# EMPLOYEE ATTRITION PREDICTION USING LOGISTIC REGRESSION AND SELECTKBEST: A CASE STUDY ON THE IBM HR ANALYTICS EMPLOYEE ATTRITION & PERFORMANCE DATASET

Putra Nurhuda Makatita[1*], Gusnaldi Pramudita[1], Muhammad Roprop Al Muntaha[1], Ilham Arya Yudha[1], Rani Rahma Wulan[1]
Universitas Teknologi Bandung, Jl. Soekarno-Hatta No.378, Bandung, Indonesia
*Corresponding Author: putramakatita723@gmail.com

**ABSTRACTS**

Employee attrition is a significant strategic concern for organizations as it directly impacts overall performance, productivity, and long-term sustainability. High attrition rates can lead to increased costs in recruitment and training, a loss of skilled and experienced employees, decreased morale among remaining staff, and disruptions to critical business operations. In response to these challenges, many organizations are turning to predictive analytics to anticipate employee turnover and implement effective retention strategies. This study proposes a machine learning-based approach to predict employee attrition using the Logistic Regression algorithm. Logistic Regression is chosen due to its effectiveness in binary classification tasks and its interpretability, which is essential for human resource (HR) professionals when making data-driven decisions. To enhance the model's performance, the SelectKBest feature selection technique is applied in conjunction with the ANOVA F-test. This method allows the identification of the most influential features contributing to attrition, helping reduce noise and computational complexity while improving model accuracy. The IBM HR Analytics Employee Attrition & Performance Dataset is used in this study. The dataset contains a variety of demographic and organizational attributes such as age, monthly income, job role, tenure, and job satisfaction. The data undergoes a comprehensive preprocessing phase that includes numerical transformation, encoding of categorical variables, normalization, and the implementation of feature selection. By combining Logistic Regression with effective feature selection, this research aims to deliver an accurate and interpretable predictive model. The results are expected to help HR departments proactively identify high-risk employees and take strategic actions to reduce attrition, ultimately supporting better workforce planning and organizational stability.

*Keywords: Machine learning; Logistic Regression; SelectKBest; ANOVA F-test.*

**INTRODUCTION**

Employee attrition, a phenomenon commonly observed in organizational dynamics, refers to the condition in which an individual is no longer part of a company's workforce, whether due to personal decisions such as resignation or organizational decisions such as termination or retirement. While attrition is considered a natural aspect of a company's lifecycle to a certain extent, an uncontrolled or excessively high attrition rate can lead to a series of significant negative consequences. These include a decline in overall team productivity, reduced operational efficiency, and a substantial increase in costs associated with recruiting and training new employees as replacements (Hom, Griffeth, and Williams 2017a).

Beyond financial costs, high and unmanaged attrition has the potential to erode internal organizational stability, foster uncertainty, and may even signal deeper issues within the workplace culture—such as low employee satisfaction, lack of career development opportunities, or an unsupportive work environment (Gerhart and Rynes 2003). In today's increasingly competitive business environment, where talent is a critical asset, the ability to anticipate and mitigate attrition risks has become fundamental. As such, proactive approaches in human resource management, driven by data and analytics, are becoming increasingly vital to ensuring long-term organizational sustainability and success.

Employee attrition is more than just a statistical figure—it is a reflection of the complex interplay between individuals and their work environment. Understanding the driving factors behind attrition, both internal and external, allows organizations to design targeted and strategic interventions(Lee and Mitchell 1994). Without effective understanding and management, companies risk losing valuable talent, institutional knowledge, and innovative momentum, all of which may hinder growth and competitiveness.

This study aims to address these challenges by focusing on the development of a robust and accurate classification model. We seek to build a predictive model based on Logistic Regression, specifically designed to project the likelihood of an employee leaving the organization. The model leverages comprehensive historical data, enhanced by the implementation of an advanced feature selection technique based on univariate statistical testing. This approach is expected not only to improve the model's predictive accuracy but also to ensure efficiency in identifying the most influential variables that drive employee attrition decisions.

**METHODS**

In general, methodology in data science refers to a structured approach encompassing data collection, preprocessing, model building, and evaluation (Kelleher, Namee, and D'Arcy 2015). Data preprocessing is essential for ensuring the quality and usability of data, while feature selection helps reduce dimensionality and improve model performance (Guyon and Elisseeff 2003). Logistic Regression(Mining, n.d.), commonly used for binary classification, operates by modeling the probability that a given input belongs to a particular category (Hosmer, Lemeshow, and Sturdivant 2013). Evaluation metrics such as precision, recall, and F1-score are crucial in assessing the robustness of classification models (Sammut and Webb 2011). Tools like Jupyter Notebook enhance reproducibility and accessibility of machine learning models (Kluyver et al. 2016).

The methodology of this study consists of several key stages. First, data preprocessing is conducted, including data cleaning and the transformation of categorical variables into numerical form. Next, feature selection is

performed using a univariate statistical test, specifically the ANOVA F-test through the SelectKBest function, to identify the input variables that are statistically most significant in relation to the target variable. Once the most relevant features are identified, a classification model is built using the Logistic Regression algorithm with class-balancing parameters. The model is then evaluated using performance metrics such as the confusion matrix, precision, recall, and F1-score to assess its classification accuracy and effectiveness(Breiman et al. 1984).

Finally, the trained model is integrated into an interface developed in Jupyter Notebook, allowing users to upload new data and automatically obtain prediction results in CSV or Excel format. To clarify the steps involved in the development of the attrition prediction model, a flowchart is presented in Figure 1, illustrating the main processes from data preprocessing to the implementation of the prediction system via an interactive interface.

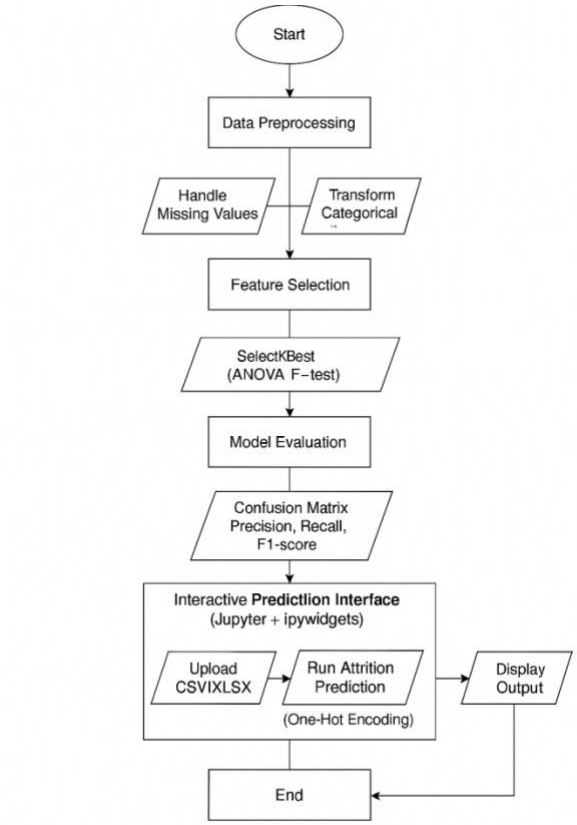


**Figure 1.** Flowchart of Employee Attrition Prediction Model Development Process.

**Dataset**

The dataset used in this study is the IBM HR Analytics Employee Attrition & Performance Dataset (IBM 2017), which consists of over 1,400 rows of data and includes various demographic and professional attributes of employees. Some of the available variables include age, monthly income, total years of employment, job satisfaction, and attrition status. This dataset is widely used in human resource analytics studies because it reflects real organizational conditions and provides relevant variables for predictive modeling purposes(Han, Kamber, and Pei 2011).

**Preprocessing**

In the initial stage, the target variable "Attrition", which was originally categorical ("Yes" and "No"), was transformed. To support classification algorithms such as Logistic Regression, this variable was converted into a binary numeric format, where employees who left the company were labeled as 1, and those who stayed were labeled as 0. The result of this conversion was stored in a new column named LeftCompany. This process aims to ensure data compatibility with the classification model and facilitate the evaluation of model performance using numerical metrics such as precision, recall, and F1-score (Raschka and Mirjalili 2019).

**SelectKBest Method**

In this study, a univariate feature selection approach was employed using the SelectKBest function from the scikit-learn library (Mozaffari et al. 2017). This technique evaluates the relationship between each independent feature and the target variable individually. The function used is f_classif, which performs the ANOVA F-test, a statistical test that measures the strength of the linear relationship between numerical features and a categorical target variable(Hom, Griffeth, and Williams 2017b).

The original dataset contains 17 features, including the following: Age, DistanceFromHome, MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager, YearsSinceLastPromotion, JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction, WorkLifeBalance, JobInvolvement, OverTime, BusinessTravel, MaritalStatus and StockOptionLevel. The dataset containing these features was then evaluated using the SelectKBest method. The scores resulting from this test can be observed in Figure 2.

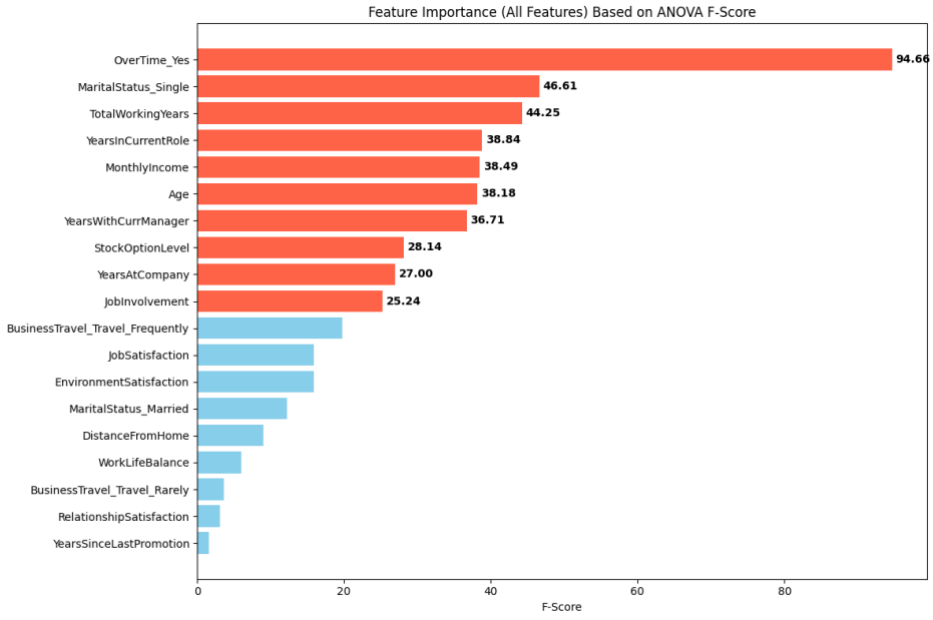


**Figure 2.** F-score diagram of all features based on ANOVA F-score.

Based on the test results, the system selected the 10 best features with the highest F-statistic values, indicating the significance of their contribution to the target variable. The 10 selected features are as follows: OverTime_Yes, MaritalStatus_Single, TotalWorkingYears , YearsInCurrentRole, MonthlyIncome, Age, YearsWithCurrManager, StockOptionLevel, YearsAtCompany, JobInvolvement. These features were then used in the model training process to ensure optimal prediction efficiency and accuracy.

**Logistic Regression Modeling**

The model was developed using the Logistic Regression algorithm, a commonly used classification approach for binary cases, such as predicting employee attrition status (leaving or staying) (Bishop 2006). This algorithm was chosen due to its simplicity, computational efficiency, and its ability to produce models that are easily interpretable by non-technical users. The parameter class_weight='balanced' was applied to handle class imbalance in the target data, while max_iter=1000 was set to ensure convergence during the optimization process(Hosmer, Lemeshow, and Sturdivant 2013).

**Evaluation**

The model was evaluated using a data split scheme of 70% for training and 30% for testing, in order to measure the model's generalization to unseen data. The model's performance was analyzed using several key classification metrics, namely:

**Confusion Matrix**

This illustrates the number of correct and incorrect predictions for each class. A detailed report of the confusion matrix is presented.

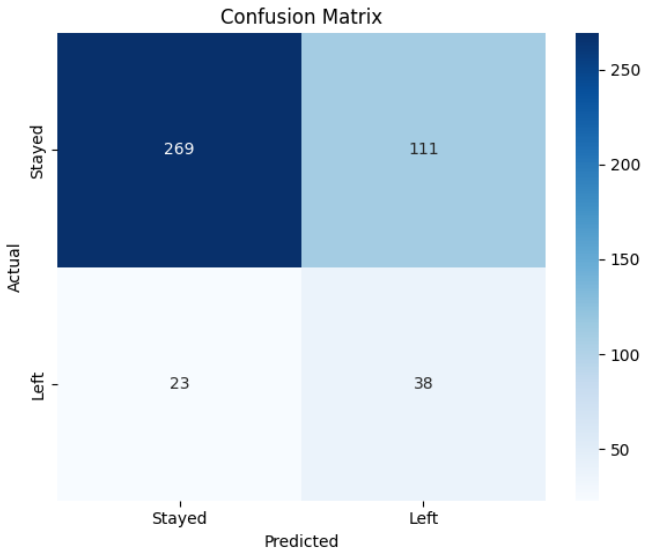


**Figure 3.** Confusion Matrix for evaluating model performance.

From Figure 3, the evaluation results can be summarized as follows: True Negative (TN) = 269, False Positive (FP) = 111, False Negative (FN) = 23, True Positive (TP) = 38. This evaluation establishes that the employee stayed, and the model correctly predicted "Stay". The employee stayed, but the model incorrectly predicted "Resign". The employee resigned, but the model incorrectly predicted "Stay". The employee resigned, and the model correctly predicted "Resign".

**Classification Report**

This evaluation aims to ensure that the model is not only accurate, but also capable of consistently detecting attrition risk and being reliable in an operational context.

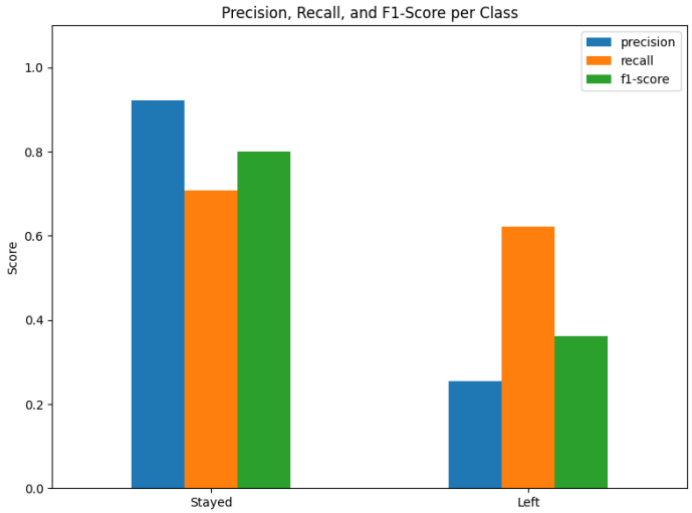


**Figure 4.** Diagram of Classification Report from Precision, Recall and F1-Score.

From Figure 4, the following detailed explanation is obtained precision, measures the proportion of correct positive predictions. Stayed: Precision is very high (~0.92), meaning the model is highly accurate when predicting someone will stay. Left: Precision is low (~0.25), meaning many employees who actually did not resign were predicted to resign (many False Positives). Recall, measures the proportion of actual positives correctly identified. Stayed: Recall is around ~0.70, meaning 70% of employees who actually stayed were successfully identified by the model. Left: Recall is better (~0.62), indicating that the model is able to capture the majority of employees who actually resigned (even though precision is poor). F1-Score, the harmonic mean of precision and recall, providing a balanced measure of model performance. Stayed: F1-score is quite high (~0.80), indicating balanced performance. Left: F1-score is low (~0.36), due to very low precision despite a fairly good recall.

## RESULTS AND DISCUSSION

The trained predictive model was saved as a file using the joblib library, allowing it to be reloaded without the need for retraining. This supports computational time efficiency and facilitates integration into a sustainable prediction system.

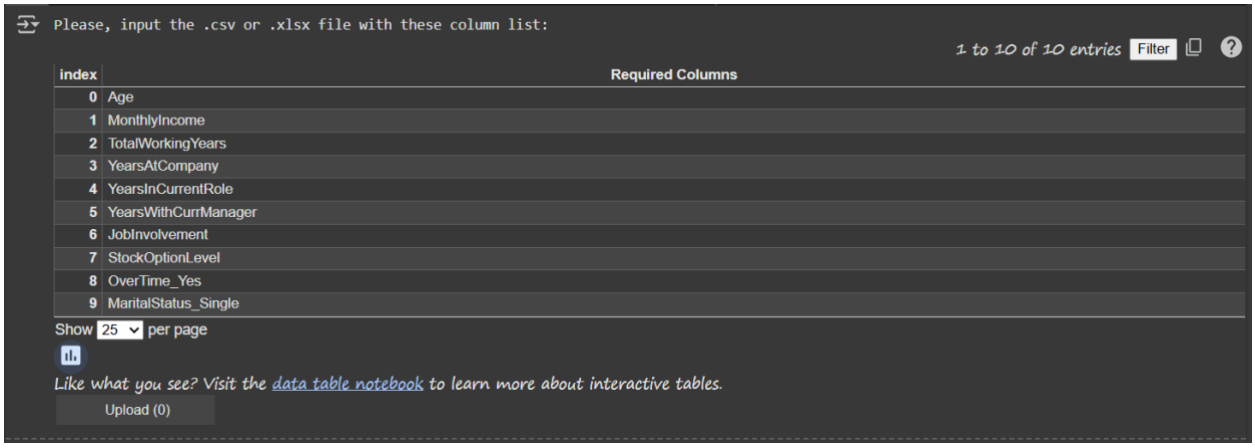


**Figure 5.** Interactive interface using ipywidgets.

For practical implementation, Figure 5 shows the interface display developed as an interactive interface based on Jupyter Notebook using the ipywidgets module (Team, n.d.). This interface allows users to upload new data

in CSV or XLSX format, which is then automatically processed to match the structure of the model. This adjustment process includes applying one-hot encoding to categorical variables and aligning dummy columns to match the features used during initial training (Géron 2019).

Once the transformation process is complete, the system performs predictions and displays the results immediately, both in numeric and descriptive labels. With this approach, the model can be used by non-technical users such as HR staff to quickly evaluate the potential risk of attrition in both new and existing employees. The prediction results are displayed in the following table.

**Table 1.** Prediction result.

| No | Prediction | Age | Monthly Income | Total Working Years | Years at Company | Years in Current Role | Years With Current Manager | Job Involvement | Stock Option Level |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Resign (1) | 41 | 5,993 | 8 | 6 | 4 | 5 | 3 | 0 |
| 2 | Tetap (0) | 49 | 5,130 | 10 | 10 | 7 | 7 | 2 | 1 |
| 3 | Resign (1) | 37 | 2,090 | 7 | 0 | 0 | 0 | 2 | 0 |
| 4 | Resign (1) | 33 | 2,909 | 8 | 8 | 7 | 0 | 3 | 0 |
| 5 | Tetap (0) | 27 | 3,468 | 6 | 2 | 2 | 2 | 3 | 1 |

Based on the prediction results, it can be seen that employees who are younger, have shorter tenure, and low job satisfaction levels tend to be predicted as likely to leave. Conversely, employees with higher salaries, longer tenure, and better satisfaction and work-life balance are generally predicted to stay with the company. Overall, the model demonstrates adequate performance in detecting potential employee attrition. Based on evaluation metrics such as precision, recall, and F1-score, the model achieves a good balance between the ability to identify employees likely to leave and minimizing prediction errors.

Furthermore, the implementation of this model in an interactive interface based on Jupyter Notebook adds value in terms of accessibility. The interface enables non-technical HR staff to upload data and directly obtain prediction results without needing to understand the complex technical processes. This makes the system a strategic tool for proactively monitoring, predicting, and mitigating the risk of employee loss. However, the use of this interface still requires running Python code in Jupyter Notebook, which constitutes one of the limitations of this study (Krishna and Sidharth 2022).

In addition, the study is limited by the absence of a comparative model or the use of alternative models. The research is restricted to the application of logistic regression and feature selection using SelectKBest. This is because the combination of these two techniques was considered sufficient for the modeling needs. The

combination of several techniques to achieve high accuracy has previously been explored in earlier studies, such as the use of the gradient boosting machine algorithm, which achieved an accuracy rate of up to 89(Mozaffari et al. 2023).

**CONCLUSION**

This study demonstrates that the employee attrition prediction model based on Logistic Regression, supported by feature selection using SelectKBest with the ANOVA F-test, is capable of effectively and efficiently identifying the risk of employee resignation. The model not only shows good classification performance based on standard evaluation metrics but also has been successfully implemented as an interactive interface that is easy to use by non-technical stakeholders in the field of human resources. The integration of the model into a Jupyter Notebook-based predictive system strengthens its practical relevance in organizational contexts, particularly in supporting strategic decision-making related to employee retention. As a future development, this model has the potential to be enhanced through ensemble learning approaches, the addition of behavior-based features, and integration with HRIS (Human Resource Information System) for real-time continuous analysis. Although the current interface is still manual and simple—albeit user-friendly for non-technical users—users still need to run code in Jupyter Notebook. It would certainly be better if, in the future, the program could be developed with a more interactive interface using other software to make it even easier to use.

**REFERENCES**

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Breiman, Leo, Jerome H Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and Regression Trees*. Chapman \& Hall.

Gerhart, Barry, and Sara L Rynes. 2003. *Compensation, Organizations, and Managerial Performance: An Integrative Approach*. McGraw-Hill Education. https://doi.org/10.4135/9781452229256.

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

Guyon, Isabelle, and André Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3:1157–82.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann.

Hom, Peter W, Rodger W Griffeth, and Leigh R Williams. 2017a. *Turnover and Retention: Current Research and Future Directions*. Cambridge University Press.

Hom, Peter W, Rodger W Griffeth, and Lisa M Williams. 2017b. "Losing Your Best: Turnover Intentions and Employee Performance." *Academy of Management Journal* 60 (1): 150–74.

Hosmer, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Wiley.

IBM. 2017. "IBM HR Analytics Employee Attrition & Performance."

Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. 2015. *Fundamentals of Machine Learning for Predictive Data Analytics*. *Igarss 2014*.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Jonathan Bussonnier, J Frederic, and Chris Willing. 2016. "Jupyter Notebooks – a Publishing Format for Reproducible Computational Workflows." In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87–90. IOS Press.

Krishna, Shobhanam, and Sumati Sidharth. 2022. "HR Analytics: Employee Attrition Analysis Using Random Forest." *International Journal of Performability Engineering* 18 (4): 275–81. https://doi.org/10.23940/ijpe.22.04.p5.275281.

Lee, Thomas W, and Terence R Mitchell. 1994. "An Alternative Approach: The Unfolding Model of Voluntary Employee Turnover." *Academy of Management Review* 19 (1): 51–89. https://doi.org/10.2307/258835.

Mining, Tanagra Data. n.d. "Logistic Regression Module."

Mozaffari, Fatemeh, Marzieh Rahimi, Hamidreza Yazdani, and Babak Sohrabi. 2023. "Employee Attrition Prediction in a Pharmaceutical Company Using Both Machine Learning Approach and Qualitative Data." *Benchmarking* 30 (10): 4140–73. https://doi.org/10.1108/BIJ-11-2021-0664.

Mozaffari, Fatemeh, Marzieh Rahimi, Hamidreza Yazdani, Babak Sohrabi, Aurélien Géron, Jupyter Development Team, Christopher M Bishop, et al. 2017. *1.13. Feature Selection*. *Academy of Management Review*. 3rd ed. Vol. 19. Cambridge University Press. https://doi.org/10.2307/258835.

Raschka, Sebastian, and Vahid Mirjalili. 2019. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2*. 3rd ed. Packt Publishing.

Sammut, Claude, and Geoffrey I Webb. 2011. *Encyclopedia of Machine Learning*. Springer.

Team, Jupyter Development. n.d. "Jupyter Widgets (Ipywidgets) Documentation.