# Letter of Transmittal

Dear Dr. Tony Martinez:

I am so grateful for the opportunity I have had to be in your machine-learning class at BYU. When I decided to take your course, I did not know anything about machine learning. It was a black box to me. I knew that computers could make educated predictions based on data. However, I never understood how. Your class was monumental in helping me understand how computers find patterns in data to make predictions. It also helped me understand where I want to go with my career. Thank you so much for making me love machine learning.

As you are aware, my group in your class is doing a project to predict students’ ACT scores. To get ahead of the game, I decided to perform some of my own research. I chose to research which machine-learning model is most effective at predicting student test scores. Using peer-reviewed papers, I gathered the opinions of various experts. I also performed some of my own empirical research using what I learned in your class. Ultimately, I have determined that the decision tree is the most optimal model for predicting student test scores. However, more research is needed to confirm my conclusion.

Thank you for everything you have taught me in CS201R. I have enjoyed the class and loved performing some of my own research. I encourage you to read this report to help me determine how valid my research has been. Also, I would love to know where I could turn to next to continue doing research in machine learning.

Sincerely,

Daniel Perkins

# Analytical Report:

**What is the most effective machine learning model for predicting student test scores?**

Daniel Perkins Brigham Young University

Writing 316: Technical Communication

### Executive Summary

This recommendation report analyzes which machine-learning model is most effective at predicting student test scores. Machine learning is a field in data science that uses mathematical models to find patterns in data. It can be used to make predictions about practically anything. Specifically, a machine-learning model can help researchers determine which factors lead to higher test scores. This report claims that a decision tree is the most effective model because it is both the most accurate and the quickest.

This technical recommendation is intended for researchers interested in learning more about how to improve student test scores. There are many ways to predict student test scores. This report will help researchers determine the most effective and accurate way to do so. Furthermore, it is intended for people interested in learning more about the fundamental principles of machine learning. It explains how two of the most common machine-learning algorithms work. Also, it discusses how to compare the results of various models.

This report first discusses how neural networks and decision trees work. Neural networks are models that resemble the human brain to make predictions. Decision trees act like a game of 20-questions to make decisions based on conditional statements. Secondly, this report discusses expert opinions about the effectiveness of each model. Both models have their advantages and disadvantages. Finally, it analyzes empirical results gathered from running the machine-learning models on a dataset about student math test scores. Based on my research I recommend using a decision tree to predict student test scores.

## Table of Contents

[Letter of Transmittal 1](#_bookmark0)

[Executive Summary 3](#_bookmark1)

[List of Figures 5](#_bookmark2)

Introduction 6

[Background 7](#_bookmark3)

[Methodology 8](#_bookmark4)

[Results 9](#_bookmark5)

[Neural Networks 9](#_bookmark6)

[Decision Trees 11](#_bookmark7)

[Empirical Research 13](#_bookmark7)

[Conclusion and Recommendation 13](#_bookmark8)

[Appendix A 1](#_bookmark8)5

[References 16](#_bookmark9)

## List of Figures

[Figure 1: A Simple Neural Network 1](#_bookmark0)0

[Figure 2: A Simple Decision Tree 11](#_bookmark1)

[Figure 3: Code for a Neural Network 15](#_bookmark2)

Figure 4: Code for a Decision Tree 15

**Analytical Report: What is the most effective machine learning model for predicting student test scores?**

Machine learning is a vastly growing field of Computer Science. It uses mathematical models to find patterns in data and make predictions. Although this field of study has been around for over fifty years, it is only now making a real impact on society. With the modern implementation of speech recognition, artificial intelligence, and self-driving cars, machine learning has become a part of our everyday lives. As research continues, this new trend will continue to change the way the world works in ways that we cannot anticipate.

There are many different types of machine-learning models. Each one can be used to make accurate predictions based on data. A machine-learning model attempts to find patterns by applying its own mathematical formulas to a specific data set. Unfortunately, because there are countless problems to solve and a large variety of datasets, there is not one model that is the most effective for every problem (Marsland, 2015). This means that to maximize the accuracy of a prediction, one must compare the results of many models.

One question that many students have is how to maximize their test scores. Before the arrival of machine learning, researchers who wanted to answer this question had to rely solely on what they could understand from common data. However, researchers are now able to use computers to decipher exactly what improves a student’s score. This allows them to find patterns that were just too complicated to recognize before. Machine learning models can make predictions based on previous data about various students’ schools, gender, age, family size, and study time. However, there is not one clear answer of which model most effectively does so.

This report describes how some of the most common machine-learning models work. It compares the results that they yield to predict students’ math test scores. Based on research, decision trees are the most effective model for this problem. However, neural networks are highly effective as well. Furthermore, each method also has its drawbacks. So, the most optimal model may depend more on the needs of the user.

### Background

Machine Learning is a complicated process with many components and steps. But, to understand how well a model runs, it is important to understand the process and vocabulary of machine learning.

To predict students’ test scores, a machine-learning model would do what is called supervised learning; it runs a model on a set dataset to find patterns (Gallant 1990). The first step is to gather the data. The data should consist of features and labels. Features are the specific characteristics of each instance. Labels are the results. They are what the model attempts to predict. For example, for a model to predict test scores, a feature could be a student’s GPA and the label is their test score.

After collecting the data, the machine-learning model uses it to create a function. This is called training. The function that it creates takes the features in as input and tries to predict their corresponding labels. Then, to use the model in a real-world setting, new (novel) data that was not originally included in the initial model is plugged into the function. The goal of the model is to use what it learned to make an accurate prediction for something that we do not know the label of (Hashim et al., 2020). The accuracy of the machine-learning model is determined by how well it generalizes (predicts) the labels of novel data.

The difficult part about machine learning is determining how accurate it really is. A trained model could accurately predict 100% of the data it is given. However, this does not necessarily mean that the model generalizes well for novel data. A model that is accurate on the training set but inaccurate on new data is known as an overfit model (Brownlee, 2019). So, for a model to be effective, it must be both accurate on the training data and accurate on new (testing) data.

### Methodology

For this report, I based my research on a variety of sources. I used peer-reviewed journals, a machine-learning textbook, a news article, and a computer science blog. I first used simple Google searches to learn more about the material. Then, after narrowing down the topic, I talked to Dan Broadbent, the computer science librarian at BYU. He encouraged me to use Scopus and Google Scholar. In each database, I used keywords such as “neural networks”, “decision trees”, “test scores”, “supervised learning”, and “accuracy” to find peer-reviewed articles. Finally, I read the abstracts of each source to determine if it was relevant to my research. After selecting the sources, I carefully analyzed each one, writing down my notes in a synthesis matrix to organize the ideas from the various authors.

In addition to literary research, I also performed some empirical research. With the help of the UC Irvine Machine Learning Repository (Cortez 2014), I found a data set used to predict student performance on high school math tests. The dataset contained information about students’ gender, age, family size, parental education, study time, extracurricular activities, and more. I imported the data set into my computer and used the Python Scikit-learn module to run the machine learning algorithms on the data set. For each model, I measured the following:

* Training Mean Absolute Error (MAE): The average distance between the predictions and true labels on the data used for training.
* Training Mean Absolute Error (MAE): The average distance between the predictions and true labels on the data never seen before.
* Temporal Complexity: The time it takes for the algorithm to run.

The sources I found helped me determine the general advantages and disadvantages of various machine learning models. The research I conducted helped me determine which machine-learning model is more effective for the specific problem of predicting test scores.

### Results

Neural networks and decision trees are two of the most common machine-learning models used in modern-day data science. Both models can be used to solve just about any supervised learning problem. I will be analyzing each of these models. First, I will explain the underlying principles for how they work. Then, I will discuss the opinions of various experts about the effectiveness of each model. Finally, I will evaluate the results of my own empirical research to determine which model most effectively predicts student test scores.

### Neural Networks

A neural network is a machine-learning model inspired by the neurological structure of the human brain (Hardesty, 2017). A brain consists of thousands of neurons, each interconnected in a large network. When one neuron in the brain is fired, a signal is sent to all the nodes connected to it. And, if the signals hit a certain threshold, the next neuron fires. The brain learns by remembering which neurons to fire whenever a specific event happens.

A neural network is a simplified model of the brain. It has one or more layers of nodes, with each node connected to every node in the adjacent layers (Sungheetha, 2020). Figure 1 shows an example of a neural network with two layers of nodes and five instances of data. The figure is oversimplified since a typical neural network would include thousands of nodes. However, it provides a good basis for understanding.

Figure 1: A Simple Neural Network

A diagram of a machine learning

Description automatically generated

This is an example of a neural network. The input layer is the set of inputs taken from the training data. The lines represent connections between nodes. The hidden layer is a set of nodes that finds its values by taking a weighted sum of the inputs. The output layer is another set of nodes that takes a weighted sum of the results in the hidden layer to make a final prediction. Information from Marsland, S. (2015). Machine Learning: An Algorithmic Perspective (2nd ed.). CRC Press.

The machine learning model learns by taking a weighted sum from the input and outputting that into a node. Then, the values of the nodes in the next layer are found by calculating a weighted sum of the previous layer. This process repeats until the model reaches the final layer. After a prediction is made, it is compared to the true label of the data. If the prediction is incorrect, the weights used to calculate the weighted sums are adjusted, layer by layer, starting with the last layer (Hagan & Demuth, 1999). This process is repeated until the model no longer improves its accuracy.

The sources I read all share the belief that neural networks are accurate machine-learning models. In fact, Hashim et al. (2020) argued that a neural network can detect all possible interactions between variables. Furthermore, Hagan and Demuth (1999) even argued that a neural network is guaranteed to approximate practically any function if it has enough nodes. Like the human brain, a computer using a neural network can make accurate predictions based on inputs it receives.

Unfortunately, neural networks also have a few disadvantages. For example, Khemphila and Boonjing (2010) explained that although neural networks are accurate, it is not always easy to explain why the model makes the predictions that it does. Models like the multi-layer Perceptron are so complex that it is nearly impossible to trace back and understand the patterns that the computer finds. Furthermore, Marsland (2015) argues that if one is not careful, a neural network can overfit the data. An overly complex neural network can mimic the data used in training so much so that it can no longer make predictions for any new instance of data. Therefore, careful consideration must be done to ensure that a neural network generalizes new data correctly.

### Decision Trees

A decision tree is a machine-learning model that uses a tree-like structure to make predictions based on conditional statements. An example is provided below in Figure 2. Once again, this example is oversimplified since a typical decision tree could have hundreds of features.

Figure 2: A Simple Decision Tree

A diagram of a diagram

Description automatically generated

This is a model of a decision tree (that I created). This decision tree could be used to predict the shape of something given its shape of edges and edge count. The machine learning model first looks at the edge shape feature to determine which subtree to enter. Then, it looks at the edge count feature. The orange node is the root of the tree, where the first decision is made. The tan node is a decision node. And, the black nodes are leaf nodes, where the tree makes a final prediction.

A decision tree uses a hierarchical structure (Hashim et al., 2020) to make predictions. It resembles a game of 20 questions. When new data is input, the decision tree uses a series of questions for each feature, to determine which subtree to enter. This reduces the number of possible solutions. Finally, when the algorithm reaches the end, it returns the majority class of the remaining possible solutions as the prediction.

The hardest part of the machine-learning algorithm is creating the decision tree itself. The algorithm's goal is to make the smallest possible tree that accurately divides the data. So, it first determines which features most effectively divide the data. The model chooses features that differentiate the output data the most (Thomas and Galambos, 2004). This can be done using an entropy or impurity function (Marsland, 2015). This minimizes the number of subtrees needed to make a prediction.

The peer-reviewed articles I analyzed all agree that the decision tree is an effective machine-learning model. Hashim et al. (2020) explain that decision trees often converge much faster than other models. This makes them more computationally efficient than neural networks. However, the model also has its own drawbacks. According to Khemphila and Boonjing (2010), the complex nature of a decision tree makes it depend too much on the data. They argue that small changes to the data can vastly alter the tree, making it difficult for the tree to generalize new data instances. Therefore, users must take careful measures to avoid overfitting.

Furthermore, I found that experts disagree about the assumptions that the decision tree makes when it attempts to analyze data. Thomas and Galambos (2004) argue that decision trees can help users understand patterns between different variables. This is because its categorical structure divides the data into smaller groups that may have varying relationships with other variables. On the other hand, Hashim et al. (2020) believe that a decision tree is a “naïve” model because it assumes that all input features are independent of each other. In other words, decision trees are unable to detect second-order relationships between each variable used in the model.

### Empirical Research

Using Python (see appendix A), I conducted empirical research on the data set from UC Irvine (Cortez 2014). I found that the neural network model had a training mean accuracy error of 0.3357 and a testing mean accuracy error of 4.295. This means that the model was only off by about 4.3% in its prediction of students who were not in the original data set. On the other hand, the Decision Tree had a training MAE of 0.0 and a testing MAE of 3.48. So, its predictions were only off by about 3.5%, meaning that it was more accurate than the neural network.

Furthermore, the neural network model took about 9.17 seconds to run while the decision tree model only took 0.007 seconds. This means that for larger problems, the decision tree is much more efficient. But, for smaller data sets, both models are quick enough to be used in almost any setting.

### Conclusion and Recommendation

There are many different machine-learning models, each with its advantages and disadvantages. Neural networks resemble the human brain and use weighted sums to make predictions. If they have enough nodes, they can approximate almost any function. But, they often overfit the data, making the model match only the input it was trained on. Decision trees, on the other hand, act like a game of 20 questions to make predictions based on conditional statements. They are very quick models but only work well if the features are independent.

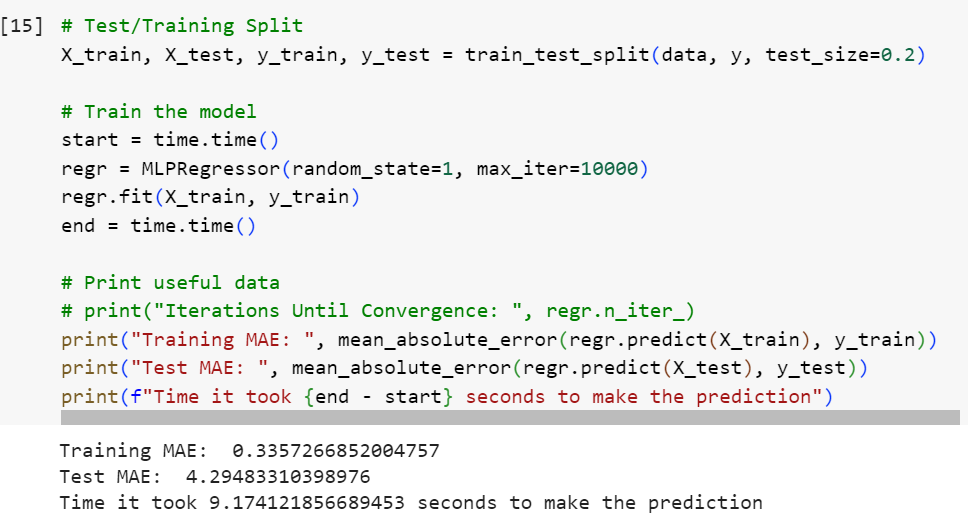
Since this research was only performed on one data set, there is not enough information to generalize the results for all data sets about student test scores. So, to solidify my research more experiments need to be done, using various types of datasets about student test scores. Furthermore, more research can be done to compare the results of these machine-learning models to other models, like k-nearest neighbors and Bayesian statistics.

In conclusion, for the specific problem of predicting student test scores, both neural networks and decision trees provide adequate results. But, based on empirical research the decision tree was both more accurate and quicker. So, I recommend that researchers of student test scores use decision trees as the basis for their machine-learning models. This approach will help them discover complicated patterns in data that no human would have been able to find before.

### Appendix A

### Python Code for Empirical Research

Figure 3: Code for a Neural Network



This is the code I used for the neural network (known as the Multi-Layer Perceptron). I first split the data into training and test sets. Then, I trained the model used the Sklearn MLPRegressor() function. Finally, I printed out the MAE and time it took to run the model.

Figure 4: Code for a Decision Tree

A screenshot of a computer program

Description automatically generated

Figure 5: This is the code I used for the decision tree. I first split the data into training and test sets. Then, I trained the model used the Sklearn DecisionTreeRegressor() function. Finally, I printed out the MAE and time it took to run the model.

### References

Brownlee, J. (2019, August 6). *How to avoid overfitting in Deep Learning Neural Networks*.

Machine Learning Mastery. https://machinelearningmastery.com/introduction-to-regularization-to-reduce-overfitting-and-improve-generalization-error/

Cortez, Paulo. (2014). *Student Performance*. UCI Machine Learning Repository.

https://doi.org/10.24432/C5TG7T.

Gallant, S. I. (1990). Perceptron-based learning algorithms. *IEEE Transactions on neural*

*networks*, *1*(2), 179-191

Hagan, M. T., & Demuth, H. B. (1999, June). Neural networks for control. In *Proceedings of the*

*1999 American control conference (cat. No. 99CH36251)* (Vol. 3, pp. 1642-1656). IEEE.

Hardesty, L. (2017, April 14). *Explained: Neural networks*. MIT News. https://news.mit.edu/

2017/explained-neural-networks-deep-learning-0414

Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020, November). Student performance

prediction model based on supervised machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 928, No. 3, p. 032019). IOP Publishing.

Khemphila, A., & Boonjing, V. (2010, October). Comparing performances of logistic regression,

decision trees, and neural networks for classifying heart disease patients. In *2010 international conference on computer information systems and industrial management applications (CISIM)* (pp. 193-198). IEEE.

Marsland, S. (2015). *Machine Learning: An Algorithmic Perspective* (2nd ed.). CRC Press.

Sungheetha, D. A. (2020). Comparative study: statistical approach and deep learning method for

automatic segmentation methods for lung CT image segmentation. *Journal of Innovative Image Processing*, *2*(4), 187-193.

Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining student-opinion data

with regression and decision tree analysis. *Research in Higher Education*, *45*, 251-269.