Student Name: …

ADTA 5340: Discovery and Learning with Big Data

**Final Project Report**

# PART I:

**Strategy for Implementing Machine Learning in Infinite Realms Gaming (IRG)**

***1. Company Overview:***

The company's name is Infinite Gaming (IG). Infinite Gaming is a Medium-sized company with an average of 400 employees. It is in the Gaming and Virtual Reality Entertainment field in IT.

Its competitors are, however, established gaming companies that leverage cutting-edge technologies for competitive advantage.

***2. Strategy Overview:***

1. Defining Business Goals:
   1. Analyze the company's current market position and identify areas where machine learning (ML) can be used to increase the company's value.
   2. Set specific goals, such as enhancing player immersion, optimizing the game's mechanics, improving fraud detection, and increasing player retention rates.
2. Technology Selection:
   1. Choose a robust ML framework like TensorFlow by considering scalability, flexibility, and industry adoption.
   2. Leverage cloud-based ML services (e.g., Google Cloud AI, Microsoft Azure ML) for seamless integration, scalability, and cost-effectiveness.
3. Data Infrastructure Enhancement:
   1. Improve data collection mechanisms within games to capture player interactions, preferences, and behavior patterns.
   2. Implement a scalable data warehouse architecture to facilitate the processing large volumes of structured and unstructured gaming data for ML model training.
4. Human Capital Development:
   1. Assess the current skill sets of the company's workforce in ML, data analytics, and also software engineering.
   2. Conduct regular training programs and certifications to upskill employees in ML algorithms, data preprocessing, and model deployment.
   3. Foster a culture of innovation and continuous learning by encouraging collaboration among teams.
5. ML Application Areas:
   1. Develop AI-driven systems for personalized game content recommendations, adaptive difficulty levels, and real-time player feedback analysis.
   2. Utilize reinforcement learning algorithms to dynamically adjust game mechanics, balance gameplay, and introduce new challenges based on player behaviors.
   3. Implement ML-based anomaly detection models to identify cheating, unauthorized access, and fraudulent activities within games.
   4. Build predictive churn models to segment player cohorts and design targeted retention campaigns to enhance player loyalty.
6. Monitoring and Evaluation:
   1. Define key performance indicators (KPIs) aligned with business objectives, such as player engagement metrics, fraud detection accuracy, and revenue growth.
   2. Deploy real-time monitoring dashboards and analytics tools to track ML model performance and derive actionable insights for decision-making.
   3. Conduct regular audits and reviews of ML models to ensure transparency and compliance with regulatory standards.
7. Collaboration and Partnerships:
   1. Forge strategic partnerships with AI technology providers, gaming platforms, and data analytics firms to access advanced ML tools, expertise, and resources.
   2. Establish collaborations with universities and research institutions to stay at the forefront of ML innovations.
   3. Explore joint ventures with other gaming companies to share knowledge and mutual growth opportunities.

# PART II: *# Report on 2018-2019 Daily Attendance Dataset*

The name of this Dataset is 2018-2019 Daily Attendance. Downloaded from Kaggle. The link to the Dataset is <https://www.kaggle.com/datasets/sahirmaharajj/school-student-daily-attendance>.

The Dataset is composed of 6 columns :

1. school DBN (Admission Number)
2. date
3. total-enrollment
4. number of students absent
5. number of students present
6. number of students released that day.

***Quality of the data:***

***Missing Values:***

The Dataset has a shape of (227152, 6).

This means that the Dataset has 227252 rows and 6 columns.

Using the duplicate() function, the Dataset has no duplicated rows or records.

Also, using the isnull() function, the Dataset has no occurrences of missing values.

***Handling Missing Values***

Since there were no missing or null values, there was no need to drop rows with missing values. The Dataset was clean, to prove the same, I dropped records having duplicate rows and missing values and exported the new data to a new CSV file called Cleaned\_2018\_2019\_Daily\_Attendance.csv file.

I created a summary report comparing the length(number of records) of both CSV files, and the result was that both had 277153 records.

***Summary of Machine Learning Project***

This Dataset can be used to analyze educational trends, attendance patterns, and their correlations with academic performance and other socio-economic factors. Analyzing these patterns over time or across different schools can help in identifying critical issues such as chronic absenteeism, the effectiveness of attendance policies, and the impact of external factors on student attendance.

The target variables being

1. Student Present
2. Student Absent

The feature variables are Date, Enrolled, Absent, or Present (depending on the target variable chosen), and released feature.

# PART III (Abalone) `Linear Regression`:

***1. Preprocessing the Dataset***

The dataset name is called abalone.

The Dataset has 4177 records, with 9 columns.

Although the Dataset has no missing values or duplicate values, it needs preprocessing for the following reasons:

1. Lack of column names: The original CSV file had no column names, so I had to name the columns, during importing, for better modeling, feature encoding, and target variable declaration.
2. Handling abnormal values: Preprocessing the data by replacing values like '3+' with other numeric counterparts.
3. One-hot encoding: The 'Sex' column is one-hot encoded to convert categorical data into numerical format for better modeling.

***2. Model Selection***

This model used is Linear Regression. This is because

1. Linear Regression takes/assumes a linear relationship between the features and target variable, which is a reasonable assumption for some datasets.
2. The model is also simple, easy to understand, and easy to interpret.

***3. Building the Model***

The model was constructed using the Linear Regression algorithm.

It aims to find the best linear relationship between the features X and the target variable y.

Here, the Rings column was dropped from X, which is for the features, and used in y, which is for the target variable y. This means that using feature X, the model would be trained to get variable y.

The features used are Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight, Sex\_I, and Sex\_M.

***4. Training the Model***

To train the model, we used a test size of 0.2, meaning that 80% of the data for training the model. 20% reserved for evaluating the model's performance on unseen data.

The model is trained with the training data, to learn patterns and relationships between the features (X) and the target variable (y).

It uses a model. fit(X\_train\_scaled, y\_train) to train the Linear Regression model.

***5. Making up Two New Records***

Two new data records with the attribute values are created to demonstrate how the model predicts the age from the records.

The attribute names include Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight, Sex\_I, and Sex\_M.

***6. Predicting the Age***

The model predicts the age of abalones for the new records based on the trained Linear Regression model.

The predicted Age for New Records was 14.84 years and 9.18 years respectively.

***7. Evaluating the Model***

Explanation: The model is evaluated using metrics like Mean Squared Error, Accuracy, and R-squared on the test set to assess its performance and predictive capability.

1. The Mean Squared Error is : 4.89
2. The Accuracy: 0.5482
3. The R-squared (Coefficient of Determination): 0.55.

***8. Interpretation of Prediction Results***

1. The MSE (4.89) indicates that the squared difference between the actual ages of abalones and the predicted ages by the model is around 4.89 years.
2. The R-squared value of 0.55 shows that the model explains 55% of the variance in the age of abalones based on the features included in the Dataset.

# PART IV (Adult Salary) `Logistics Regression`:

## EDA: Data Discovery

The dataset name is Adult Salary.

The Dataset has a shape of (48842, 15) which means that the Dataset has 48841 records and 15 columns. The dataset columns include, Age, Employment type,Education and Education\_num, which is the highest level of education in numerical form. Other features include marital, Occupation, relationship, race, sex, Capital gain and loss, Weekly\_hours, number of hours worked per week, Country of origin and lastly Income.

Data Types of the features include : int64 for Age,Fnlwgt,EducationNum, Capital\_gain, Capital\_loss and Weekly-hours. The rest of the features are objects.

The Dataset also has 29 duplicate values.

## EDA: Preprocessing Data

The Dataset is not clean, and needs preprocessing as it has 29 duplicate records/rows. I dropped rows with duplicate rows.

The Dataset has no null values. However, there are instances where empty fields are represented by ?. There are 1049 instances, where ? is in the Dataset across several features. To remove these null values, I opted to replace the character "?" with NaN. This ensures that clean data is used in training the model.

Categorical variables were encoded using the LabelEncoder preprocessing import. For each categorical column, label encoder was applied to transform into numerals/numerical values.

Features to be used were selected based on their importance to the target variable y(Income feature).

Chosen features include

1. Age
2. employment\_type
3. Education
4. Marital\_status
5. Occupation
6. Relationship
7. Race
8. Sex
9. Capital gains and Losses,
10. weekly hours
11. Country.

## Model Planning

The model selected is Logistic Regression.

Logistic Regression is faster to train as compared to other models for example SVM, especially for large datasets and high-dimensional feature spaces It also provides reasonable performance for binary classification tasks.

## Model Building

The model was constructed with the following parameters:

1. Random State=42 (To ensure consistent results across different runs of the model).

Standard Scaler was applied to the features before training the model. This step ensures that all features are on the same scale, which is important for many machine learning algorithms.

The Dataset was split into training and testing sets using 80/20 split ratio. I.e 80% of the data for training the model. 20% reserved for evaluating the model's performance on unseen data.

## Reporting Results

The model (Logistic Regression) was evaluated using the accuracy and cross validation scores.

The model achieved a testing accuracy of approximately 55.18%, indicating its ability to predict income levels correctly on unseen data.

Using the new records used to test the data predicts that income for both records is less than 50K.

The model's cross-validation scores across 10 folds are :

[0.55006402 0.55339309 0.54980794 0.54852753 0.55288092 0.55288092 0.54699104 0.54647887 0.54443022 0.54775928]

The mean accuracy across these folds is approximately 0.549.

# PART V (Car\_evaluation) `KMeans Clustering`:

## EDA: Data Discovery

The dataset name is car\_evaluation.

The Dataset's shape is (1728, 7), which means that the Dataset has 1728 records, and 7 columns. Initially, the Dataset had no column names, so I had to allocate feature names to the columns. The dataset features include, Price, Maintenance, Doors, Passengers, Luggage, Safety and Evaluation.

Data Types of the features are objects.

## EDA: Preprocessing Data

Apart from having no column names, the Dataset required no cleaning, as it had no duplicate or missing values in the records.

This was determined by calculating the sum of missing values and duplicate values. In addition to this, I dropped missing values and duplicate values, and exported the cleaned Dataset to a new csv file called cleaned\_adult\_salary.csv file. I calculated the length of the records. Both had 1728, showing that there were no missing values.

Categorical variables were encoded using the LabelEncoder preprocessing import. For each categorical column, label encoder was applied to transform into numerals/numerical values.

## Model Planning

The model used is KMeans Clustering.

KMeans is suitable to identify natural groupings within the Dataset as it is able to split the data into several distinct clusters. The number of clusters was set to 4.

## Model Building

The model was constructed with 4 clusters and a random state of 42. Dimensionality reduction was conducted using t-SNE inorder to reduce the number of features while retaining most of the important information. t-SNE also transforms original into a lower dimensional space allowing us to visualize the data in a violin plot provided the matplotlib and the seaborn import .

## Reporting Results

The KMeans clustering model successfully identified 4 cluster centers showing distinct characteristics for price, maintenance, doors,Passengers, luggage and safety features.

The KMeans clustering model predicted the cluster for the new records as [1 0], assigning both records to cluster 1.

The silhouette score for the model is approximately 0.114, suggesting that the clusters might not be well-separated.

# PART VI `Linear Regression vs. Decision Tree (CART) Regression`:

## EDA: Data Discovery

The dataset name is income-education. The data was downloaded from Kaggle with the following being its location <https://www.kaggle.com/datasets/sheydaayati/average-personal-income-by-level-of-education>.

The Dataset has a shape of (425040, 10) which means that the Dataset has 425040 rows, and 10 columns. The dataset columns include: id,YEAR,Geography,Type of work,Wages,Education level,Age group,Both Sexes,Male,Female.

Data Types of the features are as follows :

1. float64 - Both Sexes,Male,Female.
2. Int64 - id,year
3. object- Geography,Type of work,Wages,Education level,Age group.

## EDA: Preprocessing Data

The Dataset chosen was a clean one as it had no duplicate nor null/missing values. To prove this, I calculated the number of duplicates and null values. In addition, I dropped rows having duplicates and missing values, exported the cleaned Dataset to a file called income-education\_cleaned.csv , and calculated the length of the records for both uncleaned Dataset and cleaned Dataset. They both had a total of 425040, indicating that no record had been dropped.

Categorical variables were encoded using the LabelEncoder preprocessing import. For each categorical column, label encoder was applied to transform into numerals/numerical values.

## Model Planning

Two models were used, Linear Regression and Decision Tree Regression. The goal was to compare the appropriateness of the model to the income-education dataset from its size, complexity etc..

The Wages feature was dropped from the features(X), in order to be used as target variable (y).

## Model Building

The DecisionTree Regressor model was constructed with a random state of 42.

R Squared scores of the both models were calculated after the data was fitted in the models to check the accuracy of the models.

## Reporting Results

I noted that the Decision Tree Regression model had a higher R-squared score(r2) of 0.637 as compared to Linear Regression which has 0.028, indicating that it explains more variance in the target variable (Wages).

A new data set with three records was introduced. Predictions for Linear Regression are : [3.44658169 3.04010716 3.8829657], while predictions for Decision Tree Regression being [4. 4. 4.].

The predictions from the Decision Tree Regression model are all rounded to 4, indicating a consistent prediction for the new introduced records.

# PART VI Logistic Regression vs. K-Nearest Neighbors

## EDA: Data Discovery

The dataset name is income-education. The data was downloaded from Kaggle with the following being its location <https://www.kaggle.com/datasets/sheydaayati/average-personal-income-by-level-of-education> .

The Dataset has a shape of (425040, 10) which means that the Dataset has 425040 rows, and 10 columns. The dataset columns include: id,YEAR,Geography,Type of work,Wages,Education level,Age group,Both Sexes,Male,Female.

Data Types of the features are as follows :

1. float64 - Both Sexes,Male,Female.
2. Int64 - id,year
3. object- Geography,Type of work,Wages,Education level,Age group.

## EDA: Preprocessing Data

The Dataset chosen was a clean one as it had no duplicate nor null/missing values. To prove this, I calculated the number of duplicates and null values. In addition, I dropped rows having duplicates and missing values, exported the cleaned Dataset to a file called income-education\_cleaned.csv , and calculated the length of the records for both uncleaned Dataset and cleaned Dataset. They both had a total of 425040, indicating that no record had been dropped.

Categorical variables were encoded using the LabelEncoder preprocessing import. For each categorical column, label encoder was applied to transform into numerals/numerical values.

## Model Planning

Two models were used, Logistics Regression and K-Nearest Neighbors. The goal was to compare the appropriateness of the model to the income-education dataset from its size, complexity etc..

The Wages feature was dropped from the features(X), in order to be used as target variable (y).

## Model Building

The LogisticRegression model was constructed with a random state of 42, while KNeighborsClassifier was constructed with "n\_neighbors=3".

R Squared scores of the both models were calculated after the data was fitted in the models.

## Reporting Results

Logistic Regression model outperforms the K-Nearest Neighbors (KNN) model in terms of accuracy for this particular classification task with accuracy score for Logistic Regression being approximately 0.500, while for KNeighborsClassifier being approximately 0.315.

A new dataset with three records was introduced. The features were encoded and fed into the models.

Predictions for LogisticRegression is : [4 4 4], while predictions for KNeighborsClassifier is [4 4 4].

Both Logistic Regression and KNN models predicted the same label for all three new records, indicating consistency in their predictions.