

# Dropout Prediction: A Systematic Literature Review

SLR Dropout

---

Pedro Sobreiro, José Garcia Alonso, Javier Berrocal

Webinar, 26 June 2020

University of Extremadura

*pdealexa@alumnos.unex.es*

# Summary

Research goals

Introduction

Methodology

Results

Conclusion

## Research goals

---

What is the current state of machine learning research studies to predict dropout in contractual settings?

# Introduction

---

## Data mining

- Customer analysis is fundamental to develop business and marketing intelligence (Sheth, Mittal, & Newman, 1998), supporting the understanding of historical data identifying trends and patterns (Berry & Linoff, 2004)
- This process is also known as data mining, the extraction of knowledge from data (Han & Kamber, 2006)

Allen, I. E., & Seaman, J. (2011). Going the Distance: Online Education in the United States, 2011. Em Sloan Consortium (NJ1). Obtido de <https://eric.ed.gov/?id=ED529948>

Martinho, D. S. (2014). O Ensino Online nas Instituições de Ensino Superior Privado. As perspectivas: Docente e discente e as implicações na tomada de decisão institucional.

McCarthy, S. A. (2009). Online Learning as a Strategic Asset. Volume I: A Resource for Campus Leaders. A Report on the Online Education Benchmarking Study Conducted by the APLU-Sloan National Commission on Online Learning. Em Association of Public and Land-grant Universities. Obtido de <https://eric.ed.gov/?id=ED517308>

# Methodology

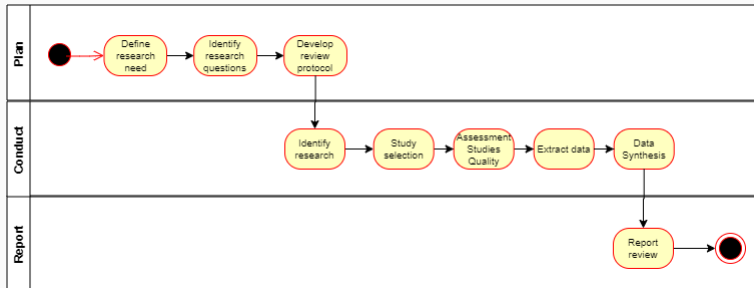
---

- Was developed a Systematic Literature Review (SLR) developed in three stages (Kitchenham & Charters, 2007): Plan, Conduct and Report;
- Plan: definition of the research need, identification of the research questions and the development of the review protocol;
- Conduct: research identification, study selections, quality assessment, data extraction, finishing with the data synthesis;
- Report: stage that develops the activity report review.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing structural literature reviews in software engineering (pp. 1–26) [Joint technical report]. Australia: Keele Univ., and Empirical Software Eng., Nat'l ICT.



# Systematic Literature Review Phases



**Figure 1:** SLR phases based on Kitchenham and Charter (2007)

# Research questions

What is the current state of machine learning research studies to predict dropout in contractual settings? Based in this question were identified the following questions:

- RQ1: What studies have been published?;
- RQ2: Which algorithms have been used to predict the dropout?
- RQ3: What are the more relevant features related to predicting customer dropout?
- RQ4: When the dropout occurs?
- RQ5: What is the accuracy of the machine learning algorithms to predict dropout?

# Population, Intervention, Comparison, Outcomes and Context

**Table 1:** PICOC criteria

PICOC	Description
Population	Research papers about dropout with contractual settings
Intervention	Machine learning algorithms to predict dropout
Comparison	Studies addressing machine learning algorithms to predict dropout
Outcome	Synthesis identifying research questions, gaps in the research domain and also best practices identified
Context	Academia and industry

Note: Context (PICOC) as suggested Kitchenham and Charters (2007) and proposed by Petticrew and Roberts (Petticrew & Roberts, 2006) to support the development of the search string.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing structural literature reviews in software engineering (pp. 1–26) [Joint technical report]. Australia: Keele Univ., and Empirical Software Eng., Nat I ICT.

Petticrew, M., & Roberts, H. (2006). Systematic reviews in the social sciences: A practical guide. Malden, MA ; Oxford: Blackwell Pub.

- Search string: ((“customer dropout”) OR (“customer churn”) AND “machine learning” AND (“contractual” OR “membership”));
- Applied to the title, abstract, and keywords in the search period between January 2000 and June 2020
- The exclusion criteria were Books, Non-English articles, patents, and thesis
- The selection process was developed using ASReview (ASReview Core Development Team, 2019) creating a dataset of the identified articles, providing five relevant papers and five irrelevant papers to train Machine Learning model Naïve Bayes;

Kitchenham, B., & Charters, S. (2007). Guidelines for performing structural literature reviews in software engineering (pp. 1–26) [Joint technical report]. Australia: Keele Univ., and Empirical Software Eng., Nat I ICT.

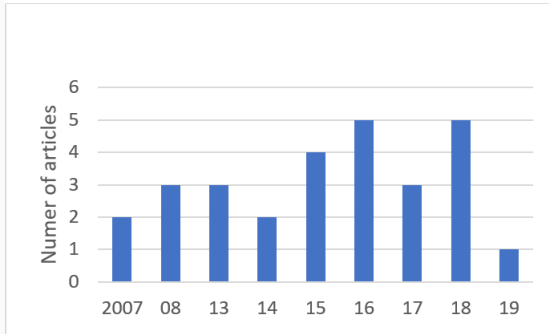
Petticrew, M., & Roberts, H. (2006). Systematic reviews in the social sciences: A practical guide. Malden, MA ; Oxford: Blackwell Pub.

## Results

---

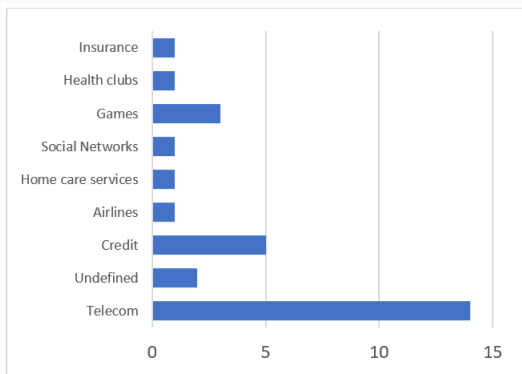
- 218 studies were found in the first step of the conduct (Identify research);
- 24 duplicates were removed
- 166 were removed after ASReview
- 1 paper rejected during data extraction
- 28 papers selected

## RQ1. What studies have been published?



**Figure 2:** Articles per year after quality assessment

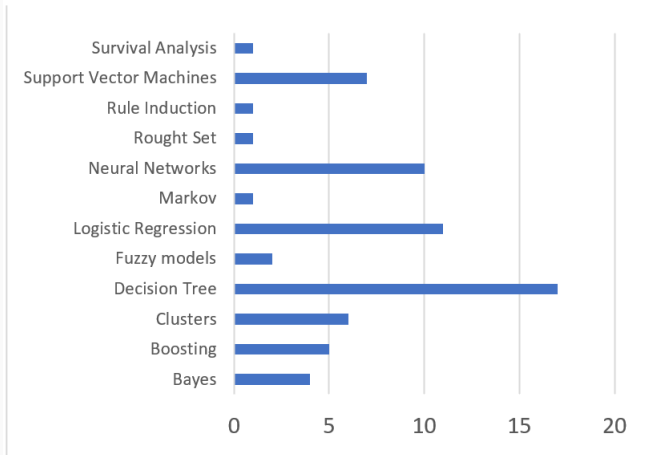
## RQ1. What studies have been published?



**Figure 3:** The number of studies per business context

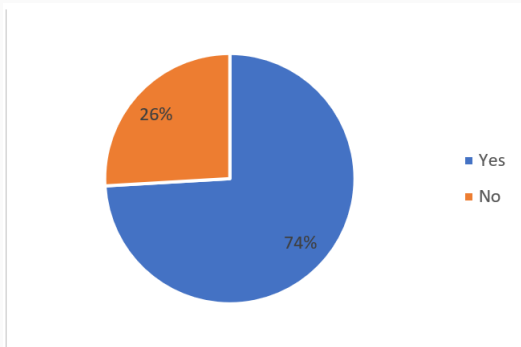


## RQ2: Which algorithms have been used to predict the dropout?



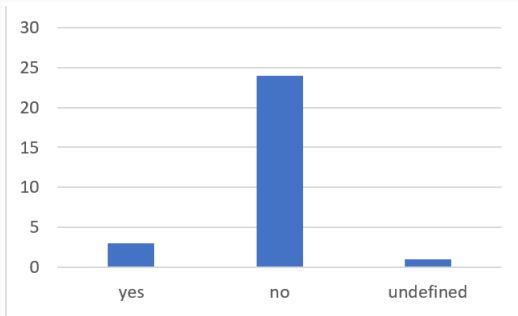
**Figure 4:** Main algorithms used in the analysed papers

### RQ3: What are the more relevant features related to predicting customer dropout?



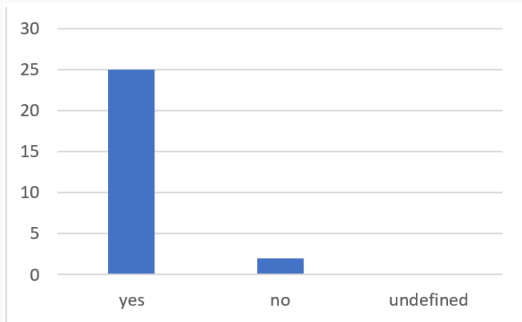
**Figure 5:** Percentage of studies identifying the relevant features

## RQ4: When the dropout occurs?



**Figure 6:** The number of studies addressing the dropout timings

## RQ5: What is the accuracy of the machine learning algorithms to predict dropout?



**Figure 7:** Number of studies identifying the prediction accuracy

# Conclusion

---

# Conclusion

- The telecommunications sector is the area where are being developing most of the studies, which identifies some business areas that need to be addressed;
- Algorithms to predict dropout using also survival analysis approaches is an area under researched, only three research papers, however considering the number of citations these approaches getting the attention (Perianez et al., 2016);
- The use of algorithms to explore the timings when the dropout will occur is an approach that could complement the dropout prediction, supporting the development of actions considering both the probability and when should be developed countermeasures to avoid the customer dropout

Perianez, A., Saas, A., Guitart, A., & Magne, C. (2016). Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 564–573. doi: 10/ggtgjh

# Thanks!

End

Start where you are. Use what you have. Do what you can.

Arthur Ashe