MCO3 Machine Learning - Gender Prediction Based on Academic Performance and Demographics

Jan Uriel A. Marcelo, Trisha Gail P. Pelagio, Bryce Anthony V. Ramirez
and Danielle Kirsten T. Sison

De La Salle University, Taft Manila Philippines

1 Introduction

It is traditionally perceived that tests and exams is a measure of a child's academic performance. Getting high marks from these tests indicate the child's natural intelligence and hardworking ethics, yet studies have shown that there are other external factors that can affect academic performance. In a study by Byoung-suk, K. (2012), children require a safe, healthy, and stimulating environment to learn. In every society, parents are the agents involved in this to raise their children in the right environment. This is why it is considered that the family is a vital part in the process of socialization.

In creating this environment, careful and meticulous planning is needed to optimise the support for education. According to Ezewu (2003), family background can especially affect children with regards to their academic performance. Children coming from high socio-economic families with both parents finishing a degree are more likely to succeed than those coming from low income households. The school environment is one of the paramount factors to be considered as well as it is a place where growth and development is influenced.

Gender has been mentioned in literature as well to have considerable effects on a child's academic performance. The socio-cultural differences of girls and boys leads the importance of examining the correlation and likelihood of children succeeding in school. Some environments have professions and vocations regarded as men's work, this can be exhibited by parents assigning menial work to girls while the son is given the complex and demanding stuff and through this, they gain more opportunities to develop their logical reasoning.

With this knowledge in mind, it is interesting to try to predict the gender of a student based on their performance in school and other factors mentioned before. To enforce this, Machine Learning is used to categorise a database of students as accurately as possible without being too biased with the outcome. Machine Learning is a category of algorithm that allows software programs by predicting the outcomes without explicitly being programmed. The basic premise of Machine Learning is to build an algorithm model that can receive input data and use statistical analysis to map out an output while more inputs are received by the model. There are four machine learning methods that are used in the field, the most adopted ones are supervised learning and unsupervised learning. In this study, a supervised learning was used to perform a binary classification on the data set selected by the researchers, meaning the researchers determine the input and the output and the goal is to find what maps the input to the output.

2 Data Set

The dataset that was used for this specific project was taken from the United States. A survey was conducted in the year 2018 with a population that consists of several high school students from different schools. This was performed in order to correlate several personal, social, and economic factors or demographics such as gender, race/ethnicity, parental level of education, meal consumption, test preparation, and different proficiency levels or scores in areas such as math, reading and writing. These are the subjects that were concentrated by the research due to how the subjects generalize the different proficiency skills of a student namely technical, problem-solving and analytical through Mathematics and the macro skills of reading and writing for communication purposes. Due to its simplicity in terms of the classification of the several classes and straight-forward approach in terms of interpreting the data gathered, the dataset was deemed to be preferable by the researchers.

2.1 Features

Gender

Gender primarily serves as the feature through which the classes are defined. Specifically, there are only 2 of which that can be obtained from the feature which fits the requirement of the project. In this way, the column of which is separated and identified as the *y* value or the output from the machine learning process. The project aims to identify the value of the feature in evaluating its samples.

Race/Ethnicity

This feature refers to the racial category which can be classified under groups. These are standardized by the Office for National Statistics if United Kingdom which can also serve as an international reference. Nonetheless, the classification method is used by the United States which is the sample population of the data set used. Through this, the research will likewise utilize the same standard.

Parental Level of Education

This feature of the demographics refers to the education attainment of the parents of the student. This was considered to be a significant feature to be added in the survey questions due to the environmental influence that could be a possibility for an impact towards a child as stated by Maccoby (2000).

Lunch

The lunch feature refers to the meal consumption of a student in the academic institution. In this way, the habits or diet of a student can be understood as schools in the United States provide different food options or meal choices. These could vary upon various factors such as depending on the type of school, family income level, or based on the discretion of the student.

Test Preparation Course

This feature of the dataset refers to the preparatory program that a student took such as tutorials or other additional courses outside of their standard and usual classes that are required to be taken in school. This would allow the students to have better proficiency in certain areas aforementioned in the research that will be used as part of the dataset such as Math, Reading and Writing.

Math Score

The Math Score feature would distinguish the average score of students in their Mathematics exams for the whole school year.

Reading Score

The Reading Score represents the students and their corresponding average score to reading subjects.

Writing Score

This feature comprises the mean score of students surveyed during 2018 regarding their writing courses.

The other features other than gender were used as the X value or the input of the machine learning process to determine the output (gender).

2.2 Labels

Gender

The gender that is represented in this feature would refer to the biological classification of the population. This would consist of the value ‘Male’ to refer to a man and ‘Female’ to refer to a woman. This would be the two main classes that would be the primary data that the machine will try to predict.

Race/Ethnicity

The Office for National Statistics has enumerated a total of 17 possible categories in terms of ethnic classification (Ethnic Category Code, 2001). Such information is primarily identified through groups. However, the data set only consisted of 5 based on the survey conducted on American high school students. ‘Group A’, ‘Group B’, and ‘Group C’ are identified under the White category while ‘Group D’ and ‘Group E’ are presented as mixed. Their corresponding values are provided in the following table.

Table 1. Ethnicities to corresponding Group Label

Code	**Ethnicity**
Group A	British
Group B	Irish
Group C	Any other White background
Group D	White and Black Caribbean
Group E	White and Black African

Parental Level of Education

In the dataset, the parental level of education was represented in 6 different values. This would denote the highest level of educational or progress that the parent of the student attained. The labels of education attainment could be seen in the table below.

Table 2. Educational Attainment Labels with Corresponding Meaning

Label	**Meaning**
Some College	The parent stopped in the middle of an undergraduate’s college program.
Associate’s Degree	This is an undergraduate degree awarded in the United States after a course of post-secondary study lasting two to three years.
Bachelor’s Degree	The parent had completed an undergraduate academic degree awarded by colleges and universities upon completion of a course of study.
Master’s Degree	The parent had an academic degree awarded by a university or college by completing a course of study in a

	high-order overview of a specific field of study or area of professional expertise.
High School	The parent had completed high school and stopped going to school afterwards.
Some High School	The parent had stopped going to school in the middle of high school.

Lunch

For the purposes of research, the gathered data from American high school students regarding the lunch feature is only classified into two categories- standard and free/reduced. ‘Standard’ refers to the meal option in which the students spend their individual money or balance for food in an academic institution. In other words, these are not funded by any scholarships or provisions from institutions whether in school or government. However, the ‘Free/Reduced’ value would refer to the meal option offered by the school through their cafeteria. In this way, the students do not spend on their food, or only allocate a portion of it.

Test Preparation Course

The test preparation course feature included 2 labels in terms of the responses of students towards obtaining extended classes apart from their schools. ‘Completed’ would denote that a student has undergone such preparation, remedial, or additional classes in high school. As such, ‘None’ would denote otherwise in which the students only attended sessions from their respective schools.

Math, Reading and Writing Scores

The values for these features would all range from 0 to 100 denoting the average scores of the students from the specific subjects.

2.3 Class Distribution

The dataset chosen exhibited an almost balanced distribution for the two main classes of the project with the percentages of 48.2% and 51.8% for the male and female respectively. This balanced nature would have implications on the hyperparameters that was used in the final system.

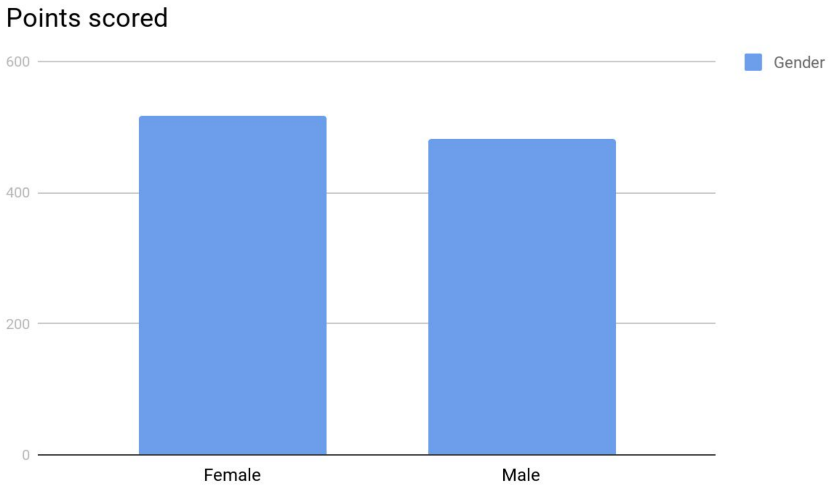

Figure 1. Distribution of Gender in the Data Set

3 Methodology

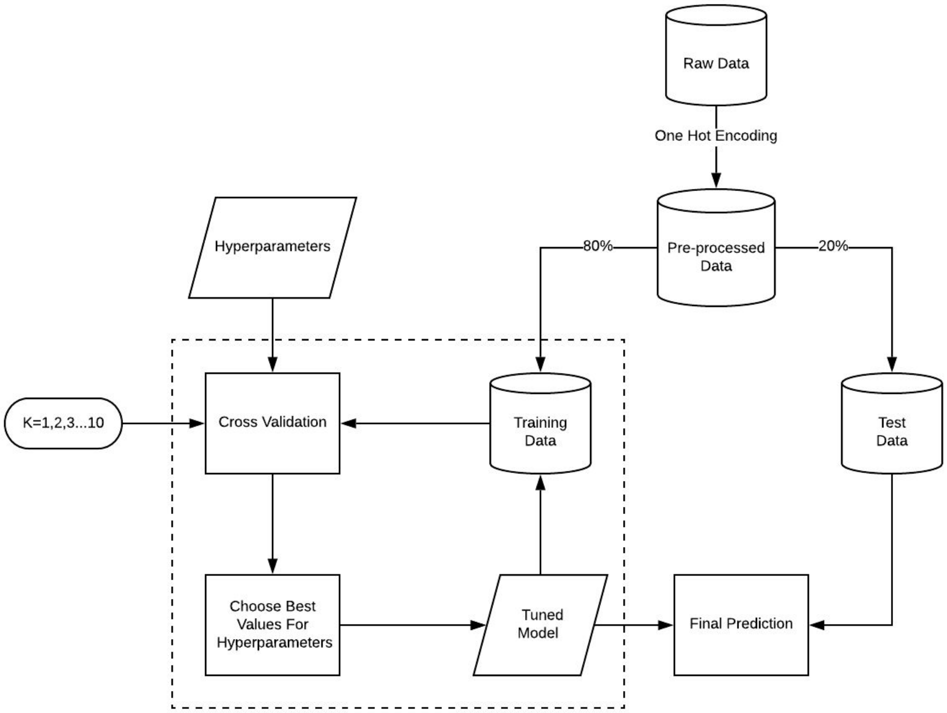

Figure 2. Gender Prediction System Pipeline

3.1 PRE-PROCESSING

1. **One Hot Encoding Method**

One of the disadvantages of the dataset chosen is that it uses multivalued attributes represented as Strings for the values of the features. This was considered to be a problem in terms of the categorization as the classifiers may not support non-numerical values thus, a strategy was used in order to have the data normalized such that it will not be categorized per feature in a non-biased manner. This is done through the representation of each label for a feature with multi-valued non-numeric data as other classes and each instances of a sample which contains the label for the feature will be labelled as 1 and 0 (meaning it does or does not have the feature) for the samples which does not contain the certain label. This is done through the *get_dummies()* function which has the dataset as its parameter which automatically detects and converts categorical data variables into dummy or indicator variables.

3.2 PROCESSING

1. Splitting of Data into Training and Test Data

In order to simulate how the trained model would perform in an unforeseen scenario which was not considered in the Training Phase, the pre-processed data was split into two data sets namely the Training Data, which will be utilized for the cross validating and model tuning, as well as the Test Data, which will be utilized for the final testing phase which will determine the actual performance of the system. This is done by an 80% to 20% ratio between the data that will be used to training as well as the data that will be used for the final testing. This was the division chosen due to the minimal amount of samples that are incorporated in the dataset that was chosen. Since there are only exactly 1,000 students who participated in the survey, the final testing data would need more samples for it to more reliable, this would result to 200 samples being utilized for the final testing phase and the remaining 800 samples for the validation.

2. Cross Validation

After splitting the data into its respective portions, the training data undergoes its validation process. Specifically, the system utilizes a cross-validation within the training data. The K-Fold cross validation iterator is used in order to further separate the data into sub-portions that execute K number of times. The system implements 10 folds of validation in this process. In this way, each fraction of the data is simulated as input through which its results are stored in the validation process. Specifically, the outputs are corresponded towards the 4 metrics such as accuracy, precision, recall, and f1 scores. These are stored within separate arrays in which a value is appended for each fold. Moreover, the mean of the aggregate values for each metric is obtained and displayed in order to appropriately define the proper classifier to implement for the data set.

3. Hyperparameter Value Selection / Model Tuning

This section of the pipeline is determined by the performances and correlation of the hyperparameters chosen to the dataset. The dataset was analyzed for it to be related to the best hyperparameter that will be used. Since there are multiple classifiers that are applied in the system, the hyperparameters for each classifier was determined through its appropriation with the dataset. The results of the cross validation will be exploited in this section and several modifications will be made to the model until the most suitable hyperparameters are already chosen which will then be the Final Tuned Model that will be used for the final testing phase of the pipeline.

4. Final Prediction

The final prediction process implements the classifier which best fits the defined data set. This is determined through the previous procedures or operations of the processing phase. Thus, these also incorporates the need to obtain the values for the significant metrics which are accuracy, precision, recall, and f1 score. In this way, the expected results of the data set are presented.

3.3 Input

The main input of the system are the dataset that will be fed from the pre-processed data. This is mainly consisted of the demographics and the proficiency scores of the students in Math, Reading and Writing. The analysis of these data will then be assisted by the several hyperparameters that will be used to make the model. These hyperparameters would also serve as an input for the cross validation phase of the whole analysis process.

3.4 Output

The system incorporates a number of processes in order for it to effectively classify samples with respect to their gender. With this, there are several results of which expected to be produced in order to achieve the defined objective. Primarily, a model or classifier must be identified through a series of data training and validation. Specifically, the most appropriate among the given options is to be selected based on validation results. Moreover, this would also refer towards the hyper parameters that are used to tune the model. The values of which must be distinctly enumerated based on its effectiveness in increasing the performance of the system. Lastly, the prediction results are also expected to be presented. These are in a form of pair values that match the predicted label towards its true label. Such information prove as evidence of the values obtained by the metrics.

3.5 Classifiers Used

One of the major components of this project is the classifiers that will be used by the system for training as well as the final prediction of the input data into the most accurate results that the system could provide. Due to the simplicity and binary nature in terms of classes of the problem, the researchers have decided to utilize two classifiers which are deemed to be efficient in terms of the chosen dataset. This would be the Decision Tree Classifier as well as the Multilayer Perceptron Classifier. These classifiers would then be discussed further in this section.

Decision Tree Classifier

The decision tree classifier, according to Panigrahi and Borah (2019), are presented similarly with a flow chart. It consists of a tree structure where instances are classified according to their feature values. Each node would represent an instance and the outcomes would be represented by branch. Each leaf node would then epitomize a class label. This classifier is preferable for the dataset since this would involve the process of recursively splitting a node into two which would give a straightforward approach to the problem. Though some problems with this chosen classifier would involve the probability of creating a biased tree is some of the classes dominate a certain feature.

Decision Tree Classifier Hyperparameters

The hyperparameters that was manipulated in the program in order to make several models that was used in the validation phase of the pipeline are criterion, splitter and average. These are the chosen hyperparameters as these parameters were deemed to be of most significance by the researchers in terms of how the dataset is defined.

The criterion refers to the function which measures the quality of the split and this hyperparameter could have two possible values which are Gini Index and Shannon's Entropy. Both of which would calculate the impurities of a dataset which would then assist the machine's capability of splitting a node based on the most preferable split. Shannon's entropy would involve logarithms in its calculations while Gini Index is calculated by subtracting the sum of the squared probabilities from each class.

The splitter parameter would then refer to the strategy that is used to split each node based on the features available. This parameter involves two possible values which are 'Best' and 'Random'. In using the 'Best' value for this parameter, the tree will be split to the most relevant feature. This could best be visualized in a scenario of a dataset which includes the skin color of a person and the system is supposed to determine the gender of a person. The attribute would seem to be irrelevant to the output, thus using 'Best' option would be preferable. As for the 'Random' value, this would allow the program to choose randomly among all the features for splitting. This would be best for scenarios wherein it is known that all features are relevant in determining the answer to a problem.

The final hyperparameter that was chosen for the validation phase is the value of the parameter average in the computation of precision, recall and f-score. This was considered to be a hyperparameter due to the significant changes it brings to the results of the computations. The purpose of this parameter is to determine how to compute all of the scores per feature in the dataset. Changing this feature would be based on how biased the data is. The relevant values that was considered in this research are only two namely: 'Macro' and 'Micro'. Since the dataset involves males and females, this would entail several values for True Positives and False Positives for each class, thus inducing 4 several classifications. These 4 classifications would involve True Positive for both male and female as well as False Positive for both male and female.

The value 'Micro' for the average parameter would calculate metrics in a global manner by taking into account the true positives, false negatives and false positives for each class. This is done by adding all of the predictions that the system answered correctly by the total number of samples. On the other hand, the ‘Macro’ value would involve calculating the metrics for each label, which is the male and female, then finding the unweighted mean of the precision for each class. The issue with this approach is that if the dataset is biased, then this would undermine the precision values for the class with a higher precision.

Multilayer Perceptron Classifier

The Multilayer Perceptron Classifier represents the Neural Networks application in the system. The advantages can certainly be derived from its capabilities to learn non-linear models which significantly corresponds to the data set (Neural Network Models, 2019).

Multilayer Perceptron Classifier Hyperparameters

The research considered a total of 2 varying hyperparameters for this classifier. These are in terms of the activation function and the learning rate of the model. Another aspect of which can significantly affect the metrics would be in terms of solver. However, in the case of the system, this has been set to its default feature which is the Adam solver.

The *solver* parameter of the classifier indicates 3 possible values which are Limited-memory Broyden-Fletcher-Goldfarb-Shanno represented as *lbfgs*, Stochastic Gradient Descent as *sgd*, or Adam. Through this, the Adam solver was utilized for weight optimization feature. This is primarily because it proves to be effective towards large data sets, which are sample ranging from a thousand or more inclusively, towards training time and validation score. Nonetheless, the Stochastic Gradient Descent was also utilized in order to provide further considerations in the learning rate.

Activation function would represent the functions used in neural networks to compute the weighted sum of inputs and biases (Nwankpa, Ijomah, Gachagan, & Marshall, 2018). These data are further used towards validating the firing or activating process of a specific neuron. In this way, activation functions manipulate presented data through the use of gradient processing (gradient descent) to subsequently produce outputs for the neural network. Specifically, Scikit includes a total of 3 of which that can be denoted within its *activation* parameter. The Rectified Linear Unit Function represented by *relu* is deemed to be the most widely used activation function. It represents a nearly linear function which preserves the properties of linear models that made them easy to optimize with gradient-descent methods. Through this, it is also the default activation function for the classifier. Moreover, Linear or Identity Activation Function is represented by *identity* in which the output of the functions are not restricted between any range. The Hyperbolic Tangent Function presented as *Tanh* is also included as a possible activation function. Lastly, *logistics* refers to the Logistic Sigmoid Function.

The learning rate is a significant hyper-parameter that controls the amount of adjustments conducted in the weights of the neural network with respect to the loss gradient (Zulkifli, 2018). Scikit provides a total of 3 of which as possible values for a parameter. *Constant* is set as the default value of the parameter which corresponds to a constant learning rate. Furthermore, *inverse scaling* corresponds to a gradual decrease in the learning rate. Lastly, *adaptive* learning rate is also inclusive as a value for the parameter.

4 Results and Analysis

The first figure shown in this section represents the several trials of the k-way validation with the hyperparameters used for the trials. The results of the changes in hyperparameters would be presented in the second table of the section that would present several metrics such as the accuracy score, precision score, recall score and f-score. In order to further visualize and examine the behavior of the results per model, a bar graph is presented for the third part of results per classifier used. Afterwhich, the classifier which provided the best results were selected to be implemented in the system. This is also supplemented with the appropriate parameters based on its effects in maximizing the metrics enumerated. Moreover, its results are also further discussed to analyze its effectivity.

4.1 Decision Tree Classifier

Table 3. Summary of Hyperparameter Changes in Decision Tree Classifier

	Trial 1	**Trial 2**	**Trial 3**	**Trial 4**	**Trial 5**	**Trial 6**	**Trial 7**	**Trial 8**
Criterion	Gini	Gini	Entropy	Entropy	Gini	Gini	Entropy	Entropy
Splitter	Best	Best	Best	Best	Random	Random	Random	Random
Average	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro

Table 4. for Comparison of Value for Trials 1 to 8

Metric	**Trial 1**	**Trial 2**	**Trial 3**	**Trial 4**	**Trial 5**	**Trial 6**	**Trial 7**	**Trial 8**
Accuracy Score	0.8135603219253008	0.8135603219253008	0.8010595405532115	0.8010595405532115	0.7697751601812783	0.7697751601812783	0.7685247694952336	0.7685247694952336
Precision Score	0.8135603219253008	0.8146198890522542	0.8010595405532115	0.8054842746029873	0.7697751601812783	0.7715838689966672	0.7685247694952336	0.770554303605435
Recall Score	0.8135603219253008	0.8134418384112386	0.8010595405532115	0.8002789421700071	0.7697751601812783	0.7690963344880114	0.7685247694952336	0.7681803503100934

F-Score	0.8135603219253008	0.8127835127567542	0.8010595405532115	0.799597658488506	0.7697751601812783	0.7685452912106279	0.7685247694952336	0.76724933258428754

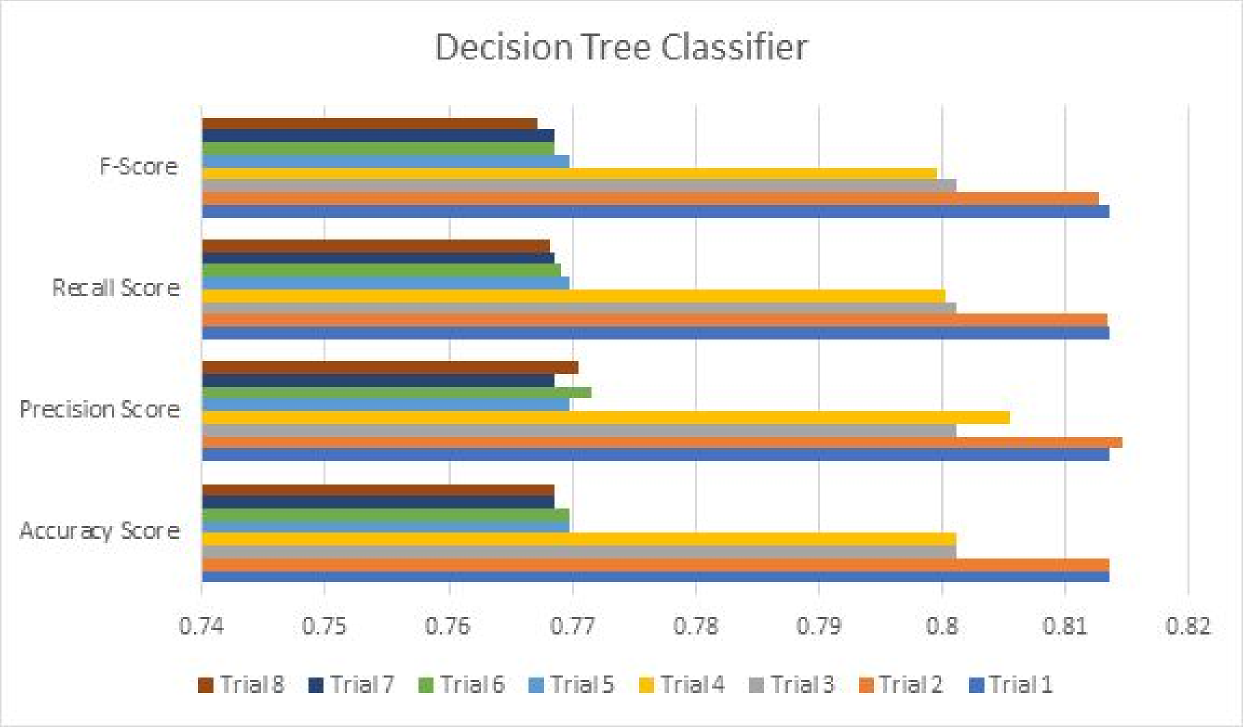

Figure 3. Summary of the Results Gathered from Varying Metrics in the Decision Tree Classifer

As mentioned before, each trial used different combinations of hyperparameters. From this, it could be deduced that the best combination of hyperparameters used is from trial 1. This would involve using *Gini* as the hyperparameter for the criterion, *Best* for the splitter parameter and lastly *Micro* for the average. This produced the best F-score of 0.81356. In totality the explanation of this result could also be reflected from the concept of how each hyperparameter affects the results as aforementioned above.

The Gini Index Loss Function, while having lower values overall for this information gained, proved to be more accurately as it was consistently better than Shannon's Entropy. Nonetheless, this is due to the behavior in which the values are obtained from both the loss functions. Gini Index is preferable in this case since the classes are perfectly mixed and Gini Index, as stated by Walter (2017), is more suitable to minimize misclassification since the curve of the Gini Index would be more symmetric to 0.5. Therefore, such results were to be expected from the trials conducted. For the splitter, *Best* was overall better due to the fact that some of the attributes may play more into a role of predicting the gender than just using *Random* features for splitting. Most of the features included in the research are also relevant features which adds value towards achieving the results in a more accurate manner. Finally, *Macro* average was better than using *Micro* average except on precision, which may be due to the fact that *Macro* can have bias on higher precision values than *Micro* if the dataset is imbalanced.

It can also be observed in the results that the trial which gained the worst scores among the trials is the trial 8 which used Shannon's Entropy, Random and Macro for its hyperparameters. These hyperparameters are the opposite of all the hyperparameters that are chosen to be the best suited for this model. This further proves the points stated above as to why the hyperparameters chosen would be the best suited hyperparameters for this dataset and model.

4.2 Multilayer Perceptron Classifier

Table 5. Summary of Hyperparameter Changes in Multilayer Perceptron Classifier

	Trial 1	**Trial 2**	**Trial 3**	**Trial 4**	**Trial 5**	**Trial 6**	**Trial 7**	**Trial 8**
Solver	Adam	SGD	SGD	Adam	Adam	SGD	Adam	SGD
Activation Function	Identity	Identity	Identity	Logistic	Tanh	Tanh	Relu	Relu
Learning Rate	Constant	Inverse Scaling	Adaptive	Constant	Constant	Inverse Scaling	Constant	Adaptive

Table 6. for Comparison of Value for Trials 1 to 8

Metric	**Trial 1**	**Trial 2**	**Trial 3**	**Trial 4**	**Trial 5**	**Trial 6**	**Trial 7**	**Trial 8**
Accuracy Score	0.8949111189248319	0.8586425613377091	0.873706633849039	0.8973489998437255	0.8936456868260665	0.7902459368651352	0.8960831770589154	0.8686749882794187
Precision Score	0.894814334284584	0.8053072745741522	0.8761761603375529	0.8940513561213586	0.8970419729540824	0.7344649554617909	0.8969565964031256	0.8724562431629943
Recall Score	0.8901688933206682	0.8561105250820441	0.8699562431629942	0.8989017956582217	0.8896134506599624	0.6798480231286139	0.8992354787030429	0.8699408110642288
F-Score	0.8927178123373855	0.8337968823253634	0.8736915924363181	0.8954264359013335	0.8904485878144035	0.7830870057821534	0.8943473315410803	0.864908774808564

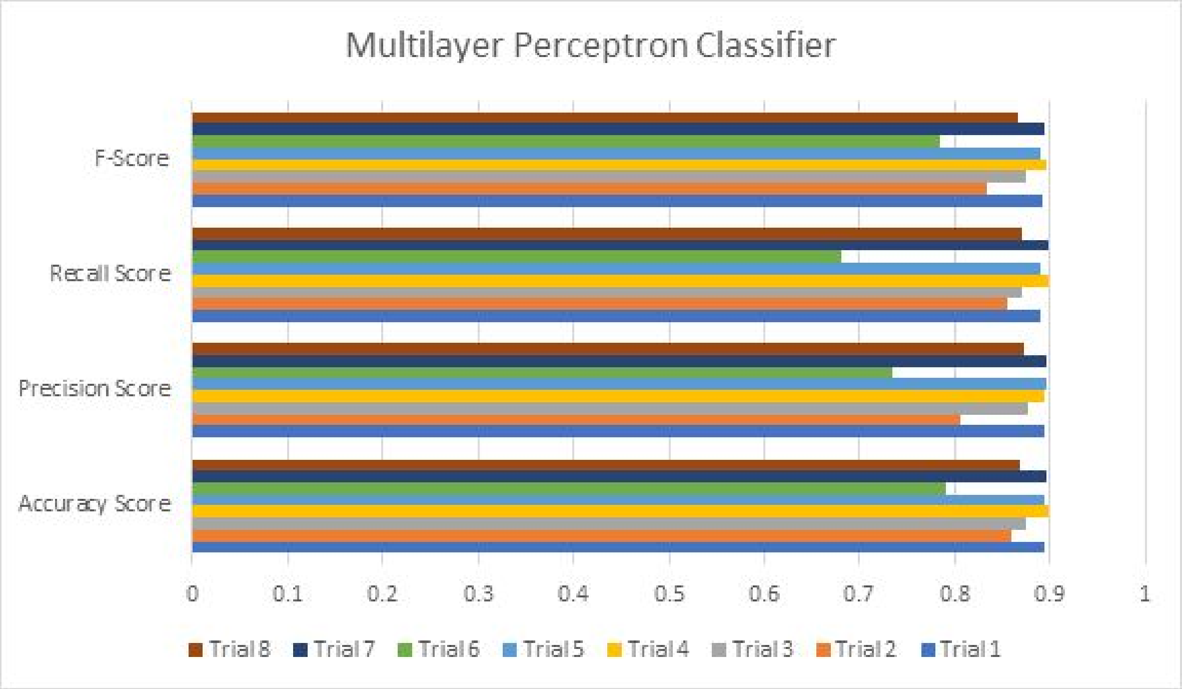

Figure 4.Summary of the Results Gathered from Varying Metrics in the Multilayer Perceptron Classifier

Similar with the Decision Tree Classifier, each trial used different combinations of hyperparameters. From this, it could be deduced that the best combination of hyperparameters used is from trial 4. This would involve using *Logistic* as the hyperparameter for the activation function and *Adaptive* for the Learning Rate. This produced that best F-score of 0.8954264359013335. Generally, each of the trial significantly differed from their values in terms of the metrics considered in the research. An evidence of which would be the selected configuration, trial 4, through which it did not necessarily performed the best in terms of precision score. Nonetheless, among all these trials, the execution with Logistic and *Adaptive* have proved to result with the highest values.

A primary reason for which can be derived from the Activation function selected from the hyperparameters. Generally, this feature of neural networks should utilize non-linear functions, as doing so would prove to be similar towards linear regression. With this, the system had implemented of which that belongs to the non-linear category. Specifically, the Logistic activation function, which is also synonymous with Sigmoid activation function, was utilized in the model. A distinct characteristic of which would be its capability to significantly affect y values with minimal changes in x. Moreover, its behavior when the values are plotted provides an *S* pattern (Neural Network Concepts, n.d.)). Therefore, this aspect of the MPC is often suited towards cases which require binary classification (Tiwari, 2018). As such, this significantly corresponds to the research in terms of classifying gender into male or female.

4.3 Comparison and Analysis

After the validation phase results, the final confusion matrices as well as scores were computed using the hidden test score which yielded the following scores below.

```
[80, 15]
[20, 85]
```

Figure 5. Decision Tree Classifier Confusion Matrix

```
[87,  8]
[ 9, 96]
```

Figure 6. Multilayer Perceptron Classifier Confusion Matrix

Table 7. Metric Comparison of the Decision Tree Classifier and Multilayer Perceptron Classifier

	Decision Tree Classifier (DTC)	**Multilayer Perceptron Classifier (MPC)**
Accuracy Score	0.825	0.915
Precision Score	0.825	0.9146634615384616
Recall Score	0.825	0.9150375939849624
F-Score	0.825	0.9148275257396227

The confusion matrices of both classifiers shows how the models perform in getting true positives and true negatives for each feature in the given. The significant numbers here are the true positives and true negatives in the top left and top right of each matrix (where top right and bottom left predict false positives and false negatives respectively). Nonetheless, the metrics were likewise considered such as the F-score which provides the harmonic mean of precision and recall sores. On average, The Multilayer Perceptron Classifier (MPC) performed better than the decision tree classifier. This is also reinforced with the scores given when evaluating both models using the researchers' data. The MPC performed better on every score evaluated considered within the system. Specifically, all of the scores of which from the Decision Tree Classifier only obtained values of 82.5%, while the MPC reached the scores of around 91% when rounded up with 2 decimal places. In this way, the Multilayer Perceptron Classifier performed 8.98275257396227% better compared to the Decision Tree Classifier in terms of the F-scores. Through these evidences, it is then appropriate to implement MPC as the classifier towards the model of the system.

The approach of the Multilayer Perceptron Classifier performs better than the Decision Tree Classifier as it would give better classification by using non-linear and differentiable boundaries such as the logistic activation function. There are also some regularizations being done in Neural Networks which would reduce the instances of overfitting data. Though the time complexity of the Neural Networks is significantly higher than that of the Decision Tree, it is much preferable due to its accuracy in terms of its use of a nonlinear activation function which would allow stacking multiple layers which could be used for deep learning that is the very nature of why Neural Networks are utilized. This will allow the hidden networks to output and analyze complex data with high accuracy.

4.4 Results Summary

Classification accuracy score is an intuitive way to evaluate the performance of the classifier being used, but the score is often misleading and prone to biases especially if data is asymmetric. In the case of the data set used in this study, it is symmetric.

To further evaluate the model's performance, the precision and recall measure must be taken into account. As noticed, the relationship of the precision and recall scores is inverse–when the other is higher, the other is tugged down to an extent. A higher precision score provides a highly accurate classification of the data but is susceptible to miss a larger instance that is difficult to satisfy. Meanwhile, a higher recall score provides more false positives being identified instead of the true positives–leading to most positives not being predicted. The classifiers with a higher precision and recall scores are likely to perform the best.

The F score is a harmonic balance of the test's precision and recall, indicating the preciseness of the classifier through which refers to the number of correctly classified instances, and the robustness of the classifier which refers to the capability of the model to not miss significant instances. Due to this derivation, a higher F score means the better the performance of the model.

As first the upper bound of the data shown in Fig. 3 and Fig. 4 depicts that Multilayer Perceptron Classifier (MLP) is incomparable to the Decision Tree Classifier in every performance measures taken. As through MLP Trial 4, the model can reach an accuracy as high as 89% and F-score of 89% to as low as 79% accuracy and 78% F-score in Trial 6.

5 Conclusion and Recommendations

The choice of a classification model can prove to be a challenging task when no algorithm is bounded to work well on a problem. As a result, the model is trained with different algorithms and different hyperparameters and select which classifier can predict the best. In this research, the DecisionTreeClassifier proved to be fitting for the gender-classification problem the ML application is dealing.

On the hyperparameters of the DecisionTreeClassifier, the use of *Random* splitter proved made the model underperform as it does not split on the basis of relevant features. The lack of need for logarithmic functions made the *Gini* index to be less computationally intensive and, in practice, has been the ideal impunity metric for binary classification. Micro-averaging was also the ideal metric to use to check the performance due to high precision output it showed.

Meanwhile, the *Identity* and *Adaptive* hyperparameters are the ideal hyperparameters to use when using the Multilayer Perceptron Classifier given the high F-score at 89.56% in the third trial of the experiment. This is primarily due to the constant learning rate it maintains throughout the execution of the program for every data fed to the MLP model and adjusts accordingly to the data. The *Identity* activation function is ideal when the problem domain deals with regression such as the one dealt with in this study.

The researchers recommend to test out other classifiers such as SVC and Naive Bayes using the data used in this research and see if the accuracy and F-score get optimal. These classifiers are recommended as well with the scikit guides with classification problems that have text data in them. Experimenting with the other hyperparameters in the DecisionTreeClassifier and Multilayer Perceptron Classifier is encouraged to explore more combinations and possibilities that can possibly increase or decrease the metric scores observed by the researchers.

With regards to the precision and recall tug-of-war, it is not really required to get a perfect balance. Throughout the development of the Machine Learning application, it is sometimes more important to have the data be classified correctly rather than classifying all at once with a huge margin of error. Sometimes classifying most of the data can be more important than correctly getting them. It all depends on the domain of the problem being solved.

References

Byoung-suk, K. (2012). LandscapePerformanceResearch; SchoolEnvironment& Students' Performance, Paper from Landscape Architecture Foundation.

Ethnic Category Code. (2001). Retrieved from National Health Service Data Dictionary: https://www.datadictionary.nhs.uk/data_dictionary/attributes/e/end/ethnic_category_code_de.asp

Ezewu E. (1994) Sociology of education. Lagos: Longman Group Ltd Hongkong

Neural Network Concepts. (n.d.). Retrieved from https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/.

Panigrahi, R., & Borah, S. (2019). *Classification and Analysis of Facebook Metrics Dataset Using Supervised Classifiers*. Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/pii/B9780128154588000013

Pedregosa et al., . (2011). *Neural network models (supervised)*. Retrieved from Scikit-learn: Machine Learning in Python: https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Nwankpa, C. E., Ijomah, W., Gachagan, A., & Marshall, S. (2018, November 8). *Activation Functions: Comparison of Trends in Practice and Research for Deep Learning*. Retrieved from arXiv Cornell University: https://arxiv.org/pdf/1811.03378.pdf

Tiwari, S. (2018, January 29). *Activation Functions in Neural Networks.* Retrieved from https://www.geeksforgeeks.org/activation-functions-neural-networks/.

Usaini, M. I., & Bakar, N. A. (2015). The Influence of School Environment on Academic Performance of Secondary School Students in Kuala Terengganu, Malaysia

Walter, C. (2017). *Data Science*. Retrieved from StackExchange: https://datascience.stackexchange.com/questions/10228/when-should-i-use-gini-impurity-as-opposed-to-information-gain

Zulkifli, H. (2018, January 22). *Understanding Learning Rates and How It Improves Performance in Deep Learning*. Retrieved from Towards Data Science:

https://towardsdatascience.com/understanding-learning-rates-and-how-it-improves-performance-in-deep-learning-d0d4059c1c10

Appendix A. Contribution of Members

Name	Contributions
Marcelo, Jan Uriel A.	Program Design, Documentation
Pelagio, Trisha Gail P.	Program Design, Documentation
Ramirez, Bryce Anthony V.	Program Design, Documentation
Sison, Danielle Kirsten T.	Program Design, Documentation