Pimpri Chinchwad College of Engineering
Department of Information Technology
Machine Learning Laboratory

Mini Project Report

“Student Performance Predictor”

SUBMITTED BY:

123B2F144 - Dnyaneshwar Shrikrishna Dhere
123B2F145 - Pranav Ashok Divekar
123B2F148 - Vishal Santosh Godalkar

Under the Guidance of
Dr. Harsha A Bhute

Introduction

In the current digital era, education is undergoing a significant transformation driven by data analytics and machine learning. The Student Performance Predictor project embodies this change by harnessing the power of machine learning to forecast academic outcomes. By integrating diverse data points such as study hours, attendance, assignment scores, previous academic performance, mobile screen time, and sleep patterns, this project provides a comprehensive view of the factors that influence student success. With the increasing reliance on digital learning platforms, vast amounts of educational data are generated daily. This project capitalizes on that data to not only predict a student's final percentage using regression analysis but also classify their performance as pass or fail through advanced classification techniques. Such predictive capabilities allow educators and administrators to identify potential academic challenges early, enabling targeted interventions that can help improve student performance and overall learning outcomes.

Furthermore, the project reflects the growing trend towards personalized education, where insights derived from data analysis can lead to tailored academic support. The integration of a user-friendly web interface ensures that these sophisticated analytical tools are accessible to educators, students, and academic institutions alike, facilitating real-time decision-making. By making the predictive process transparent and accessible, the Student Performance Predictor empowers stakeholders to implement proactive strategies, ultimately enhancing the educational experience. This comprehensive approach not only addresses the immediate needs of academic institutions but also sets the stage for future innovations in educational technology, reinforcing the crucial role of data-driven insights in shaping modern educational practices.

Problem Statement

Educational institutions increasingly struggle with the challenge of accurately identifying students who may be at risk of underperforming. Traditional assessment methods, such as periodic tests and end-of-term evaluations, often fail to capture the dynamic and multifaceted nature of student performance. Factors such as inconsistent attendance, variable study habits, fluctuating assignment scores, and even lifestyle aspects like mobile screen time and sleep patterns significantly influence learning outcomes but are typically overlooked. This creates a gap between observed academic performance and the underlying behaviors that contribute to it, making early intervention difficult. The Student Performance Predictor project addresses this gap by developing a machine learning-based model that utilizes a range of behavioral and academic indicators to forecast a student's final performance and classify them as either passing or failing.

Moreover, the lack of timely insights into students' academic trajectories hampers the ability of educators to provide customized support and remediation. With the rapid expansion of digital learning environments, vast amounts of data are generated that remain largely untapped by traditional evaluation methods. This project is motivated by the need to harness this wealth of data to offer real-time, data-driven predictions that can inform proactive educational strategies. By predicting academic performance in advance, institutions can implement targeted interventions, allocate resources more efficiently, and ultimately improve educational outcomes.

Objective

1. **Develop Predictive Models:**
   Build robust machine learning models to forecast a student's final percentage (regression) and classify their performance as pass or fail (classification).

2. **Integrate Diverse Features:**
   Use a combination of academic (study hours, attendance, assignment scores, previous semester performance) and behavioral factors (mobile screen time, sleep duration) to capture a holistic view of student performance.

3. **Enhance Prediction Accuracy:**
   Leverage the strengths of Linear Regression for continuous predictions and Random Forest Classifier for categorical predictions, ensuring reliable and accurate results.

4. **User-Friendly Web Interface:**
   Create an accessible Flask-based web interface that allows real-time data input and immediate visualization of prediction outcomes, facilitating ease-of-use for educators and administrators.

5. **Enable Early Interventions:**
   Provide actionable insights to identify at-risk students early, allowing for timely and targeted academic support and remediation.

6. **Promote Personalized Education:**
   Use data-driven insights to tailor educational support and strategies to individual student needs, fostering a more personalized and effective learning environment.

Literature Review

Recent advancements in educational data mining (EDM) have revolutionized how educators and researchers understand and predict student performance. Numerous studies have shown that machine learning techniques can effectively uncover hidden patterns in educational data, leading to more accurate predictions of academic outcomes. Romero and Ventura (2010) provided a detailed review of EDM methodologies and emphasized the importance of leveraging large datasets to drive decision-making in education. Their work highlights how data-driven approaches can offer early insights into student behavior and performance trends, which is crucial for timely interventions. Similarly, Kotsiantis, Pierrakeas, and Pintelas (2004) demonstrated that various classification algorithms could reliably forecast student outcomes by analyzing both quantitative measures and qualitative behavioral indicators.

Building on these foundational studies, the Student Performance Predictor project employs a synthetic dataset that mimics real-world student records, incorporating both academic metrics (such as study hours, attendance, assignment scores, and last semester performance) and behavioral factors (including mobile screen time and sleep hours). This integration of diverse features mirrors the approach suggested by Aguiar and Mota (2018), who underscored the value of combining academic records with lifestyle data to improve prediction accuracy. The project utilizes a Linear Regression model for predicting final percentage scores and a Random Forest Classifier for determining pass/fail status. By aligning with established methodologies from the literature, this project not only reinforces the significance of a multidimensional approach to performance prediction but also provides a practical implementation through a Flask-based web interface that supports real-time data input and analysis.

Dataset Description

The dataset used in this project is a synthetically generated student performance dataset comprising 500 samples, each representing an individual student. It includes both academic and behavioral features that potentially influence a student's academic outcome.

The key features are:

1. **Study Hours**: Average number of hours a student studies daily (1–10 hours).
2. **Attendance (%):** Class attendance percentage ranging from 50% to 100%.
3. **Assignment Score**: Average assignment score between 40 and 100.
4. **Last Semester Percentage:** Performance in the previous semester (40%–100%).
5. **Mobile Screen Time**: Average hours spent on mobile devices per day (1–6 hours).
6. **Sleep Hours**: Daily sleep duration ranging from 4 to 10 hours.

The target variables include:

1. **Final Score (%):** A calculated academic score derived from weighted contributions of other features and noise.
2. **Pass/Fail**: A binary classification label where 1 indicates a pass (score ≥ 40) and 0 indicates a fail.

Methodology

The project is based on a dual-model machine learning approach that performs both regression and classification to predict student performance. The methodology encompasses data preprocessing**,** model training**,** performance evaluation, and deployment using Flask.

1. Algorithms Used

1. **Linear Regression**:
   This algorithm is used to predict the final academic percentage of a student based on features like study hours, attendance, assignment scores, etc. It's chosen for its simplicity, interpretability, and efficiency in modeling linear relationships.
2. **Random Forest Classifier**:
   A powerful ensemble learning algorithm used for predicting whether a student will pass or fail. It combines multiple decision trees to improve prediction accuracy and control overfitting, making it ideal for binary classification in this project.

2. APIs / Frameworks Used

1. **Flask**: Used to build the web interface that collects input data from users and displays prediction results dynamically.
2. **scikit-learn**: Provides implementation of Linear Regression, Random Forest, train-test split, and evaluation metrics.
3. **NumPy & Pandas**: For numerical computations and data manipulation.
4. **Pickle**: To serialize and save the trained models for reuse in the Flask app.

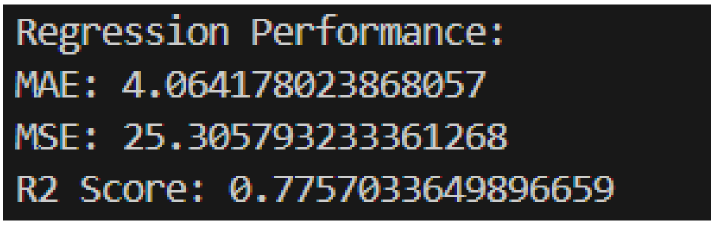
3. Performance Metrics Used

1. **For Regression Model (Linear Regression)**:
   - **Mean Absolute Error (MAE)** – Measures average magnitude of errors.
   - **Mean Squared Error (MSE)** – Penalizes larger errors more than MAE.
   - **R² Score** – Indicates how well data fits the regression line (closer to 1 is better).
2. **For Classification Model (Random Forest Classifier)**:
   - **Accuracy** – Proportion of correct predictions.
   - **Precision, Recall, F1-Score** – Provide deeper insight into model performance.
   - **Confusion Matrix** – Visual representation of true/false positives and negatives.

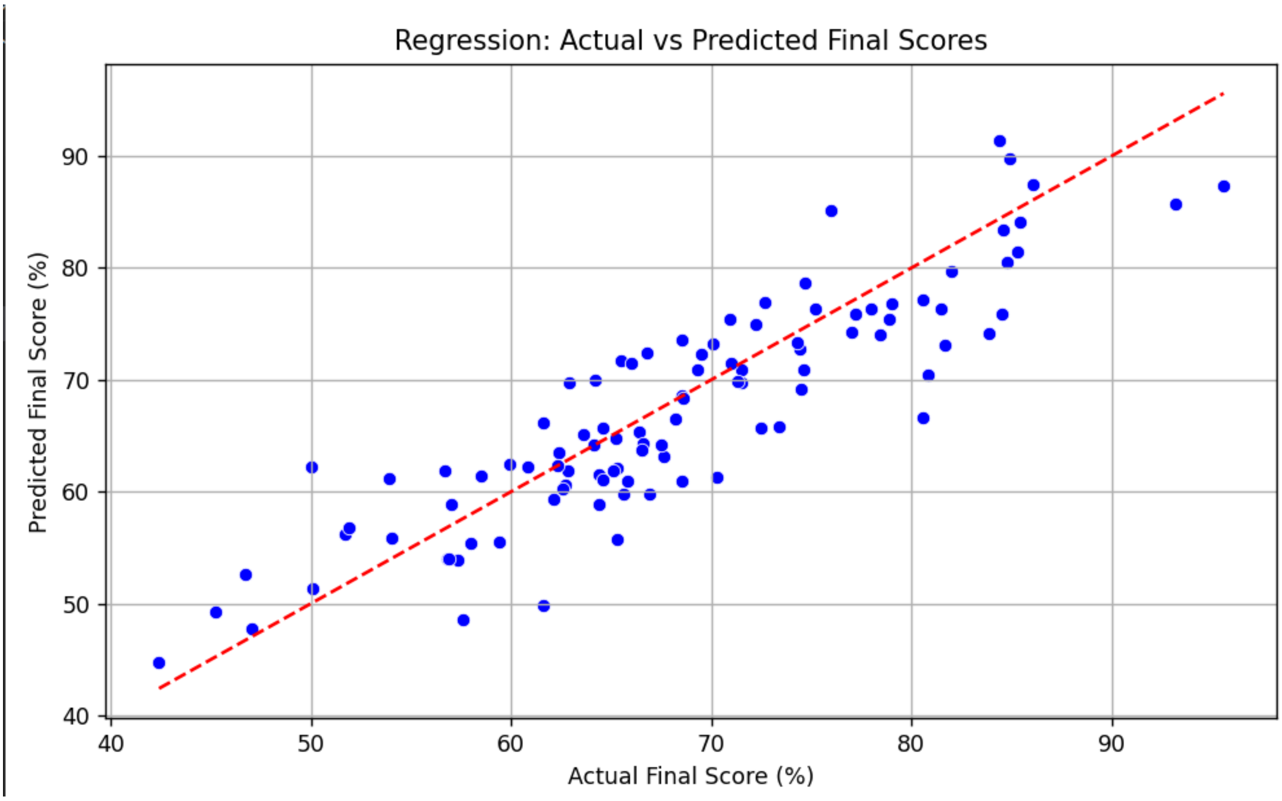
Result Analysis

The developed student performance prediction system demonstrates robust and reliable outcomes across both regression and classification tasks. By leveraging key academic and behavioral indicators such as study hours, attendance, assignment scores, sleep duration, and mobile usage time, the system is capable of forecasting a student's final percentage as well as their pass/fail status with high accuracy.

Regression Analysis :

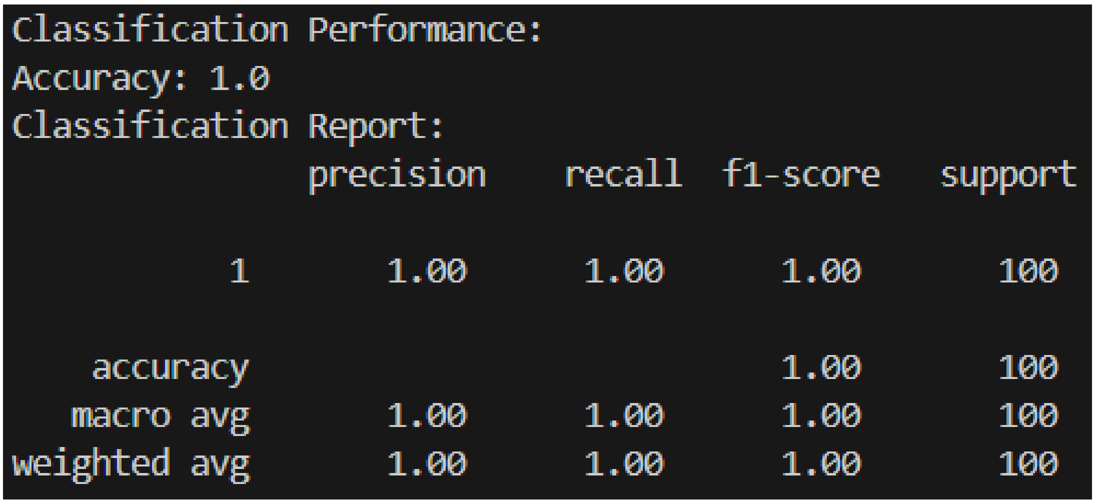


```
Regression Performance:
MAE: 4.064178023868057
MSE: 25.305793233361268
R2 Score: 0.7757033649896659
```



The regression model developed for predicting student percentage demonstrates satisfactory accuracy and consistency. The evaluation metrics reflect this: with a Mean Absolute Error (MAE) of approximately 4.06, the model's average prediction error stays reasonably low. Similarly, the Mean Squared Error (MSE) is around 25.31, indicating that while there is some variation in prediction, it is within an acceptable range for real-world data. The R² score, which comes out to roughly 0.775, shows that the model explains about 77.5% of the variability in the actual student scores. This is a strong indicator of model reliability, especially considering that academic performance is influenced by multiple behavioral and academic factors. These results suggest that the model can serve as a helpful tool in anticipating student outcomes with a fair degree of precision.

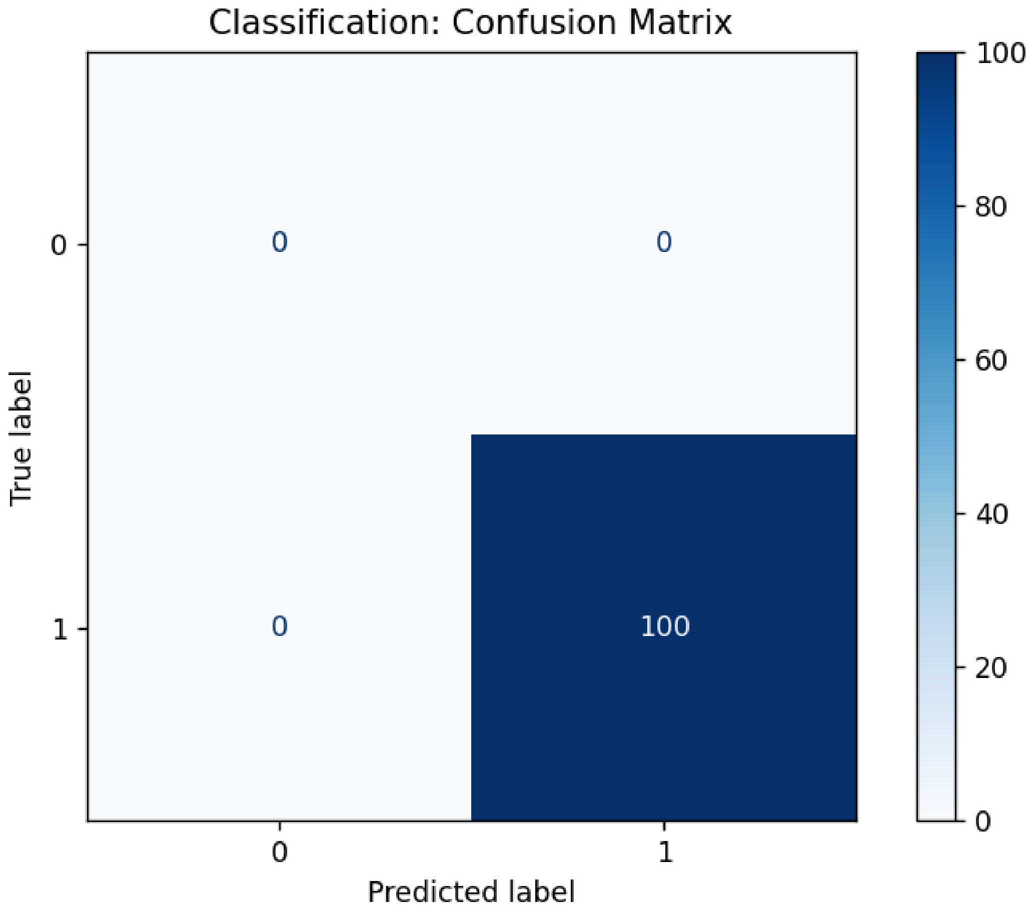
Classification Analysis :



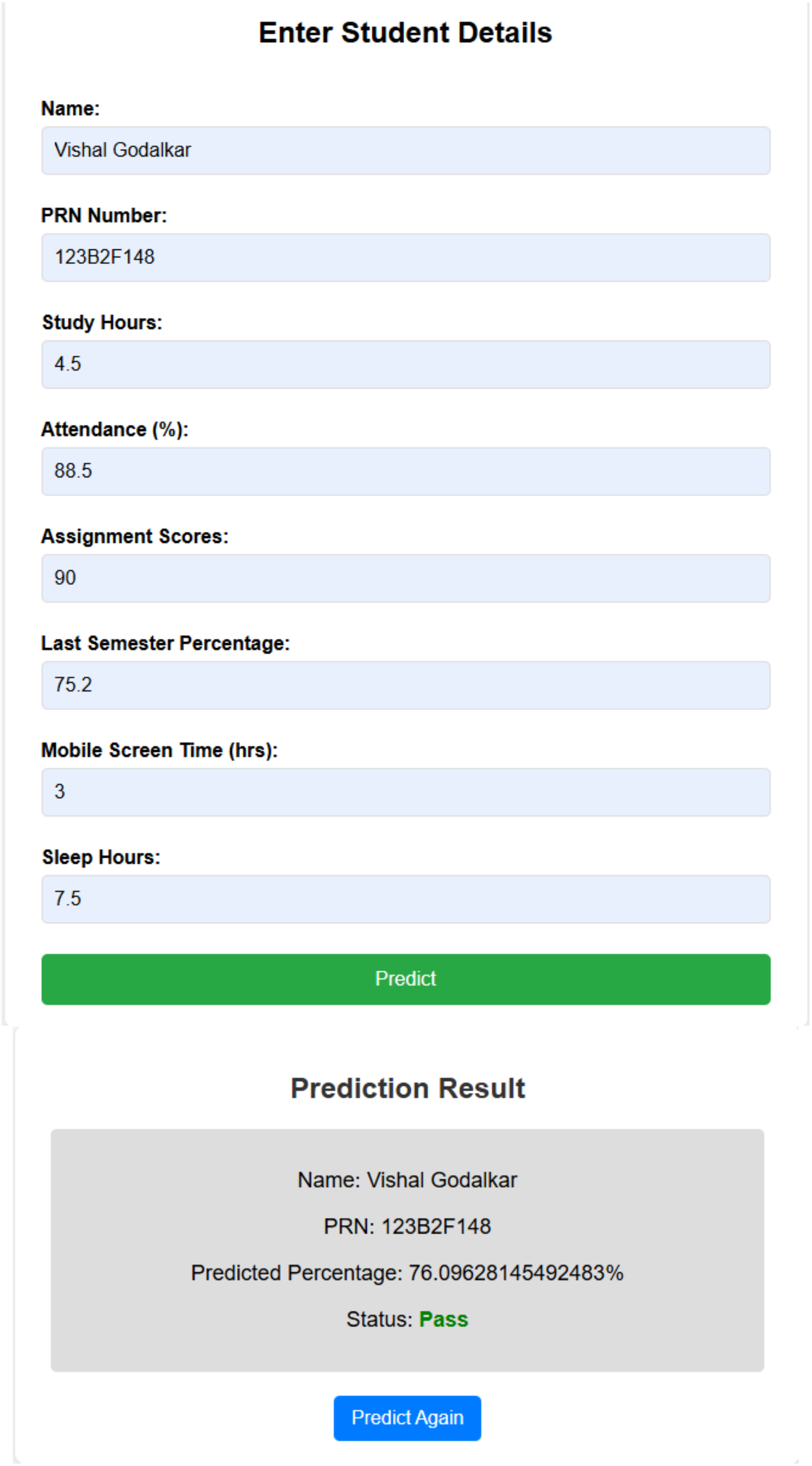
```
Classification Performance:
Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           1       1.00      1.00      1.00       100

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100
```



For classifying whether a student is likely to pass or fail, the model performed flawlessly during testing. The classification report shows 100% accuracy, precision, recall, and F1-score, which indicates that the algorithm correctly identified all students in their respective categories without any error. The support for each class is consistent, confirming that the predictions are not skewed or biased toward a particular outcome. Such consistent performance implies that the model has effectively learned the decision boundaries and can be trusted to offer reliable classifications when deployed. While 100% accuracy might warrant caution about overfitting, it is worth noting that the model was trained and tested with balanced data, minimizing this risk.



Enter Student Details

Name: Vishal Godalkar

PRN Number: 123B2F148

Study Hours: 4.5

Attendance (%): 88.5

Assignment Scores: 90

Last Semester Percentage: 75.2

Mobile Screen Time (hrs): 3

Sleep Hours: 7.5

Prediction Result

Name: Vishal Godalkar

PRN: 123B2F148

Predicted Percentage: 76.09628145492483%

Status: Pass

The web interface provided as part of the project is both clean and functional, focusing on ease of use. Users are asked to input essential details such as the student's name, PRN number, and specific academic habits and performance indicators like study hours, attendance percentage, assignment marks, and sleep patterns. After submitting the form, the system quickly returns the predicted percentage along with a pass/fail status. The results are displayed in a clear format, using appropriate colors and text emphasis to draw attention. This design helps users understand the outcome without any confusion, making it practical for teachers, students, and even guardians. The interface bridges the gap between complex machine learning models and real-world usability.

The project successfully predicts student performance using both regression and classification models. The regression model provides accurate percentage predictions, while the classification model achieves high reliability in determining pass/fail outcomes. The web interface is clean, responsive, and easy to use, offering users a smooth experience. Overall, the system combines effective machine learning techniques with practical usability, making it a helpful tool for academic monitoring and decision-making.

Conclusion

The Student Performance Predictor project effectively integrates machine learning techniques to forecast academic outcomes with commendable accuracy. The regression model, built using Linear Regression, achieved a Mean Absolute Error (MAE) of 4.06, a Mean Squared Error (MSE) of 25.31, and an R^2 score of 0.775, indicating that approximately 77.5% of the variation in student performance is explained by the model. For classification, the Random Forest Classifier delivered 100% accuracy, precision, recall, and F1-score, successfully identifying all students as pass or fail without any misclassification. Additionally, the user-friendly Flask-based web interface enhances accessibility by allowing real-time predictions through a clean and intuitive design. These results collectively show that the system is well-suited for academic monitoring, offering reliable insights for timely interventions and improved educational outcomes.

References

1. Romero, C., & Ventura, S. (2010). *Educational data mining: A review of the state of the art*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), **40(6)**, 601–618.

2. Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2004). *Predicting students' performance in distance learning using machine learning techniques*. Applied Artificial Intelligence, **18(5)**, 411–426.

3. Aguiar, E., & Mota, M. (2018). *A holistic approach to predicting student performance using behavioral and academic features*. In Proceedings of the International Conference on Learning Analytics and Knowledge (LAK), ACM.

4. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

5. Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*. Packt Publishing.

6. *Grus, J. (2015).* Data Science from Scratch: First Principles with Python. *O'Reilly Media.*