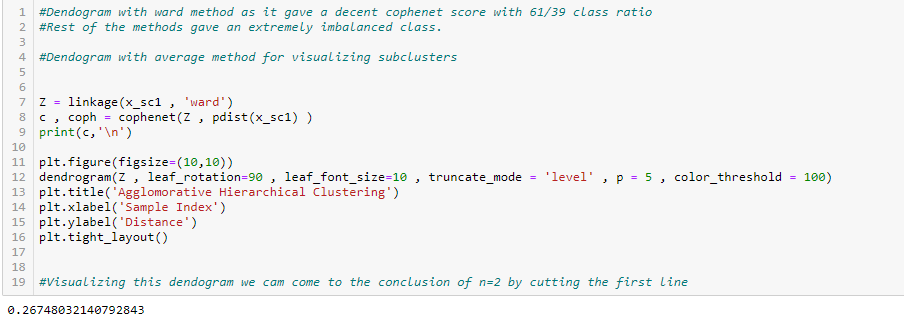
**Clustering algorithms**



**Graphical user interface, text

Description automatically generated**





**Principal Components Analysis**

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, application, table

Description automatically generated**

**Text

Description automatically generated**

**Chart

Description automatically generated**

**Graphical user interface

Description automatically generated**

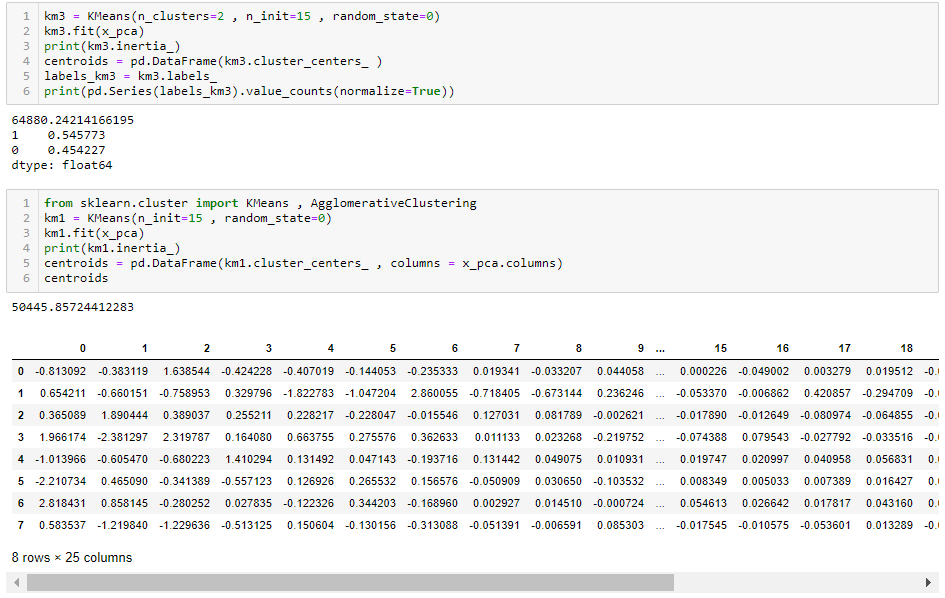
**Graphical user interface, text, application, email

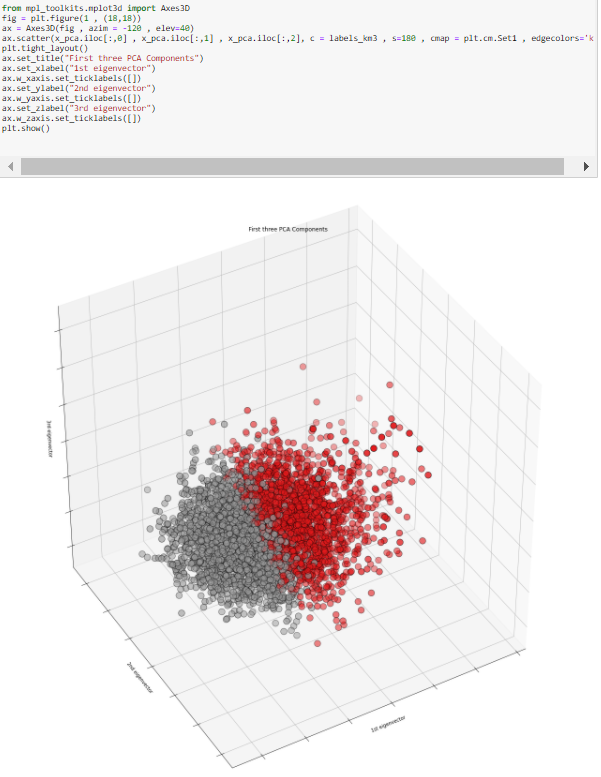
Description automatically generated**

**BIPLOT OF PCA**

**Graphical user interface, text, application

Description automatically generated**

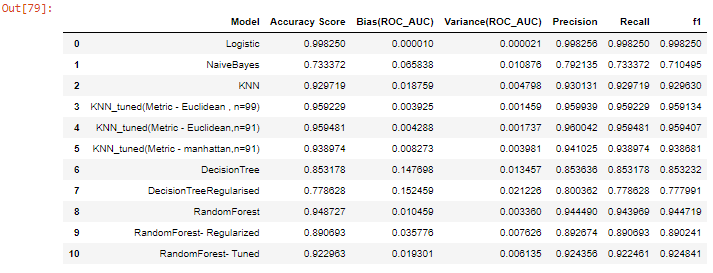




**Classification Model Evaluation Using PCA Components**

**Text

Description automatically generated**



**Base Models - Classification on Initial Threshold**

**Text

Description automatically generated**

**Text

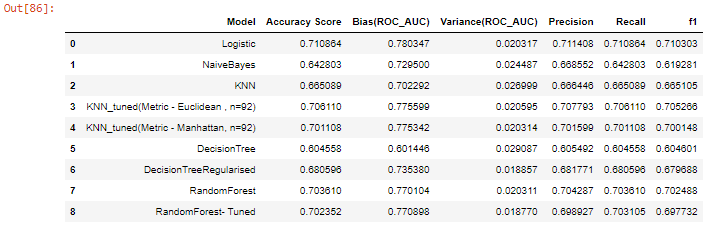
Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

**Text

Description automatically generated**



# All Base Models have been applied. Now we will change the threshold for Salary and apply Oversampling.

**Graphical user interface, text

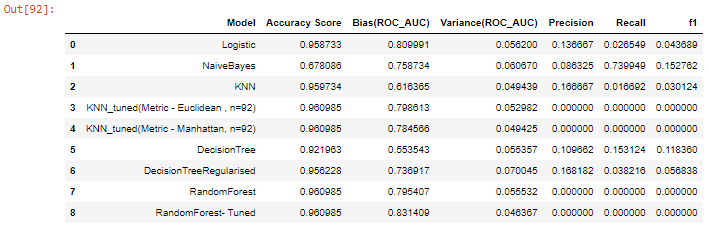
Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**



**Chart, box and whisker chart

Description automatically generated**

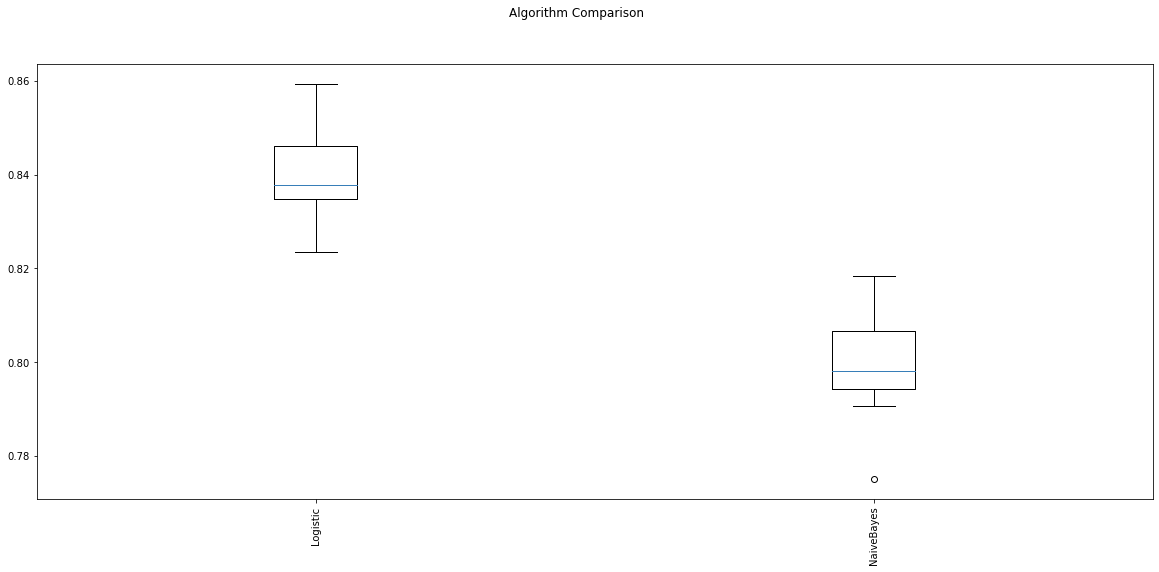
# The precision and recall for minority class falls below the acceptance range for all the models. We will have to treat the imbalanced data and run our models again

**Graphical user interface, text, application, email

Description automatically generated**

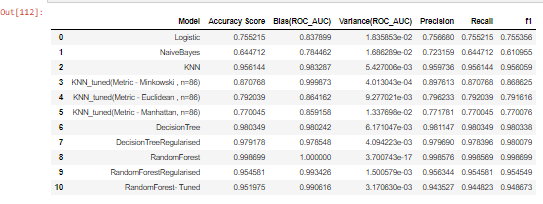
**Text

Description automatically generated**

****

**Text

Description automatically generated**



# **SMOTE**

**Graphical user interface, text, application, email

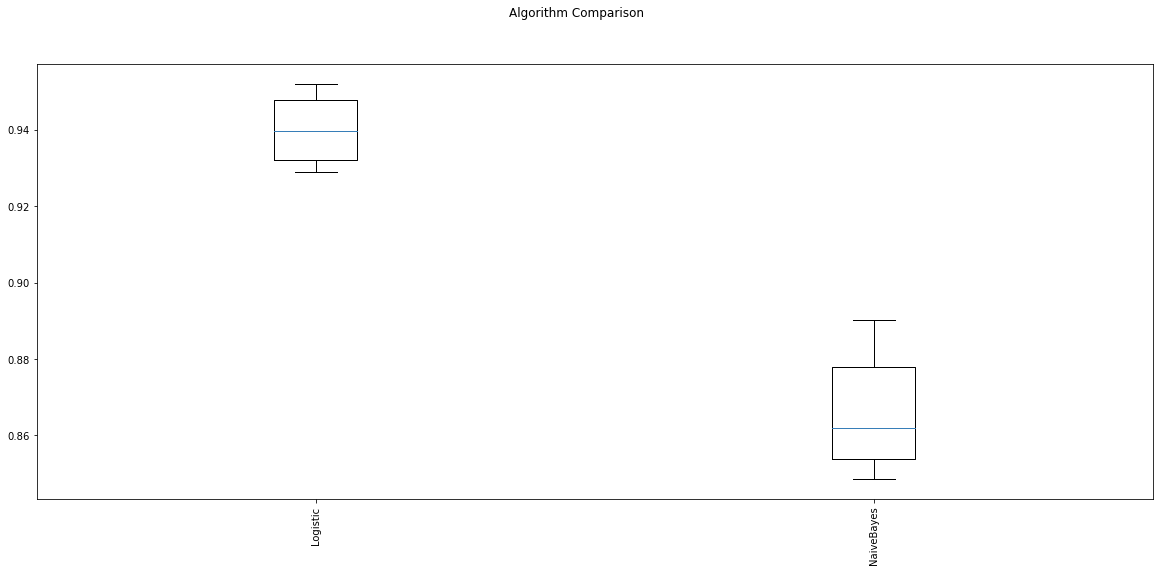
Description automatically generated**

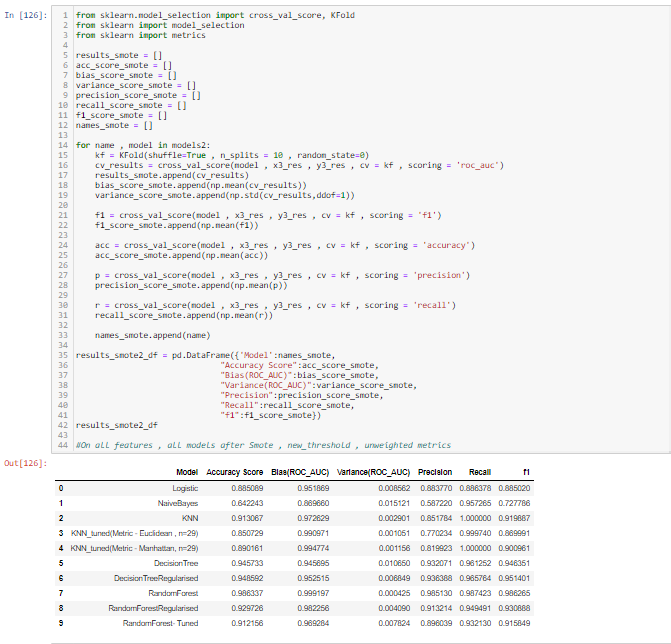
**Text

Description automatically generated**

**Text

Description automatically generated**

****



**Text

Description automatically generated**

