# Predicting Students' Exam Score using ANN

TEAM SAPHIRE
Samvid Pundir(230150021)
Bhavika Pandya(230150006)

## Contents

# 1 Introduction

This project involves the analysis of student performance data from two datasets: `student-mat.csv` and `student-por.csv`. The datasets contain information about students' demographic, social, and academic attributes. Our goal is to build a deep learning model to predict final grades and perform detailed exploratory data analysis (EDA) to extract insights.

# 2 Data Description

The dataset used in this project consists of two files: `student-mat.csv` (Math course) and `student-por.csv` (Portuguese language course). Each file contains detailed attributes about student performance and background information. Below is a description of each attribute:

- **school** - student's school (binary: `GP` - Gabriel Pereira or `MS` - Mousinho da Silveira)
- **sex** - student's sex (binary: `F` - female or `M` - male)
- **age** - student's age (numeric: from 15 to 22)
- **address** - student's home address type (binary: `U` - urban or `R` - rural)
- **famsize** - family size (binary: `LE3` - less or equal to 3 or `GT3` - greater than 3)
- **Pstatus** - parent's cohabitation status (binary: `T` - living together or `A` - apart)
- **Medu** - mother's education (numeric: 0 - none to 4 - higher education)
- **Fedu** - father's education (numeric: 0 - none to 4 - higher education)
- **Mjob** - mother's job (nominal: `teacher`, `health`, `services`, `at_home`, `other`)
- **Fjob** - father's job (nominal: same as Mjob)
- **reason** - reason to choose this school (nominal: `home`, `reputation`, `course`, `other`)
- **guardian** - student's guardian (nominal: `mother`, `father`, `other`)
- **traveltime** - home to school travel time (1 - ¡15 min. to 4 - ¿1 hour)
- **studytime** - weekly study time (1 - ¡2 hrs to 4 - ¿10 hrs)
- **failures** - number of past class failures (0 to 4)
- **schoolsup** - extra educational support (binary: yes or no)
- **famsup** - family educational support (binary: yes or no)
- **paid** - extra paid classes within the course (binary: yes or no)
- **activities** - extra-curricular activities (binary: yes or no)

- **nursery** - attended nursery school (binary: yes or no)
- **higher** - wants to take higher education (binary: yes or no)
- **internet** - Internet access at home (binary: yes or no)
- **romantic** - with a romantic relationship (binary: yes or no)
- **famrel** - quality of family relationships (1 - very bad to 5 - excellent)
- **freetime** - free time after school (1 - very low to 5 - very high)
- **goout** - going out with friends (1 - very low to 5 - very high)
- **Dalc** - workday alcohol consumption (1 - very low to 5 - very high)
- **Walc** - weekend alcohol consumption (1 - very low to 5 - very high)
- **health** - current health status (1 - very bad to 5 - very good)
- **absences** - number of school absences (numeric: 0 to 93)
- **G1, G2, G3** - first, second, and final period grades (numeric: 0 to 20)

**Note:** There are 382 students that appear in both datasets. These duplicate entries can be identified by matching all non-grade-related attributes as specified in the accompanying R file.

# 3 Libraries Used

- `pandas`, `numpy` for data handling
- `matplotlib`, `seaborn` for visualization
- `scikit-learn` for preprocessing and evaluation metrics
- `tensorflow.keras` for deep learning

# 4 Data Visualizations

The following visualizations were generated to explore the relationship between various student attributes and their final grades:

- **Urban & Rural vs Student Count** – A bar plot comparing the number of students from urban and rural areas.
- **Grade Distribution vs Gender** – A distribution plot showing how grades vary between male and female students.

- **Absences vs Final Grade** – A scatter plot to analyze how the number of absences impacts final grades.
- **Study Time vs Final Grade** – A boxplot or scatter plot to observe how study time correlates with final academic performance.

# 5 Data Preprocessing

- Merged the two datasets (Math and Portuguese) to larger dataset, thereby increasing the training data volume for improved model learning.
- Dropped Redundant Columns like school, activities, internet availability etc.
- Normalized features using `MinMaxScaler`
- Removed records with zero or missing final grades, as they were considered outliers and could negatively impact model training
- The following table shows the mapping used to convert categorical values into integers during preprocessing:

| **Category** | **Encoded Value** | **Category** | **Encoded Value** |
|---|---|---|---|
| M (sex) | 1 | F (sex) | 0 |
| U (address) | 1 | R (address) | 0 |
| GT3 (famsize) | 1 | LE3 (famsize) | 0 |
| A (Pstatus) | 1 | T (Pstatus) | 0 |
| father (guardian) | 2 | mother (guardian) | 1 |
| other (guardian) | 0 | course (reason) | 0 |
| home (reason) | 1 | reputation (reason) | 2 |
| `at_home` (job) | 0 | health (job) | 1 |
| services (job) | 2 | teacher (job) | 4 |
| yes | 1 | no | 0 |

Table 1: Integer Encoding of Categorical Values

# 6 Model Architecture

- Dense neural network with dropout and LeakyReLU activations
- Batch normalization for regularization
- Two models compared:
  1. Model with linear output layer
  2. Model with softmax output layer (multi-class classification)

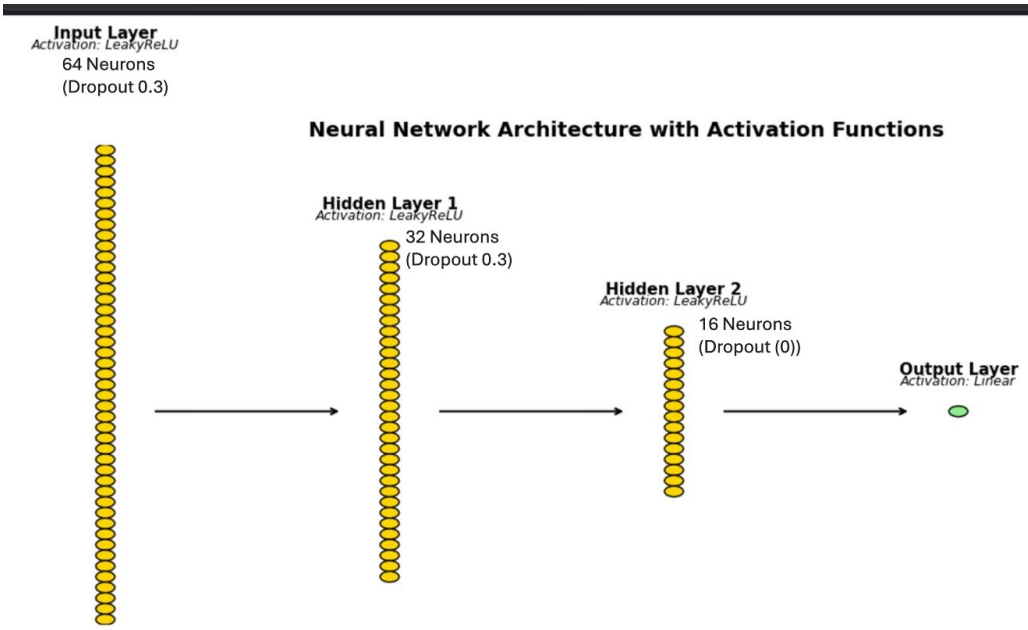


Figure 1: ANN with Linear Output Layer

# 7 Evaluation Metrics

- **Accuracy** – The proportion of total correct predictions out of all predictions made.
- **Precision** – The ratio of correctly predicted positives to all predicted positives (how precise your positive predictions are).
- **Recall** – The ratio of correctly predicted positives to all actual positives (how well you found all the real positives).
- **F1-Score (Macro Average)** – The harmonic mean of precision and recall averaged equally across all classes, regardless of class imbalance.
- **Mean Squared Error (MSE)** – The average of the squares of the errors between predicted and actual values, penalizing larger errors more.
- **Mean Absolute Error (MAE)** – The average of absolute differences between predicted and actual values, treating all errors equally.
- **R-squared ($\mathbf{R^2}$)** – Represents the proportion of variance in the target variable explained by the model.
- **Adjusted R-squared** – $R^2$ adjusted for the number of predictors, helping prevent overfitting.
- **Confusion Matrix** – A table showing true vs. predicted classifications to analyze performance across all classes.

# 8 Results and Analysis

## Model 1: Linear Output

Achieved baseline regression performance. Metrics:

| Metric | Value |
|---|---|
| Accuracy | 0.4020 |
| Error | 0.5980 |
| Precision | 0.2730 |
| Recall | 0.2650 |
| F1 Score | 0.2585 |
| Mean Squared Error (MSE) | 1.1055 |
| Mean Absolute Error (MAE) | 0.7538 |
| R-squared ($R^2$) | 0.8678 |
| Adjusted R-squared | 0.8470 |

Table 2: Model 2 Evaluation Metrics

## Model 2: Softmax Output

Classified students into 21 grade classes. Metrics:

| Metric | Value |
|---|---|
| Accuracy | 0.3618 |
| Error | 0.6382 |
| Precision | 0.2934 |
| Recall | 0.2622 |
| F1 Score | 0.2614 |
| Mean Squared Error (MSE) | 1.4121 |
| Mean Absolute Error (MAE) | 0.8693 |
| R-squared ($R^2$) | 0.8312 |
| Adjusted R-squared | 0.8045 |

Table 3: SoftMax Model Evaluation Metrics

## Comparison

Model 1 employed a linear output layer and approached the problem as a regression task. This model is more suitable when the target variable is continuous, allowing it to predict exact grade values. It generally offers smoother learning and better generalization for problems involving ordered outputs like student grades.

In contrast, Model 2 utilized a softmax output layer and treated the problem as a multi-class classification task by discretizing grade outcomes into distinct classes. While this method is effective in scenarios where precise class boundaries are important, it can struggle

with class imbalance and might require more preprocessing such as data balancing. Classification models like Model 2 may provide sharper and more confident predictions but can be sensitive to how labels are defined. Therefore, Model 2 may be more appropriate when the goal is to categorize students into performance levels or grade brackets.

# 9 Conclusion

The deep learning approach successfully modeled student academic outcomes. Social and academic factors like study time, family support, and alcohol consumption significantly impacted final grades.

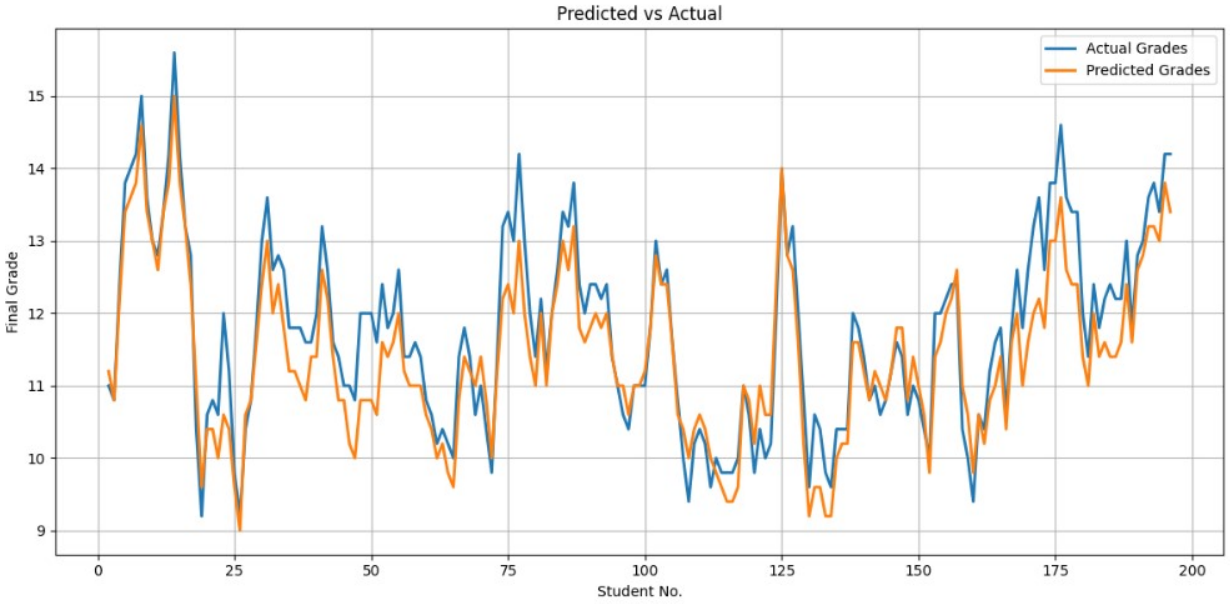


Figure 2: FINAL RESULT MODEL VS ACTUAL

# 10 References


- Cortez, P., & Silva, A. (2008). *Using Data Mining to Predict Secondary School Student Performance.* UCI Machine Learning Repository - Student Performance Data Set.
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media.
- Chollet, F. (2017). *Deep Learning with Python.* Manning Publications.
- Scikit-learn Documentation: `https://scikit-learn.org/`
- TensorFlow Documentation: `https://www.tensorflow.org/`