

# Understanding and Mitigating Data Bias in Machine Learning and Data Science

Hariti Bhatia, Shivam Lahoti, Vijeth Reddy

## 1 Title and Research Question

**Title:** "Understanding and Mitigating Data Bias in Machine Learning and Data Science"

This title encapsulates the core objective of the research, which is to understand how data bias manifests in machine learning and data science, and to explore strategies for mitigating it. The focus is on uncovering the challenges posed by biased data and identifying practical solutions to ensure fairness and accuracy in model outcomes.

**Research Question:** "How does data bias impact machine learning models, and what methods can be employed to identify and mitigate selection bias effectively?"

This research question is twofold:

1. **How does data bias impact machine learning models?** The first part seeks to explore the consequences of biased data on model performance. Bias in data can cause models to make inaccurate predictions or favor certain demographic groups over others, leading to skewed outcomes. For example, a biased loan approval model might disproportionately reject applicants from lower-income or minority groups, despite them having similar financial profiles to those approved.
2. **What methods can be employed to identify and mitigate selection bias effectively?** The second part of the question aims to investigate techniques to detect and address bias in datasets. This includes both identifying sources of selection bias (e.g., certain groups being over- or under-represented in the data) and applying strategies to correct or mitigate the bias. Techniques such as resampling, propensity score matching, and fairness-aware algorithms are potential solutions to ensure the model treats all groups equitably.

**Relevance:** Data bias poses significant risks in many real-world applications of machine learning. When models are trained on biased data, they may yield outcomes that are not only inaccurate but also unethical. In fields such as healthcare, finance, and criminal justice, biased models can lead to unfair treatment of individuals, reinforcing societal inequalities. By understanding how

data bias affects machine learning models and exploring methods to mitigate it, we can develop more reliable, accurate, and fair models. This research ultimately contributes to creating ethical AI systems that promote equity across different demographic groups.

## 2 Theory and Background

### 2.1 Theoretical Foundation

#### 2.1.1 1. Overview of Bias in Data Science

- **Definition of Bias:** Bias refers to systematic errors that can skew data analysis and model predictions. It can arise from various sources, including data collection methods, the sampling process, and the inherent characteristics of the data itself.
- **Impact on Model Predictions:** Bias can severely affect model accuracy. For example, if a model is trained on data that overrepresents a particular demographic, it may not generalize well to underrepresented groups, leading to poor performance in real-world applications.
- **Generalization Issues:** Generalization is the model's ability to perform well on unseen data. Bias directly impacts this ability, as models trained on biased data often fail to recognize patterns in more diverse datasets.

### 2.2 Literature Review

#### 2.2.1 Notable Research on Data Bias

##### Key Studies:

- **Barocas and Selbst (2016)** explored how bias in algorithmic decision-making can lead to unfair outcomes, emphasizing the importance of fairness in model training.
- **Dastin (2018)** discussed how biased training data in a hiring algorithm led to the exclusion of certain demographics, raising concerns about the ethical implications of automated decision-making.

#### 2.2.2 Focus on Selection Bias

- **Selection Bias Defined:** Selection bias occurs when the data used to train a model is not representative of the overall population. This can lead to models that are ineffective or discriminatory.
- **Real-World Implications:** Research has shown that selection bias can lead to significant disparities in outcomes, such as in predictive policing algorithms that disproportionately target certain communities based on biased historical data.

## 2.3 Core Concepts

### 2.3.1 1. Definitions of Various Types of Bias

- **Sampling Error:** Sampling error refers to the difference between the sample statistics and the actual population parameters due to a non-representative sample. It is a natural occurrence in statistical analyses but can lead to misleading conclusions if not accounted for.
- **Variation:** Variation in data can arise from numerous factors, including environmental influences and changes over time. Models must be robust enough to handle this variation to remain accurate in dynamic settings.
- **Selection Bias:** This specific type of bias occurs when certain groups are systematically excluded from the sample. For example, if a medical study primarily includes participants from one demographic, its findings may not apply to the broader population.

### 2.3.2 Examples of Bias

- **Facial Recognition Systems:** A prominent example of selection bias is seen in facial recognition technology, which often performs poorly on individuals from underrepresented racial or ethnic backgrounds due to a lack of diverse training data.
- **Healthcare Algorithms:** In healthcare, predictive models that rely on historical data may propagate biases that disadvantage minority groups if those groups are underrepresented in the training data.
- **Loan Approval Systems:** In loan approval models, bias can arise when the training dataset disproportionately contains more applicants from high-income or majority demographic groups. For example, if the dataset overrepresents wealthy individuals or applicants from urban areas, the model may learn to favor these groups while disadvantaging applicants from low-income or rural backgrounds. This leads to skewed predictions, where minority groups are less likely to be approved for loans, even if they possess similar financial profiles. The lack of diverse representation in the training data can result in unfair outcomes that reinforce existing economic disparities.

## 3 Problem Statement

**Detailed Description:** Data bias occurs when certain groups or variables in a dataset are overrepresented or underrepresented, leading to skewed outcomes in data analysis or machine learning models. Two common forms of data bias are sampling bias and selection bias.

- **Sampling Bias:** Sampling bias occurs when the method of selecting data for analysis results in a dataset that is not representative of the population. This can happen if a certain group is under-sampled or over-sampled, which leads to misleading model predictions. For example, if a survey only includes responses from people in urban areas, rural perspectives are not represented, leading to biased insights.
- **Selection Bias:** Selection bias happens when the selection process used to gather the data favors certain outcomes. This could be due to the inherent characteristics of the data-collection process. For instance, if you're building a predictive model to assess job performance and only collect data from high-performing employees (leaving out low performers), the dataset won't represent the full spectrum of job performance, skewing the model toward predicting only high performance.

In machine learning, biased data can lead to biased models, which may:

- Undervalue certain features or variables.
- Produce incorrect predictions, particularly for underrepresented groups.
- Have poor generalization when applied to real-world scenarios with diverse populations.

**Input-Output Format:** To illustrate how data bias impacts model performance, we define a sample problem using a dataset that could be prone to bias. Here's a detailed example of a dataset format and expected outcomes:

- **Problem Scenario:** Suppose you're tasked with building a machine learning model to predict loan approval decisions. The dataset contains demographic information, such as age, gender, income, and education level, alongside the loan application outcomes (approved or rejected).

Input: A biased dataset containing demographic information:

- Age: A continuous variable representing the age of the applicants.
- Gender: A categorical variable, possibly biased toward males.
- Income Level: A continuous variable.
- Education Level: A categorical variable representing different levels of education.

Output:

- The under-representation of female applicants.
- The over-reliance of the model on the features related to males (gender bias).

- Suggestions for mitigating this bias, such as resampling techniques (over-sampling the minority class or applying SMOTE) or collecting more data to balance the representation of gender.

**Sample Inputs and Outputs:**

Sample Input (Biased Dataset Example):

ID	Age	Gender	Income	Employment Status	Credit Score	Loan Amount	Loan Approved
1	25	Male	50000	Full-Time	700	10000	Yes
2	35	Female	40000	Part-Time	650	5000	No
3	29	Male	55000	Full-Time	710	7000	Yes
4	40	Female	30000	Unemployed	620	4000	No
5	31	Male	60000	Full-Time	720	15000	Yes
6	45	Female	25000	Unemployed	600	3000	No
7	28	Male	52000	Full-Time	680	8000	Yes
8	32	Male	65000	Full-Time	730	12000	Yes
9	33	Female	42000	Part-Time	640	6000	No
10	38	Female	35000	Part-Time	625	4500	No

Table 1: Sample Biased Dataset for Loan Approval Prediction

In this dataset, there are more male than female entries, and the loan approval outcomes tend to favor males. The representation of female applicants is low compared to male applicants, which could lead to the model learning biased decision patterns.

**Sample Output (Bias Report):**

• **Bias Detection:**

- Gender Distribution: 80% male, 20% female.
- Loan Approval Bias: 75% approval rate for male applicants, 25% for female applicants.
- The model may predict higher loan approval for males due to the disproportionate representation.

• **Areas of Concern:**

- Gender Bias: The model might not perform well when tested on a balanced gender dataset, as it hasn't learned well from the female applicant data.
- Over-reliance on Gender: Since most approved applicants are male, the model might heavily weigh gender in its decision-making process.

• **Mitigation Suggestions:**

- Resample the data to increase the representation of female applicants.
- Use fairness-aware algorithms that balance the predictions across demographic groups.

- Monitor model outputs for fairness metrics, such as demographic parity or equal opportunity, to ensure fair treatment of all groups.

**Conclusion:**

- The biased loan approval dataset reveals significant gender and income bias, where males and higher-income individuals are more likely to get loans approved. By applying fairness-aware mitigation techniques such as resampling and fairness constraints, the dataset can be balanced, and the model can provide more equitable predictions. Ensuring fairness and reducing bias is essential to build trustworthy and accurate predictive models, especially in socially sensitive areas like loan approvals.
- This process emphasizes the importance of identifying and mitigating bias early in the model development life-cycle to prevent unfair outcomes.

## 4 Problem Analysis and Solution Explanation

### 4.1 Constraints: Dataset Limitations and Algorithmic Fairness

#### 4.1.1 Dataset Limitations

- **Incomplete Data:** Incomplete data occurs when some records, features, or observations in a dataset are missing or not fully captured. This introduces bias if certain groups are underrepresented.
- **Example in Healthcare:**
  - Ethnic minorities may not visit hospitals frequently due to access barriers or distrust, leaving gaps in patient data.
  - Rural patients may skip follow-up appointments, leading to incomplete information.
- **Impact on Model Performance:**
  - Overfitting to Majority Groups: Models trained primarily on data from overrepresented groups may struggle to generalize for underrepresented groups.
  - Skewed Predictions: Predictions become less reliable for groups missing from the dataset.
- **Solutions:**
  - Data Imputation: Use statistical methods to fill in missing values.
  - Balanced Data Collection: Collect data equally across groups using stratified sampling.

- Active Learning: Query underrepresented groups to enrich datasets during training.
- Resampling Techniques: Use oversampling (duplicating underrepresented samples) or undersampling (reducing overrepresented samples) to balance the dataset.

## 4.2 Small Sample Size for Minority Groups

In machine learning models, particularly in contexts like hiring or healthcare, the presence of imbalanced datasets—where minority groups are underrepresented—poses significant challenges.

### 4.2.1 Impact on Models:

- **Reinforcement of Bias Towards Majority Groups:** When there are fewer data points for minority groups, models tend to overfit the majority group and make biased decisions.
- **Skewed Accuracy Metrics:** A model may achieve high overall accuracy but perform poorly for minority classes.
- **Propagating Historical Inequities:** Models trained on biased data can replicate societal biases.

### 4.2.2 Solutions for Addressing Small Sample Size:

- Oversampling and Undersampling to balance minority and majority classes.
- Data Augmentation for fair representation of minority groups.
- Bias Mitigation Tools and Metrics such as Aequitas, Equalized Odds, or Demographic Parity.

## 4.3 Bias from the Data Collection Process (Historical Bias)

Historical bias arises when data used to train machine learning models reflects systemic inequalities from the past.

### 4.3.1 Causes of Historical Bias:

- Underrepresentation in Data Sources: Certain populations may be underrepresented due to past norms.
- Embedded Social Inequities: Institutional practices and policies during the data collection period often reflect societal biases.

#### 4.3.2 Impact of Historical Bias:

- Reinforcing Inequality: Predictive models trained on biased data replicate discrimination.
- False Predictive Validity: Models trained on historical data often fail in new contexts.

#### 4.3.3 Solutions:

- Auditing and Updating Datasets.
- Rebalancing and Reweighting Data.
- Fairness-Aware Algorithms.

### 4.4 4. Imbalanced Classes

#### 4.4.1 Impact of Imbalanced Classes:

- Overrepresentation of one class can lead to biased predictions.
- Example: A credit scoring model might favor affluent applicants and reject lower-income groups, reinforcing economic disparities.

#### 4.4.2 Solutions for Imbalanced Classes:

- Balanced Data Collection using stratified sampling.
- Cost-Sensitive Learning: Penalizing models for misclassifications in minority classes.

### 4.5 5. Algorithmic Fairness

#### 4.5.1 Why Fairness Matters:

- **Healthcare:** Algorithms predicting healthcare outcomes can lead to disparities in care.
- **Finance:** Biased models can reject loan applications from certain demographic groups, reinforcing economic disparities.
- **Criminal Justice:** Predictive models for recidivism may disproportionately affect minority groups, leading to biased sentencing and parole decisions.

#### 4.5.2 The Need for Fair Algorithms:

- Pre-processing to remove bias from datasets.
- Fairness-aware algorithms that adjust predictions for fairness.
- Transparency and accountability to identify and address sources of bias.



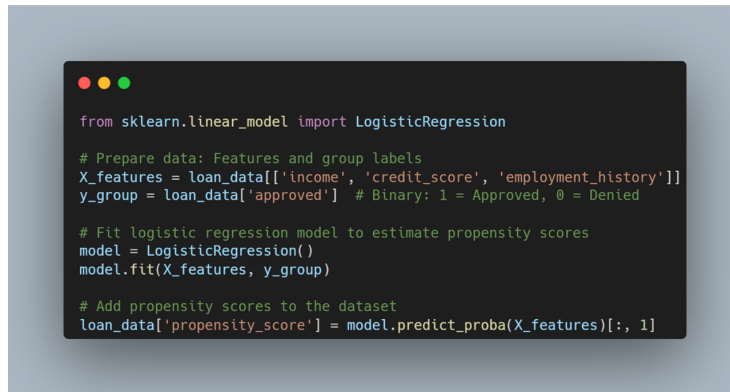
## 4.6 Logic and Approach: Detecting Bias Using Propensity Score Matching (PSM)

### 4.6.1 What is Propensity Score Matching (PSM)?

- Propensity scores represent the probability of a sample belonging to a group based on characteristics such as age, income, and education level.
- The goal is to reduce selection bias by comparing individuals from different groups who share similar characteristics.

### 4.6.2 How PSM Works in Practice:

1. **Estimate Propensity Scores:** A logistic regression model predicts the probability of an individual being in a group based on characteristics.



```
from sklearn.linear_model import LogisticRegression

# Prepare data: Features and group labels
X_features = loan_data[['income', 'credit_score', 'employment_history']]
y_group = loan_data['approved'] # Binary: 1 = Approved, 0 = Denied

# Fit logistic regression model to estimate propensity scores
model = LogisticRegression()
model.fit(X_features, y_group)

# Add propensity scores to the dataset
loan_data['propensity_score'] = model.predict_proba(X_features)[:, 1]
```

Figure 1: Python code example

2. **Match Individuals:** Matching methods such as nearest neighbor matching are used to pair individuals with similar propensity scores from different groups.
3. **Analyze Outcome Differences:** Compare outcomes between matched groups.
4. **Visualize Propensity Scores:** Visualizing propensity score distributions before and after matching helps check for balance.

## 4.7 Logical Reasoning or Proof of Correctness

### 4.7.1 Logical Reasoning:

- The core idea behind bias detection and mitigation methods, such as Propensity Score Matching (PSM), is that when groups are not directly comparable due to differing characteristics, the observed outcomes may

be skewed. By calculating propensity scores and matching individuals from the treatment and control groups who share similar characteristics, we effectively create a balanced comparison. This ensures that any differences in outcomes are not influenced by these underlying differences, thus providing fairer and more accurate results.

- The approach ensures that models are less likely to propagate bias by focusing on the key principle that, once matched, the treatment (e.g., loan approval) is applied uniformly across groups with comparable attributes, making the groups more comparable. This helps in eliminating the bias introduced by historical data or over-represented groups.

#### 4.7.2 Proof of Correctness:

- **Mathematical Basis of Propensity Scores:** Propensity Score Matching (PSM) works under the assumption that, given the same observable characteristics, the outcomes (such as loan approval) should be comparable across groups. The key proof of correctness lies in the fact that after matching, the treatment and control groups should exhibit no systematic differences in their observable characteristics, thus reducing bias. This is validated through statistical measures of balance (e.g., comparing means between the groups before and after matching).
- **Reduction of Bias in Outcomes:** By comparing outcomes between matched groups, any differences can be more confidently attributed to the effect of the treatment (e.g., loan approval decision) rather than differences in demographic or financial characteristics. This leads to a more equitable evaluation of applicants.
- **Visual Validation:** The visual comparison of propensity score distributions before and after matching (e.g., through histograms) further proves the correctness of the approach. If the distributions are well-aligned after matching, it indicates that the groups are balanced, and any bias introduced by differences in characteristics is minimized.
- **Correctness in Fairness Evaluation:** By applying fairness metrics such as Demographic Parity and Equal Opportunity after the bias mitigation process, we can verify whether the algorithm's decisions are fair across different demographic groups. These metrics ensure that the model does not favor one group over another, proving the algorithm's fairness and correctness in real-world applications such as loan approval.

## 5 Results and Data Analysis

This section presents the outcomes of the exploratory data analysis, preprocessing, and bias mitigation efforts applied to the loan approval dataset. The key

objective is to address class imbalance and demographic biases to improve the fairness and reliability of model predictions.

## 5.1 1. Exploratory Data Analysis (EDA)

- **Class Imbalance:**

- The dataset had a significant imbalance, with approximately 69% of loan applications approved and 31% denied.
- This imbalance could result in a biased model that favors the majority class (approved loans) and performs poorly for the minority class (denied loans).

- **Demographic Imbalances:**

- **Gender:** Around 82% of the applicants were male, indicating a potential gender bias.

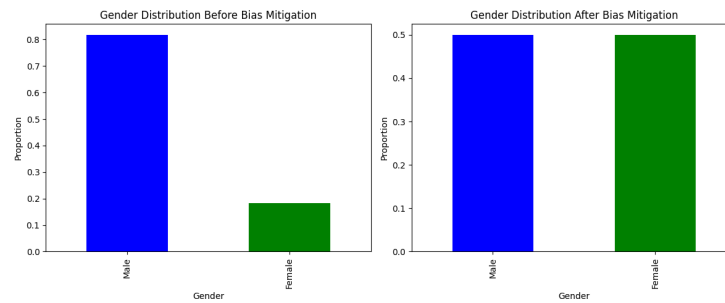


Figure 2: Impact of Oversampling on Class Distribution

- **Marital Status:** Approximately 65% of applicants were married, further suggesting an imbalance in demographic representation.

### Income and Loan Amount Correlation:

- A strong positive correlation was found between ApplicantIncome, CoapplicantIncome, and LoanAmount, indicating that income is a key factor in loan approval.

### Visualization Results:

- Bar plots revealed differences in approval rates between male and female applicants.
- Box plots showed higher median loan amounts in urban areas, with some outliers across property areas.

## 5.2 2. Data Preprocessing

- **Missing Value Imputation:**

- *Loan Amount*: Imputed using the median to handle outliers.
- *Loan\_Amount\_Term and Credit History*: Imputed using the mode to ensure categorical consistency.
- *Gender and Marital Status*: Filled using the mode to maintain demographic completeness.

**Categorical Encoding:** Categorical variables, including Gender and Property Area, were encoded into numerical values using LabelEncoder for compatibility with machine learning models.

**Feature Scaling:** StandardScaler was applied to ApplicantIncome, CoapplicantIncome, and LoanAmount to ensure all features contribute equally to the model.

**Outlier Detection and Capping:** The IQR method was used to cap outliers in income and loan amounts, reducing their impact on model performance.

## 5.3 3. Bias Mitigation Using Oversampling Techniques

- **Random Oversampling:**

- Random oversampling was used to balance gender and marital status distributions, ensuring equal representation of male/female and married/unmarried applicants.

- **SMOTE (Synthetic Minority Oversampling Technique):**

- SMOTE was applied to the training dataset to generate synthetic samples for denied loans, addressing the class imbalance between approved and denied loans.
- The class distribution was visualized before and after SMOTE, confirming that the technique effectively balanced the dataset.

## 5.4 4. Class Distribution Analysis

- **Before and After Comparisons:**

- Prior to oversampling, the dataset was heavily imbalanced, with a significant number of approved loans compared to denied loans.
- After applying SMOTE, the class distributions were balanced, allowing the model to learn equally from both classes.

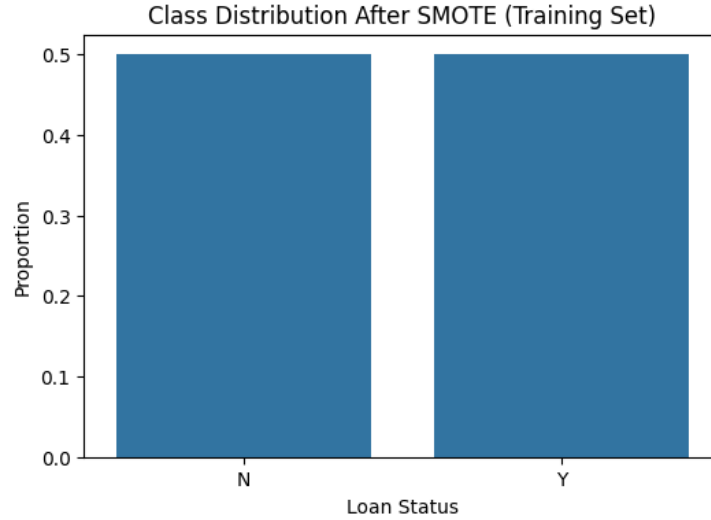


Figure 3: Impact of Oversampling on Class Distribution

- **Bias Mitigation Success:** The oversampling techniques (random oversampling and SMOTE) successfully balanced class and demographic distributions, reducing bias in the model.

- **Impact on Bias Detection:**

- Gender and marital status imbalances were reduced, minimizing the potential for biased predictions based on these demographic factors.
- The increased representation of denied loans allows the model to make more accurate predictions for minority cases.

## 5.5 5. Key Findings and Implications

- **Bias Mitigation Success:** The oversampling techniques (random oversampling and SMOTE) successfully balanced class and demographic distributions, reducing bias in the model.
- **Fairness and Equity:** The bias mitigation efforts ensure fairer predictions by treating applicants from all demographic groups equally.
- **Improved Model Generalization:** Addressing class imbalance improves the model's ability to generalize to new data, enhancing its reliability.
- **Influence of Income:** Despite the improvements in demographic fairness, income-based factors remained dominant, suggesting the need for further exploration of bias mitigation techniques.

## 5.6 Conclusion

The analysis demonstrates the importance of addressing class imbalance and demographic bias through oversampling techniques. While the results show improvements in class distributions and fairness, future work could explore additional mitigation strategies, such as class weighting or fairness-aware algorithms, to further enhance model equity. These efforts are essential to building fair and unbiased machine learning models, particularly in socially sensitive areas such as loan approval.

## References

- [1] Barocas, S., Selbst, A. D. (2016). *Big Data’s Disparate Impact*. California Law Review, 104(3), 671-732. <https://doi.org/10.15779/Z38BG31>. This paper explores how data bias can lead to disparate impacts in algorithmic decision-making and emphasizes the importance of fairness in model training.
- [2] Dastin, J. (2018). *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*. Reuters. Retrieved from: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>. This article discusses how Amazon’s AI-powered hiring tool exhibited gender bias, illustrating the real-world implications of biased training data in recruitment.
- [3] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021). *A Survey on Bias and Fairness in Machine Learning*. ACM Computing Surveys, 54(6), 1-35. <https://doi.org/10.1145/3457607>. This survey provides a comprehensive overview of different types of bias, fairness definitions, and mitigation strategies in machine learning systems.
- [4] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S. (2016). *On the (Im)possibility of Fairness*. ACM FAT/ML. Retrieved from: <https://www.fatml.org/resources/research>. This paper discusses the difficulties of achieving fairness in machine learning models and how different fairness criteria may conflict with each other.
- [5] Zliobaite, I. (2017). *Measuring Discrimination in Algorithmic Decision Making*. Data Mining and Knowledge Discovery, 31(4), 1060-1089. <https://doi.org/10.1007/s10618-017-0506-1>. This article investigates discrimination in algorithmic decision-making and proposes methodologies for detecting and addressing bias.
- [6] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16, 321–357. <https://doi.org/10.1613/jair.953>.

- [7] Friedman, J., Hastie, T., Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY.  
<https://doi.org/10.1007/978-0-387-21606-5>.
- [8] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.