**PREDICTING HND CERTIFICATES IN POLYTECHNICS IN GHANA – AN APPLICATION OF DISCRIMINANT ANALYSIS.**

**Buckman Akuffo1\***

1Department of Applied Mathematics and Statistics

Faculty of Applied Science and Technology

Koforidua Polytechnic, Ghana.

\*E-mail for correspondence: akuffobuckman@yahoo.com

**ABSTRACT**

It is important that to help students in their course of studies as to what will help them graduate with good certificates. Analyzing the contributory factors that shape students graduation is very imperative. This is even more important at a time when many students want to further their education after the Higher National Diploma. This paper uses age of students, type of certificate used in gaining admission, grade point average, scores obtained in Probability Theory, Element of Mathematics, Statistical Computing of 360 students from the Applied Mathematics Department to classify them on the type of certificate they would receive at graduation. Discriminant analysis was briefly reviewed and adopted to successfully classify students into satisfactory certificate (First class, second class upper and second class lower divisions) and unsatisfactory certificate (Third class or fail). The results revealed that the variables so selected adequately classify the students into receiving satisfactory or unsatisfactory certificate. It is recommended that instructional intervention be carried out on the courses selected since they form the basis of the type of certificate students eventually graduate with. Age of students and entry certificate were found not to have statistically difference among the two groups.

**Keywords**: Classification, confusion matrix, linear discriminant function, variance-covariance

**INTRODUCTION**

Students’ academic performance hinges of a number of factors. When these factors are carefully analyzed, the benefits to both the students and educational institutions are enormous. There have been many studies in the prediction of academic performance in higher educational institutions with mixed approaches and hence results. According to Paliwal M. et al. (2008) the critical issue is that, most of these studies use hard data such as the overall grade average from previous schooling, admission test scores, work experience, age, and gender. Again, sometimes studies on prediction of students’ academic performance include goal statement, references, personal interview among others (Pritchard M. E., and Wilson G. S., 2003).

Most of the early research and even current research involved academic prediction in terms of class, chances for admission. However, it is important to help students to identify whether they are graduating with a satisfactory class, ***Sat-HND***, (First class, second class upper or second class lower) or unsatisfactory class of certificate, ***Unsat-HND***, (Third class or Fail). But one influential difference among humans is personality. With the quest of students wanting to further their programs to obtain higher qualifications, it is important to classify students based on some identified factors the kind of certificate they may potentially get at graduation within the first two years of study. This will then help potential lower graduating students to back up and spur those doing well to work harder.

Of the many categories of students that are admitted to polytechnics nowadays, there is an increase in those who are enrolled through the access programme. Through this, candidates are taken through a three-month access course after which examination is organized and those who pass successfully are then admitted to the main higher national diploma (HND). Over the years, some lecturers have express reservation about the potential of these students and their potential of graduating with a satisfactory certificate or unsatisfactory certificate.

This paper will show how discriminant function can be used to help determine what variables have relationship with performance, and an illustration of using the discriminant function to predict graduating students’ class of higher national diploma in a polytechnic. Consequently, the objective of this paper is to develop a discriminant function that will discriminate among the classes of HND diploma and further give a generalization of this function as a tool for classifying an individual student to a satisfactory certificate, ***Sat-HND***, (First class, Second upper and second lower divisions) or unsatisfactory certificate, ***Unsat-HND***, (Third class or Fail). This is actually based on the individual profile of scores on the set of predictor variables in the future.

**MATERIAL AND METHODS**

**Justification of discriminant analysis**

Several statistical techniques are available to analyze the differences between groups. The traditional approach has been to use simple measures of correlation. That the correlation between the various factors may be computed and those with less, negative or positive correlations grouped. However such measures are often misleading as they give no information about the relative effects of multiple measures. Again, the correlations do not provide a means of assessing the cumulative effects of several variables.

In other breath, multiple regressions are used. That is we determine a linear relationship between a set of predictors and a criterion. This technique provides a measure of association in terms of the amount of variation accounted for by each variable, along with an estimate of their combined effect on the criterion or groupings. One major inappropriateness of linear multiple regressions in this case is the fact that most variables used in educational intervention are not continuous but discrete. To avoid a spurious regression or analysis, a more appropriate technique under the current circumstances is the discriminant analysis.

**Methods**

Discriminant analysis assumes linear relations among the independent variables. Suppose you

have data for  groups, with N observations per group. Let  represent the total number of observations. Each observation consists of the measurements of  variables. The  observation is expressed as. Again let  represent the vector of means of these variables across all groups and the vector of means of observations in the group. Let us define three sums of squares and cross products matrices, as  as follows:

 … … … … … … … (**1**)

and

… … … … … … … (**2**)

then

 … … … … … … … … … (**3**)

A discriminant function is a weighted average of the values of the independent variables. The weights are selected so that the resulting weighted average separates the observations into the groups. High values of the average come from one group; low values of the average come from another group. The problem reduces to one of finding the weights that, when applied to the data, best discriminate among groups according to some criterion. The solution reduces to finding the eigenvectors,, of . The canonical coefficients are the elements of these eigenvectors.

A goodness – of – fit parameter, Wilks’ Lambda is defined as:

 … … … … … … … … (**4**)

Where the eigenvalue corresponding to the eigenvector described above and *m* is the minimum value of *K – 1* and *p.* the canonical correlation between the *j th* discriminant function and the independent variable related to these Eigen values as given by:

 … … … … … … … … … (**5**)

Then the families of linear discriminant functions (*LDFk*) can be defined as:

 … … … … … … … … … (**6**)

Where

 … … … … … … … … … (**7**)

The interest in predicting in the application of discriminant analysis has overshadowed the fact that there are two (2) set of techniques in discriminant analysis. These are the predictive discriminant analysis (PDA) and Descriptive discriminant analysis (DDA). In the predictive discriminant analysis, the focus is on classifying subjects into one of several groups or predicate group membership, whereas in descriptive analysis the focus is on revealing major differences among groups (Stevens, 1996). Also according to Huberty and Barton (1989) the purpose of the two analyses are different but there are some feasibility of combining the two techniques for corroborating results. They however stated that research questions are either of predictive or descriptive types and must be view as such. According to Hurberty (1994) discriminant analysis for the past first four decades focused on prediction of group membership. However, descriptive discriminant analysis usage did not appear until the 1960s but its usage has been very limited to apply research settings over the past two decades

Predictive discriminant analysis is appropriate when the researcher is interested in assigning units or individuals to groups based on composite scores on several predictor variables. PDA focus is its prediction and the accuracy of hit rates and is interested in determining a classification rule and assessing its accuracy.

In many ways, discriminant analysis parallels multiple regression analysis. The main difference between these two techniques is that regression analysis deals with a continuous dependent variable, while discriminant analysis must have a discrete dependent variable. The methodology used to complete a discriminant analysis is similar to regression analysis. You plot each independent variable versus the group variable. You often go through a variable selection phase to determine which independent variables are beneficial. Then one has to conduct a residual analysis to determine the accuracy of the discriminant equations.

**Data**

The data for this study was chosen from records for the first and second year Higher National Diploma (HND) students drawn from the evening session programme in the Applied Mathematics department during the 2012/2013, 2013/2014, 201/2015 and the 2015/2016 academic years in Koforidua Polytechnic. The students were grouped into two groups as to the class of certificate they were to receive. The groups were either satisfactory, ***Sat-HND***, (first class, second class upper, second class lower) or unsatisfactory, ***Unsat-HND***, (third class).

The overall grade point average (GPA), age of students, highest level of education certificate obtained by students (JSH cert., SHS cert, Vocational certificate and those with no formal qualification), and scores obtained in Element of Mathematics (STA 101), Probability Theory (STA 107) and Statistical Computing (STA 207) were found to be the predictors with significant independence and combined contributions after a forward stepwise analysis was conducted. In order to confirm that these variables are the best subsets of the predictor variables, all possible subsets were ran but yielded the same results as affirmed by Huberty (1989) and Thompson (1995). A factor analysis was also conducted on several variables.

**RESULTS AND DISCUSSION**

Discriminant analysis normally brings a lot of outputs. The most silent and relevant outputs are presented and discussed in line with the objective of this paper.

The analysis is presented as preliminary analysis and detailed analysis for clarity of expression and target reading.

**Preliminary analysis**

Table 1 shows the group statistics for the independent variables for the two groups of students we wish to classify.

**Table 1:** Summary Statistics

|  |  |  |  |
| --- | --- | --- | --- |
| ***Class*** | ***Mean*** | ***Std Deviation*** | ***Valid N(listwise)*** |
| ***Unweighted*** | ***Weighted*** |
| Satis HND GPA STA 101 STA 107 STA 217 Age Cert | 3.0564.7655.9357.8324.602.00 | 0.5967.074.539.8364.3870.518 | 180180180180180180 | 180180180180180180 |
| Un-satis HND GPA STA 101 STA 107 STA 217 Age Cert | 1.92443.0240.5853.4324.682.083 | 0.51011.208.5810.3404.3030.332 | 180180180180180180 | 180180180180180180 |
| Total GPA STA 101 STA 107 STA 217 Age Cert | 2.48851.9750.7154.3324.642.041 | 0.7916.938.4412.9994.3400.436 | 360360360360360360 | 360360360360360360 |

From Table 1, it can be seen that, as expected the mean GPA for the group with satisfactory certificate of 3.05 is higher than the 1.92 recorded for the other group. Again, scores obtained in STA 101, and STA 107 for the satis-HND are all higher than the scores obtained by the other group with unsatisfactory certificates. For example, a score of 43.02 in STA 101 for the unsatisfactory group implies the candidate fails in that course. In essence students who failed in the courses were predicted to have receive unsatisfactory certificate except STA 217.

Again it is worth noting that the average ages for both groups are basically at 24.60 and 24.68 years with the prospective unsatisfied HND holders slightly older than the other groups. The total average of the 360 cases analyzed shows an average GPA of 49.68 approximately 50, which is the minimum pass mark for the courses. However, the total average GPA score of 2.488 falls in with the minimum satisfactory grade point of second class lower division.

The test of equality of group means for each of the independent variable is presented in Table 2.

**Table 2:** Test of Equality of Group Means for the classes of HND

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***Wilks’ Lambda*** | ***F*** | ***df1*** | ***df2*** | ***Sig.*** |
| GPASTA 101STA 107STA 217AgeCert | 0.4910.6010.5430.6291.0000.991 | 371.150237.725259.553213.0470.0333.303 | 111111 | 358358358358358358 | 0.0000.0000.0000.0000.8560.070 |

The results from Table 2 shows that whereas, grade point average (GPA) and scores obtained in Element of Mathematics (STA 101), Probability Theory (STA 107) and Statistical computing (STA 217) are statistically different the students, age of students (AGE) and type of certificate used for admission (Cert) do not differ statistically for the two groups we intend to discriminant or classify.

Again, the lambda of 1.00 for the ages of the students shows that the mean ages for the two groups are not statistically different. The lambda value of 0.991 for the type of certificate used by the two groups also implies they are not statistically different for the two groups. However, the lambda of 0.491 and 0.601 recorded in the grade point average (GPA) and scores obtained in Element of Mathematics (STA 101) respectively show that there are higher proportional variance that exist between these variables for the two groups we seek to classify. Indeed the p-value of 0.856 and 0.070 for age of students and type of certificates respectively means that age and certificate are not statistically different for the two groups.

The implication of this result is that inasmuch as there are variables which can show clear distinction between the two groups, that is scores obtained in Elements of Mathematics (101), Probability Theory (STA 107) and Statistical Computing (STA 217) as well as the grade point average (GPA), there are also factors that show no clear difference between the two groups in respect to the preliminary analysis. This leads us to further analysis and the relevance of discriminant theory in achieving the objective of this paper.

**Further analysis**

In further analysis, we present confusion matrix for actual and predicted categories of class of HND obtainable by the students, and related validated confusion matrix for discussion.

Table 3 we presents the observed categories for the class of certificate in the rows and columns totals are the predicted categories.

**Table 3:** Confusion matrix for actual and predicted categories of class of HND.

|  |  |  |
| --- | --- | --- |
| **Class of HND** | **Predicted Group Membership** | **Total** |
| ***Sat-HND*** | ***Unsat-HND*** |
| 1st Class2nd Class upper2nd Class lower3rd Class | 361011513 | 0512178 | **36****106****27****191** |
| **Total** | **165** | **195** | **360** |

According to Table 3, 165 students were predicted to obtain satisfactory certificate (first class, second upper and lower divisions) whereas 195 students were predicted to have unsatisfactory certificates (third class). From the confusion matrix, it can be seen that of the 165 students who were predicted to have satisfactory certificates, (first class, second upper or second lower division), actually a total of 152 were able to do so. This represents a Hit ratio of 92.12%. Also, of the 195 students predicted to obtain unsatisfactory certificate, a total of 178 students were able to obtain same representing a Hit Ratio of 91.28%.

Table 4 shows the classification of groups and the corresponding cross validation with its percentages.

**Table 4:** Classification Results b, c

|  |  |  |
| --- | --- | --- |
| Class | **Predicted Group membership** | **Total** |
| Satis HND | Unsatis HND |
| Original | Count Satis HND Unsatis HND | 10010 | 20110 | 120120 |
| % Satis HND Unsatis HND | 83.38.3 | 16.791.7 | 100100 |
| Cross-validateda | Count Satis HND Unsatis HND | 9810 | 22110 | 120120 |
| % Satis HND Unsatis HND | 81.78.3 | 18.391.7 | 100100 |

| a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case. |
| --- |
| b. 87.5% of original grouped cases correctly classified. |
| c. 86.7% of cross-validated grouped cases correctly classified. |

The classification results (Table 4) reveal that 87.5% of students were classified correctly to receive satisfactory certificate and 86.7% were classified correctly to receiving unsatisfactory certificate. Again from Table 4, the success in identifying students who will graduate with satisfactory certificate was 83.3% which has been validated as 81.7%. Again from Table 3, it was realized that the success of classifying students with unsatisfactory certificate was 91.28% and this again has been validated in Table 4. A look at the Hit Ratio in Table 3 and the validation results in Table 4 shows that the results are essentially the similar.. Hence this shows that the classification result of the historical sample was not biased upward.

The conducted discriminant analysis is a two group analysis hence one discriminant function was performed successfully. This makes it very possible to discriminant between students who will graduate with satisfactory HND and those to graduate with unsatisfactory HND certificates. Table 5 shows the evaluation of the function according to the means of canonical variables.

 **Table 5:** Means of canonical variables.

|  |  |
| --- | --- |
|  | **Canonical variable means** |
| Satisfactory HND certificatesUnsatisfactory HND certificates | 0.84-0.81 |

From Table 5 it is evident that students predicted to graduate with satisfactory certificate contribute to the canonical function to a higher degree than their counterparts who have been predicted to graduate with unsatisfactory HND certificates.

The significance of the discriminant function is presented in Table 6.

**Table 6:** Chi-square test

|  |  |
| --- | --- |
| **Measure** | **Parameter** |
| Eigen valueCanonical RWilks’ LambdaChi – squareDfp - value | 1.2550.610.444191.07760.0000 |

The Wilks’ Lambda value of 0.444 indicates the existence of a difference between the groups, that is, the effect of the variables of the model upon the discrimination between groups. An Eigen value of 1.255 shows the significance of dimensions in the independent variables in the classification. It can also be seen that there is a canonical value of 0.53 on the basis of which it can be concluded that there is a correlation between the discriminant function and the two groups, that those with satisfactory certificate and those with third class certificate, unsatisfactory.

It is important to indicate the relative importance of the predictors in any discriminant analysis. Table 7 provides an index of the importance of each predictor just like we have in standardized regression coefficients (betas). Finally, we create the discriminant function (equation) by presenting the canonical discriminant function coefficient in Table 7.

**Table 7:** Canonical discriminant function coefficients

|  |
| --- |
| Canonical Discriminant Function Coefficients |
|  | **Function** |
| **1** |
| GPASTA 101STA 107STA 217AgeCert(Constant) | 1.3550.4100.0130.095- 0.012-0.035-3.796 |

From Table 7, it can be seen that the overall grade point has much weight in the overall function. The age of students and type of certificate used for admission have little bearing on the model. It is also evident that Element of mathematics has a higher weight than the other courses. The model is presented as:



**Conclusion**

The consistent high hit rates for both the analysis sample and hold-out sample, i.e., the overall percentage of correct classifications which is 87.5 and 88.2%, as seen in the confusion matrices (Tables 2 and 3), for this study, which is a measure of predictive ability shows that discriminant analysis can be used to predict students’ graduating class of degree from knowledge of variable(s) that have relationship with performance. The use of discriminant analysis in this manner that is, conducting discriminant analysis for predictive purpose enables us to identify the students who might be termed at risk; these are students that will graduate with unsatisfactory HND certificate, that is third class. The study also identifies Element of Mathematics (STA 101), Probability Theory (STA 107) and Statistical Computing (STA 207) as having a booster effect on final graduating Cumulative Grade Point Average (CGPA). However, the weight of Element of Mathematics is higher than the other two courses in the model. The difficulties in understanding the concept of these identified courses were therefore brought to the fore. Therefore there is need for the department to paid special attention to the teaching and learning of these courses. That is to say there is the need for an instructional intervention in these courses. In conclusion, this study shows that discriminant analysis provides results that are both more interpretable and statistically sound, in addition to being a statistically correct procedure for prediction purpose than traditional measures.

**Reference**

Ardila A. Predictors of University Academic Performance in Colombia. International Journal of educational research. 2001; 35: 411 – 417.

Garson G. D. Discriminant Function Analysis, Statnotes: Topics in Multivariate Analysis. 2008

Koforidua Polytechnic Student Hand Book. (2014)

Paliwal M. Kumar U. A. A study of academic performance of business school graduates using neural network and statistical techniques. Expert Systems with Application. (2008).

Pritchard M. E. Wilson G. S. Using emotional and social factors to predict students’ success. Journal of College Student Development. (2003) 44(1): 18 – 28.