UNIVERSITÀ DEGLI STUDI DI MILANO

Predictive factors of occupational change

AUTHORS

Casalingo - 28865A
Coaro - 41208A
El Kouch - 31045A
Zanone - 33927A

Contents

1 Introduction **1**

2 The Dataset **2**

2.1 Structure 2

2.2 Challenges of the Dataset 2

2.3 Data preparation 3

3 Data Analysis **4**

3.1 Outlier Treatment 4

3.1.1 Cook's Distance 4

3.1.2 MCD - Minimum Covariance Determinant and Mahalanobis Distance 5

3.2 Unsupervised Learning 7

3.2.1 Principal Component Analysis and K-means 7

3.2.2 Model Based Clustering 8

3.2.3 Factor analysis of mixed data 10

3.2.4 Robust clustering: Trimmed K-means and T-clust 12

3.3 Supervised Learning 14

3.3.1 Logistic Regression 14

3.3.2 Logistic Regression with HC1 Covariance Type 16

3.3.3 Random forest 16

3.3.4 Mixed Effect Model 16

3.3.5 Evaluation of the models 17

4 Summary and Conclusions **19**

4.1 Unsupervised Learning 19

4.2 Supervised Learning 20

4.3 Supervised Learning 2 20

List of Figures **I**

1 Introduction

Occupational change represents a dynamic aspect of workforce behavior, influenced by a a variety of factors, from individual preferences to broader economic conditions. This study's aim is to employ multivariate statistical techniques to explore a possible predictive nature of the data with respect to the changes in the respondents' careers.

A significant aspect of our analysis is the predominance of categorical variables within the dataset. Categorical data, while rich in contextual information, present unique analytical challenges. Unlike continuous variables, categorical data requires careful encoding and interpretation to ensure that their influence on occupational change is accurately captured. Therefore a consistent approach with respect to these variables will represent a signficant part of this project.

In addition to the challenges posed by categorical data, the presence of outliers can significantly impact the robustness of statistical analyses. To tackle this issue, robust statistical methods were employed, enhancing the reliability of our findings.

The analytic pipeline of this project is meticulously structured to address these challenges and extract meaningful insights. The project starts with an initial clustering process, which groups similar observations and reduces the dimensionality of the dataset. This step facilitates the identification of underlying structures and patterns that might not be immediately apparent. Following clustering, a logistic regression was implemented to model the probability of occupational change based on the identified predictors. To further account for the hierarchical nature of the data and potential random effects, a mixed-effect model is incorporated into the analysis. This layered approach allows for a comprehensive exploration of the factors influencing occupational transitions.

With minimal missing data, the dataset appears to provide a solid foundation for the desired multivariate analysis. This study not only highlights the methodological considerations essential for handling categorical variables and outliers but also demonstrates the efficacy of a structured analytic pipeline in uncovering the key determinants of occupational change. By integrating clustering, logistic regression, and mixed-effect modeling, this report aims at offering a robust framework for predicting and understanding career transitions over the initial dataset.

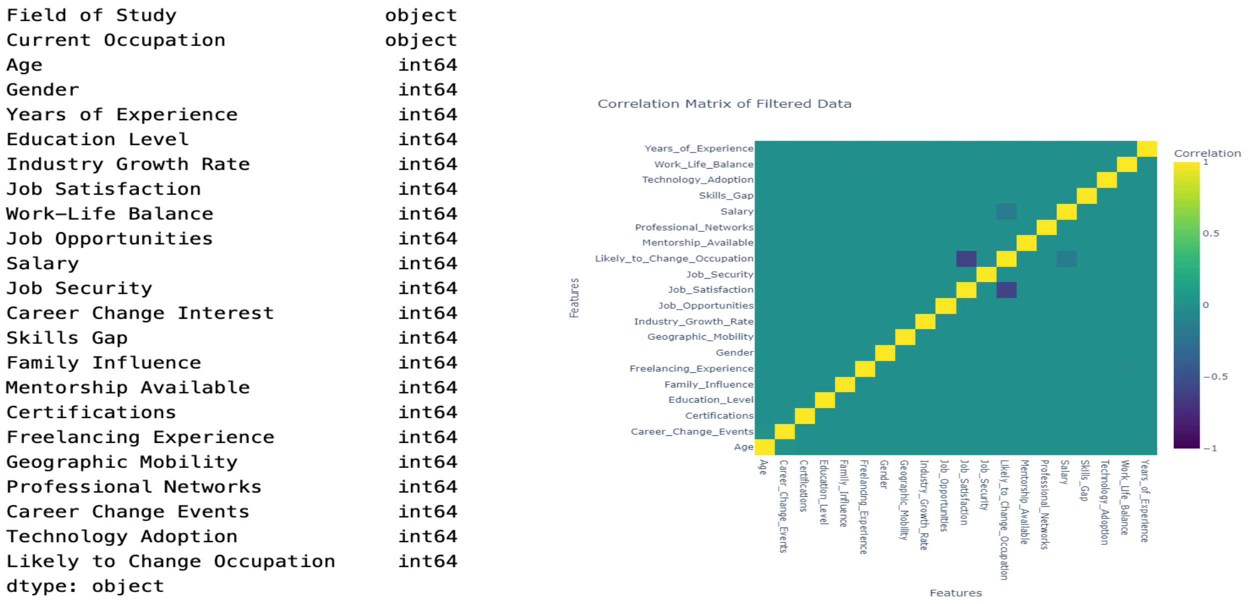
2 The Dataset

2.1 Structure

The dataset, which was originally retrieved from Kaggle, contains a variety of information and records about the current employment of the respondents, as well as an array of other statistics and metrics relative to the statuts of the respondents. Among them, some of the most relevant are:

- **Field of Study**: describes the type of degree obtained by the respondent.
- **Current Occupation**: actual job position held by the respondent.
- **Job Satisfaction**: describes a respondent's contentment with respect to their employment status.
- **Likely to Change Occupation**: a binary variable reflecting the inclination of the respondent toward a possible career change.

The structure of the data set is detailed in the following images:



```
Field of Study                object
Current Occupation            object
Age                            int64
Gender                         int64
Years of Experience            int64
Education Level                int64
Industry Growth Rate           int64
Job Satisfaction               int64
Work-Life Balance              int64
Job Opportunities              int64
Salary                         int64
Job Security                   int64
Career Change Interest         int64
Skills Gap                     int64
Family Influence               int64
Mentorship Available           int64
Certifications                 int64
Freelancing Experience         int64
Geographic Mobility            int64
Professional Networks          int64
Career Change Events           int64
Technology Adoption            int64
Likely to Change Occupation    int64
dtype: object
```

(a) Column Type

(b) Column influence

Figure 1: Exploration of Variables

2.2 Challenges of the Dataset

The dataset contains many categorical variables that have proved challenging to work with. Categorical data poses many issues when it comes to applying clustering or other data reduction techniques.

Specifically, clustering techniques rely on distance metrics such as the Euclidean one, which are designed for numeric variables, and do not perform as well with categorical ones. This implies that simply encoding categorical variables as numerical, might not be enough to ensure proper treatment of the data.

To address this issue, the devised approach was initially to implement a data reduction technique that worked with both categorical and numerical data. The approaches used are of various nature and intent, and range from using dummy variables in a Principal Component Analysis to performing FAMD (Factorial Analysis of Mixed Data). These, among other methods, were used to try and extract knowledge from a seemingly impervious dataset. For this reason, the next section will explain the different methodologies and the obtained results in depth.

Moreover, there are some variables in the dataset that are worthy of attention due to either their nature or their behavior with respect to the other variables involved. It is the case of

- **Field of Study**: Text variable that might prove to be relevant to the final analysis, but conversely shows to be not easy to keep within the scope of the research due to it being a text variable
- **Job Satisfaction**: Significantly more correlated with the final result than most variables.
- **Career Change Interest**: A mix of the two above.

2.3 Data preparation

To effectively work with the dataset, certain modifications are necessary to better align it with the requirements of our analysis. The first crucial step is categorical mapping in order to convert qualitative data into quantitative formats. The "Education Level" variable was encoded in ascending order of education such as "High School" represented as 1, "Bachelor’s" as 2, "Master’s" as 3 and "PhD" as 4, reflecting an ordinal relationship. The same procedure was applied to other variables, such as the "Industry Growth Rate," "Family Influence," and "Gender" variables, with distinct mapping dictionaries tailored to their respective contexts. The resulting numerical representations not only facilitate computational efficiency but also enable algorithms to process the data effectively. The target variable chosen is the likelihood of changing occupation, signaling its central role in the forthcoming chapter about predictive analysis . To focus on pertinent data, columns deemed less relevant, such as ’Field of Study’, ’Current Occupation’, and ’Career Change Interest’, are excluded. To further enhance consistency and usability, the column names are standardized by replacing spaces and hyphens with underscores, enabling smoother handling in subsequent operations.

3 Data Analysis

3.1 Outlier Treatment

Performing robust data analysis on the now filtered dataset requires finding ways to properly handle the outliers that could potentially distort the results. Thanks to the techniques described in the following subsections reliability and robustness of the results will be enhanced.

3.1.1 Cook's Distance

Cook's Distance is a statistical tool used to identify observations in a dataset that have a particular influence on the process of fitting a linear regression model to the data. Differently to other methods used to identify outliers, Cook's Distance strenght lies in its ability to highlight the influence of each individual data point on the regression model. The Cook's distance is defined as follows:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{p \times MSE} \tag{1}$$

Where:

- D_i: Cook's Distance for the i^{th} observation
- Σ: The sum symbol, indicating that we sum up the squared differences for all n observations
- $j = 1$ to n: Indicates that the sum goes from the first observation to the n^{th} (last) observation
- $\hat{Y}_j$: The predicted value of the dependent variable for the j^{th} observation, using all data points
- $\hat{Y}_{j(i)}$: The predicted value of the dependent variable for the j^{th} observation, when the i^{th} observation is removed
- p: The number of predictor variables in the model, including the constant term
- MSE: Mean Squared Error, an average of the squares of the residuals for the model

There are two main interpretations that can be attributed to the calculated distance, the first being:

Cook's Distance (D_i)	**Influence**
$D_i < 1$	Low influence
$0.5 < D_i < 1$	Moderate influence
$D_i > 1$	High influence

Table 1: Interpretation of Cook's Distance Values

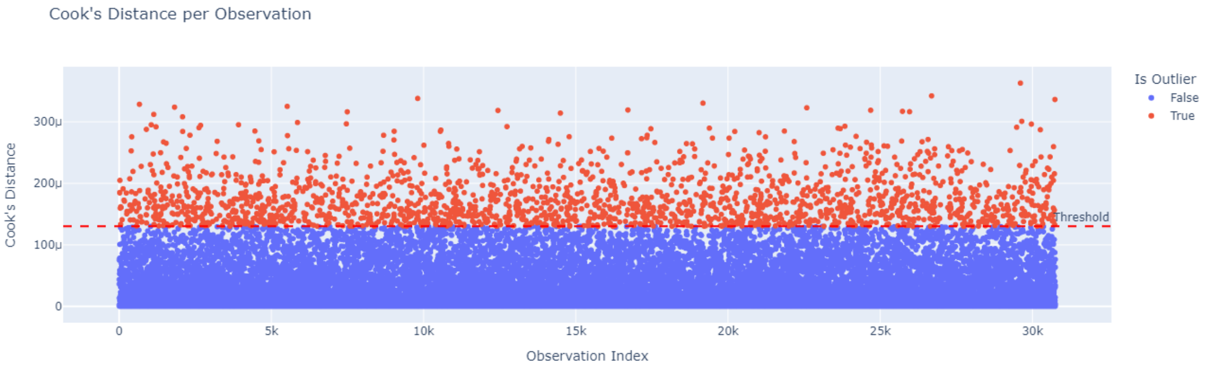


Figure 2: Cook's Distance and threshold

The second possible rule for the identification of outliers is to use $D_i < 4/n$, with n as the total number of observations in the dataset. The chosen approach is that using $4/n$ takes the size of the dataset into consideration, whereas setting arbitrary did not. The results of the analysis based on Cook's Distance is visible in the above plot.

Regardless of how valuable of a tool Cook's Distance is, it still comes with some limitations. The first is the measure's sensibility the sample size and the amount of predictors in the model. Specifically, if those two dimensions are large, difficulties arise in calculating the threshold to use for the identification of influential information. Secondly, the Distance does not provide information about the direction of the influence, or in other words whether an observation pulls the regression estimates towards itself or pushing them away from it.

3.1.2 MCD - Minimum Covariance Determinant and Mahalanobis Distance

The Minimum Covariance Determinant (MCD) method is a robust statistical approach that identifies outliers. The method estimates a robust mean μ_{MCD} and covariance matrix Σ_{MCD} by minimizing the determinant of the covariance matrix for a subset of size h (where $n/2 \leq h \leq n$) of the data. Once the robust estimates are obtained, the MCD-based Mahalanobis Distance is computed as:

$$D_{MCD}(x_i) = \sqrt{(x_i - \mu_{\text{MCD}})^T \Sigma_{\text{MCD}}^{-1} (x_i - \mu_{\text{MCD}})}. \tag{2}$$

Here, μ_{MCD} represents the robust center of the data, and Σ_{MCD} is the robust covariance matrix derived from the most central subset. By using these robust estimates, the MCD method ensures that the influence of outliers is minimized, leading to a more reliable identification of abnormal observations in multivariate data. As a result, the MCD-based distance offers a significant advantage when data contamination is present.

This ensures that the resulting location and scatter estimates are not unduly influenced by outliers, in contrast to traditional methods like the classical covariance matrix that can be heavily skewed. By focusing on the most "central" portion of the dataset, MCD effectively highlights points that deviate significantly from the core distribution. Unlike Cook's distance, which is primarily designed for regression-based influence assessment, MCD operates directly on the multivariate distribution of the data. Consequently, MCD is well-suited for general outlier detection in high-dimensional settings, whereas Cook's distance is specialized for diagnosing influential points within a linear regression context.

Mahalanobis distance provides a way to measure how many standard deviations a point is away from the multivariate mean of the data, taking into account the covariance structure. The Mahalanobis Distance (MD) for an observation x_i is defined as:

$$D_M(x_i) = \sqrt{(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)}, \tag{3}$$

where μ is the mean vector of the data, and Σ is the covariance matrix. This distance measures how far a point x_i is from the center of the distribution while accounting for correlations between variables. However, when outliers are present, the mean μ and covariance Σ can be significantly distorted, making the Mahalanobis Distance unreliable. By incorporating the covariance matrix, Mahalanobis distance can effectively capture relationships among variables and highlight observations that lie abnormally far from the data cloud. This makes it particularly useful in multivariate settings where dimensions are correlated. However, unlike the Minimum Covariance Determinant (MCD), which uses a robust covariance estimate resistant to outliers, Mahalanobis distance relies on the classical covariance matrix and may be more sensitive to extreme values.

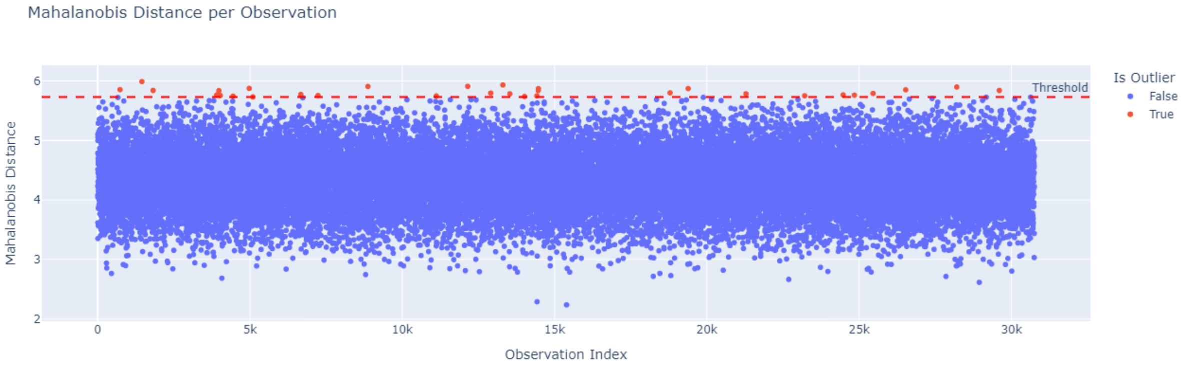


Figure 3: Mahalanobis' Distance over the Dataset

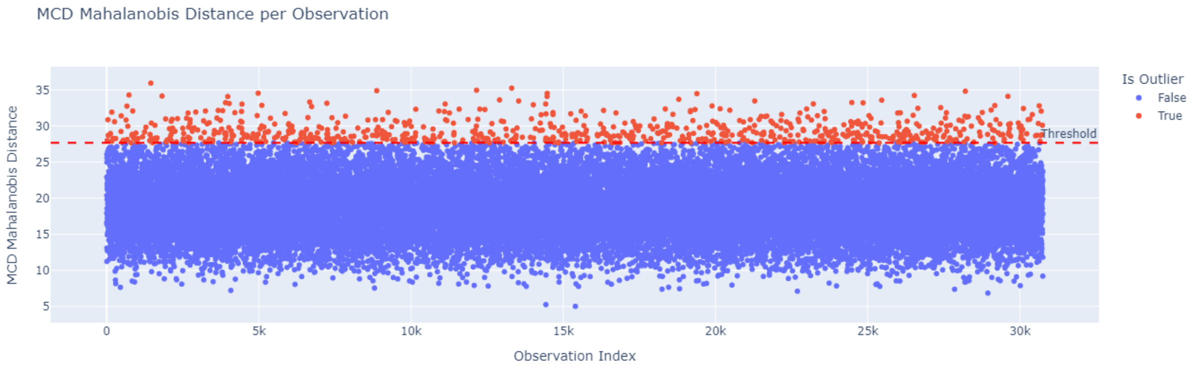


Figure 4: MCD-based Mahalanobis Distance

3.2 Unsupervised Learning

In this paragraph, we explore an unsupervised learning approach utilized for exploratory data analysis to uncover hidden patterns in the data that cannot be distinctly categorized. Data points can be grouped into sets, or clusters, based on shared characteristics. The core principle of cluster analysis lies in ensuring that objects within the same cluster are more similar to each other than to those in other clusters. Various clustering methods are available, each tailored to specific data types and the desired characteristics of the resulting clusters. Various clustering algorithms will be used to show which one is more effective in providing exploratory insights.

3.2.1 Principal Component Analysis and K-means

The Principal Component Analysis is a linear dimensionality reduction technique that takes in high-dimensional data and uses dependencies between the variables to represent it in a more tractable and lower-dimensional form, without losing too much information. In other words it extracts important information from the data table and expresses it as a set of new orthogonal variables called principal components, obtained as linear combination of the original variables. To summarize, PCA is carried out in the following steps:

- extract the most important information from the data table;
- compress the size of the data set by keeping only this important information;
- simplify the description of the data set;
- analyze the structure of the observations and the variables.

Once the PCA is performed the first clustering method employed was K-means, an iterative procedure that partitions N objects into K disjoint clusters by choosing the optimal number of centroids for cluster representation. There are three phases to complete the process:

- initialization phase which involves an iterative process where each data point is assigned to its nearest centroid using the Euclidean metric;
- update of the clusters' centroids given the partition of the previous phase;
- stop of the iterative process when no data point changes the cluster or the maximum number of iterations is reached.

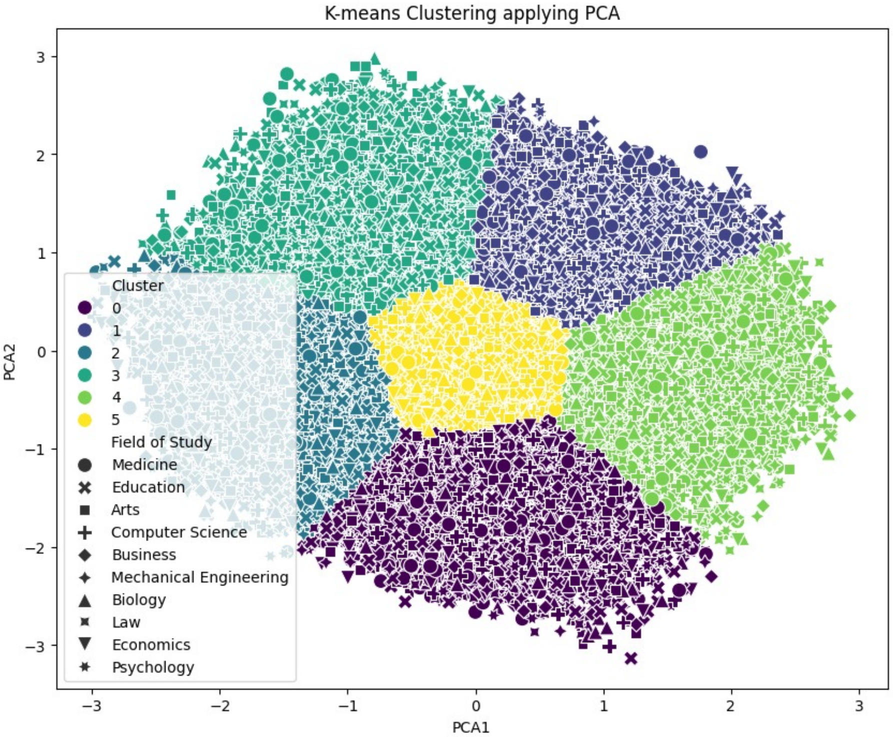


Figure 5: K-means clustering applying PCA

The optimal number of centroids selected is 6, as depicted in the clustering layout. The figure reveals that no clusters contain clearly defined attribute groups, as various fields of study are represented across all clusters. Since all possible fields appear scattered across each different cluster, the method was unable to distinguish a specific field for each cluster. Therefore, the colors do not correspond to any particular group.

3.2.2 Model Based Clustering

Model Based Clustering is a statistical approach that assumes data generated from a mixture of probability distributions and where each cluster is represented by a mixture distribution. In this case the distribution employed is the Gaussian distribution so the model is based on a Gaussian mixture model (GMM). The objective of GMM is to estimate the parameters of the Gaussian components and determine the probability of each data point belonging to each cluster, thus representing a soft-clustering process, conversely from what happened with K-means. The shape of the soft clusters obtained is an ellipsis where the mean parameters denote the position of the clusters while the co-variance matrices denote the shape and volume and orientation. This is achieved by implementing the Expectation-Maximization (EM) algorithm, which alternates between assigning data points to clusters and refining the parameters of the distributions. In the **E-step**, given the current parameter estimates π_k, μ_k, Σ_k at iteration t, the responsibilities $\gamma_{i,k}^{(t)}$ are computed as

$$\gamma_{i,k}^{(t)} = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)} \tag{4}$$

In the **M-step**, the parameters are updated using the responsibilities calculated in the E-step:

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{i,k}^{(t)}, \quad \mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{i,k}^{(t)} x_i}{\sum_{i=1}^{N} \gamma_{i,k}^{(t)}}, \quad \Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{i,k}^{(t)} (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{N} \gamma_{i,k}^{(t)}} \tag{5}$$

With

- π_k: The mixing coefficient for the k^{th} Gaussian component, representing the prior probability of a point belonging to cluster k.
- μ_k: The mean vector of the k^{th} Gaussian component, indicating the central location of cluster k.
- Σ_k: The covariance matrix of the k^{th} Gaussian component, characterizing the shape and spread of cluster k.
- $\mathcal{N}(x_i \mid \mu_k, \Sigma_k)$: The probability density function of a Gaussian distribution evaluated at the data point x_i given mean μ_k and covariance Σ_k.
- $\gamma_{i,k}^{(t)}$: Latent binary variable, which takes value 1 if x_i belongs to component k, and 0 otherwise
- N: The total number of data points in the dataset.
- K: The total number of Gaussian components (clusters).

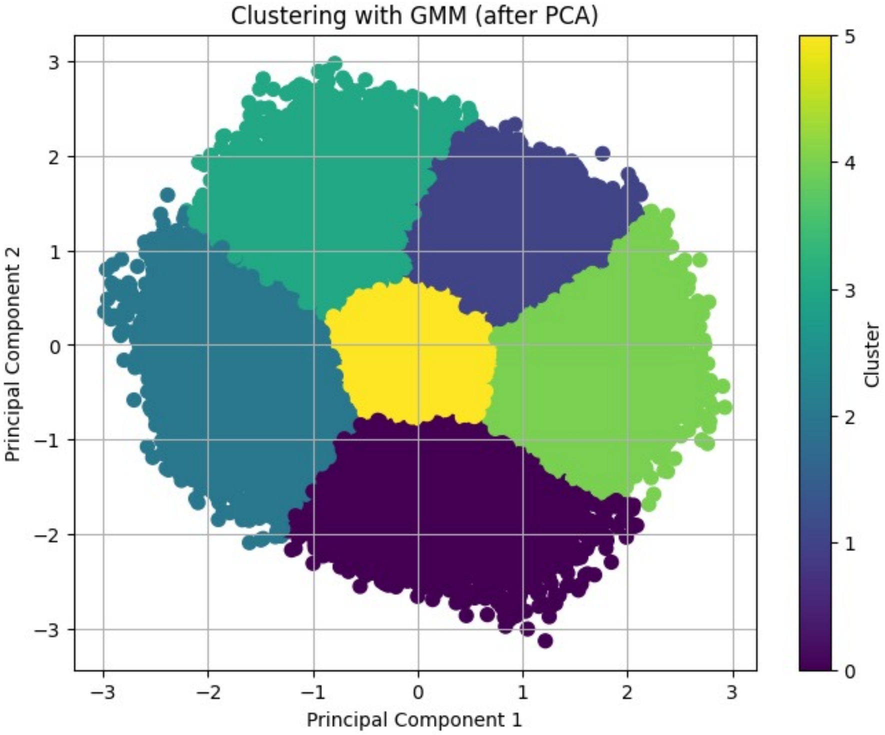
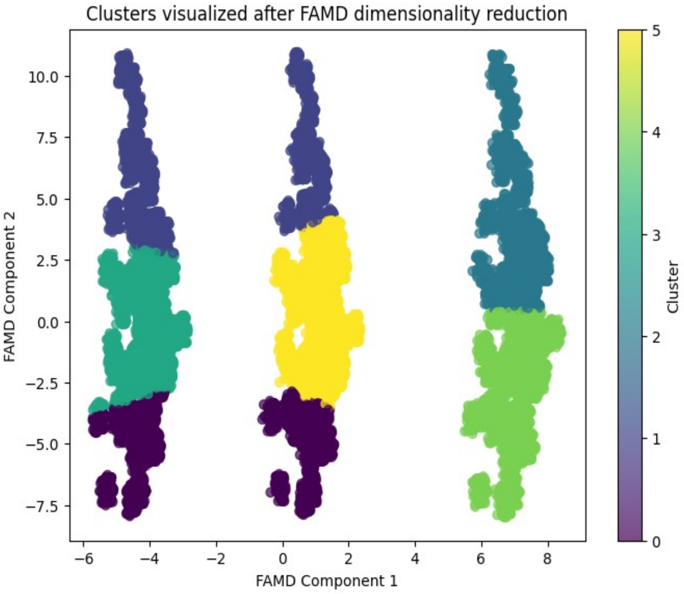


Figure 6: Clustering with GMM

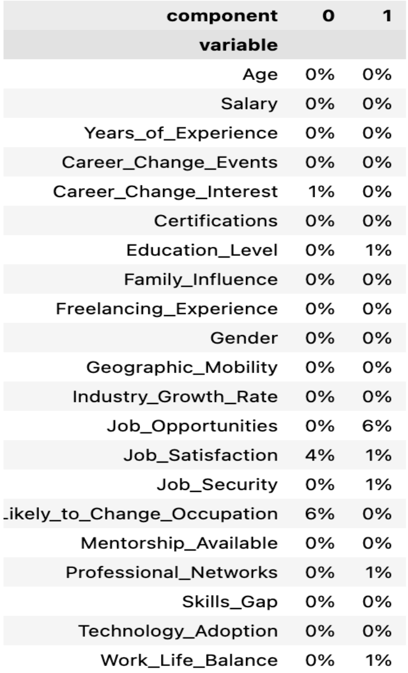
The representation of clustering with GMM confirms the results obtained in the previous K-means by showing a good identification of clusters but once again it is not possible to infer specific attributes or features that are unique to the clusters.

3.2.3 Factor analysis of mixed data

Factor analysis of mixed data is an advanced analytic method particularly used on datasets that contain a mix of categorical and continuous variables. The main goal of this technique is to model the observed variables as linear combinations of factors, which are unobserved latent variables. At the same time it performs a reduces the dimensionality of the dataset to a predetermined number of components while keeping the basic data structure and conveyed information intact. In short, FAMD combines two dimensionality reduction techniques: the Principal Component Analysis **(PCA)** and the Multivariate Component Analysis **(MCA)**. The continuous variables are observations represented as points in a Euclidean space, each one corresponding to a dimension, and the covariances between the variables describe the linear relationships between them. The categorical variables on the other hand, are treated by using the distance measure that is able to describe dissimilarities between categories. In our analysis the dataset was already prepared, therefore applying the method required simply reordering columns and tidying the data before performing the analysis.



(a) FAMD plot



component	0	1
variable		
Age	0%	0%
Salary	0%	0%
Years_of_Experience	0%	0%
Career_Change_Events	0%	0%
Career_Change_Interest	1%	0%
Certifications	0%	0%
Education_Level	0%	1%
Family_Influence	0%	0%
Freelancing_Experience	0%	0%
Gender	0%	0%
Geographic_Mobility	0%	0%
Industry_Growth_Rate	0%	0%
Job_Opportunities	0%	6%
Job_Satisfaction	4%	1%
Job_Security	0%	1%
Likely_to_Change_Occupation	6%	0%
Mentorship_Available	0%	0%
Professional_Networks	0%	1%
Skills_Gap	0%	0%
Technology_Adoption	0%	0%
Work_Life_Balance	0%	1%

(b) Column influence

Figure 7: FAMD for data reduction

The figure above shows how FAMD reduced the dataset into two components, each responsible for 2.60% and 1.49% of inertia i.e., the variance in our model. The two components represent the relationship between the variables and highlight how much each variable is able to contribute to explaining the variance. Of course, the best result could be obtained through larger percentages where the variables have very high contributions. In our case, the values together capture only 4% of the total variability of the dataset which implies that only a partial view of what is happening is enabled by this method, while higher dimensions explain the rest of the data. In short, the two components alone are not able to strongly separate and explain a large part of the overall variation. Unfortunately, we have seen that this is a common issue when it comes to datasets with many categorical features. As the figure above to the right shows, the percentages of contribution of each variable are all zero, with only the likelihood of changing occupation and the job satisfaction scoring 6% and 4%. The fact that there is hardly any contribution and no strong influence between the variables is due to the structural and qualitative issues within the dataset: the variables have little to no correlation with others.

The above plot shows the clusters obtained by FAMD. The shape of the resulting clusters is apparently defined well enough to outline possible subgroups within the real data. However, independently of how good the clusters look, they are not very informative due to the nature of the data itself. It is in fact impossible to associate the clusters to different real world group. With ideal data structure, some variables would have a much higher influence on the final result than others and it would be possible to identify different subgroups within the population to be associated to clusters. The results obtained are therefore of little meaning, as they are based upon distances calculated between sparsely correlated data. The unreliability of the result is thus to attribute more to the nature and structure of the data, rather than to the techniques employed.

3.2.4 Robust clustering: Trimmed K-means and T-clust

Trimmed K-means is a robust clustering method that is able to address the limitations of the traditional K-means by dealing with outliers and non-ideal data distributions. In fact, the traditional one assumes that all data points equally contribute to defining the clusters while the robust version trims a fraction of the data points at each iteration, allowing the algorithm to ignore a specified percentage of the farthest points when updating the cluster centroids. It's important to underline that the points that have been trimmed in the process, do not participate in centroid updates but they are re-evaluated at every iteration in case they might fit into the clusters.

- Initialization phase: Select K initial centroids using a designated method.
- Trimming phase: Calculate the distance between each point and its nearest centroid. Rank the points by proximity, and remove the top α-fraction of the farthest points.
- Iteration phase: Update the centroids and repeat the trimming process until the centroids stabilize or the maximum number of iterations is reached.

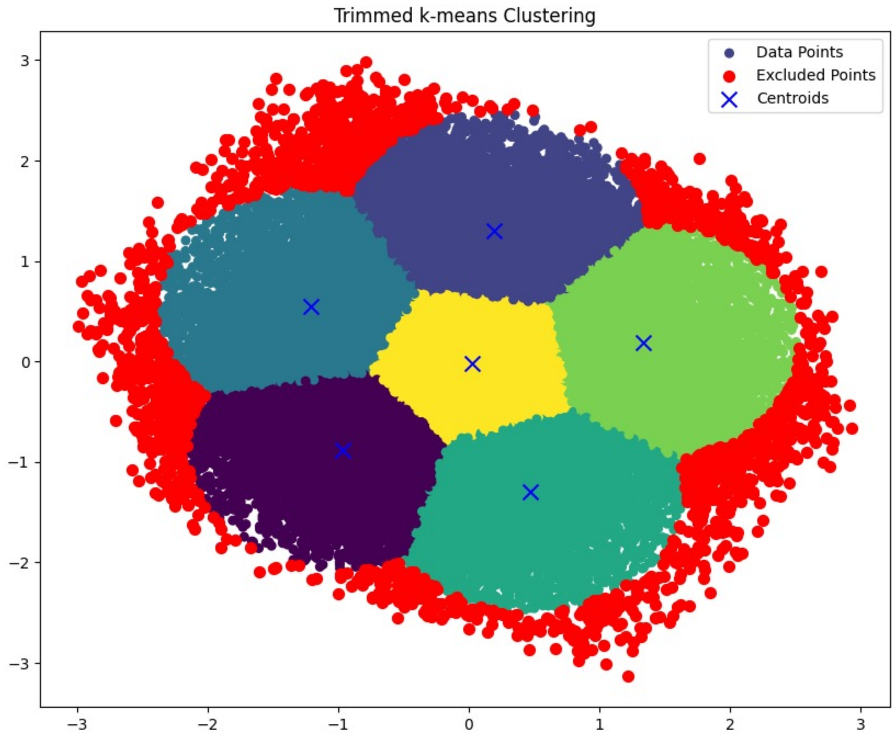


Figure 8: Trimmed K-means

The primary distinction between clustering with GMM and trimmed K-means is that the robust approach can identify and exclude outliers (shown in red) during the clustering process. In addition, a second robust clustering method, T-clust, was used in the analysis. T-clust is similar to trimmed K-means but offers greater flexibility due to its estimation of the covariance structure for each cluster. Furthermore, instead of using Euclidean distance, T-clust employs the Mahalanobis distance. While trimmed K-means tends to form circular clusters, T-clust is capable of creating elliptical clusters.

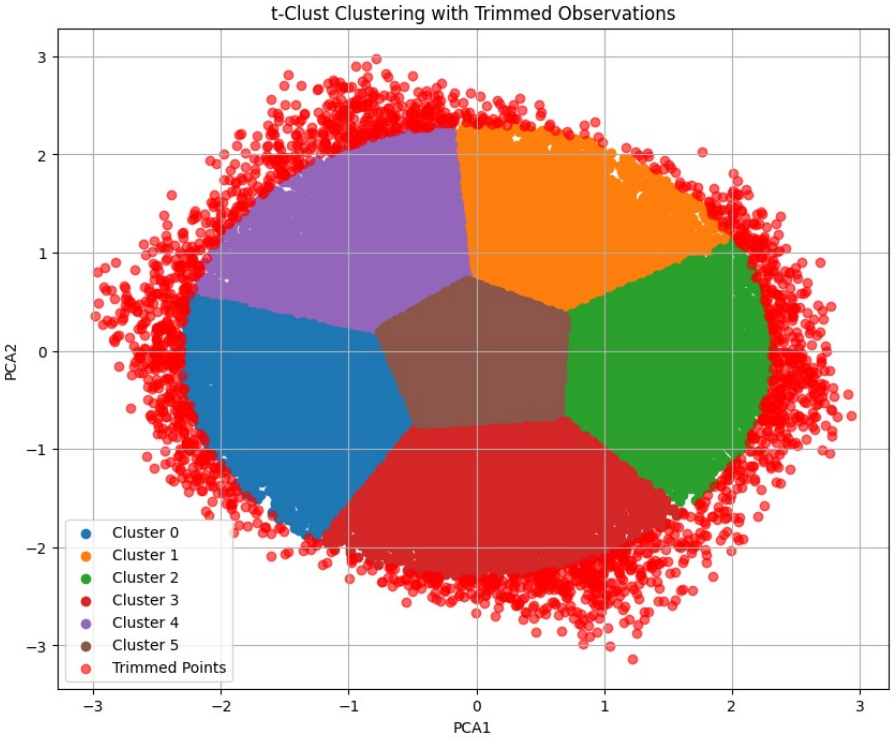


Figure 9: T-clust

3.3 Supervised Learning

In this section supervised learning techniques are employed to acquire the input-output relationship information of a system based on a given set of paired input-output training samples. Given the output regarded as the label of the input data or the supervision, an input-output training sample is also called labelled training data, or supervised data. The objective of supervised learning is to develop an artificial system capable of learning the relationship between inputs and outputs, with the ultimate of predicting outputs for new inputs. When the output consists of a finite set of discrete values representing class labels, the learned mapping performs classification of the input data. Conversely, when the output takes continuous values, the mapping corresponds to regression. The input-output relationship is often described through model parameters, which encapsulate the learning process. If these parameters cannot be directly derived from the training data, the system must undergo an estimation process to determine them.

3.3.1 Logistic Regression

Logistic regression, also called the logistic model or logit model, analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve. It is a technique used to fit a regression model, $y = f(x)$, when the response variable y is binary-coded (with 0 or 1,representing failure and success, respectively). When the response is a binary (dichotomous) variable and x is numerical, logistic regression models the relationship between x and y using a logistic curve. The logistic curve, also known as a sigmoid curve, is S-shaped. This curve begins with slow, linear growth, transitions into exponential growth, and eventually slows down to a stable rate.

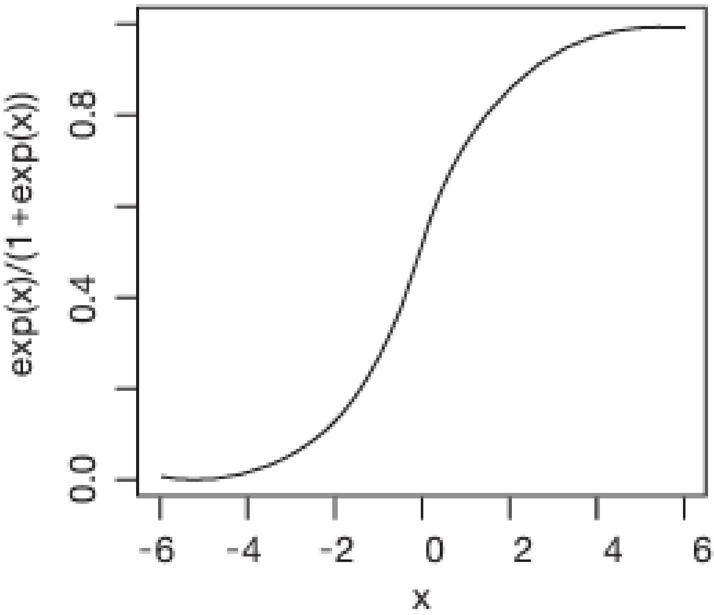


Figure 10: Logistic curve with $\alpha = 0$ and $\beta = 1$

To enhance the accuracy of the previously presented logistic regression model, several techniques were applied. Among the various algorithms tested, the most effective one in technical terms was the **Backwards algorithm**, which involves iteratively eliminating the least significant variable. The variable chosen for removal is the one with the highest p-value. After identifying this variable, two criteria had to be met for it to be removed from the model:

- the p-value had to be higher than 0.1;
- the t-value had to be smaller than 2.

With t-value being

$t = \frac{\text{coeff}}{\text{stderr}}$

3.3.2 Logistic Regression with HC1 Covariance Type

The robust logistic regression was performed with a method that computes standard errors by using heteroscedasticity consistent covariance estimators. This method provides more reliable standard error estimates by correcting for non-constant variance in the error terms. The HC1 estimator adjusts the covariance matrix by applying a correction factor to the usual estimates, based on the residuals of the model. This adjustment improves the robustness of statistical inference, ensuring that hypothesis tests and confidence intervals remain valid even when the assumption of homoscedasticity is violated.

3.3.3 Random forest

Random forests are an ensemble method that combines multiple decision trees to improve predictive performance. Each tree in the forest is trained independently using a random vector, which is sampled from the same distribution for all trees. These random vectors influence the growth of the trees by introducing randomness in both the selection of training data and the feature subsets considered for splitting at each node. The use of random sampling in both the data and features helps ensure diversity among the individual trees, which, in turn, leads to a more robust and accurate overall model. By aggregating the predictions of all the trees, random forests reduce the risk of overfitting. More specifically a random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(x, k), k = 1, \dots\}$ where the $\{k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

3.3.4 Mixed Effect Model

It's a class of statistical models that combines both fixed and random effects to analyze data with complex structures. Fixed effects represent the systematic influences or predictors that are consistent across all units in the data, such as the impact of a treatment or a specific characteristic of the population. Random effects, on the other hand, account for variability across different levels of the data, such as individual differences or group-specific deviations. By modeling both, mixed-effect models are usually able to provide a more flexible and accurate representation of the data.

3.3.5 Evaluation of the models

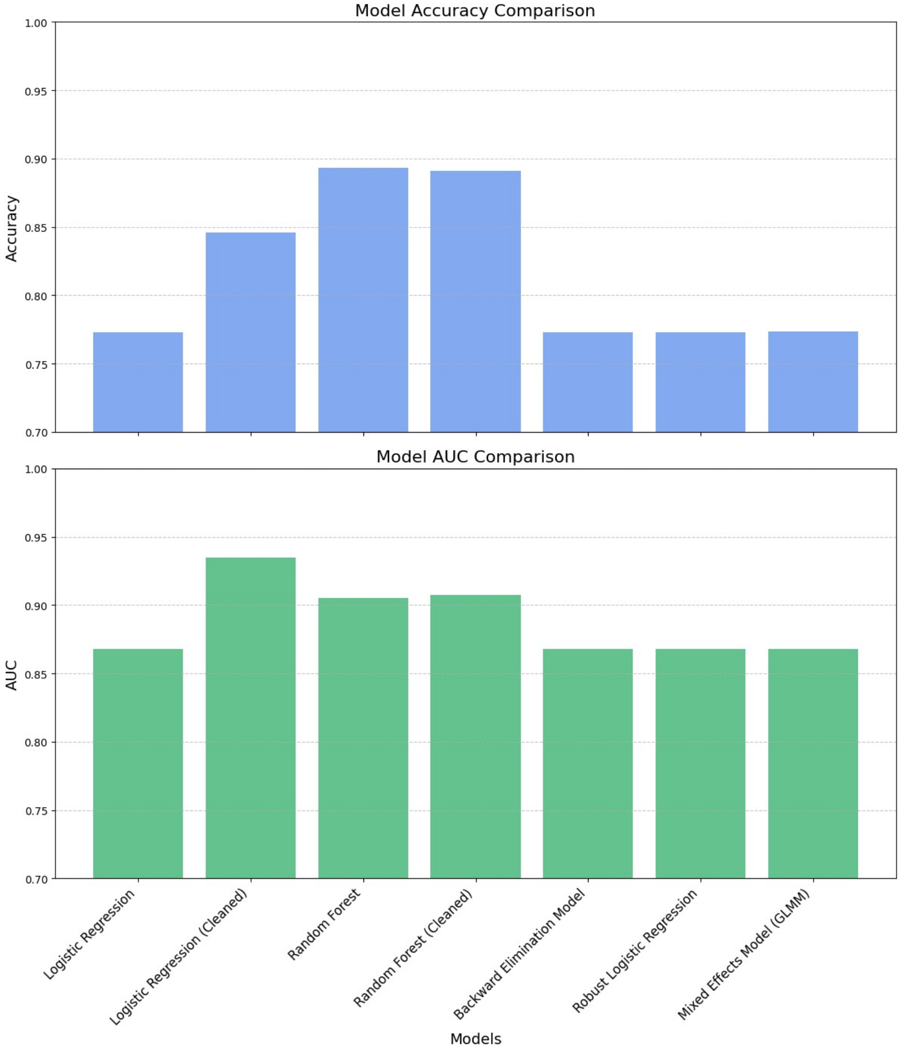


Figure 11: Comparison between employed models

The above graphs represent the culmination of the supervised analysis. In fact, the two plots represent the Accuracy and the Area Under the Curve of each of the models used. Accuracy is the percentage of correct classifications obtained by a trained model, whereas the AUC represents the probability that the model, if given a randomly chosen positive and negative example, will rank the positive higher than the negative. Moving on to the evaluation of the models used, it is possible to see that the Random Forest and Backward Elimination models both achieve the highest accuracy, nearing 0.90, while other models, such as the baseline Logistic Regression, the Mixed Effects Model (GLMM), and the Robust Logistic Regression, hover around a significantly lower 0.77–0.78. Notably, the cleaned Logistic Regression model shows a moderate improvement over its uncleaned counterpart, reaching about 0.85, highlighting the substantial impact of data preprocessing. Turning to AUC, the cleaned Logistic Regression outperforms all other models, attaining a value close to 0.94 and surpassing even the Random Forest models, which both maintain respectable AUCs of around 0.91. Surprisingly, models like the Backward Elimination and Robust Logistic Regression do not see similar gains in AUC, suggesting that robustness does not necessarily ensure improved discrimination. These insights underscore the importance of preprocessing and careful model selection: while the Random Forest provides a strong balance of accuracy and AUC, a cleaned Logistic Regression stands out for tasks that prioritize AUC, and robust methods may not always deliver the expected performance improvements.

4 Summary and Conclusions

4.1 Unsupervised Learning

Unsupervised learning techniques were applied throughout the project to uncover hidden patterns in the data, attempting at understanding the complications brought by its mixed categorical and continuous nature. The following methods were employed:

- **K-means Clustering with PCA**: Principal Component Analysis (PCA) reduced dimensionality by extracting key components. However, K-means clustering on these components failed to produce meaningful groupings. All clusters exhibited overlapping attributes, with no distinct separation of fields or job characteristics.
- **Gaussian Mixture Models (GMM)**: GMM offered soft clustering by estimating probability distributions and accounting for covariance structures. Despite its flexibility, GMM results aligned with K-means, showing no clear clusters or associations within the data.
- **Factor Analysis of Mixed Data (FAMD)**: To handle mixed data types, FAMD combined PCA and MCA. While it reduced dimensionality, the two components explained only 4% of the total variance. This limitation highlighted the weak inter-variable relationships exsiting in the dataset and its structural issues.
- **Robust Clustering**: Trimmed K-means: Trimmed K-means addressed outliers by excluding a fraction of distant points during iterations. While this improved centroid stability, the resulting clusters remained constrained by the data lack of structure.

In summary, unsupervised learning methods, while theoretically sound, provided limited insights due to the dataset's poor structure. The absence of strong inter-variable relationships prevented the identification of clear patterns. Robust methods like Trimmed K-means handled anomalies effectively but could not overcome the inherent dataset's weaknesses. However the correct application of the methods mentioned enabled for significant discoveries to be made within the data.

4.2 Supervised Learning

Several supervised learning techniques were employed during the study to evaluate the importance and relationships of the variables, as well as to identify the optimal model. An additional objective was set to deepen the understanding of robust techniques by comparing the pre-processing of variables with the application of robust logistic regression. Here is possible to find all the methods implemented:

- **Logistic Regression**: Logistic regression served as a baseline due to its simplicity and interpretability. Backward elimination was utilized to iteratively remove variables with low statistical significance, resulting in a streamlined model focused on the most impactful predictors. While the cleaned model improved performance, its accuracy was constrained by the weak structure of the dataset, highlighting the importance of robust variable selection in such scenarios.
- **HC1 Covariance-Adjusted Logistic Regression**: To address heteroscedasticity and non-constant residual variance, HC1 covariance adjustments were applied to logistic regression. This approach provided more reliable statistical inference, improving confidence intervals and hypothesis testing. However, the model's predictive capabilities remained comparable to the baseline, as the adjustments could not compensate for the lack of strong inter-variable relationships.
- **Random Forest**: Random forest, an ensemble method, leveraged multiple decision trees to capture non-linear interactions and complex patterns in the data. It achieved the highest accuracy among all models, nearing 90%. Despite this, its AUC was marginally lower than that of the backward-eliminated logistic regression, illustrating that even sophisticated methods may not always outperform simpler approaches in terms of discriminative power.
- **Mixed Effect Model**: Mixed effect models were used to account for both fixed and random effects, aiming to capture variability across potential group structures in the data. However, the dataset's lack of clear hierarchical structures limited the utility of this method, and its performance did not surpass simpler models like logistic regression.

In summary, supervised learning approaches highlighted the critical role of preprocessing, particularly feature selection, in improving model performance. Logistic regression with backward elimination demonstrated the strongest discriminative ability (highest AUC), while random forests excelled in accuracy by effectively handling the dataset's noise and complexity. Robust methods, such as HC1 adjustments and mixed effect models, provided theoretical stability but yielded only incremental improvements given the dataset's structural limitations.

4.3 Supervised Learning 2

Supervised learning techniques were employed to examine variable importance, identify optimal predictive models, and further explore robust modeling approaches, particularly when combined with careful variable preprocessing. The following models were considered:

- **Logistic Regression**: Logistic regression served as a baseline, valued for its simplicity and interpretability. Backward elimination removed statistically insignificant variables, creating a more focused model. Although the cleaned model improved performance, its accuracy was still limited by the dataset's weak structure, emphasizing the need for robust feature selection.
- **HC1 Covariance-Adjusted Logistic Regression**: To address heteroscedasticity, HC1 covariance adjustments were applied. This improved the reliability of statistical inference (e.g., more stable confidence intervals), but did not significantly boost predictive performance, indicating that refined statistical corrections cannot fully compensate for poor inter-variable relationships.
- **Random Forest**: An ensemble of decision trees, the Random Forest model captured complex patterns and interactions, yielding the highest accuracy (close to 90%). Nevertheless, its AUC remained slightly lower than that of the backward-eliminated logistic regression model, suggesting that more sophisticated methods do not always offer superior discriminative power.
- **Mixed Effect Model**: By accounting for both fixed and random effects, mixed effect models sought to capture variability related to potential hierarchies within the data. However, due to the absence of clear hierarchical structures, their performance did not surpass simpler alternatives like logistic regression.

In summary, supervised learning approaches highlighted the importance of preprocessing, particularly variable selection, in enhancing model performance. Logistic regression with backward elimination achieved the highest AUC, while the Random Forest excelled in accuracy and handled the dataset's complexity effectively. Although robust methods, such as HC1 adjustments and mixed effect models, offered theoretical advantages, they only provided marginal improvements given the dataset's inherent structural limitations.

List of Figures

1 Exploration of Variables 2

2 Cook's Distance and threshold 5

3 Mahalanobis' Distance over the Dataset 6

4 MCD-based Mahalanobis Distance 7

5 K-means clustering applying PCA 8

6 Clustering with GMM 10

7 FAMD for data reduction 11

8 Trimmed K-means 13

9 T-clust 14

10 Logistic curve with $\alpha = 0$ and $\beta = 1$ 15

11 Comparison between employed models 17