TECHNICAL UNIVERSITY OF CRETE
SCHOOL OF PRODUCTION ENGINEERING AND MANAGEMENT
MANAGEMENT SYSTEMS LABORATORY

Anticipated Emotions in Nascent Entrepreneurship: A machine learning analysis

M.Sc. Thesis

Anastasia Koufaki

Thesis Committee:
Professor Vassilis Moustakis, Thesis Supervisor
Professor Tom Kontogiannis
Associate Professor Konstantinos Kafetsios

Chania, June 2016

Αναμενόμενα συναισθήματα στα πρώιμα στάδια της επιχειρηματικότητας: Διερεύνηση μέσω μεθόδων μηχανικής μάθησης

Η εφαρμογή μεθόδων μηχανικής μάθησης με σκοπό την ανάλυση δεδομένων αποτελεί νέα τάση στο πεδίο της επιχειρηματικής ανάλυσης. Σε αυτή την διπλωματική εργασία προτείνεται ένα μεθοδολογικό πλαίσιο, το οποίο στοχεύει στην αναγνώριση σημαντικών συσχετίσεων και λειτουργικών σχέσεων ανάμεσα στα αναμενόμενα συναισθήματα που κάποιος αισθάνεται κατά την πρώιμη επιχειρηματική δραστηριότητα και στις μεταβλητές αξιολόγησης που αντιπροσωπεύουν επιχειρηματικά χαρακτηριστικά, όπως το μέγεθος της παρακίνησης που αισθάνεται για να αρχίσει να αναπτύσσει το δικό του κεφάλαιο (Επιχειρηματικές προθέσεις, Entrepreneurial Intentions - INT), τη στάση του απέναντι στην επιχειρηματικότητα (Attitude Towards Entrepreneurship - ATT) και τις διακριτές δεξιότητες που πιστεύει ότι έχει για την ανάπτυξη πρώιμης επιχειρηματικής δραστηριότητας (Αντιλαμβανόμενος Συμπεριφορικός Έλεγχος, Perceived Behavioral Control - PBC)

Τα δεδομένα που χρησιμοποιήθηκαν στην εργασία αποτελούν ένα υποσύνολο δεδομένων τα οποία συλλέχθηκαν για προηγούμενη μελέτη. Εκείνη η μελέτη στόχευε στην εξέταση των προσδοκιών που έχουν φοιτητές σχετικά με την καινοτομία και την ανάπτυξη κεφαλαίων. 1160 φοιτητές πανεπιστημίων από όλη την Ελλάδα απάντησαν σε ερωτηματολόγιο από το οποίο έγινε εξαγωγή των τριών προαναφερθέντων επιχειρηματικών χαρακτηριστικών και των αναμενόμενων συναισθημάτων που νιώθουν σχετικά με την έναρξη επιχειρηματικής δραστηριότητας.

Για την υλοποίηση του προτεινόμενου μεθοδολογικού πλαισίου χρησιμοποιήθηκαν επιβλεπόμενες (supervised) και μη-επιβλεπόμενες (unsupervised) μέθοδοι μηχανικής μάθησης.
Στο **πρώτο** μέρος της ανάλυσης δεδομένων χρησιμοποιούνται αλγόριθμοι από το πεδίο της μη-επιβλεπόμενης μάθησης. Έγινε αυτή η επιλογή προκειμένου να πραγματοποιηθεί διερευνητική ανάλυση δεδομένων, για τον εντοπισμό κρυφών δομών στα δεδομένα με σκοπό να συγχωνευθούν παρόμοια συναισθήματα σε ομάδες.
Στο **δεύτερο** μέρος της προτεινόμενης μεθοδολογίας χρησιμοποιούνται αλγόριθμοι από το πεδίο της επιβλεπόμενης μάθησης, με σκοπό να δημιουργηθούν και να εκπαιδευθούν ταξινομητές (classifiers) που θα επιτρέψουν να προβλεφθούν οι κατηγορίες συναισθημάτων (θετικά - αρνητικά) με βάση χαρακτηριστικά που αντιστοιχούν σε επιχειρηματικές προθέσεις.
Το **τρίτο** μέρος της μεθοδολογίας στοχεύει στην ορθή αξιολόγηση της απόδοσης των ταξινομητών (classifiers) μέσα από μετρήσεις ακρίβειας (accuracy), ευαισθησίας (sensitivity) και ειδικότητας (specificity). Στο βήμα αυτό αναδεικνύονται τα ζητήματα που προκαλούνται από την ανισορροπία των δεδομένων, η οποία μπορεί να οδηγήσει σε λανθασμένη ερμηνεία της απόδοσης της μεθόδου ταξινόμησης.
Το **τέταρτο** μέρος της ανάλυσης περιλαμβάνει επιλογή χαρακτηριστικών, προκειμένου να προσδιορίσει ποια από τα τρία επιχειρηματικά χαρακτηριστικά συνδέονται σε σημαντικό βαθμό με την προβλεπόμενη κατηγορία συναισθήματος.
Το προτεινόμενο πλαίσιο υλοποιήθηκε στη γλώσσα R.

Το βασικό σημείο που προκύπτει από τα αποτελέσματα της παραπάνω μεθοδολογίας είναι ότι μέθοδοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν με επιτυχία για την πρόβλεψη θετικών ή αρνητικών συναισθημάτων που βασίζονται σε μεταβλητές αξιολόγησης (επιχειρηματικά χαρακτηριστικά). Επιπλέον, η στάση απέναντι στην επιχειρηματικότητα (ATT) αναγνωρίζεται ως το πιο σημαντικό επιχειρηματικό χαρακτηριστικό που προκαλεί θετικά ή αρνητικά συναισθήματα στα πρώιμα στάδια επιχειρηματικότητας.

Όσον αφορά μελλοντικές εργασίες, αφού παρατηρείται ότι η ανισορροπία των δεδομένων αποτελεί σημαντικό παράγοντα που περιορίζει την απόδοση της ταξινόμησης, θα μπορούσαν να εφαρμοστούν μεθοδολογίες ειδικές για ανισορροπία, όπως τυχαία υπό-δειγματοληψία. Τέλος , νέα σετ δεδομένων, κατά προτίμηση σετ δεδομένων που δεν πάσχουν από ανισορροπία, θα μπορούσαν να επεξεργαστούν με το ίδιο μεθοδολογικό πλαίσιο, προκειμένου να αξιολογηθούν περαιτέρω τα ευρήματα αυτής της μελέτης.

Abstract

The application of machine learning methodologies for data analysis is the new trend in the field of business analytics. In this thesis, a methodological framework is proposed, which aims to identify significant connections and functional relationships between the anticipated emotions someone feels in nascent entrepreneurship and variables that represent entrepreneurial features i.e. the level of motivation they have to start developing their own venture (Entrepreneurial Intentions - INT), the Attitude Towards Entrepreneurship (ATT) and the perceived skills towards nascent entrepreneurship (Perceived Behavioral Control - PBC).

The data used are a subset of data collected in a previous study that aimed to examine the expectations of students towards entrepreneurship and venture growth. 1160 university students responded to a questionnaire from which the aforementioned three entrepreneurial features and the anticipated emotions that are examined in this study were extracted.

Supervised and unsupervised machine learning methods are utilized in the proposed framework. The first part of the data analysis process of this study utilizes algorithms from the field of unsupervised learning, in order to conduct an exploratory data analysis, to identify hidden structure in the data and to merge similar emotions into clusters. The second part of the proposed methodology utilizes algorithms from the field of supervised learning in order to build and train classifiers which enable us to predict emotion classes based on features that correspond to entrepreneurial intentions The third step of the methodology aims at correctly assessing the performance of the classifiers through the metrics of accuracy, sensitivity and specificity and highlights the issues caused by class imbalance in the data, which can lead to misleading interpretation of the classification accuracy. The fourth step of the analysis includes feature selection, in order to identify which of the three entrepreneurial intention features are significantly associated to the emotion class being predicted. The proposed framework is implemented in the R language for statistical computing.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Related Work

Inspiration

Our inspiration concerning data analysis using machine learning methodologies came from a paper by Meuleman & Scherer (2013) "Nonlinear Appraisal Modeling: An Application of Machine Learning to the Study of Emotion Production" [1]. In that study, the data were collected from a web-based expert system GEA (Geneva Emotion Analyst). The system asked each participant, first to describe an intense emotion event, then to choose from a list of 12 emotions which one described better the primary emotion they felt during the event and then to assess that emotion in a six-point scale with 25 appraisal variables. The sample of the study consisted of 6034 respondents. Their data analysis had three stages: (1) cluster analysis, (2) black box modeling, (3) appraisal feature selection.

First, they performed cluster analysis in order to find the common structure among the 12 emotion classes, based in the score the respondents assess in the 25 appraisal variables. They obtained mean appraisal profiles by calculating the mean value of each appraisal variable for each emotion class (a 12x25 matrix, each row represents each emotion class centroid vector). The method used was hierarchical clustering, in order to perform a non-flat examination of the structure of the emotional classes.

Second, in the black box modeling stage they selected 14 classification machine learning methodologies, 4 linear and 10 nonlinear, in order to find a model which better describes that kind of data. They modeled the 12 emotion classes (categorical data), as a function of the 25 appraisal variables (six-point scale - ordinal data). They used a training set of 1200 samples (100 respondents of each emotion class) and the rest of the sample was the validation set. The accuracy of each model was assessed only in the validation set, in order to avoid overfitting. The predictive performance of a fitted model was calculated as the bootstrapped average hit rate across the 12 emotion classes. They also performed t-tests in order to compare the top rated models accuracy formally.

In the last stage, feature selection, they performed regression analysis (LASSO – Least Absolute Shrinkage and Selection Operator) in order to find out which of the 25 appraisal variables were significant.

Data

For this study, the data were provided from a previous statistical analysis study [2]. In that previous study, 1160 university students responded to a questionnaire regarding the expectations they had towards entrepreneurship and venture growth, similar to the one in the Appendix A section. The questionnaire is divided in seven sections:
(A) demographic information (categorical data),
(B) assessment of 2 entrepreneurial intention features: Attitudes Towards Entrepreneurship (ATT) and Entrepreneurial Intention (INT), in a seven-point scale (ordinal data),
(C) assessment of the perception the respondents – students had regarding the level of difficulty when they begin to perform action in order to start their new business in a five-point scale (ordinal data),
(D) assessment of the emotions they feel when they thought of starting their own business in 67 emotion adjectives in a five-point scale (ordinal data),
(E) assessment of the approval from the immediate environment when the student mentions the thought that he or she wants to start his or her own business in a five-point scale (ordinal data),
(F) assessment of the entrepreneurial intention feature, called Perceived Behavioral Control (PBC) in a seven-point scale (ordinal data),
(G) assessment of the degree of positive attitude towards entrepreneurship depending on negative attitude (grid table).

In this study the data set is formed from sections B and F in order to obtain the score of the students in 3 entrepreneurial features and from section D to obtain the primary emotion they felt when they were thinking about starting their own business.

1.2 Thesis Outline and Innovation

Scope

The aim of this study is to perform a machine learning analysis of data that describe anticipated emotions in nascent entrepreneurship. Namely, we aim to find any significant connections between the emotions someone feels, when beginning to think of starting their own business, and their intentions in entrepreneurship.

In this research, a particular situation is examined. That is, when an individual is thinking to start a business. Each individual appraises this situation differently and forms a personal pattern. Depending on that pattern, either positive or negative emotions are elicited. This study aims to assess whether state of the art machine learning methodologies can be successfully utilized to identify an appraisal pattern which is responsible for the elicitation of, and can be used to predict, positive or negative emotions. In addition, the appraisal variables (three entrepreneurial features) which form the patterns are evaluated through feature selection in order to identify possible markers of nascent entrepreneurship.

Strategy and Innovation

This study is inspired by the methodology of Meuleman & Scherer [1]. Contrary to [1], this study does not investigate intense emotional events in general but focuses specifically to the case where the emotional event is the intention for entrepreneurship. Similar to [1] Hierarchical clustering is applied in order to find similar subgroups of emotions and then supervised learning methods are utilized to predict the emotion felt based on entrepreneurial features. However, in contrast to [1], not only LASSO is used to reduce the number of entrepreneurial features but also the features selected by LASSO are further examined using t-tests in order to see whether they differ significantly between the different emotion classes.

2. Theoretical Background

2.1 The Appraisal Theory of Emotions

Appraisal theory, namely claims that emotions are elicited by evaluations of events or situations. An individual appraises a situation and therefore an appraisal pattern of the situation is created. That pattern elicits either positive or negative emotions. For example, in response to the successful score in a University exam, some students are feeling relief, some others happiness and others have no emotion response. However, some students can also feel anger or sadness because even though they succeed, the score was not the anticipated one. Appraisal theories were developed/proposed in order to help researches solve particular problems that other emotion theories (behavioral, psychological) had difficulties to explain. [3]

In the beginning, appraisal theorists claimed that there is a strong and invariable one to one relationship between the appraisal pattern and the elicited emotion of a specific situation. In addition, researchers claimed that appraisal theories aid in developing/creating appraisal patterns which have a strong and invariable relationship with a particular emotion. In recent research, appraisal theorists concluded that emotion responses are organized and adapted in particular external and internal circumstances of a situation that triggers the emotion. [4][5]

The appraisal of emotions can be elicited from a cognitive procedure. That procedure it is not necessary automatic, usually an evaluation of a situation can be controlled by logic and the emotion outcome is biased from the attitude of each individual, for example different patterns could be created if a person is an optimist or a pessimist. [6] The evaluation of a specific situation (imaginary or not) is generating the emotion process by initiating changes in behavior, attitude and intentions as a result of a specific emotional state. [4][7][8][9] If a situation is appraised as motive inconsistent then the emotion that it elicits is more likely negative, if a situation is evaluated as motive consistent then the emotion that it elicits is more likely positive. [10]. Many researchers had accused appraisal theories that they cannot account unreasonable or involuntary emotional responses. Appraisal theorists [4][11][12][13] also mention that although appraisal involves cognitive, logical and complex processing, it also involves low level cognitive processes (unconscious thoughts also elicit emotions). These emotions are usually irrational and unrelated to the appraisal pattern of cognitive procedure when a situation occurs.

One situation is evaluated differently, each individual creates a different appraisal pattern for the same situation and these differences in evaluation elicited different emotions too. In this research the situation which is examined is nascent entrepreneurship, when someone starts to think about the possibility of creating his own business, emotions are elicited from the appraisal of the situation.

2.2 Machine Learning and Pattern Recognition

In this study we obtain an outcome measurement of categorical data (six emotion classes), that we wish to predict based on a set of features (three indicators of entrepreneurial intentions).

We have a training set of data in which we observe the outcome and the feature measurements for a set of objects (students). Using this set we build a learner (prediction model) which will enable us to predict the outcome for new unseen objects. A good learner is the one that accurately predicts such an outcome.

Supervised learning algorithms [14] solve regression and classification problems. Namely, applying a supervised learning algorithm in a classification problem helps predict (aids in predicting) the probability of a right classification of a new sample in an existing learning model.

Unsupervised learning algorithms [14] aid in finding a structure description of an existing sample of data based in characteristics features of the data. That procedure is important in mapping the relevant associations, which are formed among the data, depending a similarity function.

2.3 Clustering Methods - Hierarchical Clustering

First, we aim to observe the similarities among the mean score in the 3 entrepreneurial features of the 6 emotion classes. It is significant to show how this 6 emotion classes are related based only on the mean score of the sample in the 3 entrepreneurial features without the suggestion which is proposed from the appraisal emotion theory. (unsupervised learning problem)

Hierarchical Clustering [14] [15] is a method, that finds hierarchical structure (grouping) in the data, in a non-flat way. Namely, when clusters have sub-clusters, which have sub-clusters and so on. In order to better understand the utility of this method, think of a biological taxonomy. Biologists classify a particular organism in a hierarchical structure in order to analyze the relations of that organism with others in the ecosystem. The results of a hierarchal cluster analysis usually they are presented by a dendrogram.

There are two approaches for the hierarchical clustering procedures: (a) agglomerative, and (b) divisive. The case of agglomerative clustering is based on a bottom-up approach where each single observation is initially considered to be a single cluster and in each subsequent step the two “most similar” clusters are merged until a predefined number of cluster remains, usually a single cluster. Then, the whole process of iteratively merging similar clusters can be visualized using a dendrogram. Three different similarity criterions for clusters are used in practice: (1) single-linkage (nearest neighbor approach) which tends to produce elongated clusters (chaining), (2) complete linkage (furthest neighbor approach) which avoids chaining but might violate the closeness property i.e. members of the merged cluster might lie closer to members of other clusters than members of their own cluster and (3) average linkage which is a compromise between single and complete linkage. The case of divisive clustering consists of a top down approach, where in the first step all data points are considered to belong to a single cluster and the hierarchy is formed by subsequently splitting clusters.

Hierarchical Clustering procedures are among the best known of unsupervised methods, because of their conceptual simplicity. But, they have lower efficiency compared to other methods of unsupervised learning. That low efficiency comes from the computational complexity $O(n^2)$ of hierarchical clustering algorithms, which makes them impractical for large data sets. Nevertheless, they consist a very popular tool of exploratory data analysis due to their ability to efficiently visualize similarities and possibly latent hierarchies of subgroups within the data, especially if the data to be analyzed do not correspond to “big data”, where computational complexity would be an issue.

Ward's Method

The Ward’s method [16] [17] is a criterion applied in hierarchal agglomerative cluster analysis. Ward 's method, called also minimum variance clustering, is a popular algorithm which minimizes the total variance within a formed cluster. In each step, it selects the merge with the smallest Residual Sum of Squares (RSS), a measure of how well the centroids represent the members of their cluster. the squared distance of each vector from its centroid summed over all vectors

$$RSS_k = \sum_{x \in \omega_k} |x - \mu(\omega_\kappa)|^2 // \mathrm{RSS} = \sum_{k=1}^{k} RSS_k.$$

The merge criterion in Ward's method is an objective function which is optimized when of all individual distances from the centroid are minimum. Namely, it addresses it as a problem of variance. In this research, the Euclidean distance is used to define the initial cluster distances.

It is important to mention that algorithms in the extended family of agglomerative clustering such as single linkage and complete linkage implement recursively the Ward's criterion in each step in order to optimize the analysis.

2.4 Classification Methods

Second, we aim to train a classifier in order to be able to automatically distinguish samples belonging to positive and negative emotions based on the scores of three entrepreneurial features. This problem of assigning discrete labels to unknown samples, based on the known labels of some observation is called classification and is a problem of supervised learning. In order to find a model that maximizes predictive performance on the available data, several classifiers were assessed: two kernel based algorithms (SVM-linear and SVM-RBF), one non parametric method (K Nearest Neighbors-KNN) and one methodology based on ensembles of decision trees (Random Forest).

Support Vector Machine – Linear (SVM)

The Support Vector Machine (SVM) [14] classifier is a maximum margin classifier which finds the separating hyperplane with the maximum margin between the classes, where the margin is defined as the distance between the separating hyperplane and the closest sample(s) of each class to the hyperplane. In the original SVM algorithm the classes are considered to be linearly separable, however the algorithm has been extended to handle the non linearly separable case using slack variables. Moreover, one of the reasons that the SVM classifier is so popular is that it can be extended using kernel functions to efficiently classify samples in cases where the classes are not linearly separable. The main idea behind using kernels (kernel trick) is that while the data are not linearly separable in the input space (original space), through the kernel function they are mapped to a higher dimensional space where they could be linearly separable. Depending on the choice of the kernel function, the higher dimensional space might even have infinite dimensions. One advantage of SVMs is that the determination of the model parameters corresponds to a convex optimization problem, and so any local solution is also a global optimum - although the training process involves nonlinear optimization, the objective function is convex, and so the solution of the optimization problem is relatively straightforward.
As stated above, in support vector machines the decision boundary is chosen to be the one for which the margin is maximized. The maximum margin solution can be motivated using computational learning theory, also known as statistical learning theory.

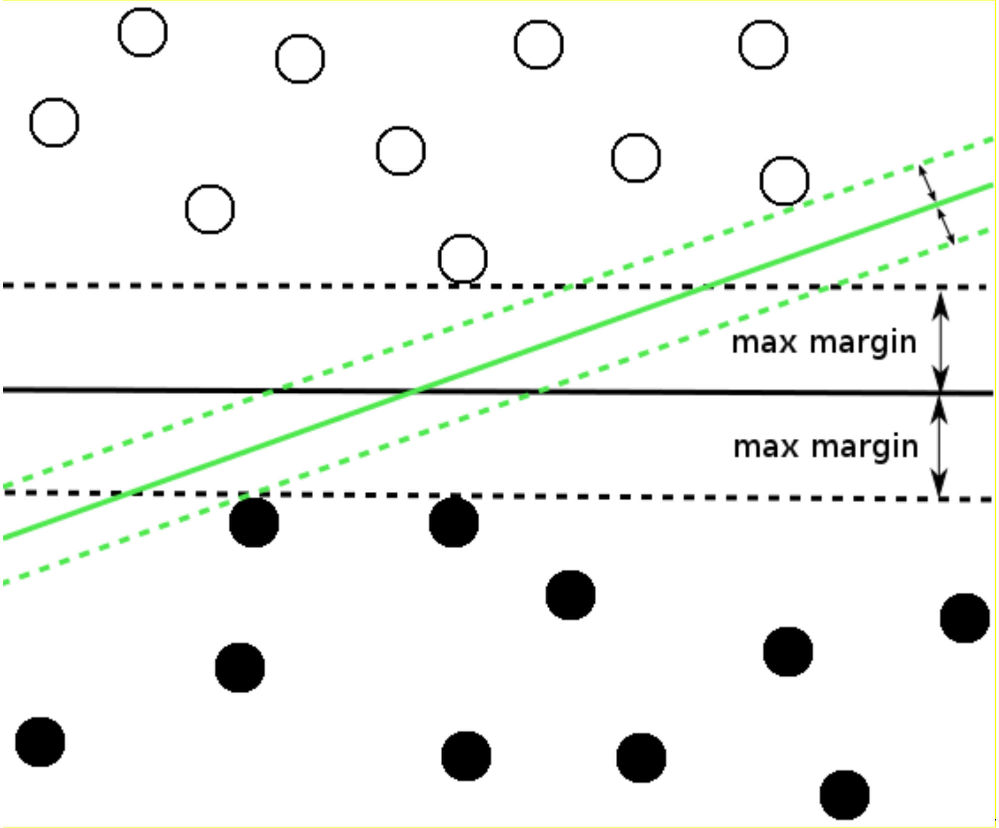

Figure 2.1: Graphical illustration of the maximum margin of SVM [21]

Support Vector Machine – Non Linear (SVM-RBF)

In the linear approach of SVM it is observed that a line divides the point of the two classes. If the data which are under examination cannot be effectively distinguished by a line a different kernel can be used in order to optimize the performance of the classification.
A very popular non-linear kernel frequently used in conjunction with the SVM is the (Gaussian) Radial Basis Function (RBF) [14]. As discussed above, when the SVM with RBF kernel is implemented, the separation of the data samples is performed in a higher dimensional space (of infinite dimensions) where the data may be linear separable. The structure of the RBF kernel is in essence a Gaussian function without the normalization term, which is not necessary since the output is not a probability. That is,

$$K(\boldsymbol{x},\boldsymbol{x}') = e^{-(\frac{||x-x'||^2}{2\sigma^2})},$$

where σ is a hyper parameter which can be tuned using cross validation.

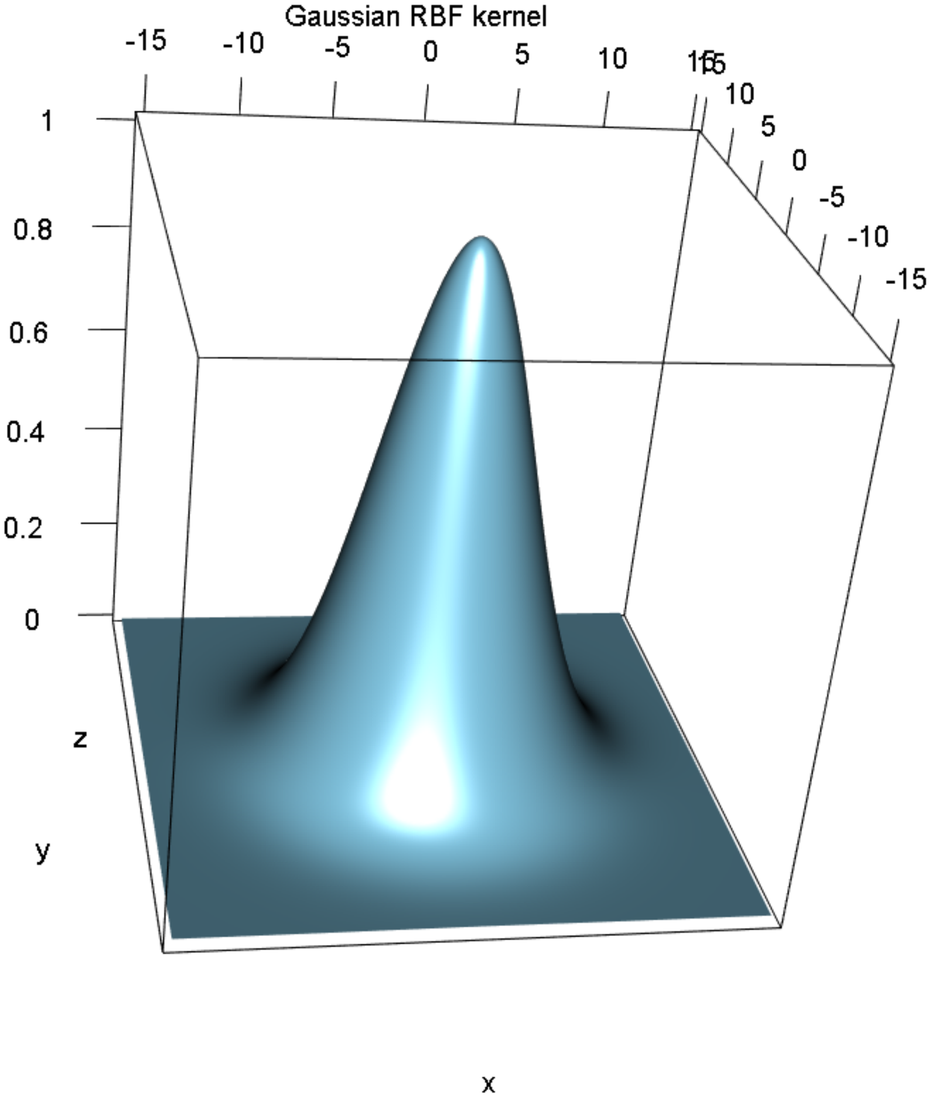

Figure 2.2: Graphical illustration of the Gaussian RBF Kernel

K – Nearest Neighbor (KNN)

The K-Nearest Neighbor classifier [14] [15] is a very straightforward non-linear classification method which simply assigns a class label to each unknown sample based on the majority vote of it's K nearest neighbors. The nearest neighbors are selected according to a distance metric, which is usually the Euclidean distance in the input space. The number of nearest neighbors K is a hyper parameter that can be tuned using cross validation. It is a positive integer, typically small. In the case of K=1 the unknown data point is simply assigned to the class of it's nearest neighbor. Despite it's conceptual simplicity it can achieve good results and it can be mathematically proven that the error rate or the 1-NN classifier is not worse that two times the optimal Bayes rate. The main drawback of the K-NN classifier is it's $O(n^2)$ computational complexity (where n the number of training samples), which makes it impractical for very large datasets. Moreover, in order to achieve good results, K-NN requires "sufficient" density of training data, especially close to the class border.

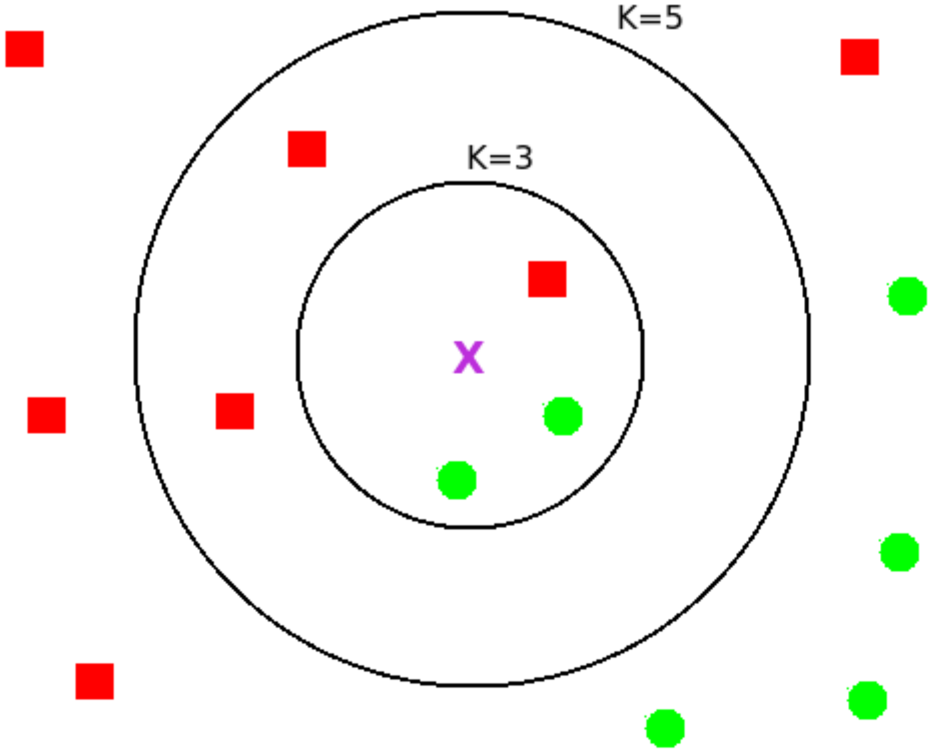

Figure 2.3: Graphical illustration of the KNN classification method [21]

Random Forest (RF)

The Random Forest (RF) classifier [14] is a very popular ensemble learning classification method which is one of the top performers in many classification tasks. It is based on the idea of bootstrap aggregating (bagging), where a committee of low bias high variance classifiers, in this case decision trees, is trained on a set of datasets generating from bootstrap resampling (with replacement) from the original dataset and aggregating the results (majority voting), which results in improved classification performance. Another advantage of Random Forests is that they also further enhance the classification performance by reducing the correlation between individual trees in the forest through selecting a subset *m* of all features before splitting the node of each tree during the training process. The subset *m* of features to be randomly selected before each split is a hyper parameter that can be tuned through cross validation.

K-fold cross validation (K-fold)

K-Fold Cross Validation [22] is a technique used in assessing the predictive capability of a classification method while avoiding extraction of misleading performance estimates. That is, if all the data are used for training and the same data are subsequently used for testing the model, then the predictive performance observed will be too optimistic since the model will be tested on data it has seen before. K-Fold Cross Validation for K=10 is illustrated in figure 2.4. In this study 10-fold cross validation was used since K=10 is frequently used in literature and is also recommended by [22]. K-Fold Cross Validation splits the dataset into K different parts (folds) of approximately the same size. The method then proceeds to iteratively use k-1 of the folds for training the model and the remaining 1 fold for testing, using a different fold for testing during each of the K iterations. In the end, all of the samples have been classified and k different performance metrics have been extracted (one for each fold). The average performance over all folds is then calculated. If αk is the accuracy achieved in each of the folds, the overall accuracy is just the average over the folds:

$$\bar{a} = \frac{1}{K} \sum_{k=1}^{K} a_k$$

Typical values used for K are K=3, 5 or 10 or setting K equal to the number of samples in the dataset (Leave One Out Cross Validation). If the ratio of the different classes in each fold is the same as the ratio of different classes in the original dataset, then the process is called stratified K-Fold Cross Validation.

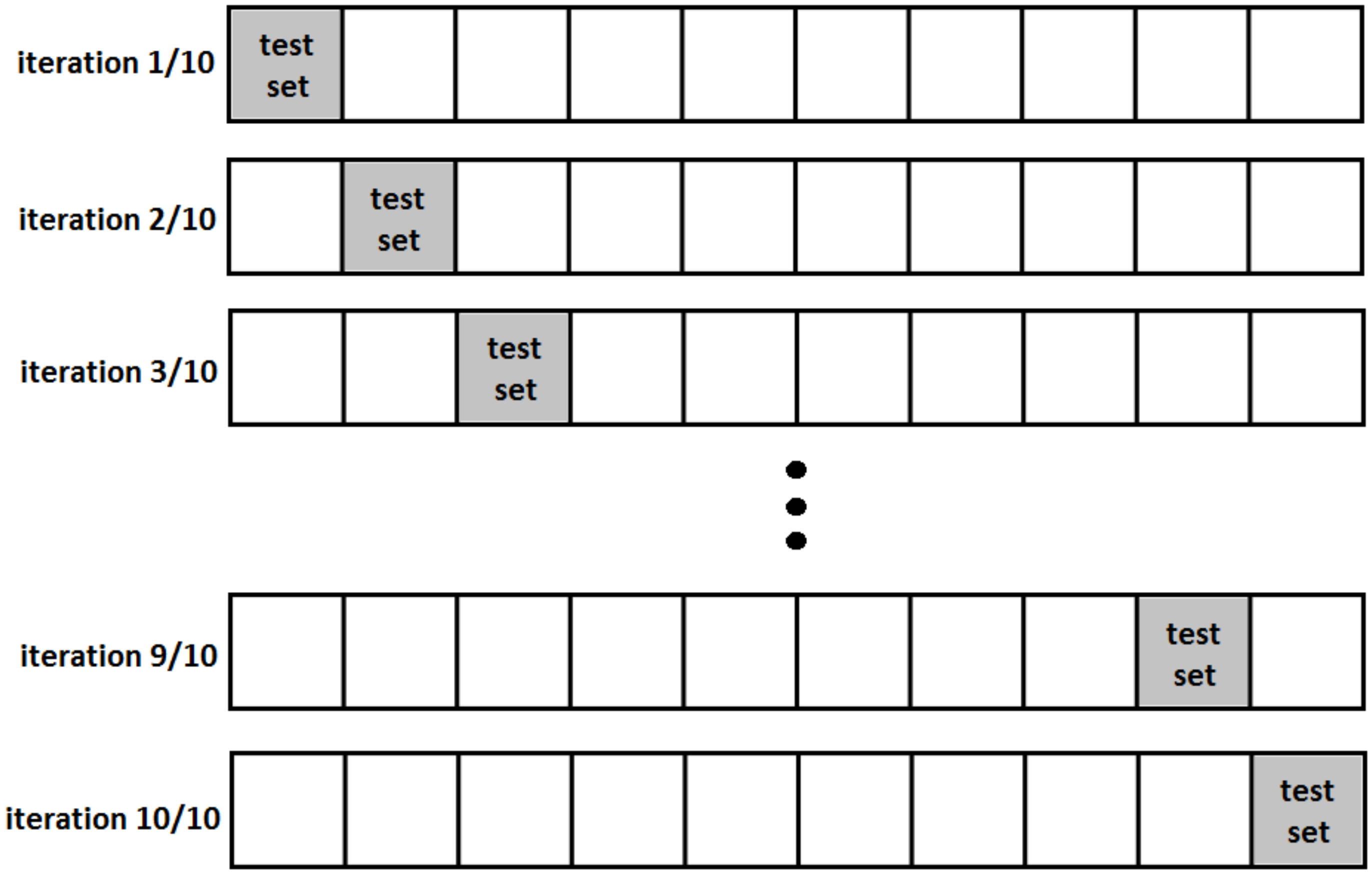

Figure 2.4: Illustration of 10-fold cross validation

2.5 Reliability assessment of the Classification Methods

A common mistake that frequently occurs when it comes to the assessment of classification results, is naive interpretation of the classification accuracy [24]. For example, when the classification of a data set which contains 80% positive and 20% negative samples gives a result of 80% accuracy if all of the samples are just assigned to the 'positive' class. The 80% accuracy achieved in that case may seem high but it has no practical significance. The reason is that such an outcome shows that the classifier classifies every sample in the positive class and none in the negative, resulting in a true negative rate of zero.

In order to avoid confusion and to interpret the results in a meaningful way, it is necessary in each classification method to calculate not only the accuracy of the model but the sensitivity (true positive rate) and the specificity (true negative rate), as well. Those measures can be obtained from the confusion matrix extracted after each classifier has been applied on the data. The confusion matrix summarizes the performance of a classifier by displaying how many samples of a given class have been correctly assigned to it, and how many samples of the same class have been incorrectly assigned to another class. An example of a confusion matrix for a binary classification problem can be seen in table 2.1.

Classified as		
Positive	Negative	Actually is
True Positive (TP)	False Negative (FN)	Positive (P)
False Positive (FP)	True Negative (TN)	Negative (N)

Table 2.1: A confusion matrix extracted after a classifier has been run on the data

Sensitivity or True Positive Rate (TPR) [23] is the percentage of the correctly identified positives of a data set after a classification method has been applied. It is computed from the extracted confusion matrix as,

$$TPR = \frac{TP}{(TP + FN)} = \frac{TP}{P}$$

Specificity or True Negative Rate (TNR) [23] is the percentage of the correctly identified negatives of a data set after a classification method has been applied. It is computed from the extracted confusion matrix as,

$$TNR = \frac{TN}{(TN + FP)} = \frac{TN}{N}$$

2.6 Feature Selection with LASSO and t-test

Finally, we aim to find which of these 3 entrepreneurial features is the most significant that aids the most in the classification procedure in comparison to the other two. (supervised learning problem)

LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) [14] [18] is a machine learning method that embeds feature selection in the regularization term of the statistical model. That is, unlike traditional l2-normalization where many variables are set to small values but not exactly zero, LASSO forces exactly zero weight to the features that are “irrelevant” to the prediction of the outcome. LASSO was originally developed for regression problems but has also been extended to cover the classification case. Another characteristic of the LASSO is that it tends to select only one “representative” feature out of a subset of correlated features and exclude the rest from the model.

t-test

The t- test [19] is a statistical hypothesis test which assesses whether the observed difference of the means of two samples is statistically significant. That is, given two samples a and b, with means μ_a μ_b and variances σ_a σ_b, the t-test assesses whether the difference is statistically significant or is observed probably due to chance. That is, given that the Null Hypothesis $H_0: \mu_a = \mu_b$, the t-test calculates the probability that the Alternative Hypothesis $H_1: \mu_a \neq \mu_b$ (two-tailed) is observed due to chance. In the one-tailed case the alternative hypothesis is either $H_1: \mu_a > \mu_b$, or $H_1: \mu_a < \mu_b$. The probability that the alternative hypothesis is observed merely by chance is the p-value. If the p-value is sufficiently small (e.g <0.05), then one can 'reject' the Null Hypothesis in favor of the alternative. That is, given the observed data the analyst is sufficiently convinced that the two samples arise from two distributions with different means.

Given two classes (positive/negative), feature selection on the basis of the t-test can be performed by selecting only one of the features and performing the t-test (two-tailed) on the two subsets of the data set (positive and negative). The main idea is that the more the selected feature separates the two classes, the smaller the corresponding p-value for that feature will be. So the importance of different features can be compared by directly comparing their p-values.

3. Methodology

3.1 Methodology Overview

In order to gain insight into data, the field of machine learning and pattern recognition provides us with statistical learning tools, that aid us in extracting information and possibly knowledge from a data set. These tools can be classified as supervised or unsupervised. Supervised learning involves building a model for predicting, or estimating, an output based on one or more inputs (features). In the case of unsupervised learning, we observe only the features and have no measurements for the outcome. That is, there are no class labels available for any of the samples and the goal is to find some hidden structure in the data (clustering).

The first step of the data analysis process of this study utilizes algorithms from the field of unsupervised learning, in order to find the hidden structure in the emotional classes, depending on the 3 entrepreneurial intention features. Hence, related work from bibliography [1] suggest that we should perform hierarchical clustering analysis. In order to perform this kind of analysis we preprocessed the data to create a 6x3 matrix, where each row represents each emotion's class centroid vector by calculating the mean value of each entrepreneurial intention feature for each emotion class.

The second step of the data analysis process utilizes algorithms from the field of supervised learning in order to build classifiers (output estimation model), which enable us to predict the outcome for data with unknown labels. In general, a classifier aims to generate a function, given a set of labeled samples. We have an outcome measurement (class label) of categorical data (emotion classes), that we would like to estimate based on a set of features (3 indicators of entrepreneurial intentions), for a sample of 1160 students. In order to perform a more meaningful classification analysis we characterize the 6 emotional classes as positive or negative, according to the results of the hierarchical clustering. As result we have 2 class labels:

1. positive emotions: by merging interest and enjoyment
2. negative emotions: by merging anger, fear, distress and surprise.

Hence, related work from bibliography [1] recommends that we should apply non-linear methods for classification to analyze the data. Support Vector Machines (linear and non-linear), KNN, Random Forests are the suggested methodologies for classification. They are popular, state of the art methodologies due to their good prediction performance in real data proposed by the literature.

In the third step of the analysis, the performance of each classifier is assessed. The performance of each method is assessed through 10Fold Cross Validation. However, the choice of the performance metric is also important since, especially in the case of imbalanced datasets, looking at accuracy alone can be very misleading. As a result, the sensitivity and specificity of each classifier are considered along with accuracy.

The fourth step of data analysis includes feature selection through regularized regression, in order to find which of the 3 entrepreneurial intention features are significantly associated to the outcome (emotion class). The LASSO method embeds feature selection in the training process and eliminates the features that are not significantly associated to the outcome. In addition to the LASSO, follow-up t-tests were performed in order to verify and validate the results of LASSO. Moreover, the t-test provides an alternative method to assess the importance of difference features, through their corresponding p-values. An overview of the proposed methodology can be seen in figure 3.1

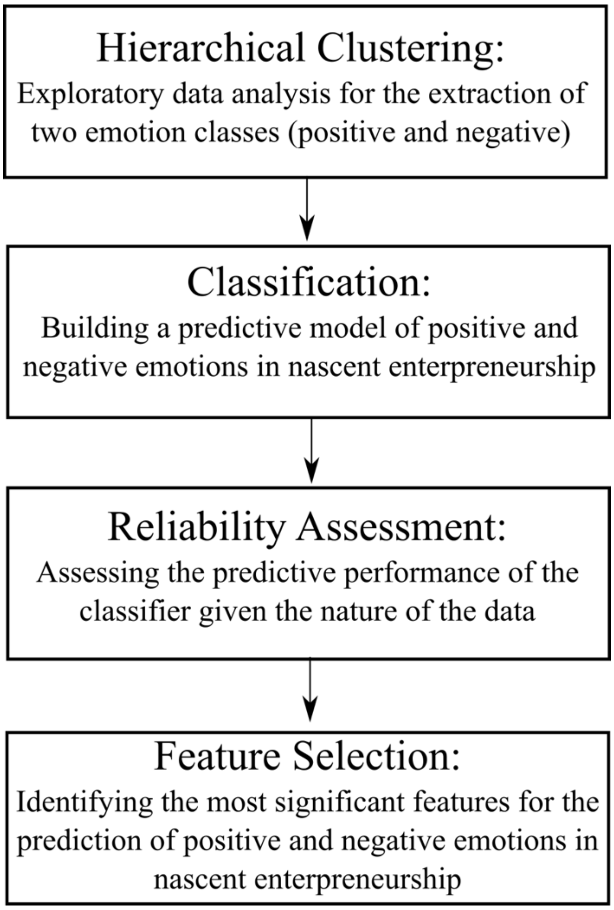
Hierarchical Clustering:
Exploratory data analysis for the extraction of two emotion classes (positive and negative)
Classification:
Building a predictive model of positive and negative emotions in nascent enterpreneurship
Reliability Assessment:
Assessing the predictive performance of the classifier given the nature of the data
Feature Selection:
Identifying the most significant features for the prediction of positive and negative emotions in nascent enterpreneurship

Figure 3.1: Steps of the proposed methodology

Preprocessing of the Dataset

In order to apply the selected data analysis methodology, it is necessary to pre-process the data.
The main reason for the pre-processing of the data set is the fact that the extracted data are selected for a different analysis approach. Therefore, it is important to form the data in a way that can fit the selected methodological approach, a machine learning analysis in entrepreneurial intentions.

For the hierarchical clustering analysis, we have to obtain one matrix that presents the average performance in each entrepreneurial feature of each emotion class. That table presents the mean score of the entrepreneurial intention someone has depending on the emotion that someone feels in the nascent entrepreneur phase-stage.
For the classification and the feature selection analysis, we have to obtain two matrices, one class label vector, which represents the primary emotion of each student - respondent in 6 emotion classes (enjoyment, fear, distress, anger, interest, surprise) and then according to the results of the hierarchal clustering analysis in 2 emotion classes (positive, negative). And another data matrix, which represents the mean score of each student in 3 appraisal features of entrepreneurial intentions (ATT, INT, PBC).

The data, which are analyzed in this research were selected for a previous research project. 1160 students answered a questionnaire (Appendix A) about their entrepreneurial intentions and the emotions they feel when they put themselves in the stage of nascent entrepreneurship. These data were collected for a different analysis approach, hence in order to perform a machine learning analysis these data must be pre-processed. As stated in chapter 1.2, for this analysis will be used a part of those data in a specific form.

The class label vector is extracted from section D (Appendix A). The answers to the 67 emotion queries from the 1160 students are related to 6 emotion classes (enjoyment, fear, distress, anger, interest, surprise). The score of 67 emotion queries is grouped in 6 emotion classes (Appendix B), inspired by the Differential

Emotion Scale (DES) [20] designed by Izard in 1991 (Table 3.2) and applied in the questionnaire which collected the data [2] (Appendix A). The overall score of each emotion class for each student is extracted from the mean value of the group queries of each emotion class. Hence, a 1160x6 matrix is created, where each row is a vector that represents each student's evaluation according to each emotion class. In order to find the class label vector, we found the position of the max value in each row vector in the 1160x6 matrix.

Emotions	Adjectives		
Enjoyment	Delighted	Joyful	Happy
Fear	Fearful	Afraid	Scared
Distress	Distressed	Sad	Discouraged
Anger	Mad	Enraged	Angry
Interest	Alert	Attentive	Interested
Surprise	Astonished	Amazed	Surprised

Table 3.2: DES for the six emotion classes, 3 adjectives that describe each emotion class [2]

The data matrix is extracted from sections B and F (Appendix A). The answers from the 1160 students represent the 3 appraisal features for entrepreneurial intentions. The 3 features are:
(1) Attitudes Towards Entrepreneurship (ATT) [2]: reflects the attitude (emotional) towards the prospect of starting his/her own business as an entrepreneur, (By definition tents to reflect the emotions someone feels about the possibility of starting a new business)
(2) Entrepreneurial Intention (INT) [2]: reflects a tendency to invest time and resources in order to become an entrepreneur and his/her own business,
(3) Perceived Behavioral Control (PBC) [2]: reflects how one perceives the skills related to start a new business.

The overall ATT, INT and PBC score for each student is calculated as the mean value of the corresponding question group. In order to create the data matrix, we merge the vertical vectors of the ATT, INT and PBC score of each student. Hence, we have a 1160x3 data matrix, where each row is a vector that represents each student's evaluation according to the 3 appraisal features of entrepreneurial intentions.

These matrices, the class label vector and data matrix, allow us to obtain a matrix that represents the mean entrepreneurial intention profile for each emotion class (Hierarchical Clustering data – HC-Data), in order to perform cluster analysis. Namely, they are used in order to create a 6x3 matrix, where each row represents each emotion's class centroid vector by calculating the mean value of each entrepreneurial intention feature for each emotion class.

HC-Data	ATT	INT	PBC
Enjoyment	4.44	2.96	3.18
Fear	3.30	2.86	2.61
Distress	3.60	2.43	2.70
Anger	3.69	3.18	2.75
Interest	4.12	2.60	2.99
Surprise	3.81	2.97	3.17

Table 3.1: Input data for hierarchical clustering analysis

Moreover, the class label vector and data matrix, are used as the main input for the majority of machine learning classification, regression and feature selection methods.

3.2 Exploratory Data Analysis – Hierarchical Clustering

Hierarchical Clustering is an informative structure which presents all the clusters that are formed until every input ends up in one cluster. The advantage of the use of this method is that there is no need to specify the number of clusters from the beginning and as an output is a figure presents a hierarchy of clusters until one is formed.

For the purpose of this research, hierarchical clustering is selected in order to observe the possible structures that are formed among 6 emotion classes based in the mean score each class has in the 3 appraisal entrepreneur features. It is, according to literature, the most appropriate method to use if the data are grouped in subsets, because the agglomerative hierarchal clustering algorithms have low efficiency when it comes to analyze large data sets. For that reason, the sample was grouped in 6 emotion classes and the mean value of the 3 entrepreneurial features of each emotion class was calculated. The purpose is to observe if there is similarity among them and more specifically if each vector of the mean value of the 3 entrepreneurial intention aid in the form of clusters that can be interpreted by the appraisal theory of emotions.

Ward's Method

Ward's method is proposed by literature, because it optimizes the objective function in each step of the cluster analysis. This criterion in each step selects the merge with the minimum Residual Sum of Squares (the squared distance of each vector from its centroids summed). By repeating this process until only one group remains, the complete hierarchical structure and a quantitative estimate of the loss associated with each stage in the grouping can be obtained.

Moreover, the pvclust package for R also offers the possibility to observe the p-values of each merge that is formed with the Ward function with bootstrap. That function executes the algorithm 10 times and presents the overall result in a dendrogram. The only difference with the previous algorithm is that it shows how high-strong-significant is the similarity in each cluster.

3.3 Classification – Predicting Positive or Negative Emotions of Nascent Entrepreneurship

Classification aids in the research when 2 groups of positive and negative emotions are formed from the hierarchal clustering step. In this step of the analysis, classification is performed in order to obtain how positive and negative emotions correlate with the 3 entrepreneurial features. By visual inspection of the 3D representation of the data (Figure 3.2) it is obvious that there is no clear linear boundary separating the two classes of positive and negative emotions, so non linear methods for classification are expected to be more suitable to the task and perform better. In this regard, four methods of classification were selected: SVM – Linear, SVM – RBF (non linear), KNN and Random Forest.

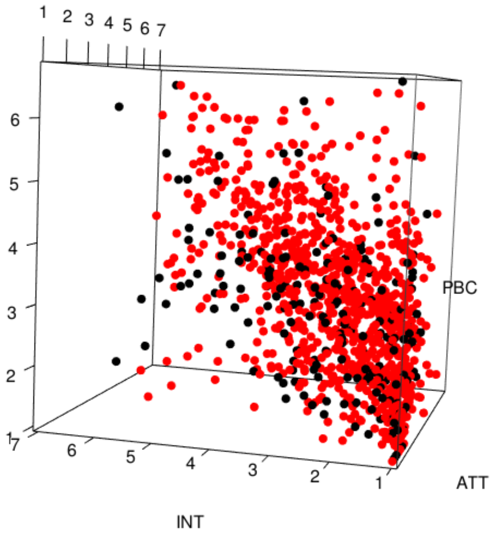

Figure 3.2: 3D representation of the data samples, the red dots are the positive and the black dots are the negative, in the axes are the 3 entrepreneurial features (ATT, PBC, INT)

Another interesting fact that is observed is that the majority of the samples are positive (82.1%) and a small minority are negative (17.9%). That arises a discussion about avoiding overfitting and misinterpretation of high but misleading accuracy of the model. (Unit: 3.4 Reliability assessment of the Classification Methods - Confusion Matrices)

3.4 Reliability assessment of the Classification Methods

In this research the data set used for classification is imbalanced because 82.1% of the sample is feeling positive emotions and 17.9% negative. In order to avoid misinterpretation of the classification results [24] it is crucial to assess not only the accuracy of the prediction model but the the true positive(TPR) and the true negative rate(TNR). The output of the classifiers in R programming is a confusion matrix which allows us to calculate the sensitivity(TPR) and specificity(TNR) of each method. Since the negative samples of the data set are the fewest the assessment of the true negative rate is the one to define the method which performs better than the others.

3.5 Feature Selection with LASSO and t-test

Another interesting question is: “Which of these 3 entrepreneurial features is the most significant and aids the most in the classification procedure in comparison to the other two?”. LASSO and t-test are the proposed methodologies from the literature to answer that question. LASSO selects the most representative features and sets all the others to zero, characterizing them as irrelevant. The t-test assesses whether the observed difference of a given feature among two different groups is statistically significant or is observed merely by chance, resulting in a corresponding p-value. The lower the p-value the lower the chance that the difference in the average value of the feature happens by chance, hence the feature is more important. Finally, to aid in the qualitative visual assessment of the quantitative results of the above methods, the differentiation of each feature among the different emotion classes can be displayed in a corresponding box plot.

4. Results

In this chapter, the results of the proposed methodology are presented. First, the extraction of the emotional clusters (2 emotional classes) using the exploratory data. Then, the performance of the classification models and the assessment of the predictive performance of the classifiers given the class imbalance of the data set. In the end, the identification of the most significant features for the prediction of positive and negative emotions in nascent entrepreneurship.

4.1 Hierarchical Clustering Results

The data matrix used for clustering the emotions into similar subgroups can be seen in table 4.1. Each of the 6 emotions (Enjoyment, Fear, Distress, Anger, Interest, Surprise) has three corresponding values of the entrepreneurial features (ATT, INT, PBC), which correspond to the mean of each feature for the given emotion. The 6 emotions are clustered hierarchically using Ward's method with and without bootstrap.

HC-Data	ATT	INT	PBC
Enjoyment	4.44	2.96	3.18
Fear	3.30	2.86	2.61
Distress	3.60	2.43	2.70
Anger	3.69	3.18	2.75
Interest	4.12	2.60	2.99
Surprise	3.81	2.97	3.17

Table 4.1: The input data for the hierarchical clustering analysis, represent the mean score of the entrepreneurial features in each emotional class. (The rating scale is 1-7)

Ward's Method Results

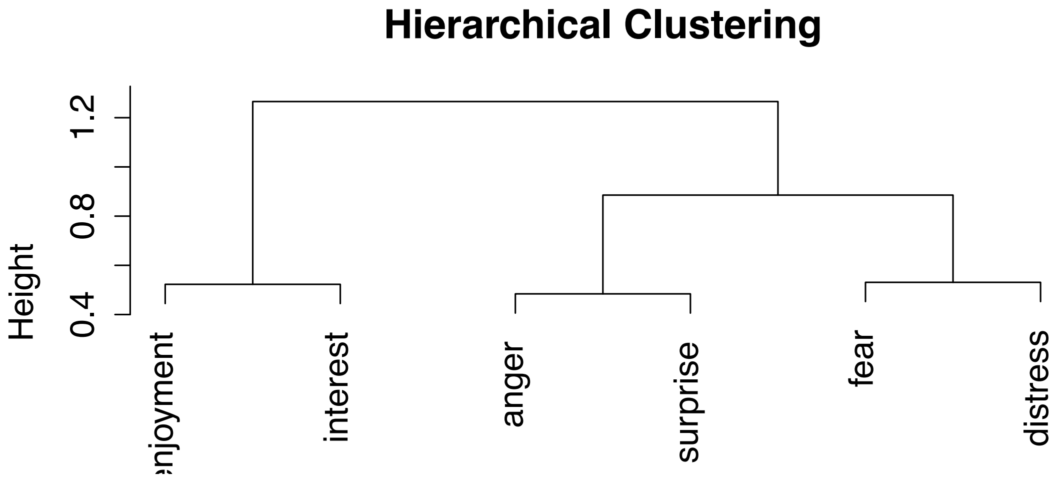

Figure 4.1: The graphic result of hierarchical clustering analysis with Ward's method

Ward’s Method with Bootstrap Results

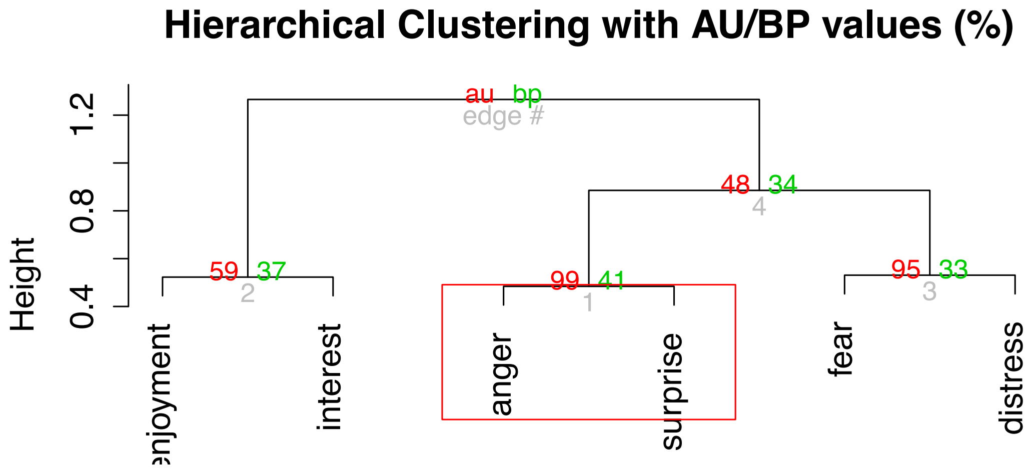

Figure 4.2: Graphical illustration of hierarchical clustering analysis with Ward’s method with Bootstrap.

By qualitative assessing the quantitative results of the hierarchical clustering, it is observed that there is strong distinction in 2 main emotion categories, positive and negative. These results are based on the mean values of 3 entrepreneurial features in the 6 emotion sub-categories. Clustering enjoyment and interest into the group of positive emotions and the rest (anger, surprise fear, distress) into negative emotion is easily interpreted qualitatively, as well. Moreover, by taking a closer look in the negative cluster it is observed that there are two subgroups within the negative cluster (anger, surprise) and (fear, distress), which also make sense qualitatively. Furthermore, the above grouping also arises a discussion on whether the emotion of surprise should be grouped as negative when it comes to entrepreneurial intentions. In that case, an assumption was made that the emotion of surprise can be interpreted as ignorance of the business field leading to a feeling of uncertainty and only in that context it is possible to group it in the negative emotion class.

4.2 Classification Results

First, by observing the representation of the data set in 3D illustration in figure 4.3 it can be concluded that probably a non-linear classifier is required for the successful classification of the data set. Since there is a high degree of overlap in the 3D space of the samples belonging in the two classes. As a result, there seems to be no linear boundary that efficiently separates the two classes.

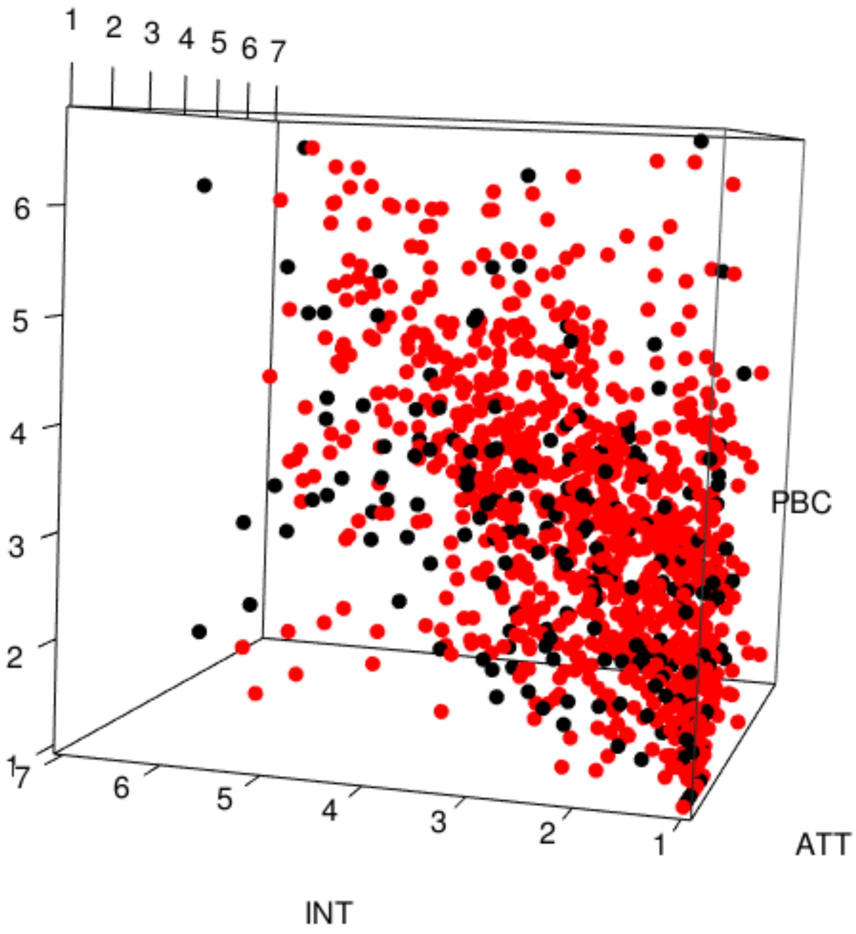

Figure 4.3: 3D representation of the data is used for the classification analysis

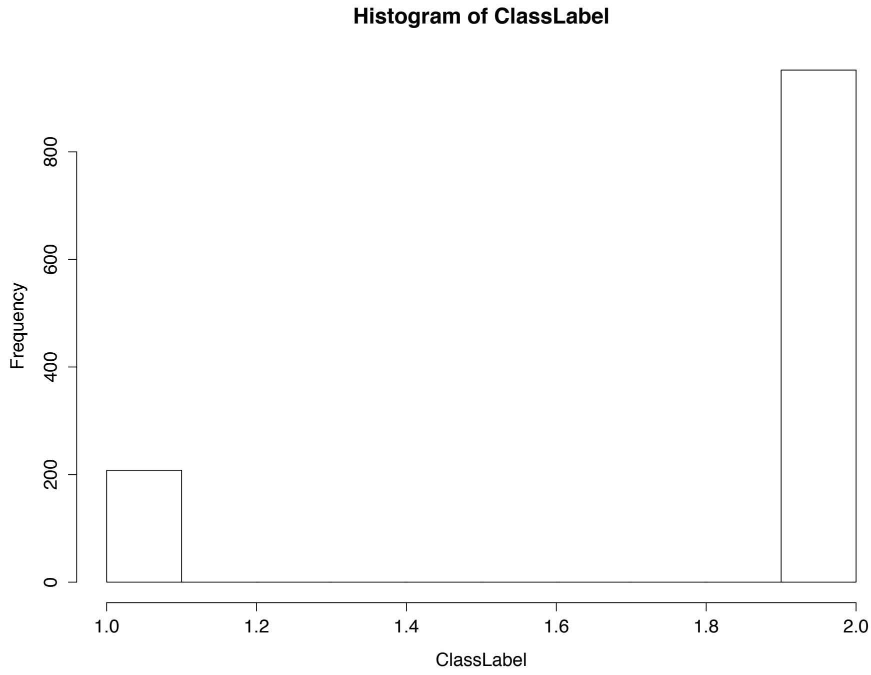

Figure 4.4: The histogram of the class labels (negative-1, positive-2) of the data set.

The histogram illustrates the class imbalance of the data set between positive and negative samples. The 82.1% of the sample is feeling positive emotions and 17.9% negative.

In the next table (Table 4.2) the metrics of the classification performance can be observed. The classification algorithm was executed for ten iterations and the mean value of the performance metrics was calculated. As it is mentioned in previous chapters (2.5, 3.4), the imbalanced data sets need to be assessed not only with the accuracy performance metric but with true positive and negative rates, as well. Since the class imbalance is caused because the negative samples are only 17.9% of the set, the true negative rate (TNR) is the most significant performance metric (since we only have very few negative samples). The highest the specificity (TNR), the highest the possibility to correctly predict the negative samples as negative and not as false positives. Even though the accuracy of the SVM-linear classifier is the highest, the true negative rate (TNR) is almost zero (1.6%). On the contrary, Random Forest has the lowest accuracy but the highest specificity (TNR). That shows higher prediction performance also in the negative samples compared to the other classifiers. Summarizing, Random Forest is the best prediction model for this data set considering all performance metrics and keeping in mind the class imbalance present. The performance of the classifiers is also illustrated in figure 4.5.

Mean value of 10 iterations	SVM Linear	SVM – RBF	KNN	RF
Accuracy – ACC	81.79 %	80.83 %	80.50 %	77.67 %
Sensitivity – SEN (TPR)	99.37 %	96.87 %	96.62 %	91.78 %
Specificity – SPC (TNR)	1.30 %	7.40 %	6.73 %	13.07 %

Table 4.2: The classification results of all methods, in terms of accuracy, sensitivity and specificity.

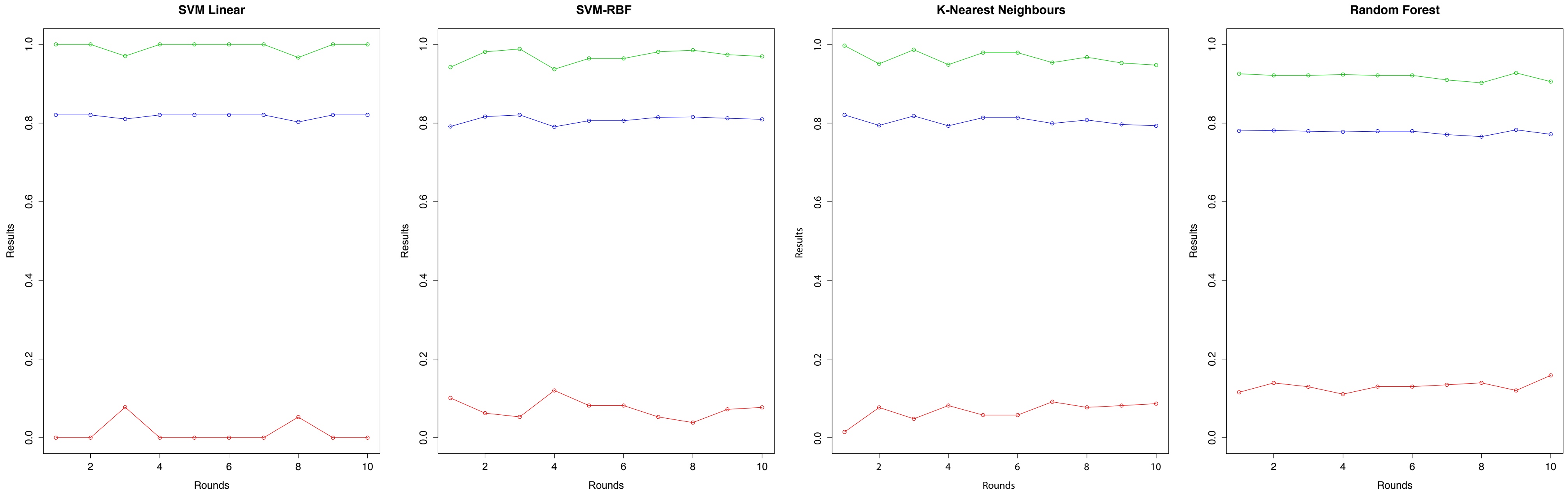

Figure 4.5: Graphical illustration of the results of all classification methods (Accuracy ------, Sensitivity ------, Specificity ------). The x-axis corresponds to 10 independent rounds of 10-fold cross validation, while the y-axis corresponds to the metrics of classification performance.

4.4 Feature Selection Results

In the following table (4.3) one can see the weights assigned to each of the entrepreneurial features by the LASSO method. As a reminder, LASSO only keeps the features which are considered important for the prediction of the outcome, while all other features are considered redundant and set to zero. According to the results of LASSO the only significant feature is ATT while the others are considered redundant.

LASSO	
ATT	0.026
INT	0
PBC	0

Table 4.3: LASSO feature selection results

A t-test was performed to assess the statistical significance of the difference of the average value of each of the features between the two classes of positive and negative emotions. The lower the p-value of the test, the less likely the observed difference in the average difference of the feature happens by chance. That is, the smaller the p-value the more significant the feature. The results of the t-tests performed can be found in table 4.4. According to the observed p-values ATT is the most significant feature, while PBC is also considered to have statistically significant difference between the two emotion classes.

t-test	p-values
ATT	$8.31\ 10^{-08}$
INT	$7.83\ 10^{-01}$
PBC	$9.16\ 10^{-03}$

Table 4.4: t-test p-values results for feature selection

Both methods agree that the most significant feature is Attitude Towards Entrepreneurship (ATT). On the other hand, PBC is rejected by LASSO while deemed significant according to the t-test. The rejection of PBC by LASSO is probably a mistake and can be explained by its high correlation (0.83) to the ATT attribute, since LASSO only selects one feature in a group of highly correlated features [14]. The rejection of INT by LASSO is probably correct since it cannot be explained the same way given that it has a very low correlation to ATT (0.067). So ATT it is correctly rejected by both LASSO and t-test. The average differences of each feature between the emotion classes which assessed by the t-tests are also visualized in the box plots of figure 4.6.

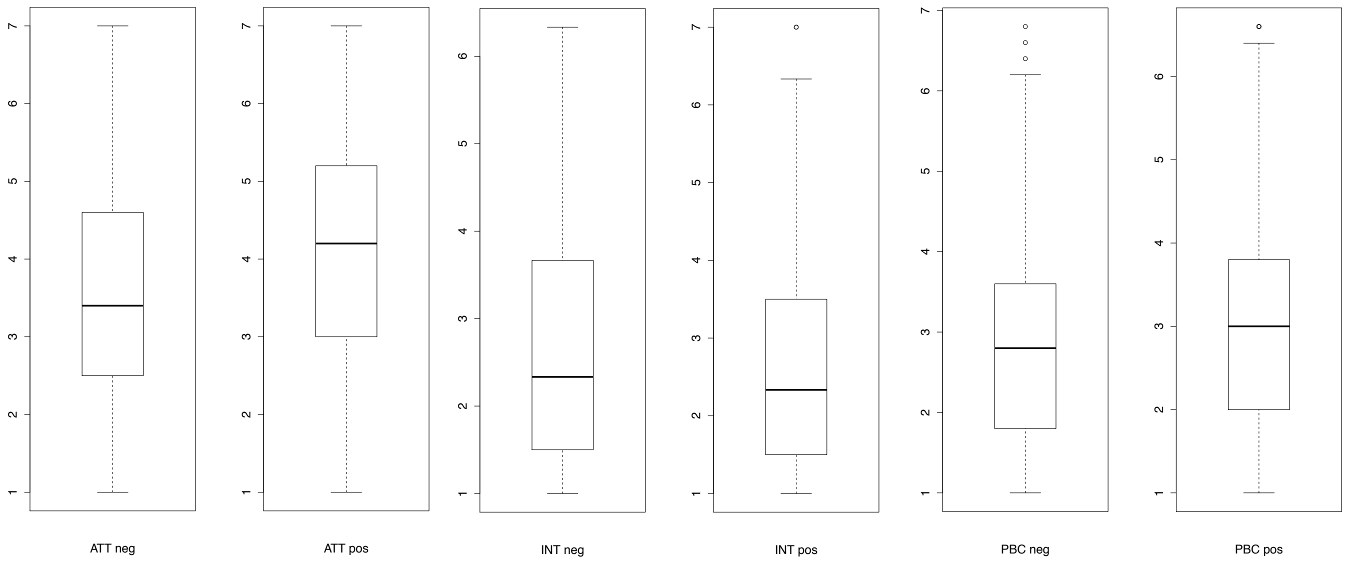

Figure 4.6: Box plots of the 3 entrepreneurial features

5. Conclusion

In this study, a quantitative analysis was conducted on data related to anticipated emotions in nascent entrepreneurship. The hypothesis of appraisal theory states that the effect of one appraisal variable on emotions can be affected by another appraisal variable [25]. In this regard, the aim of the proposed methodology was to utilize state of the art machine learning methodologies in order to identify any significant connections between the emotions someone feels when beginning to think of starting their own business and their intentions in entrepreneurship.

In the first step of the study, the quantitative results of supervised learning (Hierarchical Clustering) supported the qualitative assessment of grouping six emotion classes (enjoyment, interest, fear, anger, distress and surprise) in two classes i.e. positive (enjoyment, interest) and negative (fear, anger, distress and surprise). Furthermore, the above grouping also arises a discussion on whether the emotion of surprise should be grouped as negative when it comes to entrepreneurial intentions. In that case, an assumption was made that the emotion of surprise can be interpreted as ignorance of the business field leading to a feeling of uncertainty and only in that context it is possible to group it in the negative emotion class.

In the second and third steps of the analysis which involved supervised learning (classification) methods, the predictive capability of the classifiers was characterized by low specificity due to class imbalance in the dataset. When not only accuracy but also sensitivity and specificity were considered non linear methods performed best, as expected, since there seems to be no linear boundary that efficiently separates the two classes. Among the non-linear classifiers Random Forest performed the best, achieving double sensitivity while maintaining similar specificity to the other methods. The findings once again highlighted the problem of class imbalance in classification problems, which can be easily overlooked if accuracy is used as the only performance metric. For example, Linear Support Vector Machines classified almost all the samples into the positive class, which lead to the best overall accuracy but almost zero sensitivity, so looking at accuracy alone can be misleading.

In the fourth and final step, feature selection was performed using the LASSO and t-test methods in order to identify the most significant of the entrepreneurial features. As supported by both feature selection methods, Attitude Towards Entrepreneurship (ATT) was the most important feature. That entrepreneurial feature reflects the attitude (emotion) of a person towards the prospect of starting his/her own business as an entrepreneur. This quantitative result can also be explained qualitatively, since by definition ATT reflects the emotions someone feels about the possibility of starting a new business, so the high correlation between that feature and the appraised emotion towards entrepreneurship is to be expected. Another entrepreneurial feature: Entrepreneurial Intention (INT) was also deemed important by statistical analysis (t-test) but is not characterized as important by LASSO. This is a known drawback of LASSO, since it only selects one feature out of a group of correlated features. Since INT is strongly correlated to ATT, LASSO only selected one of the two as important.

The key point that arises from the results of the aforementioned methodology is that state of the art machine learning methodologies can be successfully utilized to predict positive or negative emotions based on appraisal variables (entrepreneurial features). Moreover, ATT is identified as the most important appraisal pattern that elicits positive or negative emotions in the case of nascent entrepreneurship. In terms of future work, since class imbalance is proven to be a major factor limiting classification performance, methodologies that specifically counter class imbalance, such as random under-sampling, could be utilized. Furthermore, additional datasets, preferably datasets not suffering from strong class imbalance, could be processed in order to further validate the findings of this study.

6. References

1. Meuleman, Ben, K.R. Scherer, “Nonlinear appraisal modeling: An application of machine learning to the study of emotion production.” Affective Computing, IEEE Transactions on 4, no. 4, pp.. 398-411, 2013.
2. L.A. Zampetakis, M.Lerakis, K. Kafetsios, V. Moustakis, "Anticipated emotions towards new venture creation: A latent profile analysis of early stage career starters", The International Journal of Management Education, vol. 14, no. 1, pp. 28-38, 2016
3. K.R. Scherer, A. Schorr, T. Johnstone, “Appraisal Processes in Emotion: Theory, Methods, Research”, Oxford University Press, Series in Affective Science, no. 1, pp. 3-19, 2001
4. R.S. Lazarus, “Emotion and adaptation”, New York: Oxford University Press, 1991b.
5. C.A. Smith, “The self, appraisal, and coping”, In C.R. Snyder & D.R. Forsyth (Eds.), “Handbook of social and clinical psychology: The health perspective”, pp. 116-137, New York: Pergamon Press, 1991
6. M.S. Clark, A.M. Isen, “Toward understanding the relationship between feeling states and social behavior”, in A. Hastorf & A. M. Isen (Eds.), Journal of Cognitive Social Psychology, Elsevier North Holland, New York, 1982.
7. I.J. Roseman, “Cognitive determinants of emotion: A structural theory”, in P. Shaver (Ed.), “Review of personality and social psychology”, vol.5, “Emotions, relationships, and health”, pp. 11-36, Beverly Hills, 1984.
8. K.R. Scherer, “Emotion as a multicomponent process: A model and some cross- cultural data”, in P. Shaver (Ed.), “Review of personality and social psychology”, vol. 5, “Emotions, relationships, and health”, pp. 37-63, Beverly Hills, 1984a
9. C.A. Smith, “Dimensions of appraisal and physiological response in emotion”, Journal of Personality and Social Psychology, vol. 56, pp. 339-353, 1989
10. E.C. Tolman, “A behavioristic account of the emotions” Psychological Review, vol. 30, pp. 217-227, 1923
11. H. Leventhal, K.R. Scherer, “The relationship of emotion to cognition: A functional approach to a semantic controversy”, Cognition and Emotion, vol. 1, pp. 3-28, 1987.
12. L.D. Kirby, C.A. Smith, “Freaking, quitting, and staying engaged: Patterns of psychophysiological response to stress”, in N. H. Frijda (Ed.), 9th Conference of the International Society for Research on Emotion, 1996.
13. L.D. Kirby, C. A. Smith, “Toward delivering on the promise of appraisal theory”, vol. “Appraisal Processes in Emotion: Theory, Methods, Research”, Oxford University Press, Series in Affective Science, no. 1, pp. 121-140, 2001.
14. T. Hastie, R. Tibshirani, J. Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction second edition,” Springer, 2009.
15. Richard O. Duda, Peter E. Hart, David G. Stork, “Pattern Classification, 2nd edition,” Wiley, 2000
16. C.D. Manning, P. Raghavan, H. Schütze, “An Introduction to Information Retrieval”, Cambridge University Press, chapter 17, pp. 377-402, 2009
17. J.H. Ward, “Hierarchical Grouping to Optimize an Objective Function”, Journal of the American Statistical Association, vol. 58, pp. 236–244, 1963
18. R. Tibshirani "Regression, Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society. Series B (Methodological) vol. 58, no. 1, pp. 267-288, 1996

19. D.M. Diez, C.D. Barr, M. Çetinkaya-Rundel, Open Intro Statistics, Second Edition. Create Space Independent Publishing Platform, 2012.
20. Izard, E. Carroll, “The Differential Emotions Scale: DES: A Method of Measuring the Meaning of Subjective Experience of Discrete Emotions”, University of Delaware, 1993.
21. N.K. Chlis, Machine Learning Methods for Genomic Signature Extraction, M.Sc. Thesis, Technical University of Crete, 2015.
22. R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proc. 14th international joint conference on Artificial intelligence (IJCAI'95), vol. 2, pp. 1137-1143, 1995.
23. T. Fawcett. “An introduction to ROC analysis”, Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006
24. J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, “Experimental perspectives on learning from imbalanced data”, In Proceedings of the 24th international conference on Machine learning, pp. 935-942, 2007.
25. K.R. Scherer, “The Dynamic Architecture of Emotion: Evidence for the Component Process Model”, Cognition and Emotion, vol. 23, no. 7, pp. 1307-1351, 2009

Appendix A

ΕΡΕΥΝΑ ΓΙΑ ΤΙΣ ΠΡΟΣΔΟΚΙΕΣ ΕΠΑΓΓΕΛΜΑΤΙΚΗΣ ΑΠΟΚΑΤΑΣΤΑΣΕΙΣ ΕΛΛΗΝΩΝ ΦΟΙΤΗΤΩΝ

Η έρευνα που κρατάς στα χέρια σου στοχεύει στη συγκέντρωση πρωτογενών στοιχείων για τις απόψεις ελλήνων φοιτητών σε θέματα που έχουν να κάνουν με τις προοπτικές επαγγελματικής αποκατάστασης δίνοντας έμφαση κυρίως στην αυτό-απασχόληση και την επιχειρηματικότητα.

Η βοήθεια σου στη συγκεκριμένη έρευνα είναι πολύτιμη. Ζητάω τη δική σου άποψη, την οποία καταθέτεις εντελώς ανώνυμα και γι' αυτό παρακαλώ να απαντήσεις με ειλικρίνεια όλα τα ερωτήματα. **Δεν υπάρχουν σωστές ή λάθος απαντήσεις**. Τα συμπληρωμένα ερωτηματολόγια θα τύχουν ποιοτικής και ποσοτικής επεξεργασίας.

Θερμή παράκληση να μην αφήσεις ερωτήσεις αναπάντητες.

Σε ευχαριστώ πολύ εκ των προτέρων.

Α. ΓΕΝΙΚΑ ΣΤΟΙΧΕΙΑ-ΤΑΥΤΟΤΗΤΑ ΤΗΣ ΕΡΕΥΝΑΣ

1. Η ηλικία σου

....

2. Φύλο

1. Άνδρας ....
2 . Γυναίκα ....

3. Είσαι:

1. Προπτυχιακός ....
2 . Μεταπτυχιακός ....

4. Είσαι φοιτητής του τμήματος:

----

5. Έχει κάποιος από τους γονείς σου δικιά του επιχείρηση;

1. ΝΑΙ
2 . ΟΧΙ

6. Γνωρίζεις προσωπικά κάποιον επιχειρηματία;

1. ΝΑΙ
2 . ΟΧΙ

7. Έχεις εργαστεί ως υπάλληλος κατά το παρελθόν;

1. ΝΑΙ
2 . ΟΧΙ

Β. ΟΔΗΓΙΕΣ: Παρακαλώ βαθμολογήστε κάθε μία από τις παρακάτω προτάσεις ανάλογα με το βαθμό διαφωνίας (Πιο κοντά στο 1) ή το βαθμό συμφωνία σας (Πιο κοντά στο 7)

		1	2	3	4	5	6	7
1	Το να είμαι επιχειρηματίας έχει περισσότερα πλεονεκτήματα από μειονεκτήματα	**Attitudes towards entrepreneurship**						
2	Μου αρέσει να κάνω καριέρα ως επιχειρηματίας							
3	Αν είχα την ευκαιρία και τους πόρους θα ξεκινούσα τη δικιά μου επιχείρηση							
4	Το να είμαι επιχειρηματίας νομίζω ότι θα μου πρόσφερε μεγάλη ικανοποίηση							
5	Ανάμεσα σε διάφορες επιλογές που έχω θα προτιμούσα να είμαι επιχειρηματίας							
6	Προτίθεμαι να ξεκινήσω τη δικιά μου επιχείρηση στο μέλλον	**Entrepreneurial intention**						
9	Διαρκώς αναζητώ επιχειρηματικές ευκαιρίες							
11	Αποταμιεύω χρήματα για να ξεκινήσω τη δικιά μου επιχείρηση							
12	Διαβάζω βιβλία σχετικά με τις διαδικασίες έναρξης επιχειρήσεων							
14	Κάνω σχέδια για να ξεκινήσω τη δικιά μου επιχείρηση							
15	Αφιερώνω χρόνο για να μάθω πώς δημιουργείται μια επιχείρηση							

C. ΟΔΗΠΕΣ: Παρακάτω θα διαβάσεις διάφορες ενέργειες που πραγματοποιούν οι επιχειρηματίες που βρίσκονται στα αρχικά στάδια δημιουργίας μιας νέας επιχείρησης. Αφού τις διαβάσεις θα ήθελα να καταγράψεις για κάθε μια από αυτές το πόση προσπάθεια θα πρέπει να καταβάλεις για να ολοκληρώσεις κάθε μία από τις ενέργειες αυτές, στην περίπτωση που ΕΣΥ θα ήθελες να ξεκινήσεις τη δικιά σου επιχείρηση.

		Καθόλου Προσπάθεια	Μικρή Προσπάθεια	Μέτρια Προσπάθεια	Πολύ Προσπάθεια	Πάρα πολύ προσπάθεια
1	Άνοιγμα τραπεζικού λογαριασμού ειδικά για την επιχείρηση					
2	Εκπόνηση χρηματοοικονομικής μελέτης με προβλέψεις για την μελλοντική πορεία της επιχείρησης					
3	Αναζήτηση οικονομικών πόρων (πχ. από τράπεζες ή επενδυτές)					
4	Πρόσληψη υπαλλήλου					
5	Εγγραφή σε επιμελητήριο					
6	Εξεύρεση χώρου εγκατάστασης της επιχείρησης					
7	Συλλογή πληροφοριών για τους πιθανούς ανταγωνιστές					
8	Δημιουργία ιστοσελίδας					
9	Αγορά πρώτων υλών					
10	Εκπόνηση επιχειρηματικού σχεδίου (business plan)					
11	Πρόσληψη λογιστή					
12	Προωθητικές ενέργειες					
13	Αναζήτηση πληροφοριών για κανονισμούς και προδιαγραφές που ισχύουν					

D. ΟΔΗΠΕΣ: Ο παρακάτω πίνακας αποτελείται από λέξεις που περιγράφουν διαφορετικά συναισθήματα. Σκέψου τον εαυτό σου ότι έχει ξεκινήσει τη διαδικασία να λειτουργήσει τη δικιά του επιχείρηση. Δηλαδή είσαι στα αρχικά

στάδια προετοιμασίας χωρίς ακόμα να έχεις πάρεις χρήματα από τη λειτουργία της επιχείρησης. Θα ήθελα να καταγράψεις το βαθμό που περιμένεις να νιώσεις το κάθε συναίσθημα για την κατάσταση που σου περιέγραψα.

Ελάχιστα ως καθόλου	Λίγο	Μέτρια	Όχι πάρα πολύ	Πάρα πολύ
1	**2**	**3**	**4**	**5**

1. Αποθαρρημένος	1	2	3	4	5
2. Έτοιμος για δράση	1	2	3	4	5
3. Ευχαριστημένος	1	2	3	4	5
4. Νευρικός	1	2	3	4	5
5. Αναστατωμένος	1	2	3	4	5
6. Άναυδος	1	2	3	4	5
7. Άρρωστος	1	2	3	4	5
8. Σαρκαστικός	1	2	3	4	5
9. Νωθρός	1	2	3	4	5
10. Αηδιασμένος	1	2	3	4	5
11. Σεμνός	1	2	3	4	5
12. Πανευτυχής	1	2	3	4	5
13. Επαναστατικός	1	2	3	4	5
14. Εξαγριωμένος	1	2	3	4	5
15. Κατάπληκτος	1	2	3	4	5
16. Εκστασιασμένος	1	2	3	4	5
17. Μεταμελημένος	1	2	3	4	5
18. Προσεκτικός	1	2	3	4	5
19. Συνεσταλμένος	1	2	3	4	5
20. Ντροπιασμένος	1	2	3	4	5
21. Έντρομος	1	2	3	4	5
22. Περιπαικτικός	1	2	3	4	5
23. Κουρασμένος	1	2	3	4	5
24. Μοναχικός	1	2	3	4	5
25. Ξυπνητός	1	2	3	4	5
26. Ένοχος	1	2	3	4	5
27. Αλαζόνας	1	2	3	4	5
28. Φρόνιμος	1	2	3	4	5
29. Προκλητικός	1	2	3	4	5
30.Αξιοκατάκριτος	1	2	3	4	5
31. Άνανδρος	1	2	3	4	5
32. Ανήσυχος	1	2	3	4	5

35. Τρομαγμένος	1	2	3	4	5
36. Παράφρων	1	2	3	4	5
37. Σκεπτικός	1	2	3	4	5
38. Ξαφνιασμένος	1	2	3	4	5
39. Περιφρονητικός	1	2	3	4	5
40. Επιθετικός	1	2	3	4	5
41.Φοβισμένος	1	2	3	4	5
42. Ενοχλημένος	1	2	3	4	5
43.Με ενδιαφέρον για κάτι	1	2	3	4	5
44. Χαρούμενος	1	2	3	4	5
45. Πικρόχολος	1	2	3	4	5
46. Καλόκαρδος	1	2	3	4	5
47. Εχθρικός	1	2	3	4	5
48. Τρεμάμενος	1	2	3	4	5
49. Με αίσθημα αποστροφής	1	2	3	4	5
50 Στοχαστικός	1	2	3	4	5
51.Περιφρονημένος	1	2	3	4	5
52. Ανεπαρκής	1	2	3	4	5
53. Με αίσθημα σιχαμάρας	1	2	3	4	5
54. Στενοχωρημένος	1	2	3	4	5
55. Ενεργητικός	1	2	3	4	5
56. Εξοργισμένος	1	2	3	4	5
57. Ενθουσιώδης	1	2	3	4	5
58. Υπεροπτικός	1	2	3	4	5
59. Λυπημένος	1	2	3	4	5
60. Έκπληκτος	1	2	3	4	5
61. Θυμωμένος	1	2	3	4	5
62. Μετανιωμένος	1	2	3	4	5
63. Άτολμος	1	2	3	4	5
64. Συγκλονισμένος	1	2	3	4	5
65. Με αίσθημα απέχθειας	1	2	3	4	5
66. Πανικόβλητος	1	2	3	4	5

33. Εριστικός	1	2	3	4	5		67. Ευτυχισμένος	1	2	3	4	5
34. Συναισθηματικός	1	2	3	4	5							

Ε. ΟΔΗΠΕΣ: Αν αποφάσιζες να ξεκινήσεις τη δική σου επιχείρηση τα άτομα στο στενό σου κύκλο θα ενέκριναν μια τέτοια απόφαση; Η απάντηση σου να είναι πιο κοντά στο ένα αν δεν θα το ενέκριναν καθόλου και πιο κοντά στο 5 αν το ενέκριναν απόλυτα.

		1	2	3	4	5
1	Οι οικογένεια σου	Subjective norms				
2	Οι φίλοι σου					
3	Οι συμφοιτητές σου					

F. ΟΔΗΓΙΕΣ: Παρακαλώ βαθμολογήστε κάθε μία από τις παρακάτω προτάσεις ανάλογα με το βαθμό διαφωνίας (Πιο κοντά στο 1) ή το βαθμό συμφωνία σας (Πιο κοντά στο 7).

		1	2	3	4	5	6	7
1	Μου φαίνεται εύκολο να ξεκινήσω τη δικιά μου επιχείρηση και να τη διατηρήσω σε λειτουργία	**Perceived behavioral control**						
2	Είμαι έτοιμος/η να ξεκινήσω μια βιώσιμη επιχείρηση							
3	Νομίζω ότι μπορώ να αντιμετωπίσω τη διαδικασία δημιουργίας μιας νέας επιχείρησης							
4	Γνωρίζω τις απαραίτητες λεπτομέρειες για να ξεκινήσω τη δικιά μου επιχείρηση							
5	Αν ξεκινούσα μια επιχείρηση θα είχα μεγάλη πιθανότητα να επιτύχω.							

G. Γενικά, πώς πιστεύεις ότι θα ένιωθες αν κάποτε στο μέλλον ξεκινούσες τη δικιά σου επιχείρηση;
(Οδηγίες: στο παρακάτω πίνακα, βάλε Χ σε ΕΝΑ μόνο τετράγωνο, ανάλογα με το βαθμό που θα ένιωθες θετικά ή αρνητικά ή και τα δύο).

		Θετικά------>				
		Καθόλου (0)	Λίγο (1)	Μέτρια (2)	Αρκετά (3)	Πάρα πολύ (4)
Αρνητικά---->	Καθόλου (0)					
	Λίγο (1)					
	Μέτρια (2)					
	Αρκετά (3)					
	Πάρα πολύ (4)					

Appendix B

	Enjoyment	Fear	Distress	Anger	Interest	Surprise
1	3. Ευχαριστημένος	11. Σεμνός	1. Αποθαρρημένος	8. Σαρκαστικός	2. Έτοιμος για δράση	6. Άναυδος
2	12. Πανευτυχής	19. Συνεσταλμένος	4. Νευρικός	10. Αηδιασμένος	13. Επαναστατικός	15. Κατάπληκτος
3	22. Περιπαικτικός	20. Ντροπιασμένος	5. Αναστατωμένος	14. Εξαγριωμένος	16. Εκστασιασμένος	38. Ξαφνιασμένος
4	28. Φρόνιμος	21. Έντρομος	7. Άρρωστος	27. Αλαζόνας	18. Προσεκτικός	57. Ενθουσιώδης
5	44. Χαρούμενος	23. Κουρασμένος	9. Νωθρός	29. Προκλητικός	25. Ξυπνητός	60. Έκπληκτος
6	46. Καλόκαρδος	24. Μοναχικός	17. Μεταμελημένος	30.Αξιοκατάκριτος	43.Με ενδιαφέρον για κάτι	64. Συγκλονισμένος
7	67. Ευτυχισμένος	26. Ένοχος	32. Ανήσυχος	33. Εριστικός	55. Ενεργητικός	
8		31. Άνανδρος	37. Σκεπτικός	36. Παράφρων		
9		34. Συναισθηματικός	50 Στοχαστικός	39. Περιφρονητικός		
10		35. Τρομαγμένος	51.Περιφρονημένος	40. Επιθετικός		
11		41.Φοβισμένος	52. Ανεπαρκής	42. Ενοχλημένος		
12		48. Τρεμάμενος	54. Στενοχωρημένος	45. Πικρόχολος		
13		62. Μετανιωμένος	59. Λυπημένος	47. Εχθρικός		
14		63. Άτολμος	66. Πανικόβλητος	49. Με αίσθημα αποστροφής		
15				53. Με αίσθημα σιχαμάρας		
16				56. Εξοργισμένος		
17				58. Υπεροπτικός		
18				61. Θυμωμένος		
19				65. Με αίσθημα απέχθειας		