Assignment 1  
Kaggle Competition

short line

Monali Patil

Student ID: 14370946  
08/09/2023

36120 - Advanced Machine Learning Application

Master of Data Science and Innovation

University of Technology of Sydney

Table of Contents

[**1. Executive Summary 2**](#_xvtyf97iwmd0)

[**2. Business Understanding 3**](#_if9iwijz2pg3)

[a. Business Use Cases 3](#_bnjfuh5ewogw)

[**3. Data Understanding 4**](#_yg03y3w5hig9)

[**4. Data Preparation 5**](#_tj77rrmixzhi)

[**5. Modeling 6**](#_6vwh14nc44ga)

[a. Approach 1 6](#_uv8sn0ik6017)

[b. Approach 2 6](#_wtwxuxemqtvc)

[c. Approach 3 6](#_nrl8g3ri789x)

[**6. Evaluation 8**](#_mngiyx5tet6w)

[a. Evaluation Metrics 8](#_qx20m6hrk8uo)

[b. Results and Analysis 8](#_25v57zsj5m7g)

[c. Business Impact and Benefits 8](#_hkob76wu4d6q)

[d. Data Privacy and Ethical Concerns 9](#_uf1z6gbsejg6)

[**7. Deployment 10**](#_8rqux13i2o01)

[**8. Conclusion 11**](#_9hr11g79asdx)

[**9. References 12**](#_ma3y8ytjvp1y)

# 

# Executive Summary

Project Overview:

The project involves working with the statistical data of college students playing basketball. This learning task aims to predict whether a college basketball player will be drafted to join the NBA league based on their statistics. The significance of this project lies in its potential to provide insights into the factors that influence NBA draft selections, allowing teams to make more informed decisions about which players to select. Additionally, it can help sports analysts and fans anticipate which college players are likely to transition to the professional NBA league.

Problem Statement and Context:

The problem addressed in this project is predicting whether a college basketball player will be drafted into the NBA. This is a highly relevant and impactful challenge in the world of sports, as it can shape the careers of young athletes and have significant financial implications for both players and NBA teams.

The context of this project involves analyzing a dataset containing various statistics related to college basketball players. These statistics include information about the player's performance, efficiency, shooting percentages, rebounding, assists, turnovers, and much more.

By training a machine learning model on historical data, the goal is to create a predictive tool that can evaluate a college player's likelihood of being drafted based on their performance metrics. This model can be used by NBA teams, scouts, and analysts to identify promising talents and make more informed draft choices.

Achieved Outcomes and Results:

The achieved outcomes of this project include the development of a predictive model with the ability to classify whether a college basketball player will be drafted into the NBA. The results of the model, based on historical data, reveal insights into the factors that influence NBA draft selections. These insights can assist in talent scouting and player evaluation. Additionally, this model will offer valuable insights to sports commentators, fans, and scouts, aiding them in predicting individual players 'potential NBA draft prospects.

In summary, this project addresses a significant problem in the world of sports by leveraging machine learning to predict NBA draft outcomes for college basketball players. The achieved model provides a data-driven approach to talent evaluation and draft decision-making, benefiting both players and NBA teams.

# Business Understanding

The project's primary business use case is centered around predicting whether a college basketball player will be drafted to the NBA league based on their performance statistics. This prediction has several practical applications:

* Talent Scouting: Helps NBA teams identify promising college players.
* Player Career Guidance: Assists players in making informed career decisions.
* Fan Engagement: Generates excitement among basketball fans.
* Sports Commentators: Enhances sports journalism and analysis.

Challenges and Opportunities:

* Predictive Accuracy: Ensuring accurate draft predictions.
* Data Complexity: Handling diverse player and team data.
* Real-time Updates: Providing real-time draft predictions.
* Fan Expectations: Managing fan expectations responsibly.
* Ethical Considerations: Addressing data privacy and fairness.

Key Objectives:

* Accurate Predictions: Build a highly accurate draft prediction model.
* Scalability: Ensure the model can handle real-time data during the draft.
* Privacy and Fairness: Handle player data ethically and avoid bias.
* User-Friendly Interface: Create an easy-to-use interface for stakeholders.
* Stakeholder Collaboration: Work closely with NBA teams, players, and media.
* Ethical Guidelines: Establish guidelines for responsible prediction use.
* Continuous Improvement: Incorporate new data to improve predictions.

# Data Understanding

Dataset:

The dataset provided contains a wide range of features that illuminate players' performance during their college basketball season. The dataset comprises 64 players' performance attributes, including Games Played (GP), Minutes Played (Min\_per), Offensive Rating (ORtg), Defensive Rating (DRtg), Field Goals Made (twoPM), Free Throws Made (FTM), and many others offer insights into various facets of a player's playing style and contribution to their team.

Data Source and Collection Methods:

The dataset, available in CSV format, was accessible via the canvas portal, with distinct files designated for training and testing. As it was acquired through the University portal as part of student resources, there were no concerns related to copyright or privacy issues.

* metadata.csv: Metadata of the Basketball Players
* train.csv: Basketball Players Training dataset
* test.csv: Basketball Players Testing dataset

Data Limitations:

* The absence of metadata for two features, 'ftr' and 'pfr,' presents a challenge in comprehending their definitions.
* The 'ht' column indicates the player's height, with the feature values presented in a data format. Consequently, some processing is necessary to extract meaningful values.
* Additionally, the dataset includes unique identifiers such as player IDs, player numbers, and conference names, which could potentially lead to overfitting of the model.

The exploratory data analysis techniques, checking the dimension of the database, feature names, accessing the initial datapoints etc. using different pandas functions were carried out to examine and study players' information to comprehend and uncover patterns, aiming to identify prospective players for NBA draft selection.

# Data Preparation

To ensure the quality of the data to be utilized by the model analysis, conducted the below activities.

* Handing missing/null values.

A graph with numbers and symbols

Description automatically generated with medium confidence

A black background with white text

Description automatically generated

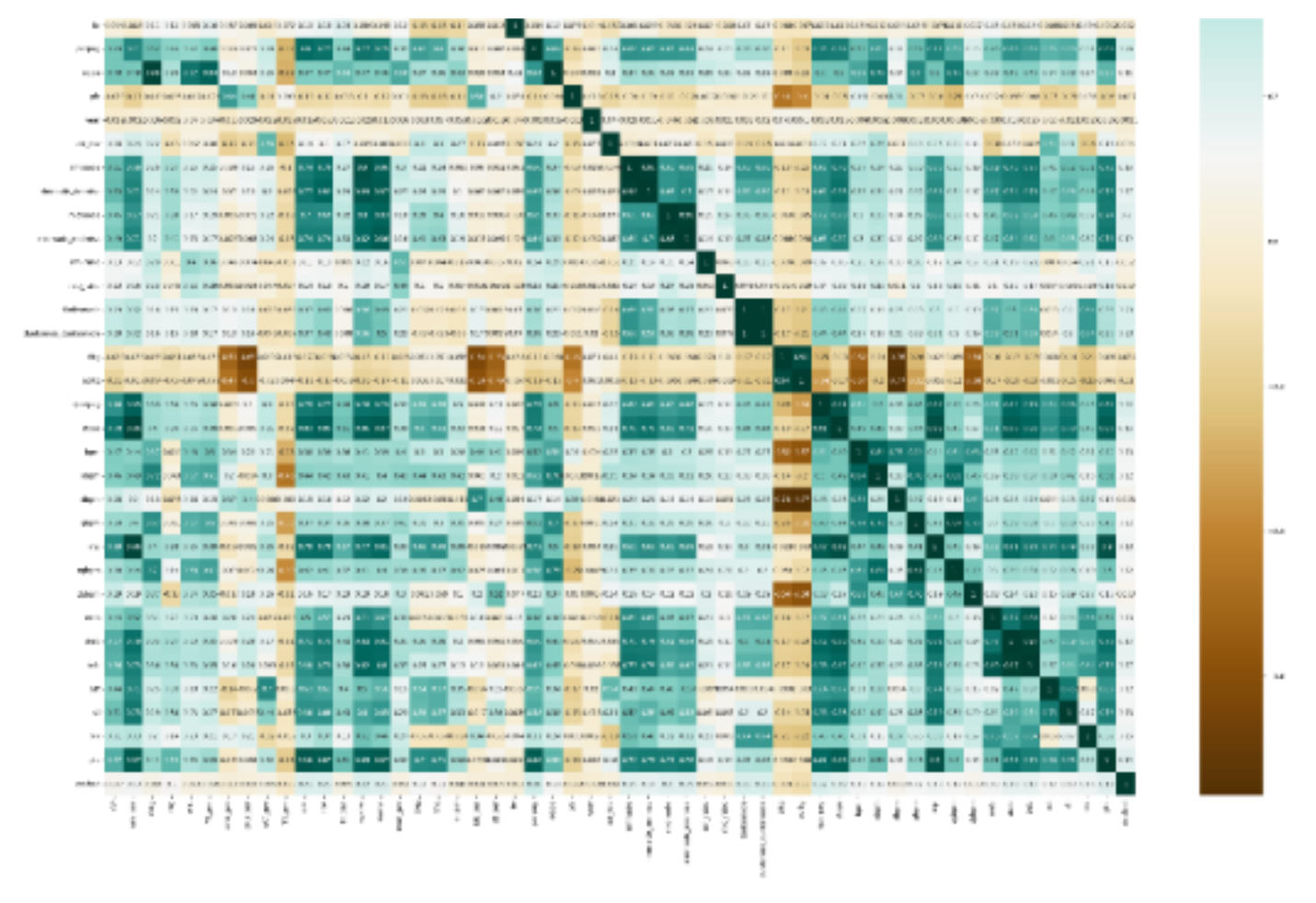
* Eliminating identifiers.

Removed unique identifiers namely 'player\_id', 'num', 'team' and 'conf' as its inclusion in the analysis can lead to overfitting, where the model fits to these specific values rather than the underlying generalized patterns in the sportsman's records.

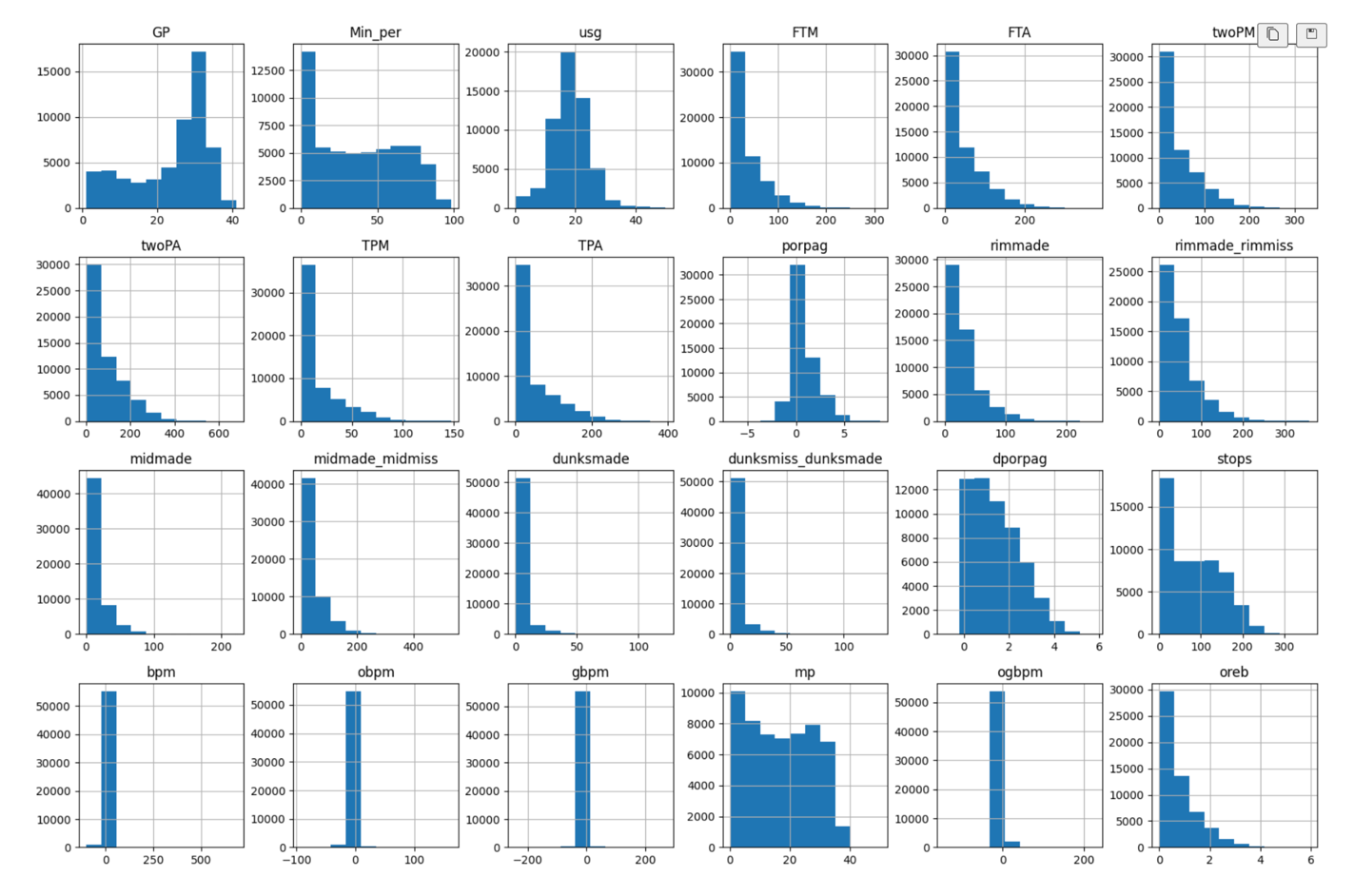
* Duplicate records.
* Feature Engineering - Processing ‘ht’ feature to derive suitable information.

The 'ht' attribute, initially in a date format, represents player height, vital in basketball. Unique value analysis revealed that 'Jun,' 'Jul,' and 'Aug' don't signify months but likely correspond to heights like 6 feet, 7 feet, etc. Consequently, these values were converted into numerical 'ht\_cm' for centimetres as numerical input for machine learning.

* Selecting appropriate features based on the correlation coefficient.

Performed correlation analysis to determine which features would be most valuable for building a predictive model for the NBA draft prediction task.

* Data distribution of various features.



* Accessing if imbalance targets classes.

A yellow circle with black text

Description automatically generated

Performed Oversampling with SMOTE (Synthetic Minority Over-sampling Technique) method to address class imbalance by generating synthetic observations for the minority class representing players who have been drafted denoted by the value 1.

* Features Scaling.

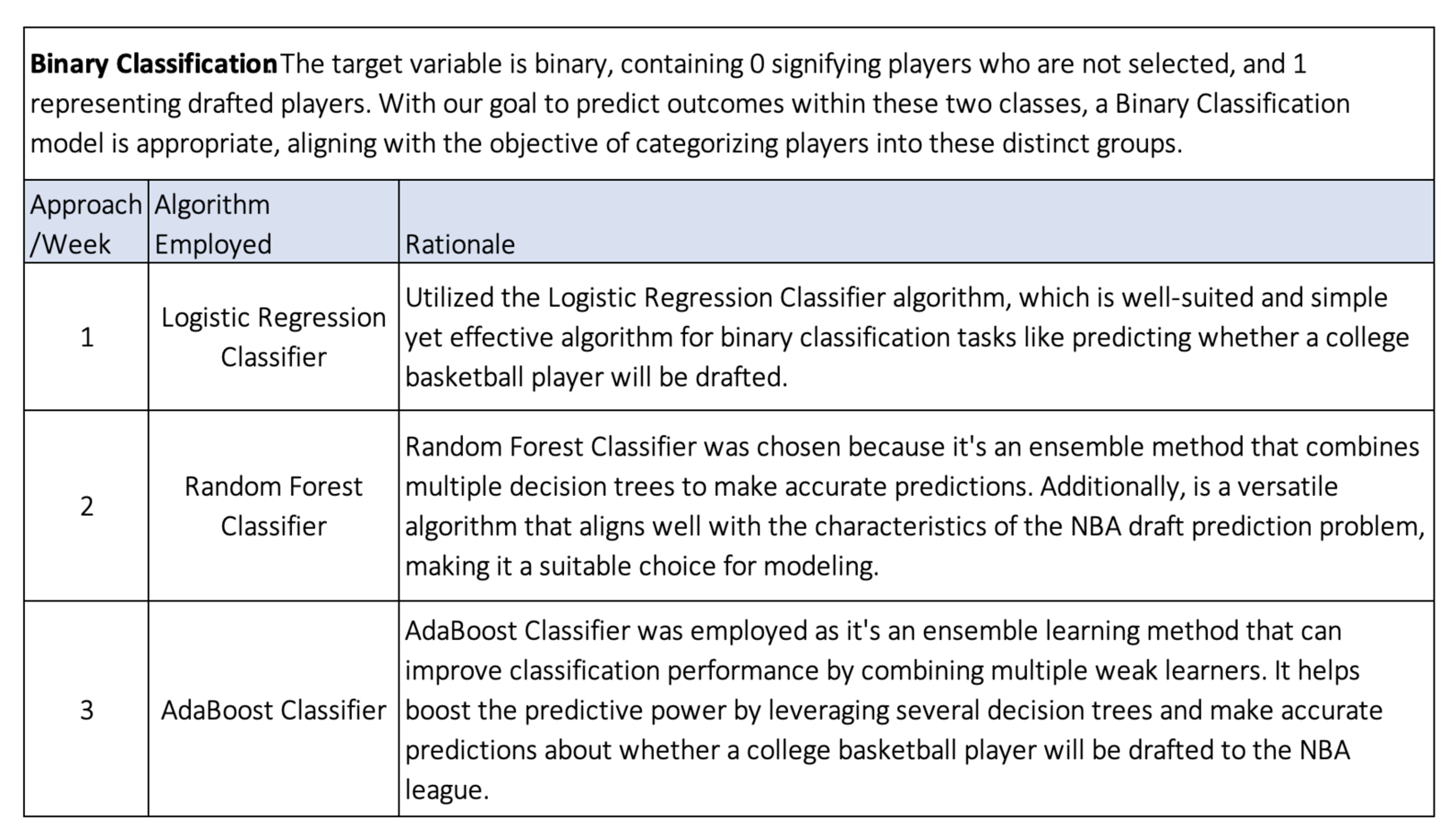
Employed StandardScaler method because it prevents the algorithm from prioritizing high-value features over other more informative ones.

# Modeling

As part of the learning process, below are the Classifier models built, trained, and tested for this binary classification problem.



Considering the business objective of determining potential players and selection, there is no good theory to map and select a suitable algorithm for this binary classification problem, so different experiments are performed to discover which algorithm and algorithm configuration results in the best performance for this binary classification task.



# Evaluation

Evaluation Metrics:

The model's performance is assessed using the AUROC (Area Under ROC) metric, which assesses a model's ability to distinguish between the two classes (drafted or not drafted) by measuring the trade-off between true positive rate and false positive rate.

The AUROC metric is relevant because it quantifies the model's capability to make accurate predictions, which directly aligns with the project goal of predicting NBA draft outcomes. A higher AUROC indicates a better ability to differentiate between drafted and non-drafted players, which is essential for making informed decisions in player selection.

Results and Analysis:

Business Impact and Benefits:

Data Privacy and Ethical Concerns:

* Data privacy was carefully considered in this project. Since the dataset was obtained from a university portal as a student, there were no concerns regarding copyright or privacy issues. However, the dataset contained unique identifiers like player IDs and numbers, which were removed to ensure the privacy of individuals associated with the data.
* Ethical concerns in this project primarily revolve around fairness and bias. When using data for predicting NBA draft selections, it's crucial to ensure that the model doesn't preserve biases related to race, ethnicity, or other factors.
* Ethical data collection and preprocessing techniques, like removing personally identifiable information, were employed to mitigate these concerns. Data was used solely for the purpose of predicting NBA draft selections, and all efforts were made to ensure fairness, transparency, and privacy in the modeling process.

## Results and Analysis

* Present the results of the model evaluation, including accuracy, precision, recall, F1-score, etc.
* Analyze and compare the performance of each model.
* Discuss the key insights gained during the experimentation phases.

Instructions: Present the results of the model evaluation, including accuracy, precision, recall, F1-score, or any other relevant metrics. Analyze and compare the performance of each model, highlighting the key insights gained during the experimentation phases. Discuss the implications of these insights on the project's goals and potential areas for further improvement.

## Business Impact and Benefits

* Assess the impact and benefits of the final model on the business use cases.
* Discuss how the model contributes to solving the identified challenges or exploiting opportunities.
* Quantify the improvements achieved and the potential value generated.

Instructions: Assess and discuss the impact and benefits of the final model on the identified business use cases. Explain how the model contributes to solving the identified challenges or exploiting opportunities. Quantify the improvements achieved and discuss the potential value generated by the model.

short dash

# Deployment

* Explain the process of deploying the trained model.
* Discuss any integration steps or considerations for real-world implementation.
* Address any challenges or considerations related to deployment.

Instructions: Explain the process of deploying the trained model, including any integration steps or considerations for real-world implementation. Discuss any challenges or considerations related to the deployment process and provide recommendations or suggestions for future deployment efforts.

short dash

# Conclusion

* Summarize the key findings, insights, and outcomes of the project.
* Reflect on the project's success in achieving its goals and meeting stakeholders' requirements.
* Discuss any future work, recommendations, or next steps based on the project's outcomes.

Instructions: Summarize the key findings, insights, and outcomes of the project. Reflect on the project's success in achieving its goals and meeting stakeholders' requirements. Discuss any future work, recommendations, or next steps based on the project's outcomes.

short dash

# References

So, A., & Tith, R. (2023, August 8). Machine Learning Engineering [Lab]. <https://docs.google.com/presentation/d/1zNzRPkJua_qJn5ZM0Eku5BBrk3EX12jI0-9BUeyADG0/edit#slide=id.gb753713d85_0_643>

So, A., & Tith, R. (2023, August 8). Custom Package [Lecture]. <https://canvas.uts.edu.au/courses/28052/pages/2-dot-0-module-2-overview?module_item_id=1465624>

So, A., & Tith, R. (2023, August 8). Custom Package [Lab]. <https://colab.research.google.com/drive/1HzN0kcWIFmApxJFDYQoiBZVurVZd9HJk#scrollTo=XH7B4qftdese>

Salvatierra, J. (n.d.). How to use pyenv to manage Python versions. Teclado. Retrieved August 21, 2023, from <https://blog.teclado.com/how-to-use-pyenv-manage-python-versions/>

Rigoulet, X. (n.d.). Your Guide to pyenv. LearnPython. Retrieved August 21, 2023, from <https://learnpython.com/blog/change-python-versions/>

Krishna, A. (n.d.). Your Guide to pyenvHow to Build and Publish Python Packages With Poetry. Freecodecamp. Retrieved August 25, 2023, from <https://www.freecodecamp.org/news/how-to-build-and-publish-python-packages-with-poetry/>

(n.d.). Packaging Python Projects. PyPA. Retrieved August 27, 2023, from <https://packaging.python.org/tutorials/packaging-projects/>

(n.d.). Create and Access a Python Package. Tutorialspoint. Retrieved August 26, 2023, from <https://www.tutorialspoint.com/create-and-access-a-python-package>

Brownlee, J. (2021, May 1). How to Develop an AdaBoost Ensemble in Python. Machine Learning Mastery. Retrieved August 27, 2023, from <https://machinelearningmastery.com/adaboost-ensemble-in-python/>

Polzer, D. (2022, August 4). AdaBoost, Step-by-Step. Towards Data Science. Retrieved August 29, 2023, from <https://towardsdatascience.com/adaboost-in-7-simple-steps-a89dc41ec4>