# Prediction of secondary school student's educational performance

**Elham Harirpoush**
eharirpo@uwo.ca
University of Western Ontario, London, ON, Canada

## Abstract

Achieving success in education could be correlated to land better job opportunities as well as higher earnings and welfare. Meanwhile, many factors can cause a better performance of students in their study.

Several researches have been competed on effective parameters on student's performance at school and how it can be correlated to their future life.

In this project a modelling method is proposed to predict student's educational performance based on a number of selected items. Genetic algorithm is used for feature selection. Neural network analysis and support vector machine techniques are applied for the task of prediction.

The most outstanding consequence of completing this study would be to plan for improvement of an educational system resulted from machine learning models.

## 1. Introduction

Children's performance in school and impacting factors have always been considered an important sociological issue since it is believed to be a fundamental item in development of societies. Children's success in school determines their success as adults. Giving equal chances to students to excel in school and life is followed up by advanced educational systems in different countries because the governments will benefit from maximized achievements of individuals as well as themselves.

Higher earnings and consequently benefitting from better life quality and welfare is the most important private benefit of a higher level education.

On the other hand, public benefits seems to be more outstanding and consequently important for the governments (Sandy Baum and Kathleen Payea). Some of the public benefits could be summarized as:

- Reduced rate of unemployment and poverty
- Lower demand for social support programs
- Higher contribution in government's tax revenues
- Increased health level and reduced medication and insurance costs
- Increased civil participation such as volunteer activities, voting and etc.

The principal aspects such as parent's cohabitation, highest education level and job, family size and quality of relationship between children and their parents have been investigated to be effective on children's future achievements. There exist a great number of databases in the field based on information usually collected via simple local questioners through educational organizations.

For instance, numerous studies demonstrate that children who grow up in single-parent families are less likely to attend in higher education levels than those who grow up with both parents. Financial status of a single parent family, less encouragement and less help with school works are recognized to be

main reasons. Other achievement indicators, including grades, attendance, attitudes, expectations and school retention are diagnosed to be affected basically (Nan Marie Astone and Sara S. McLanahan 1991).

In this study the grade of students in Mathematic and Portuguese language courses are collected from public schools in Portugal during 2005-2006.

The purpose of this study is to identify the effective values and comparing the methods which can predict student's educational achievements more effectively. Due to the supervised learning nature of the problem, grades are modeled based on classification and regression methods.

Support Vector Machine (SVM) method is applied for classification and Deep Neural Network (DNN) is used for regression model.

## 2. Data Description

A dataset of Portuguese schools is used in this study, which is composed of 33 different features per each student for two courses. The features could be roughly categorized in three main parts as student's personal information on both educational and social attitudes, their family statues and parent's characteristics and finally their grades on subjects (P. Cortez and A. Silva 2008)

The data were collected during 2005-2006 school year from two public schools in Portugal. There are two sets of data for two different subjects with 395 examples related to Mathematics and 649 examples related to the Portuguese language. Table 1 describes the details of data.

### 2.1. Preprocessing phase

In these dataset, 18 out of 33 feature columns consist of string values. Encoding technique is applied to convert them to numerical format based on number of appearance in each column.

Then, in order to speed up analysis, data are scaled in roughly a same range. Normalization method is applied using the following equation (URL 2 and 1):

$$xi := \frac{xi - \mu i}{si}$$

Where $\mu i$ is the average of all values for feature (i) and $si$ is the standard deviation.

Students achieved marks in two major courses (Mathematics and Portuguese) are modeled based on two decision making goals:

a) Pass/Fail binary classification
b) Regression (numeric output between 0 and 20).

## 3. Background: Learning Algorithms

### 3.1. Feature Selection

Feature selection helps us to determine which features best predict the student's grade, so the most irrelevant features is removed to increase the accuracy and speed of prediction and decrease the risk of overfitting. In this paper, the genetic feature selection algorithm is applied to find the relevant attributes.

Genetic algorithms are sort of domain independent search techniques in which, selection of appropriate representation and an evaluation function is the key part. (Fernando Gomez and Alberto Quesada)

In a genetic algorithm, a population (sub-set of features) is modified repeatedly. In a better word, a random combination of features (parent) will generate sub-sets of features (children).

For this purpose, sets of features (parents) exchange their subsets from a random point which is called crossover.

After each crossover stage, selection of next generation is based on higher accuracy on model.

Following up with crossover, features could be replaced by other features or re-ordered

in newly generated child which is called mutation. (Avrim L. Blum and Pat Langley)

### 3.2. Deep Neural Network

Neural Networks can help to get better results for nonlinear decision making. In a neural network, "neuron" is a computational unit that computes the weight of inputs plus bias, based on different functions (activation functions) such as Sigmoid, Tanh, Softmax and etc. Weighted neurons are called a hidden layer (shown in figure 1). Results of activated neurons are forwarded to a final output.

In a deep learning approach, several weighting stages are applied before resulting in final output, which creates several hidden layers (shown in figure 1).

In this study, multilayer perception is used with 10 hidden layers and 15 neurons.

Optimum values of weights are selected by using backpropagation. Momentum, adam and adadelta methods are applied and due to lower errors, momentum method is finally selected. Comparison of results taken from above three methods, is shown in figure 2.

To balance the fitness of model and magnitude of weights, L2 regularization is implemented. (Sebastian Raschka and URL 2)

Due to the large number of neurons in Deep Neural Network, overfitting is the most probable issue because of co-adapting between neurons and model. To address the problem, neurons are randomly dropped from the network during training. This technique is called dropout. (Nitish Srivastava et al.)

Regarding the accuracy of the model, K-Fold cross validation is applied with 5 folds. For measuring loss function, root mean square function is applied. (URL 2)

According to the small size of current data sets, validation is ignored and 20% of data are considered as test data which randomly alters in the range of data set.

### 3.3. Support Vector Machine (SVM)

Support vector machine is another powerful way of learning complex non-linear functions. SVM is a modified version of logistic regression. Unlike logistic regression, SVM does not output the probability. Instead, there is a cost function which minimizes the weights of data and predicts the outputs of zero or one. In this method, a factor “C” is calculated to minimize the cost function. In current project, SVM method is applied for binary classification of grades to figure out whether the grades will pass or fail. (URL 1)

A kernel is a function which finds similarity between two inputs. A couple of functions such as linear, rbf, poly and sigmoid are typically used. In this study among kernel functions, polynomial kernel resulted in smallest error. (Sebastian Raschka)

Accuracy of the model is evaluated by confusion matrix method which is shown in figure 3 and 4.

## 4. Analysis of results

### 4.1. Genetic Algorithm Results

Number of 28 features out of 30 features are selected as most relevant features, by applying genetic algorithm.

### 4.2. Deep Neural Network Analysis Results

Tflearn library on top of Tensorflow is used to implement deep neural network model. The final error of the model decreases in 1800 training steps to the value of rounded 7.2 (based on RMSE). The reduction of errors in consecutive steps, is shown in figure 5.

Based on the results taken from deep neural network, all grades are predicted to be equal 11.593413. Switching from Tflearn to Sklearn library with “tanh” activation function and adaptive as a learning rate results in an

error of rounded 6.0 (based on RMSE) with more reasonable predictions. For instance, some of the predictions (grade G2) are as below:

True value = 11 / Predicted value = 12.9
True value = 5 / Predicted value = 7.3
True value = 15 / Predicted value = 12.3
True value = 12 / Predicted value = 13.5

### 4.3. SVM Analysis Results

Initial C value is chosen close to zero and incremental value of 0.01 is added to the initial value. Accuracy of the model is evaluated at each increment considering the fact that, steps could go backwards in case of decreasing accuracy. The trend of C selection is shown in figure 6. Final C magnitude is 0.86 in this case.

Accuracy of the model for Portuguese dataset is 80 percent, and there are 10 miss-classified cases out of 50 classes which results in miss-classification of 20 percent. For Mathematics dataset, there are 24 out of 50 miss-classification which results in 48 percent miss-classification and 50 percent accuracy.

## 5. Visualization

With visualization techniques, large amount of data could be explored and the relationship between variables could be found. Scatter plots and bar charts applied to display the relationship between input and output variable. For example, each feature impact on Portuguese course grades (G2) is demonstrated in figure 7. This graph, clearly indicates the highest effect of parent's jobs on their children's grades. As another instance, in bar charts of figure 8 and 9, the impact of father's and mother's sort of job is illustrated.

Visualization of the DNN is shown in figure 10 which is created by Tensorboard.

Principal Component Analysis (PCA) method is applied to reduce the features from more than thirty features to two in order to visualize the miss-classification of SVM method. In figure 11 and 12, miss-classification of math and Portuguese grades is shown. In this figure, blue color stands for pass and black color is representative of failed grades. Large circle is the predicted values and small ones are true values. If small and large circles are in same colors, it verifies the classification otherwise, it is indicative of miss-classification.

## 6. Discussion and Conclusion

Individual's quality of life is associated with their highest educational level. Effectiveness of some social, demographic and school related parameters on educational performance of Portuguese students is used in this study, to predict their grades.

Two dataset of Mathematics and Portuguese language courses are encoded and normalized for modelling purpose. The most relevant features are selected using genetic algorithm in each modelling method.

Finally, two analysis methods are used to evaluate the features correlation with student's grades. Deep neural network (DNN) and support vector machine (SVM) are applied.

Despite of no overfitting when applying DNN method with Tensorflow library, this method does not yield proper results. Switching the library to Sklearn with tanh activation function, resulted in better performance and predicted the grades with more accuracy. It predicts the grades with around 0.12 per unit grade error.

SVM method is also applied to both datasets. Due to smaller range of data included in Mathematics dataset, this method was not accurate enough to predict Mathematics grades properly. However, the results taken form this method regarding the other dataset (Portuguese) sounds reasonable. A comprehensive visualized output is demonstrated in appendices.

## References:


Ashish Kumar. Learning Predictive Analytics with Python. Packt publishing. Birmingham.

Sebastian Raschka. Python Machine Learning. Packt publishing. Birmingham.

Nitish Srivastava, Geoffray Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural network from overfitting. University of Toronto.

Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. Elsevier. Artificial intelligence 97 (1997) 245-271.

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Nan Marie Astone and Sara S. McLanahan. Family Structure, Parental Practices and High School Completion. American Sociological Review, Vol. 56, No. 3 (Jun., 1991), pp. 309-320.

Douglas B. Downey. Family Size, Parental Resources, and Children's Educational Performance. American Sociological Review, Vol. 60, No. 5 (Oct., 1995), pp. 746-761.

Henderson, Anne T., Ed.; Berla, Nancy, Ed. New Generation of Evidence: The Family is Critical to Student Achievement. National Committee for Citizens in Education, Washington, DC.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. New York: Springer.

The Elements of Statistical Learning by Hastie, Tibshirani and Friedman. Expanded version of required text.

Fernando Gomez and Alberto Quesada. Genetic algorithms for feature selection in data analytics.

Trevor Hastie, Robert Tibshirani and Jerome Friedman. The elements of statistical learning. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Sandy Baum and Kathleen Payea. The benefits of Higher Education for Individuals and Society. Revised edition, 2005. WWW.COLLEGEBOARD.COM.

URL1: WWW.COURSERA.COM: Machine Learning courses

URL 2: cs231n.stanford.edu/

## Appendix A.

Data description

Table 1: Detailed description of dataset

| No. | Attribute Description (Domain) |
|---|---|
| 1 | **sex** student's sex (binary: female or male) |
| 2 | **age** student's age (numeric: from 15 to 22) |
| 3 | **school** student's school (binary: Gabriel Pereira or Mousinho da Silveira) |
| 4 | **address** student's home address type (binary: urban or rural) |
| 5 | **Pstatus** parent's cohabitation status (binary: living together or apart) |
| 6 | **Medu** mother's education (numeric: from 0 to 4a) |
| 7 | **Mjob** mother's job (nominalb close to home, school reputation, course preference or other) |
| 8 | **Fedu** father's education (numeric: from 0 to 4a) |
| 9 | **Fjob** father's job (nominalb close to home, school reputation, course preference or other) |
| 10 | **guardian** student's guardian (nominal: mother, father or other) |
| 11 | **famsize** family size (binary: <=3 or >3) |
| 12 | **famrel** quality of family relationships (numeric: from 1- very bad to 5- excellent |
| 13 | **reason** reason to choose this school (nominal: close to home, school reputation, course preference or other) |
| 14 | **traveltime** home to school travel time (numeric: 1-< 15 min., 2-15 to 30 min., 3-30 min. to 1 hour or 4 -> 1 hour). |
| 15 | **studytime** weekly study time (numeric: 1-< 2 hours, 2- 2 to 5 hours, 3-5 to 10 hours or 4 -> 10 hours) |
| 16 | **failures** number of past class failures (numeric: n if 1 n < 3,else 4) |
| 17 | **schoolsup** extra educational school support (binary: yes or no) |
| 18 | **famsup** family educational support (binary: yes or no) |
| 19 | **activities** extra-curricular activities (binary: yes or no) |
| 20 | **paidclass** extra paid classes (binary: yes or no) |
| 21 | **internet** Internet access at home (binary: yes or no) |
| 22 | **nursery** attended nursery school (binary: yes or no) |
| 23 | **higher** wants to take higher education (binary: yes or no) |
| 24 | **romantic** with a romantic relationship (binary: yes or no) |
| 25 | **freetime** free time after school (numeric: from 1- very low to 5- very high) |
| 26 | **goout** going out with friends (numeric:<br>rom 1- very low to 5- very high) |
| 27 | **health** current health status (numeric: from 1- very bad to 5- very good) |
| 28 | **absences** number of school absences (numeric: from 0 to 93) |
| 29 | **Wal** weekend alcohol consumption (numeric: from 1- very low to 5- very high |
| 30 | **Dal** daily alcohol consumption (numeric: from 1- very low to 5- very high |
| 31 | **G1** first period grade (numeric: from 0 to 20) |
| 32 | **G2** second period grade (numeric: from 0 to 20) |
| 33 | **G3** final grade (numeric: from 0 to 20) |

## Appendix B.

Figures

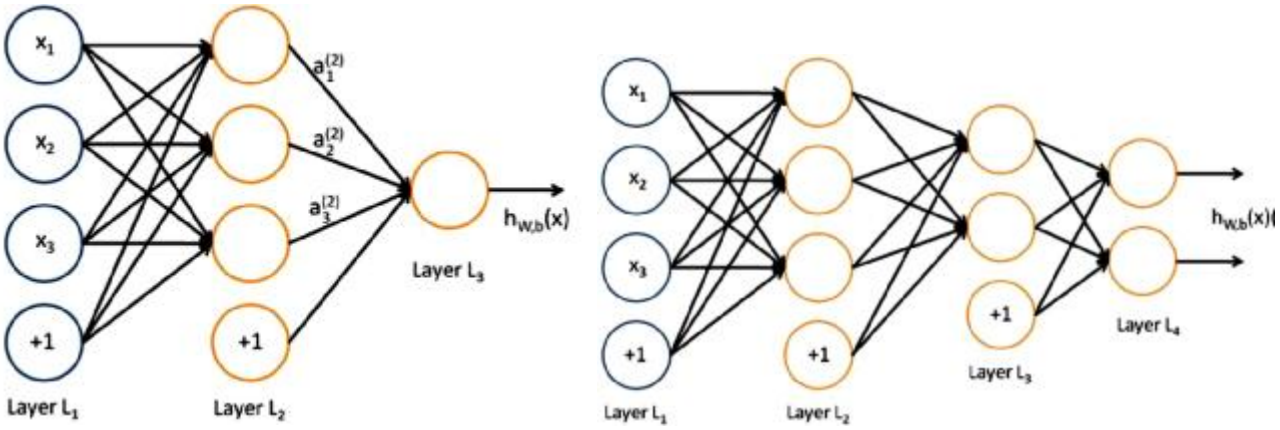



*Figure 1) Neural network model with one hidden layer (left) and multiple hidden layers (right)*

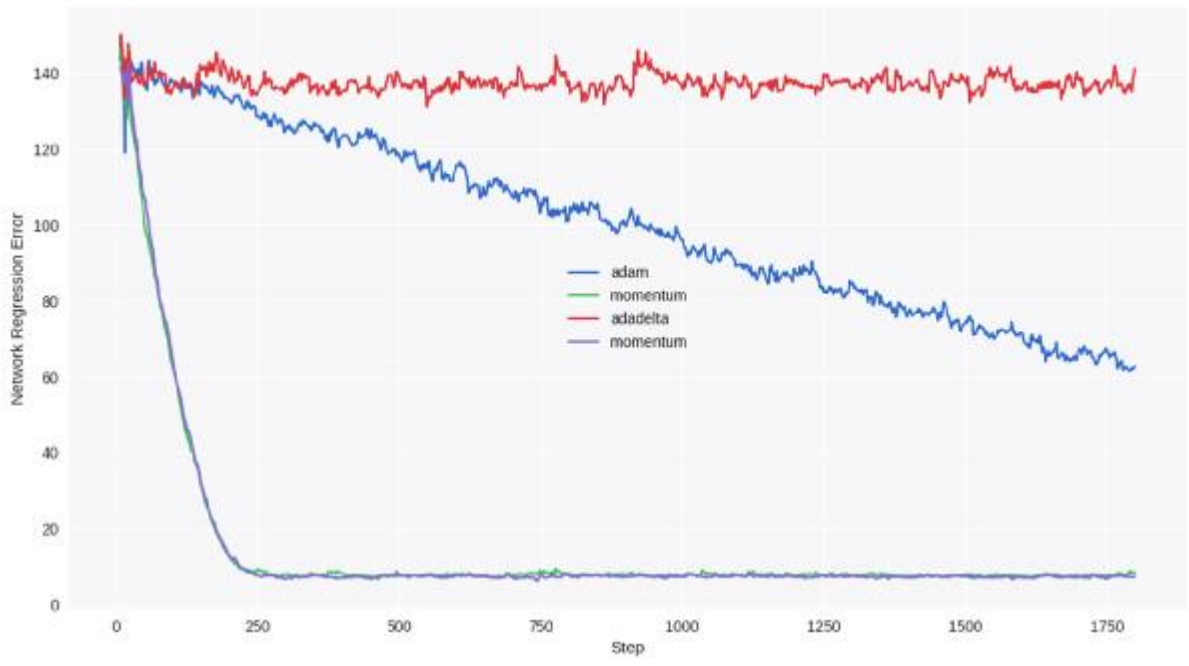



*Figure 2) Trend of different optimization algorithms*

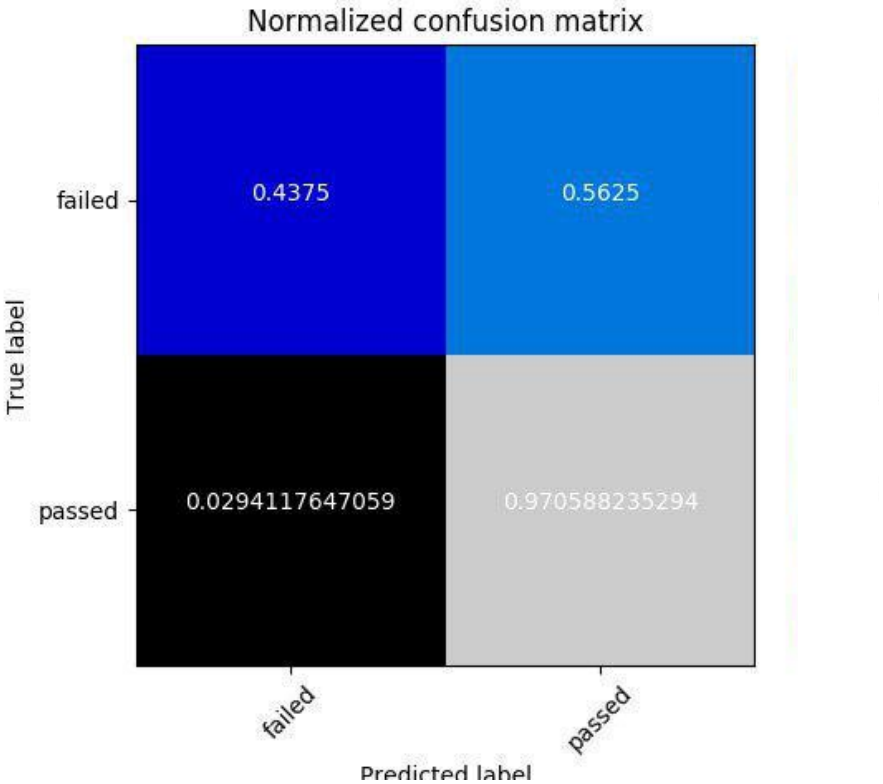



*Figure 3) Normalized confusion matrix (Portuguese dataset)*

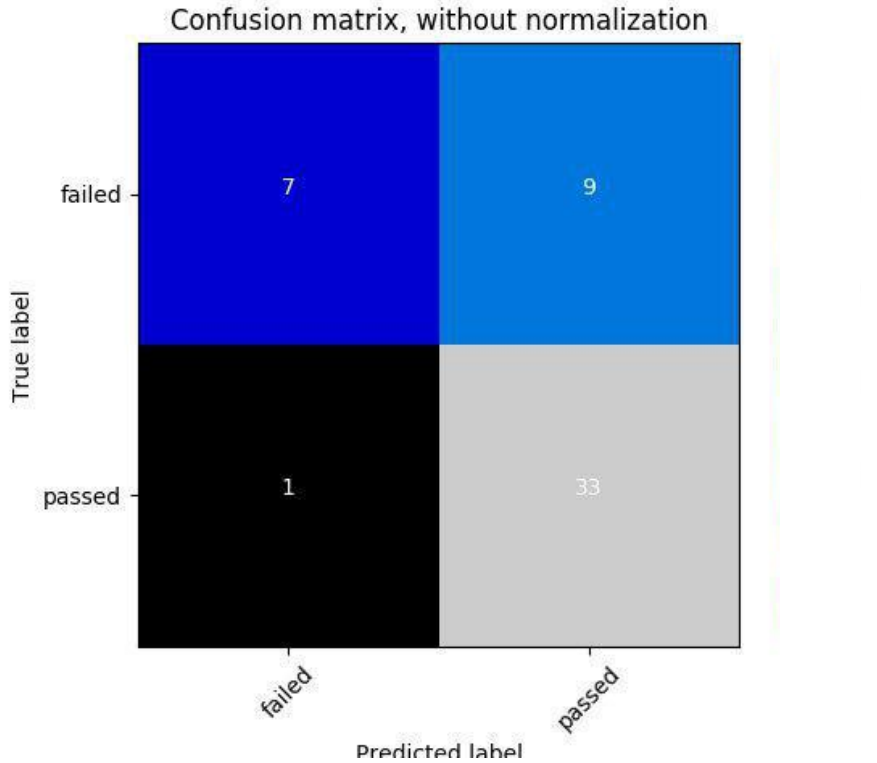



*Figure 4) Confusion matrix without normalization (Portuguese dataset)*

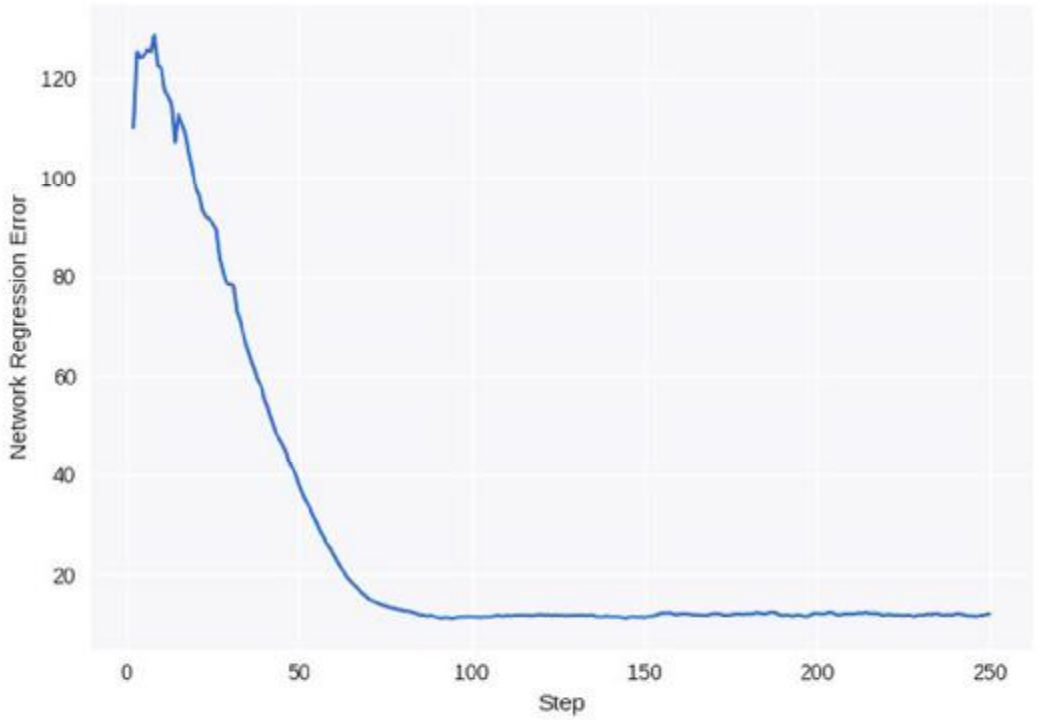



*Figure 5) Trend of error reduction with increasing steps in DNN method (Tflearn library)*

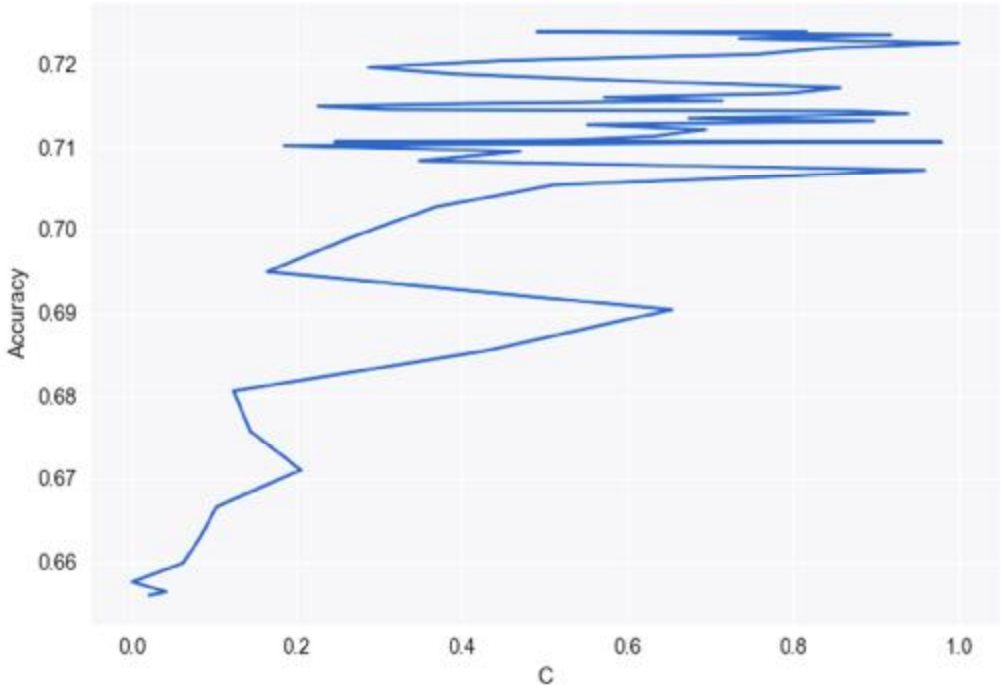



*Figure 6) Changing trend of accuracy based on "C" variation in SVM model*

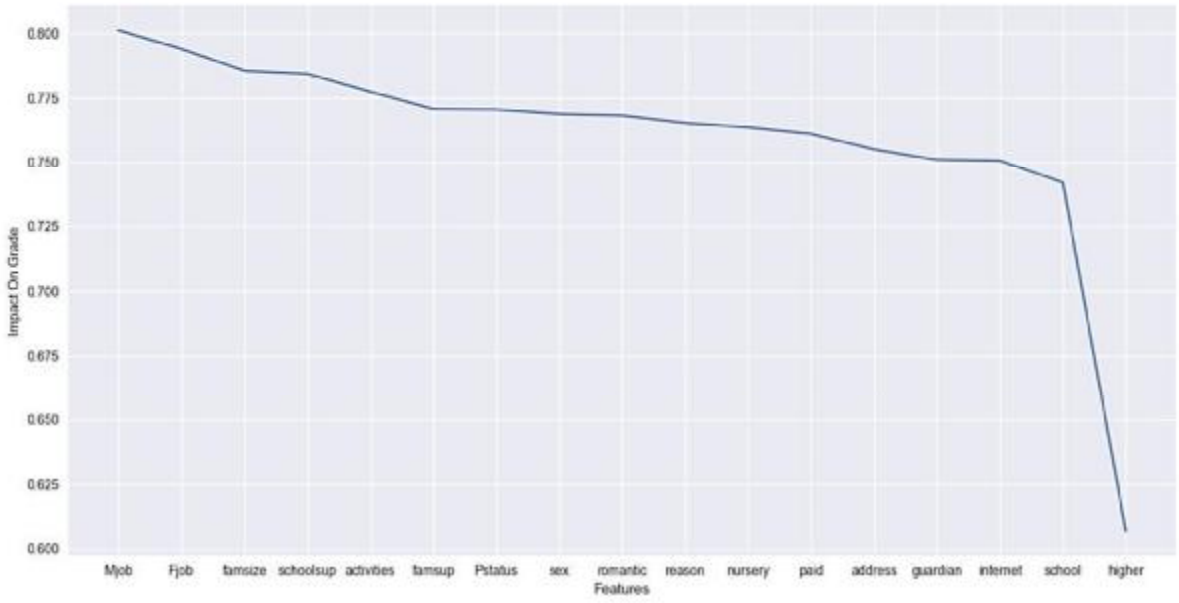



*Figure 7) Features impact on students grades in Portuguese course*

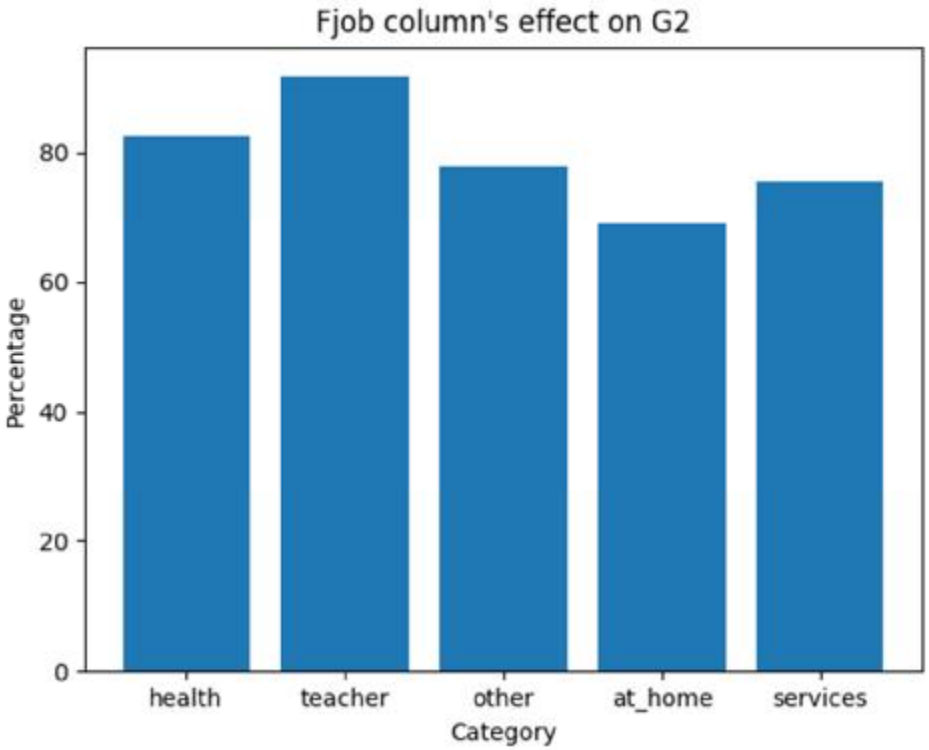



*Figure 8) The effect of father job on grades*

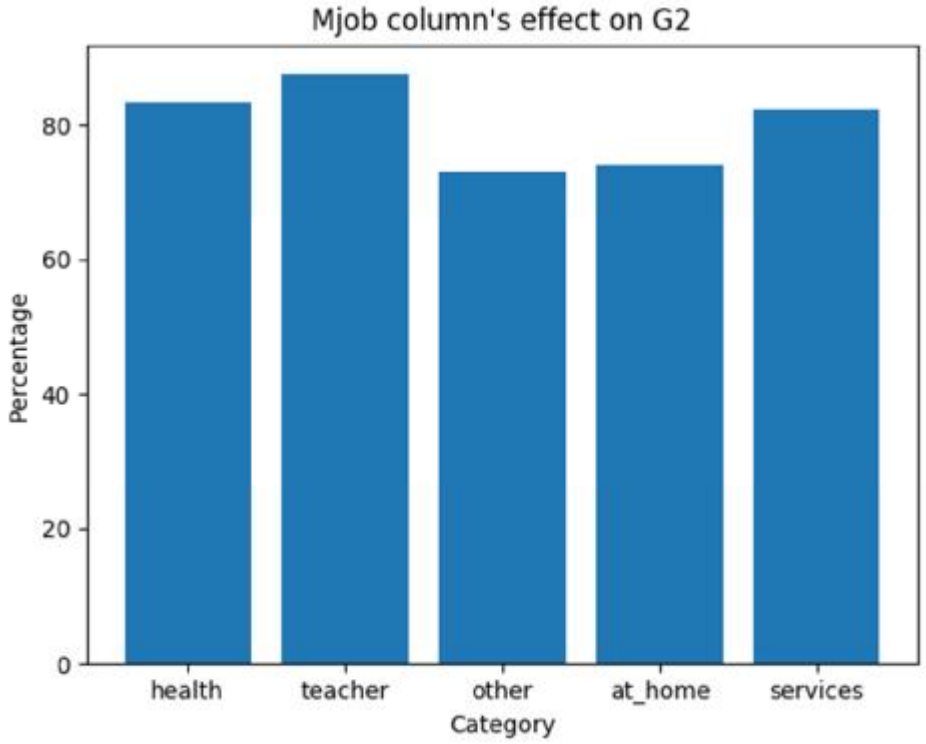
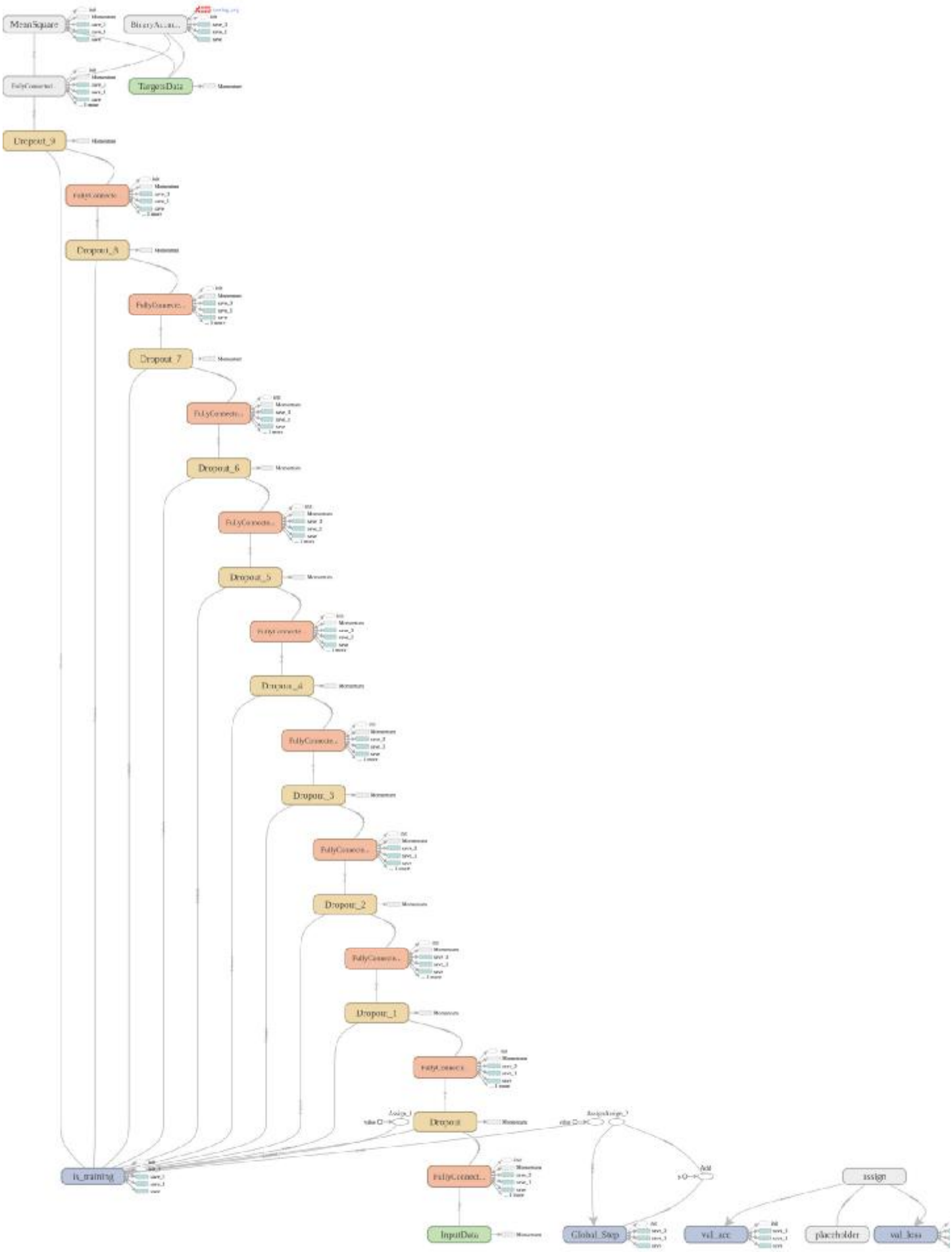



*Figure 9) The effect of mother job on grades*

*Figure 10) Topology of DNN with Tensorboard*



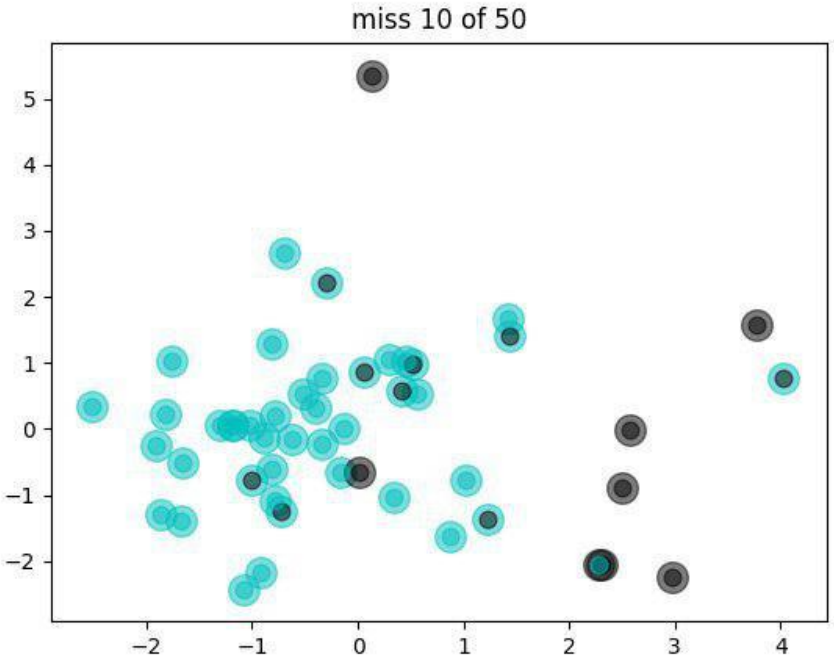



*Figure 11) Miss-classification scatter plot for Portuguese dataset (PCA)*

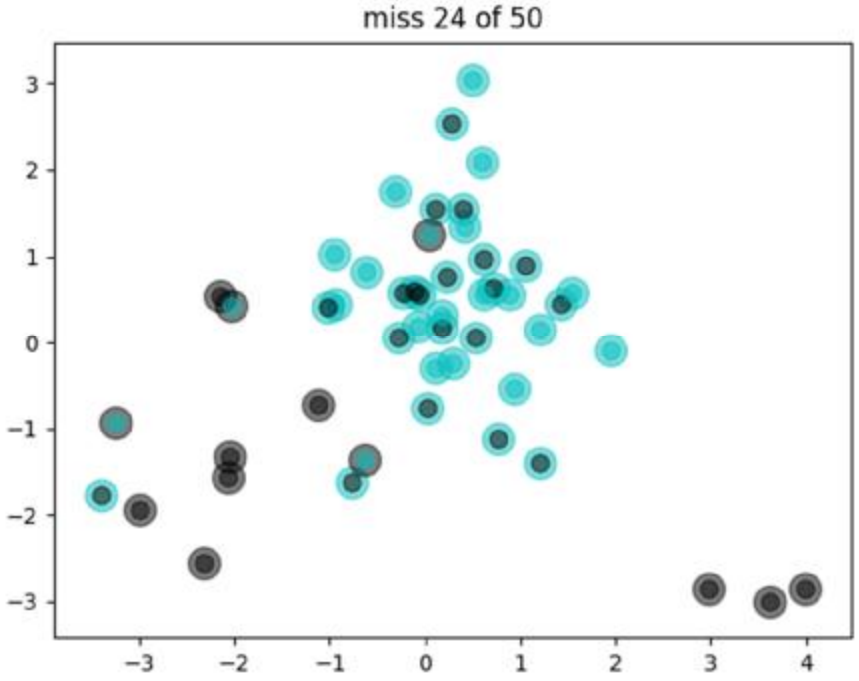



*Figure 12) Miss-classification scatter plot for Mathematics dataset (PCA)*