To appear in the *Journal of Applied Statistics* Vol. 00, No. 00, Month 20XX, 1–20

Pattern Discovery in Students' Evaluations of Professors A Statistical Data Mining Approach

Necla Gündüz^{a*} and Ernest Fokoué^b

^aDepartment of Statistics, Faculty of Science, Gazi University, Ankara, Turkey; ^bSchool of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA

(Received 00 Month 20XX; accepted 00 Month 20XX)

The evaluation of instructors by their students has been practiced at most universities for many decades, and there has always been a great interest in a variety of aspects of the evaluations. Are students matured and knowledgeable enough to provide useful and dependable feedback for the improvement of their instructors' teaching skills/abilities? Does the level of difficulty of the course have a strong relationship with the rating the student give an instructor? In this paper, we attempt to answer questions such as these using some state of the art statistical data mining techniques such support vector machines, classification and regression trees, boosting, random forest, factor analysis, kMeans clustering. hierarchical clustering. We explore various aspects of the data from both the supervised and unsupervised learning perspective. The data set analyzed in this paper was collected from a university in Turkey. The application of our techniques to this data reveals some very interesting patterns in the evaluations, like the strong association between the student's seriousness and dedication (measured by attendance) and the kind of scores they tend to assign to their instructors.

Keywords: Likert; Ordinal; Clustering; Pattern Recognition; Classification; Zero-variation.

AMS Classification codes: 62H30; 62H25

1. Introduction

The evaluation of instructors by their students has been practiced at most universities for many decades. Typically, these evaluations are administered in the form of long surveys answered by students at the end of the semester (quarter). Questions in the survey are related to such aspects as course organization, level and quality of delivery, clarity of course objectives, level of difficulty of the course, impact of the course on the student's overall university experience and goals, relevance of the course, preparedness and competency of the instructor, likeability and fairness of the instructor, overall satisfaction of the student, and overall rating of the instructor by the student, just to name a few.

This study investigates a data set which was anonymously collected in recent years from Gazi University in Ankara (Turkey). It contains a total 5820 evaluation scores provided by students for three different instructors. There is a total of 28 course specific questions (see the Appendix for a detailed list of all the questions) presented in Likert-type format, and an additional 3 attributes, namely student's perceived difficulty level of the course, attendance, number of repetitions of the course.

On the other hand, the overarching goal of students' evaluations of professors is the extraction of knowledge, patterns and information, with the finality of providing their professors with hopefully useful feedback to help them teach better and give students a

 $^{^*}$ Corresponding author. Email: ngunduz@gazi.edu.tr

richer and more effective learning experience. However, there has always been heated debates regarding the effectiveness or even the validity of such evaluations. Many scholars have wondered over the years if it is at all possible to improve education quality based on the outcomes of students' evaluations of professors. For them, they wonder if the answers are informative. Do the answers given by students provide the kind of knowledge and information that can help reshape and improve course quality and professors' teaching abilities? Typically, most university administrators such as department heads, school directors, college deans, provosts and chancellors have tended to rely on a single grand average of the questionnaire scores as a measure of the quality of an instructor.

Given the complex and multidimensional nature of the questionnaires administered, it is clearly misleading to summarize such evaluations with a single number. Besides, the averages usually relied upon are not valid, because of the non-numeric nature of the Likert-type of the evaluation responses/scores.

Indeed, since the publication of the seminal [17] paper, Likert-type scores have been extensively used in a wide variety of fields ranging from Anthropology, Psychology, Education, Sociology, Sports just to name of a few. Unfortunately, with the astronomical number of applications of the Likert measurement system, there have also been innumerable abuses, especially the misuse of Likert-type scores as real-valued scores. Authors such as [31], [8], [15] and [3] provide pointers to the uses abuses of Likert-type data. Many authors have indeed cautioned experimenters on the meaninglessness of statements made based on analyses with inappropriate techniques. To quote [2], "Nothing is wrong per se in applying any statistical operation to measurements of given scale, but what may be wrong, depending on what is said about the results of these applications, is that the statement about them will not be empirically meaningful or else that it is not scientifically significant". Along the lines of [2], many authors have written numerous articles providing guidelines as to which statistical techniques are most appropriate for Likert-type and the so-called Likert-scale datasets. [6] of instance provides a clear separation between Likert-type and Likert-scale, and strongly recommends nonparametric techniques for Likert-type and parametric techniques for Likert-scale.

(Need to correction paragraph) To avoid such pitfalls of meaningless conclusions on our data, we strive to guarantee the validity of our analyses and summaries, by using mostly Likert-type specific (or at least Likert-type compatible) techniques and tools of exploratory data analysis, cluster analysis, dimensionality reduction and pattern recognition.

The rest of this paper is organized as follows: in section 2 we present some general definitions and address important aspects of survey data such item reliability and respondent reliability. We also present empirical answers to most of the above questions using both appropriate exploratory data analysis tools and some straightforward tests of association. In section 3 we focus on the multivariate aspects of the data and answer most of the students' evaluation of instructors questions by using tools such as factor analytic and cluster analysis which both reveal very meaningful confirmation of some beliefs and perceptions about the rating of professors by their students. Section 4 uses some of the results from section 3 to perform predictive analytics on this data. We specifically apply state of the art pattern recognition techniques such as support vector machines, boosting, random forest and classification trees to predict the satisfaction level of a given

student based on their answers to the 28 questions on the survey. Section 5 provides our conclusion and discussion, along with pointers to our future work.

2. Definitions, Data Quality, Exploratory Data Analysis and Basic Tests

2.1 Definitions and data quality

The dataset is represented by an $n \times p$ matrix \mathbf{X} whose ith row $\mathbf{x}_i^{\top} \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{ip})$ denotes the p-tuple of characteristics, with each $\mathbf{x}_{ij} \in \{1, 2, 3, 4, 5\}$ representing the Likert-type level (order) of preference of respondent i on item j. Recall that a Likert-type score is obtained by translating/mapping the response levels {Strong Disagree, Disagree, Neutral, Agree, Strongly Agree} into pseudo-numbers $\{1, 2, 3, 4, 5\}$. A usually crucial part in the analysis of questionnaire data is the calculation of the Cronbach's alpha coefficient which measures the reliability/quality of the data. Let $X = (X_1, X_2, \cdots, X_p)^{\top}$ be a p-tuple representing the p items of a questionnaire. The Cronbach's alpha coefficient is a function of the ratio of the sum of the idiosyncratic item variances over the variance of the sum of the items, and is given by

$$\alpha = \left(\frac{p}{p-1}\right) \left[1 - \frac{\sum_{j=1}^{p} \mathbb{V}(X_j)}{\mathbb{V}\left(\sum_{\ell=1}^{p} X_{\ell}\right)}\right]. \tag{1}$$

For our data, we found $\hat{\alpha} = 0.992$, indicating a reliable (ie good quality) survey instrument from Cronbach's point of view. We must emphasize however, that this is just *item reliability*, which is of course of great importance, but herein contrasted with respondent reliability which we had to assess and address as a result of some patterns discovered in our data.

Definition 1 Let $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be a dataset with $\mathbf{x}_i^{\top} = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{ip})$. An observation vector \mathbf{x}_i will be called a zero variation vector if $\mathbf{x}_{ij} = \text{constant}, \ j = 1, \dots, p$. Respondents with zero variation response vectors will be referred to as single minded respondents/evaluators.

In our data set of n=5820 evaluations, we found a rather high prevalence of single minded evaluators, specifically, about half of the evaluations (2985/5820 $\approx 51\%$). In fact, zero variation responses essentially reduce a p items survey to a single item survey. As can be seen for our earlier calculation, the estimated Cronbach's α coefficient for our data is considerably high. The reason may be the high ratio of zero variation observations. It therefore became interesting to also estimate the Cronbach's α coefficient for only the non zero variation observations, which turned out to be 0.9755. Not surprisingly, the Cronbach's α value for the zero variation observations is 1, since that corresponds to perfectly reliable questionnaire.

Despite the fact that zero variation responses correspond to a perfectly reliable instrument from Cronbach's alpha perspective, it is our view that zero variation responses are an indication that the respondent did not give deep thought to each of the questions/items of the survey. One could always argue that such evaluators came in with a single rating on all the items, and that such responses are just fine, in the sense that they provide a clear and unambiguous overall assessment of the professor being evaluated. However, considering the sometimes drastically different foci of the questions, it is rather unlikely that a given instructor on a given course would perform exactly the same on all the items. On the other hand, such zero variation responses convey the impression that the

respondent rushed the answering process. Finally, from a point of view of feedback to the instructor in order to help them improve the course, such answers provide very little if any feedback at all. Authors like [20], [32], [22] and [25] have contributed extended studies and findings related to the effectiveness of students' rating of professors and have also touched extensively on aspects like biases, utility, reliability and validity. In the spirit of most of the points raised by those authors, we seriously question the effectiveness, utility, reliability and validity of a students' evaluations data with high incidence/prevalence of zero variation. [20], [21], [22], [24] and [23] has done a lot of research work highlighting the importance of adopting a multivariate view of students' evaluations of professors. Clearly, the multivariate aspect of the feedback sought is lost in the prevalence of too many single minded respondent. For all the above reasons, we deem the zero variation responses unreliable with respect to the multivariate view of the evaluation.

Definition 2 Let $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ be a dataset with $\mathbf{x}_i^{\top} = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{ip})$. Let the estimated variance of the *i*th respondent be $\tilde{S}_i^2 = \sum_{j=1}^p (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^2/(p-1)$. Let $Z_j = \sum_{i=1}^n \mathbf{x}_{ij}$ represent the sum of the scores given by all the *n* respondents to item *j*. Our respondent reliability is estimated by

$$\hat{\alpha} = \left(\frac{n}{n-1}\right) \left[1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{p} \left(\mathbf{x}_{ij} - \frac{1}{p} \sum_{j=1}^{p} \mathbf{x}_{ij}\right)^{2}}{\sum_{j=1}^{p} \left(\sum_{i=1}^{n} \mathbf{x}_{ij} - \frac{1}{p} \sum_{j=1}^{p} \sum_{i=1}^{n} \mathbf{x}_{ij}\right)^{2}}\right].$$
(2)

We use a straightforward adaptation of the Cronbach's alpha coefficient to measure and capture respondent reliability. Given a data matrix \mathbf{X} , respondent reliability can be computed in practice by simply taking the Cronbach's alpha coefficient of \mathbf{X}^{\top} , the transpose of the data matrix \mathbf{X} . Let m be the number of nonzero variation. If $m \ll p$ and m/n is very small, then respondent reliability will be very poor. Fortunately, for our data, respondent reliability is estimated at 0.996, which is very satisfactory. We think this large value is due to the fact that, despite having more than 50% zero variation respondents, we still a large enough sample. Despite this however, we will perform analyses taking into account the dichotomy between single minded respondents and their counterparts.

2.2 Exploratory Data Analysis and Basic Tests

As we said earlier, students' evaluations of instructors are administered with the goal of measuring the effectiveness (quality) of instructors and hopefully provide them (the instructors) with useful feedback to help them teach better. Clearly, such a goal is complex, and because of its complexity, there have always been heated and often very passionate debates about the validity and the appropriateness of such evaluations [20], [22], [32], [5]. As a matter of fact, many professors strongly believe and claim that students, especially undergraduate students, are neither mature enough nor knowledgeable enough nor objective enough to provide useful feedback to their instructors [10], [1], [30] and [25]. To a certain degree, such anti-students' evaluations professors do have a valid point because even with the crucial issues of maturity, knowledge and objectivity, there are very important points of concerns with students' evaluations of instructors: (a) a complex multidimensional instrument like a 28 items questionnaire should never be summarized using a single number (as it is commonly practiced around the world), because such a simplistic summarization definitely fails to capture all the niceties inherent in the com-

plex art of teaching (b) given the Likert type nature of the scores (responses), the often used grand average is at best misleading because averages computed on non-numeric variables are often meaningless [2]. It makes sense that only a multidimensional summary [20], [21], [22], [24] and [23] or better yet a functional summary (density or mass function) can meaningfully capture the pattern underlying a multidimensional instrument like the students' evaluations of instructors.

2.3 Univariate summaries

It's a very common practice among people dealing with Likert type data to use averages and standard deviations as their measures of central tendency and measures of spread (variation) respectively. Typically, students' evaluations questionnaires have **one** item aimed at measuring the overall rating of the professor being evaluated. At universities like Gazi University where the questionnaire does not have such a summarizing item, the grand mean (mean of all the means) is used as the estimate of the overall rating, namely

$$\operatorname{grandmean}(\mathbf{x}) = \frac{1}{np} \sum_{j=1}^{p} \sum_{i=1}^{n} \mathbf{x}_{ij} = \operatorname{mean}\left(\operatorname{mean}_{j=1:p}(s(\mathbf{x}_{j}))\right), \tag{3}$$

where $s(\mathbf{x}_i) = \{\mathbf{x}_{ij} : i = 1 \cdots, n\}$. When an instructor opens the website containing her/his student evaluation data, there are 28 averages, one for each questions, and then there is the average of those averages which is the grand mean representing the overall rating of the instructor. With $x_{ij} \in \{1, 2, 3, 4, 5\}$, such a grand average is at best misleading and at worst just plain invalid. In the hierarchy of data types, Likert type scores are no more than ordinal, which prohibits the use of averages. By their very nature, Likert-type observations are inherently definitely not numerical in the usual sense of interval or ratio data. Considering our motivating example of the students' evaluation of instructors, the use of the grand mean as the overall rating of the instructor misses the subtle and important information revealed by appropriate frequencies (proportions) and the corresponding bar plots. When the grand mean is used, Instructor 1 scores an average of 3.4, which of course tells us nothing about the distribution of her scores. The distribution for this instructor is skewed to the left, with a pronounced/strong mode at 4 for most of the questions/items. Although we still do not advocate the use of a single number to summarize a complex instrument like a students' evaluations of instructor, we would recommend trusting the mode rather than the mean if a single number were to be used. This led us to defining a grand mode in place of the invalid grand mean as follows: Let $s(\mathbf{x}_i) = \{\mathbf{x}_{ij} : i = 1 \cdots, n\}$ and let $\tilde{\mathbf{x}}_{ij} = \text{unique}(\mathbf{x}_{ij})$. If m_i denotes the mode of variable X_j , then for $j = 1, \dots, p$, we can readily compute the mode of the jth column as

$$m_j = \underset{\tilde{\mathbf{x}}_{ij} \in s(\mathbf{x}_j)}{\operatorname{arg}} \max \left\{ frequency(\tilde{\mathbf{x}}_{ij}) \right\} = \operatorname{mode}(s(\mathbf{x}_j)). \tag{4}$$

The set $M = \{m_1, m_2, \dots, m_p\}$ containing the modes for the p columns. Let $\tilde{m}_j = \text{unique}(m_j)$. We can find the grand mode as

$$\operatorname{grandmode}(\mathbf{x}) = \underset{\tilde{m}_j \in M}{\operatorname{arg max}} \left\{ frequency(\tilde{m}_j) \right\} = \operatorname{mode} \left(\underset{j=1:p}{\operatorname{mode}} (s(\mathbf{x}_j)) \right). \tag{5}$$

The grand mode for instructor 1 is found to be 4, which, in light of the distribution of her scores, is a more accurate summarization of her effectiveness and teaching quality. It might be tempting, given the ordinal nature of Likert-type data, to use the grand median

grandmedian(
$$\mathbf{x}$$
) = median $\left(\underset{j=1:p}{\text{median}} (s(\mathbf{x}_j)) \right)$, (6)

in place of the grand mean. From our experience, such a summarization is not as accurate as the grand mode, partly due to the floor and ceiling effect, see [8]. If instead of considering only instructor 1 we use the entirety of the data with all the n=5820 evaluations, the distribution of all the 28 course specific questions, it is then noted that most questions attain their mode at 3, and we find the grand mode to be 3. Thanks to the distributional features of the scores of the instructors in this dataset, namely the skewness to the left, we able to comment in a more complete manner on the effectiveness (or at least the students' perception thereof). With the highest frequencies being between 3 and 5, it is fair to say that the instructors evaluated here are **not** negatively perceived by their students.

2.3.1 Examining the Effect of Response Variation

Despite the ability to provide a more meaningful single summarization of the whole evaluation through the grand mode along with distributional qualifications, we still need to answer relational questions like the association between student maturity and their rating, student seriousness/dedication/objectivity and their rating. We now propose to focus on zero variation responses, as we believe that the reflect the reliability of the respondent. In a sense, we claim that a student who gives a zero variation response is providing a less objective and less mature answer to the survey. We then try to find out if there is an association between zero variation and the answers to the questions. First and foremost, it is interesting to assess the association between response variation and instructors. See Table 1.

	Zero Variation	Nonzero Variation	Total
Instructor 1	0.0789	0.0543	0.1332
Instructor 2	0.1309	0.1172	0.2481
Instructor 3	0.3031	0.3156	0.6187
Total	0.5129	0.4871	1

Table 1. Distribution of response variation among instructors

The chi-squared test of association between *Instructor* and *Response Variation* is significant, namely with $\chi^2_{\rm obs} = 28.45$, df = 2, p-value = 0.000. This means that there are some differences among instructors with respect to zero variation responses.

The lack of richness of the zero variation responses in this study is less concerning because most of such responses are either neutral or positive. In a sense, those who were single minded about their rating of the courses, were so mostly not because of dissatisfaction. Given the fact that the zero variation responses came from satisfied students (see higher percentage of Neutral, Agree, and Strongly Agree in Table (2) depicting the percentages of responses within the zero variation group), their feedback was really not needed with respect to the multidimensional aspect of the feedback sought. For that reason, we can proceed with the remaining aspects of the analysis of this data, secured that both the item reliability (measured by Cronbach's alpha) and the respondent reliability are satisfactory.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Instr 1	0.1678	0.0545	0.2288	0.2767	0.2723
Instr 2	0.1378	0.0407	0.3018	0.3084	0.2113
Instr 3	0.2086	0.0726	0.3294	0.2183	0.1712

Table 2. Proportion of each response category for answers with zero variation

2.3.2 Examining Various Important Associations

We now examine a variety of association between different important variables. Taking the view that attendance is a measure of dedication/seriousness, and therefore a decent and plausible indicator of the ability/authority of the student to correctly assess their instructor, we will now test the association of various variables with attendance. In other words, if a student is not dedicated ie not serious (as measured by attendance), their assessment should probably not be taken seriously. See [19] for a detailed account on the influence of the student's interest on their rating of their instructor. We also consider the variable difficulty, a self reported variable provided to allow the student to indicate their perception of the level of difficulty of the course. This variable is particularly important because some instructors strongly believe that students tend to give a negative feedback when they perceive the course to be difficult. Extensive studies on the impact of the difficulty level of the course being evaluated have been carried out by authors such as [28] and [9].

Association between Attendance Level and Response Variation for Instructor 3.

	Poor	Minimal	Good	Good	Excellent
nonzero	0.14912524	0.08747570	0.07220217	0.12663149	0.07470147
zero	0.22299361	0.07497917	0.05942794	0.07275757	0.05970564

Table 3. Cross tabulation of Attendance level vs Response Variation for Instructor 3.

As can be seen on Table (2.3.2), there is empirical evidence of a substantial difference between zero variation and nonzero variance respondent in the group with poor attendance. The corresponding chi-squared test of association between *Attendance level* and *Response Variation* is significant, specifically with $\chi^2_{\rm obs}=117.7398,~\nu={\rm df}=4,~p-value<2.2\times10^{-16}$.

Association between Difficulty Level and Response Variation for Instructor 3.

	Too Easy	Easy	Normal	Difficult	Too Difficult
nonzero	0.12385448	0.04498750	0.14940294	0.12968620	0.06220494
zero	0.19550125	0.03471258	0.10274924	0.09080811	0.06609275

Table 4. Cross tabulation of Difficulty level vs Response Variation given by students for Instructor 3.

In the group of those who deemed the course to be too easy, there appears to be some evidence of more zero variation respondents. More formally, the corresponding chi-squared test of association between *Difficulty level* and *Response variation* is significant, specifically with $\chi^2_{\text{obs}} = 117.7398$, $\nu = \text{df} = 4$, $p\text{-value} < 2.2 \times 10^{-16}$.

Various Tests of Association using the whole dataset

It appears that for all the evaluations provided for Instructor 3, both Attendance Level Difficulty Level are strongly associated with Response Variation. The question then arises

as to whether that association holds when all the 5820 evaluations are considered.

	Poor	Minimal	Good	Good	Excellent
nonzero	0.1273	0.0880	0.0718	0.1218	0.0782
zero	0.1995	0.0887	0.0643	0.0933	0.0672

Table 5. Cross tabulation of Attendance level vs Response Variation for all the Instructors.

The corresponding chi-squared test of association between Attendance level and Response variation is significant, namely with $\chi^2_{\rm obs} = 118.3$, $\nu = {\rm df} = 4$, $p\text{-value} < 2.2 \times 10^{-16}$

	Too Easy	Easy	Normal	Difficult	Too Difficult
nonzero	0.1065	0.0527	0.1601	0.1158	0.0519
zero	0.1718	0.0416	0.1447	0.0947	0.0601

Table 6. Cross tabulation of Difficulty level vs Response Variation for all the Instructors.

The corresponding chi-squared test of association between Difficulty level and Response variation is significant, namely with $\chi^2_{\rm obs} = 113.5$, $\nu = {\rm df} = 4$, $p\text{-value} < 2.2 \times 10^{-16}$.

Students with poor attendance give an overwhelmingly large number of zero variation answers whereas students with reasonable to excellent attendance level tend to give nonzero variation answers. This somewhat confirms or at least supports the strongly held belief that only those answers provided by dedicated/serious students should be taken into account. On the other hand, students who perceive a course as too easy and therefore boring or at least uninteresting also tend to give an overwhelming proportion of zero variation answers. Interestingly, students who think the course has a normal difficulty level tend to take time to provide varied answers to different questionnaire items.

Since we discovered interesting patterns between response variation and both attendance and difficulty level, it is interesting to examine if there might be a similar type of strong association between the courses and the response variation. Indeed many other authors have researched on the relationship between the nature of the courses taught and the rating of the instructors. See [19] and [21] for more information. In our dataset, there was a total of 13 courses included in the 5820 evaluations considered. Figure (1) depicts the barplot of the relationship between response variation and the courses. It can be seen that except for five courses, zero variation respondents are the majority. The chi-squared

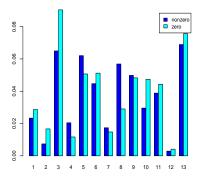


Figure 1. Cross-tabulation of response variation versus course indicator for all the instructors.

test of association between Course and Response variation was found to be significant, with $\chi^2_{\rm obs} = 150.7$, $\nu = {\rm df} = 12$, $p\text{-value} < 2.2 \times 10^{-16}$. Finally, we look at the overall relationship between attendance and the perceived diffi-

Finally, we look at the overall relationship between attendance and the perceived difficulty level of the course. Table (7) shows an overwhelming support in favor of a strong relationship, with dominance of the strength between too easy and poor.

	Too Easy	Easy	Normal	Difficult	Too Difficult
Poor	0.2263	0.0170	0.0380	0.0249	0.0206
Minimum	0.0137	0.0311	0.0653	0.0431	0.0234
Reasonable	0.0093	0.0131	0.0591	0.0393	0.0153
Good	0.0158	0.0187	0.0859	0.0679	0.0268
Excellent	0.0132	0.0144	0.0565	0.0352	0.0259

Table 7. Cross tabulation of Difficulty level vs Attendance Level for all the Instructors.

The corresponding chi-squared test of association between Difficulty level and Attendance Level is significant, with $\chi^2_{\rm obs} = 2528.06$, $\nu = {\rm df} = 16$, $p\text{-value} < 2.2 \times 10^{-16}$. The most obvious feature of this association is the astronomically high proportion of poor attendance in courses deemed too easy. No surprise here, just plain common sense. Sadly however, there is no category in which excellent attendance dominates.

Although all the 28 items in the questionnaire were carefully selected by the designers of the students' evaluation, one could make a strong case that some questions are better indicators of overall assessment than others. One such question is Q10: My initial expectations about the course were met at the end of the period or year. This question is in fact often used as the measure of the overall assessment of the course and the instructor at most American universities. The University of Central Florida students' evaluation questionnaire given in Appendix has a question phrased in a very similar way. Given its summarizing nature, we will use this question as a response/dependent variable in our supervise learning section.

3. Pattern Recognition and Association Analysis

In this section, we turn our attention to the multivariate aspects of our data set. We perform Factor Analysis to extract meaningful latent structure and cluster analysis to identify potential groups in the way students rate their professors. [28], [14], [34], [4], [29] and [33] are some of the authors who have used data mining and machine learning techniques on student evaluation data. As we said clearly in sections 1 and 2, Likerttype scores are inherently non-numeric, and applying techniques designed for numeric data on Likert-type will yield answers that are potentially meaningless or at best very difficult to interpret. When it comes to correlation analysis for instance, the default choice is the Pearson correlation measure. With Likert type data however, one wonders if Pearson correlation should ever be used. Based on recommendations by [2], [6], and [8], the correct type of correlation for Likert-type data should be Kendall-tau-B correlation or the Spearman correlation, as these are designed for (ordinal) ranked data. There have been many recent interesting contributions to the multivariate analysis of Likert-type: in her doctoral thesis, [16] provides a wide variety of univariate and multivariate tools for analyzing Likert-type data. [27] proposes the use of rough sets in the analysis of Likert-scale data. We start off by checking how different the Pearson correlation matrix would be from the Kendall-tau B correlation matrix on our data. Recall, that given two random variables X_i and X_j for which observed (realized) values $x_{1i}, x_{2i}, \dots, x_{ni}$ and

 $\mathbf{x}_{1j}, \mathbf{x}_{2j}, \cdots, \mathbf{x}_{nj}$ have been respectively gathered, the so-called Pearson sample correlation matrix is given by

$$r_{ij} = \operatorname{correlation}(\mathbf{x}_i, \mathbf{x}_j) = r(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{n-1} \sum_{\ell=1}^n \left(\frac{\mathbf{x}_{\ell i} - \bar{\mathbf{x}}_i}{s_{\mathbf{x}_i}} \right) \left(\frac{\mathbf{x}_{\ell j} - \bar{\mathbf{x}}_j}{s_{\mathbf{x}_j}} \right).$$

For a random p-tuple $X = (X_1, X_2, \dots, X_p)^{\top}$ and the corresponding data matrix $\mathbf{X} = (\mathbf{x}_{ij}), i = 1, \dots, n, j = 1, \dots, p$, the Pearson sample correlation matrix is given by

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}.$$

As we have been stressing all along, the Likert-type nature of our data makes the matrix \mathbf{R} meaningless, in the sense that the averages on which it is based may not have an interpretable meaning. If we consider two Likert type (ordered categorical) variables X and Y once again, their Kendall τ -B correlation coefficient $\tau_B(X,Y)$ is given by

$$\tau_B(X,Y) = \frac{n_c(X,Y) - n_d(X,Y)}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

where $n_0 = n(n-1)/2$, $n_1 = \sum_i t_i(t_i-1)/2$, $n_2 = \sum_j u_j(u_j-1)/2$, t_i =number of tied values in the i-th group of ties for the first quantity, u_j =number of tied values in the j-th group of ties for the second quantity, $n_c = number$ of concordant pairs, n_d =number of discordant pairs. For a p-tuple $X = (X_1, \dots, X_p)$ of p Likert type variables, the Kendall Tau-B correlation matrix is \mathbf{K} where

$$\mathbf{K} = \begin{bmatrix} \tau_B(X_1, X_1) \ \tau_B(X_1, X_2) \cdots \ \tau_B(X_1, X_p) \\ \tau_B(X_2, X_1) \ \tau_B(X_2, X_2) \cdots \ \tau_B(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \tau_B(X_p, X_1) \ \tau_B(X_p, X_2) \cdots \ \tau_B(X_p, X_p) \end{bmatrix}.$$

The empirical calculations based on our data reveal that the Pearson and the Kendall τ -B correlation matrices are so similar, in pattern and magnitude as to be almost indistinguishable (in fact, it turns out that $\hat{\mathbf{R}} \approx \hat{\mathbf{K}} + 0.05 I_p$)¹. For that reason, in our subsequent correlation-based analyses like Principal Component Analysis or Factor Analysis, we can use the Pearson in place of the Kendall τ -B correlation, since the latter is computationally very expensive.

3.1 Are there distinguishable groups among students?

A natural question that arises in the presence of data like the one we have, is whether the observations can be clustered. In other words, is there such a thing as different groups of students as far as their patterns of feedback to instructors are concerned? Can the patterns of students' evaluations of their instructors be grouped into distinct and clearly

¹This strong similarity is somewhat surprising because while Pearson is appropriate for numeric data types, Kendall τ -B is suitable for ordinal and rank like Likert data in our study.

describable categories? Now, one of the most celebrated approaches to cluster analysis is the ubiquitous kMeans clustering algorithm². Obviously, as the name suggests, it is based on the computation of averages that represent the centers of potential underlying groups. With Likert-type data, it has been stressed all along that averages are potentially meaningless because of the inherently categorical non numeric nature of such data. With the Pearson and Kendall-tau-B correlation matrices computed earlier showing strong similarities in pattern and magnitude, one might conjecture that it might not be wrong to use average-driven techniques on our data. The kMeans clustering algorithm in this case would proceed by partitioning the data into k clusters to form the optimal partitioning $\mathcal{P}^* = C_1^* \cup C_2^* \cup \cdots \cup C_k^*$ that minimizes the within-cluster sum of squares (WCSS). In other words, if \mathcal{P}^* denotes the best partitioning (clustering) of the data, we must have

$$\mathcal{P}^* = \operatorname{argmin}_{\mathcal{P}} \left\{ \sum_{j=1}^k \sum_{i=1}^n I(\mathbf{x}_i \in C_j) \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \right\}$$

where μ_j is the mean vector (center) for cluster C_j . From our kMeans clustering calculations in R, the percentage of variation explained seems to clearly suggest that one should retain three distinct clusters. Indeed, two clusters would capture a very low percentage of the variation in the data, while four clusters do not substantially improve the amount of variation captured by three clusters. We therefore retained three clusters and carefully examined both the percentage of observations in each one of them and the values of the centers. As Table 8 shows, one could venture to say that almost 60% of the students have a neutral opinion of the courses they took, and this seems to apply to almost all the 28 questions of the survey. The cluster analytic result also suggests that 17% expressed maximum satisfaction with the courses they took. Finally a third group of the students seems to be the group of very dissatisfied students, with our data showing roughly 23% of such students. These numbers apply to all the 5820 evaluations analyzed. It certainly would be more beneficial, in the interest of instructor's improvement, to extract such clustering for each course in order to help the instructor identify areas of improvement.

	Cluster 1	Cluster 2	Cluster 3
Average of Center	4.80	1.52	3.37
Number of Observations	1010	1364	3446
Percentage of observations	17.35%	23.44%	59.21%
Suggested class label	Satisfied	Dissatisfied	Neutral

Table 8. Clusters extracted using kMeans clustering.

The patterns discovered through kMeans clustering and revealed in Table (8) are interesting in their own right, but as we'll show later, we used the labels generated here to extend our analysis to supervised learning. As we mentioned earlier, the dataset [12] used here was gathered at Gazi University where there is no dedicated response variable in the questionnaire. Given this absence of response, we later use $Y \in \{\text{Dissatisfied}, \text{Neutral}, \text{Satisfied}\}$ as our response variable in both classification trees and random forest.

² In a previous analysis of the dataset used in this paper, [11] explored various of cluster analysis, including hierarchical clustering on both the raw data and transformed versions of the data for which the Jaccard distance was used. More details of that analysis can be found through the reference.

3.2 Are there meaning concepts underlying the items of the evaluation?

It goes without saying that 28 questions for a single respondent can be quite overwhelming. Besides, it's indeed very likely that many of the questions end measuring the same aspect of the perception of the student. Recall for instance that the correlation matrices calculated earlier revealed extremely large correlation values. We should therefore expect the 28 dimensional questionnaire given to students to boil down to a much lower number of latent concepts. From a factor analytic perspective, this means that the student evaluation vector $\mathbf{x}^{\top} = (\mathbf{x}_1, \cdots, \mathbf{x}_{28})$ does have a representation of the form

$$\mathbf{x} = \mathbf{\Lambda}\mathbf{z} + \boldsymbol{\epsilon}$$

where $\Lambda \in \mathbb{R}^{28 \times q}$ and $\mathbf{z}^{\top} = (\mathbf{z}_1, \dots, \mathbf{z}_q)$ for some $q \ll 28$. Factor Analysis typically assumes that the factor scores vector Z has a multivariate Gaussian (normal) distribution. Such an assumption is bound to be violated here because of the non-normality of the vector X. Many authors have performed factor analysis on Likert-type data despite this non-normality. [26] and [18] provide a detail account of the pitfalls resulting from the misuses of factor analysis on Likert-type data. It turns out that part of the problem with the use of factor analysis on Likert-type data stems from the fact that some analysts use the Pearson covariance matrix as their main ingredient. To somehow avoid the pitfalls and hope for meaningful factor analytic results, we use the Kendall τ -B correlation matrix as the basis of our factor analysis. Based on Table 9 our factor analytic seem to reveal the following facts: Questions 13 to 28 have estimated factor loadings that are all higher on factor 1 than they are on Factor 2. These 16 questions are all related to how the student rate the competence of the instructor teaching the course. We therefore name the first factor score Z_1 the "instructor rating score". Questions 1 to 12 have estimated factor loadings that are all higher on factor 2 than they are on Factor 1. These 12 questions are all related to how satisfied the student was about the course. We therefore name the second factor score Z_2 the "student satisfaction score".

(b) Revelation of Factor 1

	*			<u> </u>	
	Factor 1	Factor 2		Factor 1	Factor 2
Q1	0.376	0.781	Q13	0.753	0.558
Q2	0.495	0.767	Q14	0.794	0.517
Q3	0.567	0.689	Q15	0.791	0.514
Q4	0.475	0.770	Q16	0.705	0.611
Q5	0.505	0.793	Q17	0.827	0.391
Q6	0.497	0.776	Q18	0.762	0.541
Q7	0.465	0.819	Q19	0.790	0.517
Q8	0.456	0.815	Q20	0.825	0.475
Q9	0.545	0.699	Q21	0.844	0.447
Q10	0.524	0.791	Q22	0.846	0.446
Q11	0.564	0.680	Q23	0.756	0.564
Q12	0.486	0.751	Q24	0.713	0.593
			Q25	0.826	0.463
			Q26	0.749	0.543
			Q27	0.695	0.567
			Q28	0.811	0.452

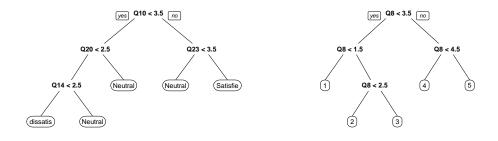
Table 9. Two-factor model from the p=28 questions on the Gazi University students' evaluation data. There was a total of n=5820 evaluations submitted by the students and used to estimate these factor loadings. It can be seen that the two factors discovered are quite straightforward.

The two factors described above captured 85% of the variation, and any attempt to generate/derive more factors resulted in very little gain along with the loss of interpretability inherent in these two factors. From a practical perspective, it seems to make sense that a student's answers would be summarized into their overall satisfaction along with some rating of the instructor who led the whole experience on the course. Clearly one could hypothesize more factors, but these two tend to intuitively capture what one would expect. To help better grasp the usefulness of the factor analytic patterns that we discovered in this data, we deem it appropriate to match the initial scores given by some students with the corresponding factor scores. For instance,

The above specific examples were provided to further show evidence that the 2-factor model extracted the latent concepts underlying the Gazi University students' evaluation instrument rather well. As we saw from all the above calculations and findings, one can readily estimate the satisfaction of a given student and their rating of their professor using two numbers, the factor scores, and in this case, the numbers are clear and unambiguous.

4. Supervised Learning Techniques

Until this point, all our analyses on this data have been entirely unsupervised. While we have discovered many interesting patterns in the data, we are now turning to supervised learning with the hope of discovering even more interesting aspects of how students rate their professors. Throughout this section, we'll concentrate on classification using trees and random forests. For all our random forest estimations we'll use 500 trees in



- (a) Classification tree with Y = Opinion.
- (b) Classification tree with Y = Q10.

Figure 2. Classification Trees with two different Response Scenarios.

the ensemble. Since the data came without a specifically dedicated response variable, we'll use Q10 as one of our response variable for reasons mentioned earlier. We will use another response variable herein denoted by Opinion, whose domain contains the labels Dissatisfied, Neutral, Satisfied generated earlier from kMeans clustering.

4.1 Classification Tree Learning

The great appeal of trees was triggered by our interest in finding out if there were some questions that drove the classification and that could therefore be considered somewhat key questions in the evaluation. Classification trees are usually highly preferred by analysts who desire an interpretable learning machine. Understanding trees is indeed straightforward as they are intuitively appealing piecewise functions operating on a partitioning of the input space. Given $\mathscr{D} = \left\{ (\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n) \right\}$, with $\mathbf{x}_i \in \mathscr{X}$, $Y_i \in \{1, \cdots, G\}$. Let $T = \bigcup_{\ell=1}^q R_\ell$ denote the tree represented by the partitioning of \mathscr{X} into q regions R_1, R_2, \cdots, R_q . Given a new point \mathbf{x}^* , its predicted response is

$$\hat{Y}_{\text{Tree}}^* = \hat{f}_{\text{Tree}}(\mathbf{x}^*) = \sum_{\ell=1}^q \left\{ I(\mathbf{x}^* \in R_\ell) \left\{ \underset{j \in \{1, \dots, G\}}{\operatorname{argmax}} \left\{ \frac{1}{|R_\ell|} \sum_{\mathbf{x}_i \in R_\ell} I(Y_i = j) \right\} \right\} \right\}.$$

The classification tree software implementation used here is taken from the R package rpart. Using the labels Dissatisfied, Neutral, Satisfied as our response levels, we get the tree depicted in Figure (2(a)), which clearly reveals variable Q10 as the root, somewhat lending support to our earlier speculation around the possibility of using Q10 as the response variable because of its apparent summarizing nature. On Figure (2(b)), we depict the classification tree generated using Q10 as the response. Bearing in mind the instability of classification trees (high estimation variance), we dedicate the last subsection of this paper to an extension of tree learning, namely the ubiquitous ensemble learning method known as Random Forest [7], which we apply to our data in the next section. One extra motivation for resorting to ensemble learning via Random Forest lies in our desire to estimate how well the rating of a professor can be predicted, but also to estimate the importance of each of the variables used in the prediction.

4.2Ensemble Learning with Random Forests

Let's consider once again the multi-class classification task as defined much earlier with labels y coming from $\mathscr{Y} = \{1, 2, \cdots, G\}$ and predictor variables $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)^{\top}$ coming from a p-dimensional space \mathscr{X} . Along the same lines of [13], let $\widehat{g}^{(b)}(\cdot)$ be the bth bootstrap replication of the estimated base classifier $\widehat{g}(\cdot)$, such that $(\widehat{y})^{(b)} = \widehat{g}^{(b)}(\mathbf{x}^*)$ is the bth bootstrap estimated class of \mathbf{x}^* . The estimated response by Random Forest is obtained using the majority vote rule, which means that the most frequent label throughout the B bootstrap replications of random subspace learning. The following algorithmic description taken from [13] captures the essential structure of the Random Forest[7] learning method.

Algorithm 1 Random Subspace Learning for Model Aggregation

```
\triangleright The Random Forest Algorithm for B trees
 1: procedure RANDOMFOREST(B)
            Choose a base learner \widehat{g}(\cdot)
                                                                                                                                      ⊳ e.g.: Trees
                                                                                                       ▷ e.g.: Recursive Partitioning
            Choose an estimation method
 3:
 4:
            for b = 1 to B do
                 Draw with replacement from \mathscr{D} a bootstrap sample \mathscr{D}^{(b)} = \{\mathbf{z}_1^{(b)}, \cdots, \mathbf{z}_n^{(b)}\} Draw without replacement from \{1, 2, \cdots, p\} a subset \{j_1^{(b)}, \cdots, j_d^{(b)}\}
 5:
 6:
                 Drop unselected variables from \mathscr{D}^{(b)} so that \mathscr{D}^{(b)}_{\mathrm{sub}} is d dimensional Build the bth base learner \widehat{\mathbf{g}}^{(b)}(\cdot) based on \mathscr{D}^{(b)}_{\mathrm{sub}}
 7:
 8:
 9:
            Use the ensemble \{\widehat{\mathbf{g}}^{(b)}(\cdot), b = 1, \cdots, B\}, the predicted label of \mathbf{x}^* is
10:
```

$$\widehat{f}^{(\mathrm{RF})}(\mathbf{x}^*) = \arg\max_{\mathbf{y} \in \mathscr{Y}} \left\{ \operatorname{freq}_{\widehat{\mathbf{C}}^{(\mathrm{B})}(\mathbf{x}^*)}(\mathbf{y}) \right\} = \arg\max_{\mathbf{y} \in \mathscr{Y}} \left\{ \sum_{b=1}^{B} \left(\mathbf{1}_{\{\mathbf{y} = \widehat{\mathbf{g}}^{(\mathrm{b})}(\mathbf{x}^*)\}} \right) \right\}.$$

11: end procedure

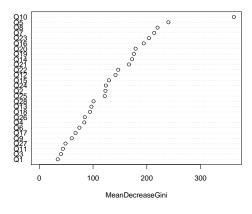
Table (10) depicts the confusion matrix of the 500^{th} random tree of the forest built using Opinion as the response. The last column is actually the training error and should not be mistaken for the average test error that measures the generalization ability of random forest. The R Package randomForest automatically gives the out of bag (OOB) error for each random tree. In the spirit of [13], we shall use the average OOB error AVOOB(\cdot), as our measure of predictive performance, namely

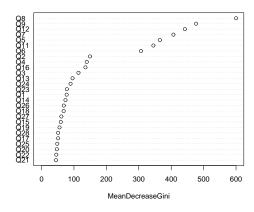
$$\text{AVOOB}(\widehat{f}^{(\text{RF})}) = \frac{1}{B} \sum_{b=1}^{B} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell(\mathbf{y}_i^{(b)}, \widehat{\mathbf{g}}^{(b)}(\mathbf{x}_i^{(b)})) \right\}, \tag{7}$$

where the observations $\{(\mathbf{x}_i^{(b)}, \mathbf{y}_i^{(b)}), i = 1, \dots, m\}$ are the $m \approx \lceil en \rceil$ observations not selected by the bootstrap sampling with replacement process.

	Dissatisfied	Neutral	Satisfied	class.error
Dissatisfied	1,219	20	0	0.016
Neutral	25	2,306	27	0.022
Satisfied	0	31	2,192	0.014

Table 10. Confusion Matrix corresponding to random forest classification with Opinion as the response variable.





- (a) Random Forest with Opinion as response.
- (b) Random Forest with Q10 as response.

Figure 3. Variable Importance plots yielded by Random Forest using different scenarios of response variable.

The average OOB error obtained from the above forest comes out as 0.02115943, meaning that the Random Forest classifier is 98% accurate. We also built a 500-tree random forest with Q10 as the response. Table (11) shows the corresponding confusion matrix, and our average out of bag error for this random forest is found to be 0.1389903, which interesting corresponds to an accuracy of 86%, the percentage of variation captured by the twofactor factor analytic model discovered earlier, and also the percentage of variation that warranted the selection of the 3 clusters solution in our kMeans clustering analysis.

	1	2	3	4	5	Class Error
1	862	63	15	1	1	0.085
2	44	594	135	14	0	0.245
3	10	109	1,521	123	8	0.141
4	1	18	138	1,248	36	0.134
5	1	4	14	63	797	0.093

Table 11. Confusion Matrix corresponding to random forest classification with Q10 as the response variable.

One of the greatest appeals of Random Forest lies in its ability to supplement excellent predictions with estimates of variable importance. For our dataset, we generated two different variable importance plots (3), one using Q10 as the response and the other using Opinion as the response. Figure (3(a)) shows the overwhelming dominance of Q10, again confirming our initial conjecture/speculation. Interestingly, out of the 5 most important variables in this case, only one, namely the 5th, is related to the professor. On the other hand, Figure (3(b)) reveals that to directly predict Q10 well based on the remaining 27 variables, the most important variable is Q8, dominating the rest substantially. Shockingly, for the prediction of Q10, virtually no professor-related variable (Q13-Q28) appear to be important. It gives the impression that student satisfaction has nothing to do with the professor. Interestingly also, Q10 is almost perfectly predicted by Q8 which has to do with the grades student had on the course. Could it be then that students don't care about who the professor is, as long as they end up with a good grade on the course?

5. Conclusion and Discussion

We have provided a comprehensive statistical analysis of a relatively large dataset containing students' evaluations of various courses at a university in Turkey. Factor Analytic results appear to reveal a very plausible two factor model suggesting that students' evaluations inherently reveal the overall satisfaction of the student at the end of the course along with impact the instructor had on their overall satisfaction. With the instructor's factor coming out as the most dominant one, it is fair to say that the instructor does play a central role in the over experience of the student. Anyone analyzing students' evaluations should be careful to consider the number of zero variation responses and examining their association with the pattern of answers provided by the students. We strongly believe that these zero variation responses somewhat determine the quality of the survey and reliability of the answers provided. We have shown evidence to support the fact dedicated students (attendance) will tend to reveal a more satisfactorily learning experience than those students who do not take their course seriously. We combined unsupervised and supervised learning techniques and were able, not only to find meaningful and interpretable groups in the data, but also identify the items in the questionnaire that appeared to be driving the students' assessment of their learning experience. The dominance of Question 10 on the three class recognition tree confirmed our intuition in the sense that it is the question that seems to measure the overall satisfaction of a student on a course, and it is re-assuring to have the tree model reveal it. From a questionnaire design perspective, it is our view that 28 questions is a bit too much for the students, and this usually large survey might be the reason why some students ended up giving zero variation responses. We would also like to suggest the use of two questions that have been found to be very revealing of the experience of student, namely (1) What is your overall rating of this instructor? (2) Would you recommend this course to any other student?. Although these two questions are inherently correlated part of our future work on this data will consist of adapting traditional classification trees to Likert-type data. This essentially boils down to using Likert-type specific loss functions for splitting the nodes of the tree. We specifically plan on deriving adaptations of the Jaccard distance as loss function or using the cross entropy measure on the tendencies of respondents. We are also planning to include the final grades of the respondents. The motivation for this is the fact that many instructors around the world have repeatedly argued that students who know (based on quiz scores, homework assignment scores, and midterm exam scores) that they will be receiving a good grade on the course tend to rate their professors very highly. We plan on finding out if there is evidence to support such a belief.

5.1 Acknowledgements

Ernest Fokoué wishes to express his heartfelt gratitude and infinite thanks to Our Lady of Perpetual Help for Her ever-present support and guidance, especially for the uninterrupted flow of inspiration received through Her most powerful intercession.

References

- [1] Abrami, P. C., W. J. Dickens, R. P. Perry, and L. Leventhal (1980). Do teacher standards for assigning grades affect student evaluations of teaching. *Journal of Educational Psychology* 72(1), 107–118.
- [2] Adams, E., R. F. Fagot, and R. E. Robinson (1965). A theory of appropriate statistics. Psychometrica 30(2), 99–127.
- [3] Allen, I. and C. A. Seaman (2007). Likert scales and data analyses. Quality Progress. 47, 417–442.

- [4] Badur, B. and S. Mardikyan (2011). Analyzing teaching performance of instructors using data mining techniques. *Informatics in Education* 10(2), 245257.
- [5] Basow, S. A. (1995, December). Student evaluations of college professor: When gender matters. Journal of Educational Psychology 87(4), 656–665.
- [6] Boone, H. and A. Boone (2012). Analyzing likert data. Journal of Extenson 50(2).
- [7] Breiman, L. (2001). Random forests. Machine Learning 45(1), 5–32.
- [8] Clason, D. and T. Dormody (1994). Analyzing data measured by individual likert type items. *Journal of Agricultural Education*. 33(4), 31–35.
- [9] Clayson, D. E. (2009). Student evaluations of teaching: are they related to what students learn? a meta-analysis and review of the literature. *Journal of Marketing Education* 31(1), 1630.
- [10] Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. Research in Higher Education 4, 69–111.
- [11] Gündüz, N. and F. E. (2013, November). Data mining and machine learning techniques for extracting patterns in students' evaluations of instructors. Technical Report 1746, Rochester Institute of Technology, Rochester, New York, USA.
- [12] Gunduz, N. and E. Fokoue (2013). UCI machine learning repository.
- [13] Gunduz, N. and E. Fokoue (2015, January). Robust Classification of High Dimension Low Sample Size Data. arXiv.org stat.AP.
- [14] Hsu, C., B. Chang, and H. Hung (2007, December). Applying svm to build supplier evaluation model - comparing likert scale and fuzzy scale. In 2007 IEEE International Conference on Industrial Engineering and Engineering Management, Singapore, pp. 6–10. IEEE.
- [15] Jamieson, S. (2004). Likert scales: How to (ab)use them. Medical Education. 38(38), 1212–1218.
- [16] Javaras, K. N. (2004, Hilary Term). Statistical Analysis of Likert Data on Attitudes. Ph.d. thesis, Balliol College, University of Oxford.
- [17] Likert, R. (1932, June). A technique for the measurement of attitudes. Archives of Psychology 22(140), 5–55.
- [18] Lubke, G. and B. Muthen (2004). Factor-analyzing likert-scale data under the assumption of multivariate normality complicates a meaningful comparison of observed groups or latent classes. Structural Equation Modeling 11 (514-534).
- [19] Marsh, H. and T. Cooper (1981). Prior subject interest, students' evaluation, and instructional effectiveness. Multivariate Behavioral Research 16, 82–104.
- [20] Marsh, H. W. (1982). Validity of students' evaluations of college teaching a multitrait-multimethod analysis. Journal of Educational Psychology 74(2), 264–279.
- [21] Marsh, H. W. (1983, February). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology* 75(1), 150–166.
- [22] Marsh, H. W. (1984, Oct). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology* 76(5), 707–754.
- [23] Marsh, H. W. (2007). The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective, Chapter Student's evaluations of university teaching: A multidimensional perspective, pp. 319–384. New York: Springer.
- [24] Marsh, H. W. and M. J. Dunkin (1992). Higher education: Handbook of theory and research, Volume 8, Chapter Student's evaluations of university teaching: A multidimensional perspective, pp. 143–233. New York: Agathon Press.
- [25] Marsh, H. W. and L. A. Roche (1997, November). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. American Psychologist 52(11), 1187–1197.
- [26] Muthen, B. and D. Kaplan (1992). A comparison of some methodologies for the factor analysis of non-normal likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology* 45(19-30).
- [27] Narli, S. (2010, March). An alternative evaluation method for likert type attitude scales: Rough set data analysis. Scientific Research and Essays 5(6), 519–528.
- [28] OConnell, D. Q. and D. J. Dickinson (1993). Student ratings of instruction as a function of testing conditions and perceptions of amount learned. Journal of Research and Development in Education 27(1), 18–23.
- [29] Ola, A. F. and Pallaniappan (2013). A data mining model for evaluation of instructors' performance in higher institutions of learning using machine learning algorithms. *International Journal of Conceptions on Computing and Information Technology* 1(2), 17–22.
- [30] Peterson, C. and S. Cooper (1980). Teacher evaluations by graded and ungraded students. *Journal of Educational Psychology* 72(5), 682–685.
- [31] Sisson, D. and H. Stoker (1989). Analyzing and interpreting likert-type survey data. The Delta Pi

- Epsilon Journal. 3(2), 81–85.
- [32] Stedman, C. (1983). The reliability of teaching effectiveness rating scale for assessing faculty performance. Tennessee Education 12(3), 25–32.
- [33] Syed, S. J., Y. H. Jiang, and L. Golab (2014). Data mining of undergraduate course evaluations. In Proceedings of the 7th International Conference on Educational Data Mining, pp. 347–348.
- [34] Wang, M. C., C. D. Dziuban, I. J. Cook, and P. D. Moskal (2009). *Quality Research in Literacy and Science Education*, Chapter Dr. Fox Rocks: Using Data-mining Techniques to Examine Student Ratings of Instruction, pp. 383–398. Springer.

6. Appendices

6.1 Appendix

Student questionnaire from Gazi University:

- Q1: The semester course content, teaching method and evaluation system were provided at the start.
- Q2: The course aims and objectives were clearly stated at the beginning of the period.
- Q3: The course was worth the amount of credit assigned to it.
- Q4: The course was taught according to the syllabus announced on the first day of class.
- Q5: The class discussions, homework assignments, applications and studies were satisfactory.
- Q6: The textbook and other courses resources were sufficient and up to date.
- Q7: The course allowed field work, applications, laboratory, discussion and other studies.
- Q8: The quizzes, assignments, projects and exams contributed to helping the learning.
- Q9: I greatly enjoyed the class and was eager to actively participate during the lectures.
- Q10: My initial expectations about the course were met at the end of the period or year.
- Q11: The course was relevant and beneficial to my professional development.
- Q12: The course helped me look at life and the world with a new perspective.
- Q13: The Instructor's knowledge was relevant and up to date.
- Q14: The Instructor came prepared for classes.
- Q15: The Instructor taught in accordance with the announced lesson plan.
- Q16: The Instructor was committed to the course and was understandable.
- Q17: The Instructor arrived on time for classes.
- Q18: The Instructor has a smooth and easy to follow delivery/speech.
- Q19: The Instructor made effective use of class hours.
- Q20: The Instructor explained the course and was eager to be helpful to students.
- Q21: The Instructor demonstrated a positive approach to students.
- Q22: The Instructor was open and respectful of the views of students about the course.
- Q23: The Instructor encouraged participation in the course.
- Q24: The Instructor gave relevant homework assignments/projects, and helped/guided students.
- Q25: The Instructor responded to questions about the course inside and outside of the course.
- Q26: The Instructor's evaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives.
- Q27: The Instructor provided solutions to exams and discussed them with students.
- Q28: The Instructor treated all students in a right and objective manner.

6.2 Appendix

Student perception of instruction items for the University of Central Florida: Source Questions Administration

- (1) Feedback concerning your performance in this course was:
- (2) The instructors interest in your learning was:
- (3) Use of class time was:
- (4) The instructors overall organization of the course was:
- (5) Continuity from one class meeting to the next was:

January 12, 2015 Journal of Applied Statistics g

- (6) The pace of the course was:
- (7) The instructors assessment of your progress in the course was:
- (8) The texts and supplemental learning materials used in the course were:

Board of regents

- (9) Description of course objectives and assignments:
- (10) Communication of ideas and information:
- (11) Expression of expectations for performance:
- (12) Availability to assist students in or outside of class:
- (13) Respect and concern for students:
- (14) Stimulation of interest in the course:
- (15) Facilitation of learning:
- (16) Overall assessment of instructor: