Applied Energy 359 (2024) 122725

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

# Subsidies for investing in energy efficiency measures: Applying a random forest model for unbalanced samples

Susana Álvarez-Diez [a], J. Samuel Baixauli-Soler [b], Gabriel Lozano-Reina [b,*], Diego Rodríguez-Linares Rey [c]

[a] *Department of Quantitative Methods for the Economy, Faculty of Economics and Business, University of Murcia, Campus of Espinardo, 30100 Murcia, Spain*
[b] *Department of Management and Finance, Faculty of Economics and Business, University of Murcia, Campus of Espinardo, 30100 Murcia, Spain*
[c] *PhD Programme in Economics, International Doctorate School (EIDUM), University of Murcia, Campus of Espinardo, 30100 Murcia, Spain*

## HIGHLIGHTS

- Public subsidies favor energy efficiency measures by reducing the up-front cost and making the investment more profitable.
- Subsidy effectiveness depends on the capacity to identify which SMEs are potential beneficiaries of energy subsidies.
- We evidence that applying a random forest approach for unbalanced samples offers greater predictive capacity and statistical power than traditional techniques.
- The most useful predictors for SMEs in the industrial sector are related to liquidity and indebtedness.

## ARTICLE INFO

*JEL codes:*
H2
M2
Q2
Q4
Q5

*Keywords:*
Energy efficiency
Public investment subsidies
SMEs
Random forest
Unbalanced samples

## ABSTRACT

Investing in energy efficiency measures is a major challenge for SMEs, both for environmental and economic reasons. However, certain barriers often make it difficult to invest in such measures. Although public financial support helps to overcome economic barriers, public bodies face the challenge of identifying which SMEs display the greatest potential to invest in energy efficiency measures. By applying a random forest technique and by using sampling balancing techniques, this paper identifies the profile of industrial SMEs that might be potential beneficiaries of public aid, thereby helping public institutions to target their calls and direct their efforts towards this group of SMEs. Specifically, liquidity and indebtedness are found to be the most useful predictors for SMEs in the industrial sector. The results are robust and reveal that applying a random forest approach for unbalanced samples offers greater predictive capacity and statistical power than applying traditional estimation techniques. By identifying potentially benefiting firms, this work helps to boost the effectiveness of public subsidies and to improve the channeling of public funds, which ultimately favors investment in energy efficiency.

## 1. Introduction

Investments in energy efficiency measures offer companies many advantages. They not only help to reduce total final energy use[1] and to improve environmental sustainability but also promote business innovation, profitability, and competitiveness [38]. Governments should promote these investments, especially in small and medium-sized enterprises (SMEs) given that the latter represent the “core” of the European industrial structure and employ around 60% of the total workforce [38], thereby making them a special focus for investments in energy efficiency. However, SMEs are less likely to invest in such efficiency measures compared to large companies [15,42]. Implementing energy efficiency measures thus remains a pending issue for SMEs, with many still reluctant to invest in them [67]. In fact, over two thirds of these businesses do not implement even the simplest rules to manage energy use [14].

* Corresponding author.
*E-mail addresses:* salvarez@um.es (S. Álvarez-Diez), samuel@um.es (J.S. Baixauli-Soler), gabriel.lozano@um.es (G. Lozano-Reina), drl@um.es (D. Rodríguez-Linares Rey).

[1] Throughout this paper, we consider “energy use” to be all the final energy supplied to industry (regardless of the energy source) [19].

https://doi.org/10.1016/j.apenergy.2024.122725
Received 27 September 2023; Received in revised form 4 January 2024; Accepted 21 January 2024
Available online 30 January 2024

Several barriers[2] make these SMEs less likely to implement energy efficiency measures. Among these obstacles, firm size is one of the main factors influencing the adoption of energy efficiency measures, thereby evidencing that such efficiency has not been a priority for SMEs –especially those that lack energy-intensive production processes [16,65]. Being smaller also means that these organizations have fewer technological options to save energy [16]. Moreover, there are other barriers related to information asymmetries as well as hidden and transaction costs, added to which capital restrictions tend to be more prevalent for SMEs when compared to larger companies [51,67]. Fleiter et al. [26] and Trianni et al. [68] highlight that economic and financial factors (specifically, lack of capital to face up-front costs) constitute important barriers that SMEs face when seeking to adopt energy efficiency measures.

Considering this context, governments and public institutions face an important challenge when designing and implementing public policies aimed at minimizing the barriers that prevent energy-efficient measures from being adopted [55]. Public financial support should thus be a priority for encouraging the implementation of energy efficiency measures, with investment subsidies being one potential driver[3] since they help to reduce the up-front cost and so make the investment more profitable [10]. In order to boost the effectiveness of these public subsidies –and thereby speed up the adoption of energy efficiency measures in SMEs– it is essential for public bodies to be able to identify which businesses are potential recipients of public aid.[4] Identifying the profile of which SMEs may be potential beneficiaries of public aid poses a major challenge since this helps to design the requirements of the calls, thus favoring efficiency in channeling public funds and preventing them from remaining unused. This is also a challenge in the sense that the proportion of businesses benefiting from this public support is often very low.

In an effort to fill this key gap, this paper aims to identify the profile of those SMEs that evidence the greatest potential to invest in energy efficiency measures. We apply the random forest approach for unbalanced samples so that –based on this identification– public institutions can direct public investment subsidies towards the group of businesses identified. For this purpose, a sample of industrial SMEs from the Region of Murcia (Spain) is used. These provide the focus of the study because the public call analyzed in this research deals exclusively with these companies since they represent over 25% of final energy consumption in Spain. It is therefore important to carry out improvements in energy efficiency in technologies and processes and in terms of implementing energy management systems in this context.

By addressing this goal, the contributions made by this paper can be grouped into the following blocks. At a methodological level, this article applies the random forest approach for unbalanced samples. The random forest model [8,9] is an automatic learning technique that allows us to identify which economic and financial factors make SMEs potential beneficiaries of public investment subsidies. However, one common issue in this context is that samples are often unbalanced –since the proportion of SMEs who might benefit from receiving public subsidies is much lower than the proportion of SMEs that might not. This may yield misleading results in that while the majority group might be correctly predicted, the minority group might not. To address this issue in a novel way, this paper applies balancing techniques to remove inefficiency from the imbalance, thereby increasing the accuracy of both the estimation and the identification of SME profile. Furthermore, this paper offers some evidence that applying a random forest approach for unbalanced samples offers greater predictive capacity and statistical power than applying traditional estimation techniques within this research field. At the context level, this paper focuses on the Region of Murcia, which allows us to carry out a homogeneous analysis of the requirements that SMEs must fulfil in order to apply for public investment subsidies, thereby making it possible to establish an accurate profile of potential SME beneficiaries. This is an advantage compared to previous studies that analyze various contexts –in which legislative, fiscal, and context aspects may vary– and which may ultimately affect the eligibility of the industrial SMEs that can opt for public subsidies (for further information about this context, see Section 2). Finally, at a practical level, public agents are provided with the profile of the industrial SMEs to which they should direct their actions (specifically, as regards designing, implementing, and disseminating their public subsidies) in an effort to therefore maximize investment in energy efficiency measures. In addition to proving useful for public institutions in the Region of Murcia, the resulting profile is also useful for contexts and sectors similar to the one analyzed in this paper.

The paper is structured as follows. After this introduction, Section 2 describes the most relevant particularities of the context of the Region of Murcia. Section 3 reviews the literature on the main barriers facing SMEs who seek to invest in energy efficiency measures, as well as the role of public subsidies as an instrument to promote these investments. In Section 4, we address all the methodological aspects (sample, data, variables) in addition to describing the random forest approach used and the techniques applied to control for unbalanced samples. The results are presented in Section 5. Finally, Section 6 offers the discussion of the findings, and Section 7 contains the concluding remarks.

## 2. The Region of Murcia

The Region of Murcia in southeast Spain is a region in the Mediterranean area of the Iberian Peninsula. Covering an area of around 11,300 km$^2$, this region combines a diversity of landscapes, ranging from the coast bathed by the Mediterranean to the inland mountains. At an economic level, the Region of Murcia has experienced significant development in recent years, with agribusinesses and the wine industry standing out in particular. In terms of energy efficiency, the Region of Murcia faces challenges common to many other regions in Spain. Diversification of the energy matrix and the transition towards more sustainable energy sources are key areas of interest. Given its location in a region that enjoys abundant sunlight, energy efficiency can benefit from promoting solar technologies and other renewable energy sources.

Opting for the Region of Murcia as the context for this paper is relevant for three main reasons. Firstly, Spain is divided into different regions whose regional policies and regulations tend to differ. This means that the public calls for investment subsidies tend to vary between regions to the extent that these calls and formalities depend directly on each regional government. In order to achieve consistency in the sample and so obtain conclusive results, we focus on a single region, thereby ensuring that the regional policy and regulation applicable to all SMEs included in the study is the same. In the specific case of the Region of Murcia, the energy subsidy analyzed seeks to compensate a specific percentage of the investment made by each SME in energy efficiency measures [19], with industrial SMEs that meet the eligibility requirements being those that can apply for said subsidy (further information about this subsidy is provided in Section 4.2).

Secondly, most energy consumption in Spain is focused on industrial sector and on electrical energy. As shown in Table 1, Spain's energy consumption amounted to €11,227 million in 2019, of which €6368

[2] A barrier is defined as "a postulated mechanism that inhibits investment in technologies that are both energy efficient and (apparently) economically efficient" ([18], p. 2), emphasizing the importance of the "cost-effective" factor [59].

[3] Drivers are defined as "factors stimulating the sustainable adoption of energy-efficient technologies, practices and services, influencing a portion of the organization and a part of the decision-making process in order to tackle the existing barriers" ([70], p. 204).

[4] Throughout this study, we consider "potential beneficiary SMEs" to be those firms that, on the one hand, comply with the eligibility conditions to access energy efficiency aid established in the different public calls and, on the other, are also more likely to make energy investments.

**Table 1**
Distribution of energy consumption in Spain and the Region of Murcia (2019).

| | Electricity | Gas | Diesel | Fuel oil | Coal and coke | Biofuels | Heat and others | Total |
|---|---|---|---|---|---|---|---|---|
| Spain | 6367.9 | 3371.3 | 678.3 | 99.1 | 104.6 | 44.7 | 560.8 | 11,227.0 |
| | 56.72% | 30.03% | 6.04% | 0.88% | 0.93% | 0.40% | 5.00% | 100.00% |
| Region of Murcia | 209.6 | 128.7 | 24.3 | 2.3 | 0.7 | 0.5 | 465.3 | 438.6 |
| | 47.79% | 29.35% | 5.53% | 0.53% | 0.15% | 0.11% | 5.09% | 100.00% |

Note: The figures are expressed in millions of euros.
Source: Spanish National Statistics Institute.

million was electrical energy and €3371 million gas. In the Region of Murcia, energy consumption came to €438.6 million in 2019, and there was a similar distribution between the different energy sources [61]. Moreover, the industrial sector represents over 25% of energy consumption [19], such that special attention must be paid to this sector in order to encourage the implementation of more energy efficiency measures. More specifically, the energy consumption based on the NACE Rev. 2 codes is shown in Appendix A, where both the relevant weight that the industrial sector has in energy consumption and the distribution of energy sources based on the NACE Rev. 2 code can be seen. Energy efficiency investments focused on electrical energy are therefore especially relevant in Spain as a whole and in the Region of Murcia in particular. In fact in 2021, Spain was the second country in the European Union in terms of generating the most electrical energy from wind and solar sources (Red [56]), with the Region of Murcia being one of the leading areas in terms of photovoltaic solar energy generation.

Thirdly, Murcia is one of the European regions with the greatest number of hours of sunshine per year. A recent study into "the Sunniest Cities in Europe" classifies the area of Murcia as the third sunniest city, with an average of 346 h of sun per month [37]. Considering that the power generation capacity of photovoltaic energy is mostly influenced by solar radiation intensity (which is quite high in the Region of Murcia, as shown in Fig. 1), investing in energy efficiency measures (especially solar) is of particular interest in the Region of Murcia. The amount of sunlight in a region like this tends to impact the attractiveness of energy efficiency projects (e.g., due to solar energy potential, economic viability, energy independence, renewable energy targets), and to influence government decisions regarding public subsidies. Regions with ample sunlight are thus likely to receive more support since they offer a more viable and efficient option for investing in energy efficiency measures (e.g., by harnessing clean and renewable energy).

In sum, the Region of Murcia offers a unique combination of geographical and economic characteristics, thus making it an interesting context in which to explore strategies and solutions in the field of energy efficiency. In this way, the regional government plays a key role in encouraging these investments through public aid –which encourages higher levels in cost reduction and in return on investment.

## 3. Literature review

Ever-increasing global competitiveness makes companies place greater emphasis on improving their efficiency and –since most of their processes are related to energy– any such improvements in energy efficiency will enhance this [51]. Investing in energy efficiency measures is thus a major concern for organizations for two main reasons. From an environmental point of view, traditional energies involve emitting harmful gases –specifically greenhouse gases (GHG)– into the atmosphere, thereby accentuating the problem of climate change. Added to this is the uncertainty surrounding the availability of these traditional energy sources [17]. Implementing more efficient energy sources therefore contributes towards environmental sustainability. In addition, from an economic point of view, implementing energy efficiency

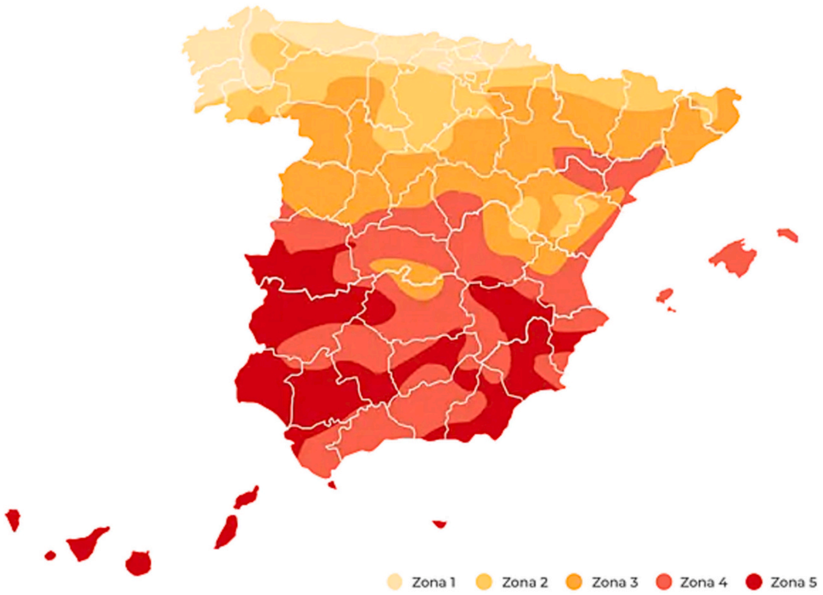

**Fig. 1.** Solar radiation map in Spain (2023).
Note: The map of solar radiation in Spain is divided into five zones according to the incidence of the sun on the surface. Types of radiation are differentiated according to the ray and its incidence, and the measuring equipment is the pyranometer.
Source: Soto [60].

measures means increased company competitiveness –which is particularly important for SMEs [38,42]– such that these measures yield cost savings and entail the modernization and innovation of firms' energy sources. Beyond the energy benefits derived from investing in these measures (such as energy savings, energy costs, environmental improvements), the non-energy benefits (such as higher productivity, better product quality, less waste, or less maintenance) also stand out [51], all of which positively impacts financial performance.

However, investments in energy-efficient technologies are still limited by the hurdles and market failures that prevent the most energy-efficient alternative from being chosen, even if this alternative is the most profitable for organizations [27]. This is particularly true for SMEs, for whom investments in energy efficiency are often considered "low priority" projects [26]. In addition, such firms must also face more difficult obstacles when investing in energy efficiency (particularly, in the financial field) [42]. All of this creates a "gap" between the theoretical opportunities for cost-effective investments in energy efficiency and the actual levels that may be achieved in practice [17,36,62]. In this context, "barrier models" have been the widely accepted framework for explaining the existence of the energy efficiency gap [3,39]. These barriers can differ substantially in nature, and several studies have stated different but related taxonomies. For instance, (i) Cagno et al. [17] establish seven categories of barriers, including *technology-related*, *information-related*, *economic*, *behavioral*, *organizational*, *competence-related*, and *awareness*; (ii) Sorrell et al. [59] classify these barriers into the following six categories, including *imperfect information*, *hidden costs*, *risk*, *access to capital*, *split incentives*, and *bounded rationality*; while (iii) Kostka et al. [42] establish three large groups of barriers, i.e. *financial*, *information*, and *organizational* barriers.

Regardless of the taxonomy used, economic-financial barriers are one of the main obstacles stated by most previous literature when seeking to implement energy efficiency measures [55], especially in an SME context [42]. Among the main economic-financial factors that prevent investment in efficient energy, the literature states technical risks [12,65], high investment costs [26,38], the existence of hidden costs [17], or uncertainty surrounding the return on investment [34,65]. In addition, access to capital proves key, given that companies often lack sufficient capital to invest in energy efficient technologies (e.g., [13,26,38,66]).

Limitations on access to capital not only refer to the external capital that companies may obtain from banking institutions (where SMEs have greater difficulty raising funds when compared to large organizations) but also to the use of internal capital and to establishing priorities among alternative investment projects [27]. Fleiter et al. [26] state that the lack of capital proves to be a major obstacle that slows down the adoption of energy efficiency measures. Similarly, Kostka et al. [42] find that companies fail to make possible investments in energy efficiency improvements because they cannot access the required investment capital at competitive prices. We therefore see how these economic-financial barriers (in particular, access to capital limitations) lie behind these low figures for investment in energy efficiency measures, and how they prevent companies (specifically SMEs) from taking advantage of the economic and energy benefits that could be derived were they to invest in them.

To counteract the above barriers or obstacles, there are several drivers that make investments in energy efficiency measures more attractive and profitable for companies. Trianni et al. [69] identify four large groups of drivers that can favor investments: *regulatory*, *economic*, *informative*, and *vocational training*. Among these, the literature has highlighted economic drivers as the ones that display the greatest potential and effectiveness for reducing economic-financial barriers [15,65,68]. Indeed, Hrovatin et al. [38] affirm that economic drivers would top the list in importance in SMEs vis-à-vis effectively promoting investments in efficient energy. These economic drivers include the cost reduction stemming from lower energy use, information about real costs, management support, public aid, or private financing [14,69].

Among these drivers, public aid plays a crucial role in promoting investment in energy efficiency measures, with government intervention proving useful for overcoming the market failures that impede the efficient allocation of resources towards energy-saving initiatives [40]. Public aid aimed at promoting investment in efficiency energy measures often varies by country –and even by region– and may take different forms, such as subsidies, tax incentives, grants, and regulatory support. For instance, production tax credits (PTCs) and investment tax credits (ITCs) are provided by the United States government to encourage the development of renewable energy projects [22]; green bonds are other public financial instruments specifically earmarked to raise funds for environmentally friendly projects [76]; feed-in tariff programs are implemented by several countries where renewable energy producers are paid a fixed rate for the electricity they generate [6]; and grants and subsidies are a common type of public aid that provide direct financial support that can be used for project development, equipment purchase, or other eligible expenses [75].

In this context, public subsidies have attracted considerable attention from academia as they are a major driver for improving energy efficiency worldwide [2,30,52]. These subsidies can make investments more attractive and economically profitable, in addition to favoring cost reduction through reduced energy use [12,14]. Public subsidies also emerge as an effective policy measure to accelerate the dissemination of energy efficiency measures in SMEs [21,26]. Given the importance of these firms for the economy, if governments are to actively encourage energy efficiency investment in SMEs then they should target their subsidy policies towards such companies [75].

Both fixed and output subsidies coexist globally [52]. While fixed subsidies offer a predetermined amount of assistance, output subsidies are linked to the quantity of goods or services produced. Governments often implement these subsidies simultaneously in order to provide financial support to various industries [52]. Such a dual approach allows governments to address different economic objectives, encourage growth, and support specific sectors through tailored subsidy mechanisms.

Moreover, these public investment subsidies are highly correlated with private financing [15] and reflect the need for companies to receive external economic support, either from public and/or private agents. In this vein, the role of governments is vital when designing and implementing certain policies or measures (such as these public subsidies) geared towards correcting market failures and reducing obstacles/barriers to investment [55]. Nevertheless, the challenges and nuances involved in the effectiveness of this public aid are also acknowledged in the literature in the sense that justifying the differing degrees of government intervention remains a controversial topic in previous studies (see [62]). For government intervention to be effective, it is necessary to direct public incentives towards those SMEs that exhibit the greatest potential to invest in energy efficiency measures. In fact, caution against blanket subsidies should be taken into account, with the argument being that poorly designed policies may lead to inefficiencies and unintended consequences [2]. It is important to carefully craft interventions that consider market dynamics and company behavior so as to maximize the effectiveness of public aid. Well-designed public policies thus prove pivotal to effectively driving the adoption of energy-efficient measures, and thereby contributing to overall sustainability goals.

Given the existing literature on energy efficiency measures, two issues need to be resolved relating to public subsidies as one of the main drivers of these measures. Firstly, despite the importance of targeted subsidies, there is no accurate information about the features of which industrial SMEs are more likely to apply for them. This failure means that public calls are often far from reality and that the channeling of this public aid proves ineffective. This also implies that the barriers which prevent the implementation of energy efficiency measures are not overcome, leading SMEs to remain reluctant to make these investments [67]. Secondly, defining the required profile is also difficult due to two methodological issues: (i) the proportion of SMEs that usually apply for public subsidies is very small, which generates an imbalance in the sample that can negatively affect the power of estimation; and (ii) there are many determinants behind the SME’s decision to apply or not for the subsidy, with traditional models simply tending to consider a limited number of indicators in their estimations. If these issues are not adequately addressed, the results may be biased, and it will not be possible to predict an adequate profile of SMEs who may potentially benefit from these public subsidies.

This paper thus seeks to narrow these gaps by offering progress in the following areas:

- Defining a profile of potential beneficiary SMEs in order to report the information to public bodies so that they can take it into account when designing, implementing, and disseminating their public policies. This profile will encourage the design of more focused calls, thereby ultimately helping to boost the effectiveness of public subsidies and the number of investments in energy efficiency measures (generating, in turn, the environmental and economic benefits mentioned at the beginning of this section).
- Moving forward by implementing more sophisticated prediction techniques that allow a number of indicators to be managed and by addressing sample imbalance, thereby offering more accurate and robust results. This is materialized by applying the random forest approach for unbalanced samples. Specifically, the first step is to apply the sample balancing technique to achieve a certain balance between the proportion of SMEs that apply for a public investment subsidy and those that do not. The random forest approach then allows several independent decision trees to be created, considering a wide array of economic and financial indicators. Based on all the decision trees, the most relevant indicators for those companies who are most likely to apply for the subsidy are then extracted and identified.

## 4. Methodology

### *4.1. Sample and data*

Our sample comprises a subset of SMEs, analysis of whom is key within the energy efficiency-related field because these organizations –in addition to representing the majority of the productive fabric worldwide [38]– tend to implement few energy efficiency measures [15,42]. For this reason, public efforts should focus on SMEs in an effort to change this undesirable trend and increase SME competitiveness. Following European Commission Regulation No. 651/2014, we consider SMEs to be organizations that meet the following three requirements: (a) businesses with less than 250 employees; (b) businesses whose annual turnover does not exceed €50 million; and (c) businesses whose assets do not exceed €43 million [24].

We specifically took SMEs included in the SABI (*Iberian Balance Analysis System*) database, but applied a double filter. Firstly, we only include SMEs located in the Region of Murcia (Spain), for the reasons highlighted in Section 2. Secondly, we only include those SMEs that meet all the requirements (or eligibility conditions) stated in the “Order of the Regional Ministry of Employment, Universities, Business, and the Environment of the Region of Murcia”, which establishes the regulatory bases of the aid program for energy efficiency actions [19]. Based on this call, we include SMEs that fall within the NACE Rev. 2 codes shown in Table 2. Basically, the industrial sector was the one for which the 2019 call for public subsidies was set up. The industrial sector comprises firms engaged in the manufacturing and production of capital goods such as construction, machinery or chemical products, among others. Any SMEs in a crisis situation were also excluded.

**Table 2**
NACE Rev. 2 codes included in the sample.

| Code | Definition |
|---|---|
| 07 | Mining of metal ores |
| 08 | Other mining and quarrying |
| 09 | Mining support service activities |
| 10 | Manufacture of food products |
| 11 | Manufacture of beverages |
| 13 | Manufacture of textiles |
| 14 | Manufacture of wearing apparel |
| 15 | Manufacture of leather and related products |
| 16 | Manufacture of wood and of products of wood and cork |
| 17 | Manufacture of paper and paper products |
| 18 | Printing and reproduction of recorded media |
| 19 | Manufacture of coke and refined petroleum products |
| 20 | Manufacture of chemicals and chemical products |
| 21 | Manufacture of basic pharmaceutical products and pharmaceutical preparations |
| 22 | Manufacture of rubber and plastic products |
| 23 | Manufacture of other non-metallic mineral products |
| 24 | Manufacture of basic metals |
| 25 | Manufacture of fabricated metal products, except machinery and equipment |
| 26 | Manufacture of computer, electronic and optical products |
| 27 | Manufacture of electrical equipment |
| 28 | Manufacture of machinery and equipment |
| 29 | Manufacture of motor vehicles, trailers, and semi-trailers |
| 30 | Manufacture of other transport equipment |
| 31 | Manufacture of furniture |
| 32 | Other manufacturing |
| 33 | Repair and installation of machinery and equipment |
| 35 | Electricity, gas, steam, and air conditioning supply |
| 36 | Water collection, treatment, and supply |
| 37 | Sewerage |
| 38 | Waste collection, treatment, and disposal activities; materials recovery |
| 39 | Remediation activities and other waste management services |

After applying this double filter –and after excluding SMEs for which there is no information available for all the years– our sample came to 1992 industrial SMEs, which are analyzed in the period between 2013 and 2018. This period was selected because the call for public subsidies was published in 2019 by the regional government of Murcia, and because compliance with the requirements for SMEs to benefit from public aid had to be accredited prior to the publication date of the call.

The data used for this study are extracted directly from the SABI database. More specifically, SABI offers information on all economic-financial predictors, in addition to providing information on which SMEs applied for an energy efficiency subsidy in 2019 (there is an “observations” section within the subsidies where the specific type of subsidy received can be coded).

### *4.2. Variables*

#### *4.2.1. Dependent variable*

Initially, we collected the data from the SABI database on whether an SME had applied or not for an energy efficiency subsidy in 2019. Based on this information –and using artificial intelligence techniques to solve a classification problem– we sought to estimate the probability that an

SME would request or not a public investment subsidy in the future, where the group of SMEs likely to request the public subsidy is really the one of interest. This variable is thus dichotomous, taking the value 1 if the SME does request a public subsidy, or the value 0 if the SME does not.

In the public call analyzed in this paper, the subsidy is calculated as a percentage of the investment costs –specifically, 30% of the investment in energy efficiency measures, considering the limits stated in the public call [19]. The investment must be directly related to energy savings and efficiency, i.e. the subsidized investment must be aimed at implementing efficient technology. The aid granted by the regional government is co-financed with contributions from the European Regional Development Fund (ERDF). Only those SMEs that previously met the eligibility requirements established in the call could apply for these public subsidies –in particular, SMEs included in any of the NACE Rev. 2 codes indicated in the call– could apply.

#### 4.2.2. Predictor variables

We include 88 economic and financial indicators (ratios) related to indebtedness, investment, personnel expenditures, dividend policy, liquid assets, and other current assets and liabilities, as shown in Table 3. In general, those indicators related to indebtedness and liabilities are expected to have a negative effect on the probability of applying for this public aid to invest in energy efficiency, since indebtedness and liabilities tend to reduce a company's investment capacity. For their part, indicators related to assets, dividends, liquidity, or size are expected to have a positive effect. Based on these 88 indicators, the dataset used in random forests consists of 528 indicators, formed as follows: (i) the corresponding value of the ratio for 2018 (R*number*); (ii) the variation rate of each ratio between 2017 and 2018 (VR*number*); (iii) the year-on-year growth rate of each ratio over the period 2013–2018 (GR*number*); (iv) industry-relative indicators in terms of the value for 2018

**Table 3**
Economic and financial indicators.

| **Indebtedness** | **Expected sign** | **Indicators** |
|---|---|---|
| *Total indebtedness* | – | **R1**: Total debt/Total equity and debt; **R2**: Total equity/Total equity and debt; **R3**: Financial debt/Total equity and debt; **R4**: Financial debt/Total debt; **R5**: Non-financial debt/Total equity and debt; **R6**: Non-financial debt/Total debt |
| *Long-term indebtedness* | – | **R7**: Long-term financial debt/Total debt; **R8**: Long-term financial debt/Long-term debt; **R9**: Long-term financial debt/Total equity and debt; **R10**: Long-term debt/Total equity and debt; **R11**: Long-term debt/Total debt; **R12**: Long-term non-financial debt/Total debt; **R13**: Long-term non-financial debt/Total equity and debt; **R14**: Long-term non-financial debt/Long-term debt |
| *Short-term indebtedness* | – | **R15**: Short-term financial debt/Total debt; **R16**: Short-term financial debt/Short-term debt; **R17**: Short-term financial debt/Total equity and debt; **R18**: Short-term debt/Total equity and debt; **R19**: Short-term debt/Total debt; **R20**: Short-term non-financial debt/Total debt; **R21**: Short-term non-financial debt/Total equity and debt; **R22**: Short-term non-financial debt/Short-term debt; **R23**: Short-term debt/Long-term debt; **R24**: Short-term financial debt/Long-term financial debt |
| *Financial expenses (and incomes)* | – | **R25:** Financial expenses/Total equity and debt; **R26**: Financial expenses/Total debt; **R27**: Financial expenses/Financial debt; **R28**: Financial expenses/Net sales; **R29**: Financial incomes/Total assets; **R30**: Financial incomes/Net sales; **R31**: (Financial incomes-Financial expenses)/Total assets; **R32**: (Financial incomes-Financial expenses)/Net sales; **R33**: (Financial incomes-Financial expenses)/Earnings before interest and taxes; **R34**: Financial expenses/Earnings before interest and taxes |
| **Investment** | **Expected sign** | **Indicators** |
| *Non-current assets* | + | **R35**: Tangible assets/Total assets; **R36**: Tangible assets/Non-current assets; **R37**: Non-current financial assets/Total assets; **R38**: Non-current financial assets/Non-current assets; **R39**: Investment property/Total assets; **R40**: Investment property/Non-current assets; **R41**: Non-current assets/Total assets |
| *Other non-current assets* | + | **R42**: Other non-current assets/Non-current assets; **R43**: Other non-current assets/Total assets; **R44**: Total equity/Non-current assets |
| **Personnel expenses** | **Expected sign** | **Indicators** |
| *Personnel expenses* | + | **R45**: Personnel expenses/Net sales; **R46** (Personnel expenses/Other operating expenses); **R47** (Personnel expenses/Number of employees); **R48**: Personnel expenses/Non-current assets |
| *Employees* | + | **R49**: Number of employees/Net sales; **R50** (Number of employees/Other operating expenses); **R51**: Number of employees/Non-current assets |
| **Dividend policy** | **Expected sign** | **Indicators** |
| *Dividends and reserves* | + | **R52**: Total ordinary dividends/Net income; **R53**: Reserves/Total equity; **R54**: Reserves/Total equity and debt; **R55**: Reserves/Net income; **R56**: Reserves/Non-current assets |
| **Liquid assets (cash holdings)** | **Expected sign** | **Indicators** |
| *Cash* | + | **R57:** Cash/Total assets; **R58**: Cash/Current assets; **R59**: Cash/Short-term debt; **R60**: Cash/Net sales |
| *Cash and cash equivalents* | + | **R61:** Cash and cash equivalents/Total assets; **R62**: Cash and cash equivalents/Current assets; **R63**: Cash and cash equivalents/Short-term debt; **R64**: Cash and cash equivalents/Net sales |
| **Other current assets and liabilities** | **Expected sign** | **Indicators** |
| *Current assets* | + | **R65**: Current assets/Total assets; **R66**: Current assets/Short-term debt; **R67**: (Current assets-inventories)/Short-term debt; **R68**: (Current assets-inventories-trade receivables)/Short-term debt; **R69**: (Current assets-Short-term debt)/Net sales; **R70**: Current assets/Non-current assets |
| *Inventories* | + | **R71:** Inventories/Total assets; **R72**: Inventories/Current assets; **R73**: Inventories/Short-term debt; **R74**: Inventories/Net sales; **R75**: (Inventories/Supplies)*365; **R76**: (Inventories+trade receivables)/Trade payables |
| *Trade receivables* | + | **R77:** Trade receivables/Total assets; **R78**: Trade receivables/Current assets; **R79**: Trade receivables/Short-term debt; **R80**: Trade receivables/Net sales; **R81**: Trade receivables/Trade payables; **R82**: (Trade receivables/Net sales)*365 |
| *Short-term financial assets* | + | **R83:** Short-term financial assets/Total assets; **R84**: Short-term financial assets/Current assets; **R85**: Short-term financial assets/Short-term debt; **R86**: Short-term financial assets/Net sales |
| *Trade payables* | – | **R87:** Trade payables/Net sales; **R88**: (Trade payables/supplies)*365 |

(IR*number*); (v) industry-relative indicators in terms of the 2017–2018 variation rate (IVR*number*); and (vi) industry-relative indicators in terms of the 2013–2018 year-on-year growth rate (IGR*number*). As can be seen, the large volume of indicators (specifically, 528 indicators for each company, implying over one million observations) does not allow for the application of regression or discriminant analysis models in which the number of parameters would eliminate the degrees of freedom of the model. In addition, linear models based on multiple discriminant analysis or logistic regression would require prior selection of the explanatory variables, which would cancel out the option of creating decision trees.

The selection technique used to identify the economic-financial indicators is preprocessing, which refers to all the transformations on the raw data into a structured data set before it is fed into the machine learning [44]. This is a crucial part of data analytics prior to modelling since data preparation affects the model's predictive capacity. After removing missing values, the preprocessing step that we followed was data reduction, which was aimed at removing redundant and irrelevant predictors. This was done carefully in order to ensure that no predictive information was lost and that no erroneous information was added to the data. This allowed us to obtain a more reduced representation of the predictors, which was smaller in volume but which produced almost the same analytical results. This dimensionality reduction helped to speed up training. We carried out a preliminary selection of predictors based on the correlation matrix, which shows the relation between two given predictors in the form of a matrix. Specifically, we used the correlation matrix to reduce dimensionality because one of the assumptions in most key machine learning models is that no variable in the model is highly correlated to any other variable. We thus built the correlation matrix for predictors to test correlation between independent variables.

### *4.3. Random forest approach for an unbalanced sample context*

We build a model to classify firms which have or have not applied for a financial subsidy to improve energy efficiency, based on their characteristics. Specifically, we develop a random forest model [8,9] on a synthetically balanced training dataset to build a classifier for our study; i.e., to classify whether a firm has applied for a financial subsidy or not to improve energy efficiency (for instance, in terms of installing energy-efficient facilities, improving technology in equipment and industrial processes, or implementing energy management systems). In this case, we face a problem with unbalanced classification, where there are few cases in the minority class or positive class (i.e., firms which have applied for a financial subsidy), while most cases belong to the majority class or negative class (i.e., firms which have not applied for a financial subsidy).

The minority class in our study is the most important, as usually occurs in learning from imbalanced data. In order to make good use of budgetary resources in informational campaigns about energy subsidies and so minimize public resource wastage and thereby provide accurate data on investment subsidies, public institutions are interested in identifying which organizations are potential beneficiaries. Consequently, correctly predicting which firms have applied for a financial subsidy is a desirable property of the classifier. Optimal deployment of this policy instrument thus depends on the direct interaction of policymaker and potential beneficiaries –represented by the minority class.

Taking this problem of imbalanced classification into account, many studies have pointed out that classification algorithms are extremely accurate for majority classes, but significantly less so for minority classes. The classification performance of standard classification algorithms is negatively affected by imbalanced datasets because minority cases cannot be identified, even though they are generally the most interesting [5,25,33,43,48]. Researchers in the field of learning from imbalanced data have focused on using data sampling algorithms or external approaches [1,11,20,23,29,32,35,53,64], modifying standard modelling algorithms or internal approaches [47,71,72,78], combining both internal and external approaches or cost-sensitive learning techniques [63], and choosing appropriate performance metrics [41,50,57].

Since we focus on evaluating model performance, we split the dataset into a training dataset (70% of the sample) to fit the model and a test dataset to validate it on an unseen dataset (30% of the sample). In order to handle the imbalance issue, we balance the number of instances across the classes in the training dataset using a data sampling algorithm –rather than a modelling algorithm– which deals with class imbalance. Specifically, we choose the Synthetic Minority Oversampling Technique (SMOTE), proposed by Chawla et al. [20] to balance the training dataset. SMOTE is an oversampling method whose key point is that it creates extra minority instances (synthetic instances) by interpolating between minority class instances which are within a neighborhood [20]. In fact, SMOTE is the most widely known data sampling method and is commonly used in practice to balance class distribution in the training dataset [28]. Fernández et al. [25] point out that the popularity of SMOTE stems from its simplicity in terms of the procedural design as well as its robustness when chosen to pre-process unbalanced data in many different applications.

Specifically, new minority instances are created as follows: for each minority instance, $k$ neighbors are randomly selected; generally, the number of nearest neighbors used to generate new instances of the minority class is $k = 5$. Depending on the amount of oversampling required, $s$ neighbors from the $k$ ones ($s \leq k$) are randomly chosen again. For example, for 300% oversampling, three neighbors from the $k = 5$ are considered. The synthetic minority instances come from multiplying the difference between the minority instance and each of its neighbors (three out of the five in 300% oversampling is considered) by a random number between 0 and 1. In sum, SMOTE carries out an interpolation among neighboring minority class instances. As such, it increases the number of minority class instances by introducing new minority class examples in the neighborhood, thereby helping the classifiers to improve its generalization capacity and, hence, the performance of the classifiers on minority class instances. Fig. 2 shows how synthetic minority instances are created.

We then fit a random forest algorithm to the synthetic training set, which performs much better on the minority class since there are almost the same number of observations in the minority class as in the majority class. We build a random forest model with a repeated $k$-fold cross-validation algorithm. Specifically, we consider 10 repetitions and $k = 10$ folds. For each repetition, the balanced training set is randomly split into 10 folds. Thus, there is a different split of the balanced training set in each repetition. Each fold is treated as a validation set or out-of-bag sample (OOB sample) to compute the performance score of a random forest model, while the remaining nine folds are used to train it. The number of trees we consider for training a random forest is 100. Given that we consider 10 repetitions –$k = 10$ and 100 trees in each training set– the overall number of trees created is 10,000. The algorithm is summarized in Fig. 3.

To create each tree, the random forest method selects several predictor variables, thereby preventing any one particular feature that may be highly influential from dominating many trees. Hence, each tree is split based on slightly different samples and different features, providing decorrelated trees and producing a more accurate predictor. Each tree's

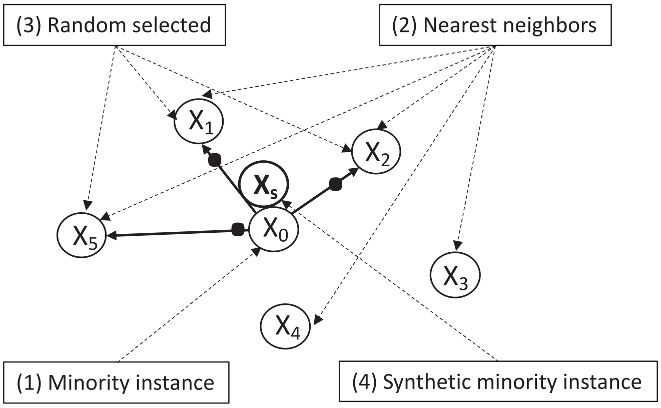

**Fig. 2.** SMOTE method to create extra minority instances.
Note: (1) $X_0$ is the minority instance; (2) $X_1$, $X_2$, $X_3$, $X_4$, and $X_5$ are the nearest neighbors; (3) $X_1$, $X_2$, and $X_5$ are the nearest neighbors randomly selected; (4) $X_s$ is the synthetic minority instance from multiplying the difference between the neighbors randomly selected and the minority instance by a random number between 0 and 1 (black dots).

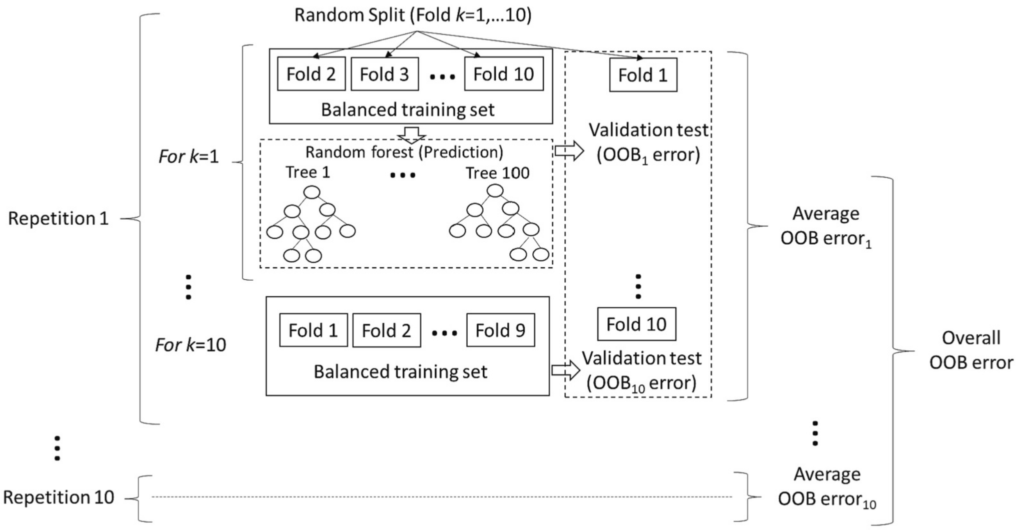

**Fig. 3.** Random forest model with a repeated *k*-fold cross-validation algorithm.
Note: repetitions are 10; $k = 10$ folds, where nine folds are used for training and one is treated as a validation set or out-of-bag sample (OOB sample); balanced training set is randomly split into 10 folds; overall number of trees created is 10,000 (100 trees in each training set, $k = 10$, and 10 repetitions); overall OOB error is the average of OOB errors in every repetition.

predictions are aggregated so as to obtain the final predictor. As $k = 10$, a random forest model is training 10 times in each repetition. At each repetition, the estimated prediction error is computed as the average of the 10 prediction errors computed in each OOB sample. The above steps are thus repeated 10 times, and the overall prediction error (or overall OOB error) is the average of OOB errors in every repetition.

Moreover, we use the random forest model trained using the balanced training set to generate predictions on the test dataset, which is 30% of the dataset we set aside at the beginning of the procedure. To do this, we use the confusion matrix (illustrated in Table 4), which summarizes information about actual and predicted classification return by the classifier. Specifically, among the minority class instances (called the "positive class" in the table), those which the classifier correctly identifies as positive are called true positives (TP), whereas those wrongly identified as negative are false negatives (FN). The majority class instances (called "negative class" in the table) that are correctly identified are true negatives (TN) and, finally, those that are incorrectly predicted are false positives (FP). Among them, TP and TN are the correct predictions, while FP and FN are the incorrect predictions.

**Table 4**
Confusion matrix design.

| | Positive prediction | Negative prediction |
|---|---|---|
| Positive class | # TP | # FN |
| Negative class | # FP | # TN |

Note: True positives (TP) and true negatives (TN) are the correct predictions. False positives (FP) and false negatives (FN) are the incorrect predictions.

**Table 5**
Performance measures: metrics.

| Measures | Definition |
|---|---|
| Recall | $\frac{TP}{TP+FN}$ |
| Specificity | $\frac{TN}{TN+FP}$ |
| Informedness | $\frac{TP}{TP+FN}+\frac{TN}{TN+FP}-1$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Inverse Precision | $\frac{TN}{TN+FN}$ |
| Markedness | $\frac{TP}{TP+FP}+\frac{TN}{TN+FN}-1$ |
| $F_1$ | $\frac{2\bullet precision\bullet recall}{precision+recall}$ |
| MCC | $\frac{TP\bullet TN-FP\bullet FN}{\sqrt{(TP+FN)\bullet(FP+TN)\bullet(TP+FP)\bullet(FN+TN)}}$ |

Note: TP equals true positives; TN equals true negatives; FP equals false positives; and FN equals false negatives.

Based on the information obtained from the confusion matrix, we assess the performance of the classifier with regard to the positive and negative classes in the test dataset, computing some widespread performance metrics, which are reported in Table 5. As regards these metrics: (i) the recall or sensitivity is the true positive rate (TPR); i.e., the fraction of true positives correctly picked up; (ii) the specificity or inverse recall is the true negative rate (TNR) and is used to know how many of the negatives are correctly picked up; (iii) informedness [54] combines recall and specificity into a single metric. This indicator ranges between −1 and 1, with 1 being reached in the event of perfect classification; (iv) positive predictive value (PPV) –also known as precision– is defined as the fraction of positively predicted instances that are truly positive: in other words, the ratio between true positives and all positives (true and false positives); (v) similarly, the negative predictive value (NPV) –also known as inverse precision– shows the relationship between true and false negatives; (vi) precision and inverse precision are combined into a single metric, ranged in the interval [−1,1] and called markedness [54]; (vii) the $F_1$ measure [77] combines precision and recall into a single metric by computing their harmonic mean; and (viii) to overcome the drawbacks of previous metrics in the event of imbalanced datasets (overoptimistic inflated results), informedness and markedness are combined into a single metric by calculating their geometric mean. This last indicator refers to the Matthews correlation coefficient (MCC) [4], which is an effective solution to overcome the class imbalance problem. MCC only produces a high truthful score if the classifier obtains good results in the four confusion matrix entries.

Finally, we measure the variable importance in random forest modelling to select the most relevant explanatory variables for predicting the target variable. For this purpose, we use the permutation feature importance method, whose main idea is to calculate the importance of predictors by randomly permuting each of them in the training set and computing the reduction in the OOB prediction error. In particular, each predictor variable is shuffled in the training set and a prediction model with the shuffled dataset is implemented. The higher the reduction in predictive performance, the greater the influence the predictor has.

## 5. Empirical results

### 5.1. Application of a traditional approach

Prior to applying the random forest approach for unbalanced samples –and in order to test whether this approach has a greater predictive capacity than traditional models– we perform some logit regressions in this section. The dependent variable is whether an SME has applied or not for a public investment subsidy (defined in Section 4.2.1). As regards the independent variables –and given the impossibility of including in the model all those described in Table 3– we choose three variables that the literature has mostly associated with requesting subsidies for energy efficiency (specifically, total indebtedness, non-current assets, and cash holdings).

Results are shown in Table 6, which contains two different panels: Panel A –showing the results related to the logistic regressions performed– and Panel B –showing the results regarding performance measures defined in Table 5. As regards Panel A, Regression (I) shows the results for the logit regression, where none of the three independent variables included are significant. For its part, Regression (II) shows the same logit regression but applying the SMOTE technique (i.e., balancing the sample, as previously explained in Section 4.3). The three independent variables are now statistically significant which, a priori, indicates better predictive capacity. Specifically, total indebtedness is negatively related to the probability of requesting an energy efficiency subsidy, while non-current assets and cash flows have a significantly positive impact. In addition –and after checking the results of performance measures in Panel B– we see how the predictive power is also greater in Regression (II), i.e., the predictive power increases when the SMOTE sample balancing technique is applied in a logit regression. These results allow us to conclude that applying the SMOTE technique

**Table 6**
Logit regression results.

*Panel A. Logit regressions*

| Variables | (I)<br>Logit regression | (II)<br>SMOTE logit regression |
|---|---|---|
| Total indebtedness | −0.693 | −0.835*** |
| Non-current assets | 4.903 | 18.909*** |
| Cash holdings | 0.000 | 0.000** |
| Intercept | −2.719*** | 0.442*** |
| AIC | 499.57 | 3625.3 |
| Chi-square (*p*-value) | 0.338 | 0.000 |
| Optimal cut-off | 0.0350 | 0.559 |

*Panel B. Performance measures*

| Metrics | Logit | SMOTE Logit |
|---|---|---|
| Recall | 0.923 | 0.230 |
| Specificity | 0.188 | 0.868 |
| Informedness | 0.111 | 0.099 |
| Precision | 0.025 | 0.037 |
| Inverse Precision | 0.991 | 0.981 |
| Markedness | 0.016 | 0.018 |
| $F_1$ | 0.048 | 0.064 |
| MCC | 0.041 | 0.042 |

Note: total indebtedness is measured as the ratio of total debt over total equity and debt; non-current assets are measured as the ratio of investment property over non-current assets; and cash holdings are defined as the ratio of cash and cash equivalents over total assets.

implies an improvement in the predictive capacity of the logistic model. This is because none of the three independent variables commonly associated with requesting subsidies for energy efficiency are statistically significant when the original unbalanced sample is considered. By contrast, when the artificially balanced sample is used, the three become relevant in terms of explaining the probability of requesting an energy efficiency subsidy.

In any case, in these logit models (including the SMOTE logit) we have had to previously select the set of variables that impact the probability of an SME requesting a public investment subsidy. However –and as discussed above– such selection is extremely difficult. An alternative is therefore the random forest approach which –based on a much larger set of predictors– allows us to determine which of these have the greatest impact on our dependent variable of interest.

### 5.2. Results of applying the random forest approach for an unbalanced sample context

As mentioned above, we face a problem with our unbalanced sample since the class distribution (majority versus minority class) is severely skewed. Specifically, there are 1919 firms in the majority class instances and only 73 in the minority class. Thus, the imbalance ratio is equal to 0.03804, which is approximately a 1:27 class ratio. This unequal distribution of classes in the dataset affects the efficient channeling and use of this public money. Optimal deployment of public policies depends on the direct interaction of policymaker and potential business beneficiaries of public investment subsidies, represented by the minority class.

After splitting the dataset into a training dataset (70% of the sample) and an unseen dataset (30% of the sample), the training set consists of information related to 1396 firms, while the number of observations of the test dataset is 596 firms. In the training set, there are 1344 majority class instances and 52 minority class instances, and in the test set, 575 and 21 observations, respectively. Consequently, both sets have a similar distribution. In the training set, 96.27% of the observations correspond to firms who do not receive a subsidy and 3.73% to firms who do. The test set consists of 96.47% of firms who do not receive a subsidy and 3.53% of firms who do. The same level of imbalance is therefore maintained (0.03804 imbalance ratio).

Once our training set has been transformed –applying the SMOTE technique to balance the sample by creating new minority instances as stated in Section 4.3– the class distribution of the synthetic training set is balanced, with 1344 observations in the majority class and 1300 (original and synthetic) observations in the minority class, as a result of duplicating the size of the minority class 24 times. In other words, since the size of the minority class in the training set is equal to 52, 24*52 = 1248 synthetic observations are generated and the final size of the minority class of the training set is (24*52) + 52 = 1300. The overall prediction error (or overall OOB error) shows a value of 0.417%, such that the balanced training set model accuracy is around 99.583%.

The confusion matrix results are illustrated in Table 7. As regards the minority (or positive) class, we see how the classifier correctly classifies all firms, such that the model is able to identify 100% of the firms that would be interested in receiving an energy subsidy. With regard to the majority (or negative) class, the classifier incorrectly classifies one firm as positive, while 574 firms are correctly classified as negatives. Given our testing set, the set of potential beneficiaries is reduced from 596 firms to 22, without any false negatives and with just one false positive in the set.

**Table 7**
Confusion matrix results.

| | Positive prediction | Negative prediction |
|---|---|---|
| Positive class | 21 | 0 |
| Negative class | 1 | 574 |

Note: True positives (TP) and true negatives (TN) are the correct predictions. False positives (FP) and false negatives (FN) are the incorrect predictions.

**Table 8**
Performance measures: scores.

| Measures | Score |
|---|---|
| Recall | 1.000 |
| Specificity | 0.998 |
| Informedness | 0.998 |
| Precision | 0.955 |
| Inverse Precision | 1.000 |
| Markedness | 0.954 |
| $F_1$ | 0.977 |
| MCC | 0.976 |

The scores related to the performance metrics are reported in Table 8. Specifically, this table shows: (i) as regards recall or sensitivity, it can be seen that the classifier achieves one in the true positive rate (TPR), which means that the model is able to identify 100% of the firms that would be interested in receiving an energy subsidy; (ii) as regards the specificity or inverse recall, the model detects 99.8% of the firms that would not be interested in receiving an energy subsidy; (iii) informedness is 0.998 –very close to 1– which means both high TPR and high TNR; (iv) as regards precision, the model is very accurate, since only 4.5% of the firms detected as potential beneficiaries are in fact not; (v) inverse precision is equal to 1, which means that true negatives correspond to all negatives (true and false negatives); (vi) for its part, markedness shows a high value –close to the best possible value– meaning that both PPV and NPV are high; (vii) $F_1$ also shows a high value, equal to 0.977 in the [0,1] interval, which could be interpreted as the model generating excellent predictions; and (viii) the MCC value is equal to 0.976 in the [−1,1] interval –very close to the best value. This reflects the high predictive capacity of the model.

Finally, Fig. 4 shows the variables ranked as being more important than the others. The most useful predictors for predicting which SMEs are potential energy beneficiaries are related to liquidity and indebtedness. The predictor in the first position of the ranking is denoted by **IVR86** (liquidity), which captures the increase in investment capacity measured in relation to the average of the sector (and defined as short-term financial assets divided by net sales). For their part, **R24**, **IVR14**, **R27** and **IVR23** are related to indebtedness. The variable **R24** represents the financial debt maturity structure of a firm (calculated as short-term financial debt over long-term financial debt); **IVR14** captures the percentage of long-term-debt (maturity of over one year) which is nonfinancial (other types of debt not associated with bank loans) measured in relation to the average of the sector; **R27** captures a firm's interest and related charges on financial debt, which are requested from banks or financial institutions; and **IVR23** is the variation ratio of a firm's debt maturity structure measured in relation to the average of the sector.

### 5.3. Traditional versus random forest approaches for unbalanced samples

Table 9 shows the comparison of performance measures between the random forest approach for unbalanced samples and traditional approaches. After verifying the changes experienced therein, we see how the increase in performance measures is very pronounced (specifically, with regard to informedness, precision, markedness, and $F_1$). This comparison reflects the greater potential of using the random forest model within this context, which makes it a suitable and accurate approach for determining which predictors affect the probability of SMEs requesting a public investment subsidy.

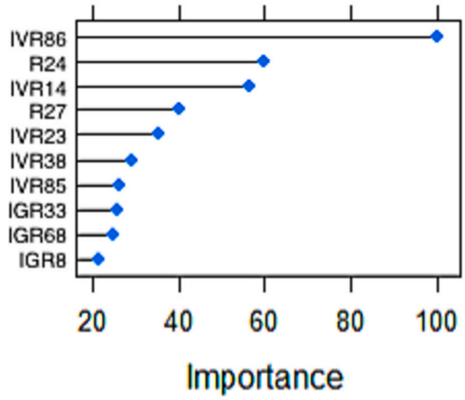

**Fig. 4.** Variables ranked as most important.
Note: **IVR86** refers to an industry variation rate of R86 (Short-term financial assets/Net sales); **R24** equals Short-term financial debt/Long-term financial debt; **IVR14** refers to an industry variation rate of R14 (Long-term non-financial debt/Long-term debt); **R27** equals Financial expenses/Financial debt; **IVR23** refers to an industry variation rate of R23 (Short-term debt/Long-term debt); **IVR38** refers to an industry variation rate of R38 (Non-current financial assets/Non-current assets); **IVR85** refers to an industry variation rate of R85 (Short-term financial assets/Short-term debt); **IGR33** refers to an industry growth rate of R33 (Financial incomes-Financial expenses)/Earnings before interest and taxes); **IGR68** refers to an industry growth rate of R68 (Current assets-inventories-trade receivables)/Short-term debt); and **IGR8** refers to an industry growth rate of R8 (Long-term financial debt/Long-term debt).

**Table 9**
Performance measures: comparisons.

| Measures | Random forest versus Logit | Random forest versus SMOTE logit |
|---|---|---|
| Recall | 0.077 | 0.77 |
| Specificity | 0.81 | 0.13 |
| Informedness | 0.887 | 0.899 |
| Precision | 0.93 | 0.918 |
| Inverse Precision | 0.009 | 0.019 |
| Markedness | 0.938 | 0.936 |
| $F_1$ | 0.929 | 0.913 |
| MCC | 0.935 | 0.934 |

Note: The figures indicated in this table refer to the difference between the performance measures regarding the random forest model for unbalanced samples and the performance measures regarding the logit models.

## 6. Discussion of the main findings

Applying a sophisticated random forest approach for unbalanced samples, this paper identifies the profile of which SMEs display the greatest potential to invest in energy efficiency measures so that governments can direct their efforts towards them, and thereby enhance the effectiveness of public policy allocation and execution. Based on this, several findings and contributions to previous literature may be highlighted and discussed, and which directly aim to cover the gaps identified. Firstly, this research provides a specific profile of which industrial SMEs are potential beneficiaries for investment in efficiency energy measures. More specifically, this profile shows the SMEs which – beyond merely meeting the legal requirements– are more likely to apply for this public aid and then make investments in energy efficiency. We find that, in general terms, aspects related to business liquidity and indebtedness are the main drivers that make SMEs in the industrial sector finally decide to apply for these public subsidies. This evidence places special emphasis on the importance of good economic-financial health, which is vital for SMEs vis-à-vis investing in energy efficiency measures, given that lack of liquidity or excessive indebtedness prevents them from undertaking projects (which chiefly affects investments in efficient energy).

Secondly, in order to predict an accurate profile it has been necessary to implement more sophisticated techniques. In fact, this research evidences the greater predictive capacity of using the random forest approach for unbalanced samples compared to traditional ones. Specifically, two key methodological challenges in this field have been addressed. On the one hand, the random forest approach [8,9] allows an accurate SME profile to be defined by handling a high volume of predictor variables and by creating several independent decision trees. This large volume of indicators (in this case, 528 indicators for each company, which implies over a million observations) does not allow the application of logit regression or discriminant analysis models in which the number of parameters would eliminate the degrees of freedom of the model. Moreover, our results allow us to affirm that applying traditional models in this field would not provide accurate results due to said models' lower predictive capacity. Added to this is the fact that the traditional model would demand prior selection of the explanatory variables, which is extremely complicated. On the other hand, a further issue is that samples in this field are fairly unbalanced (such that the proportion of SMEs who applied and who are potential energy beneficiaries is very small, even though this is the group of greatest interest to public institutions). If this lack of balance is not considered by researchers, the results obtained would be biased because the "minority class" would not really be considered in the estimation. To address this issue, sample balancing techniques are applied in this paper, thereby increasing the accuracy of both the estimation and the identification of SME profile.

The choice of the random forest model for this work is due to its innumerable advantages. On the one hand, random forest overcomes the over-fitting problem of decision trees, evidences good tolerance to noise and anomaly values, and exhibits good scalability and parallelism to the problem of high-dimensional data classification [7,8,46]. In addition, random forest is a non-parametric classification method and is data-driven. It trains classification rules by learning given samples and it does not require prior classification knowledge. On the other hand, this approach is faster than boosting, is simple, and is relatively quick to develop and can be easily parallelized [8]. In addition, it gives useful internal estimates of error, strength, correlation, and variable importance. Apart from taking into consideration these advantages, we chose random forest as the classifier because numerous articles for imbalanced data classification over the years have combined data resampling (in particular, SMOTE) and random forest in order to achieve classification goals. The balanced samples for training generated by SMOTE improve the performance of classifiers. They are used to train a random forest model to recognize the rate of samples in the minority class (e.g., [49,58,73,74]). Finally, we highlight that we have not compared the classification performance of random forest with other ensemble algorithms because of the extraordinary performance of random forest. In particular, the F-value index and MCC index used to evaluate the classification performance for the imbalanced dataset are equal to 0.977 and 0.976, respectively, as shown in Table 8.

Thirdly, identifying the profile of potential beneficiary SMEs helps governments to directly filter which SMEs might be possible beneficiaries, thereby reducing the costs (in particular, dissemination and information costs) that would be incurred by not having this profile, in addition to increasing the effectiveness of public aid allocation. Based on our evidence, public bodies should be aware of the need to specifically target their calls towards companies who are more likely to invest in energy efficiency measures. Should governments fail to pay attention to this profile and launch more generalist calls, the channeling of public subsidies becomes more difficult, which often results in a lower investment rate in energy efficiency. Beyond what are merely economic motivations, success in the design and implementation of public calls is also important in terms of boosting the commitment to sustainability, given the environmental advantages to be derived from energy efficiency measures [38]. Similar to what happens in commercial building operations, investment subsidies play a crucial role in connecting the broader goal of carbon reduction through practical, actionable steps [45]. The financial support provided by subsidies facilitates the implementation of energy-efficient technologies, thus contributing to a more sustainable and environmentally friendly approach to commercial building management.

Fourthly, although this research is focused on the specific context of the Region of Murcia, our findings can be generalized both at the country and sector level. At the country level, Spain displays a similar energy consumption across its various regions –in terms of distribution of sources– with electrical energy representing the highest cost [61]. For this reason, Spain tends to show a strong commitment to producing electrical energy from cleaner and more efficient sources [56], highlighting the relevance of investing in energy efficiency measures. Added to this is the fact that the European Union has recently agreed to end coal and gas heating by 2040, which represents an added boost at country level in terms of promoting more energy efficiency measures [31]. Going further, the Region of Murcia –with its distinctive Mediterranean climate and specific combination of socioeconomic factors– provides a representative case that resembles other European regions. This means that the results obtained here can be extrapolated to those regions which display similar features. At the sector level, the findings obtained can be generalizable to organizations that operate in the industrial sector (or similar sectors), which is where public aid has been analyzed in this article, and which represent a high proportion of both the regional and national product fabric.

Finally, this study has some limitations that may lead to future research lines. First, this study is focused on public subsidies as the main driver of investments in energy efficiency measures. Future studies may consider other drivers that alleviate other barriers (both economic and otherwise) that make it difficult for SMEs to invest in efficient energy. Secondly, this study is focused on SMEs from the Region of Murcia. Future studies could analyze to what extent the profile stated in this paper is the same for larger companies and for companies that operate in other contexts whose (political, fiscal, economic, and social) features differ slightly from those of Spain. Thirdly, future papers may delve into which specific aspects of liquidity and indebtedness within an enterprise would be valuable, thereby providing a more specific profile of SMEs that could potentially benefit from public investment subsidies. Finally, it would be interesting to explore to what extent those businesses who are potential beneficiaries of these public subsidies ultimately do actually request them and invest in said support to achieve energy improvements in their organizations. To do this, it may be useful to examine the motivations, implications, and impact that such public subsidies have on organizations.

## 7. Concluding remarks

Implementing efficient energy measures should not only be encouraged for environmental reasons but also because of the benefits in terms of business competitiveness [38], increased cost saving, and performance. Despite these motivations, many companies (especially SMEs) are still failing to implement energy efficiency measures [15,42] due to the barriers that prevent them from undertaking these investments [65]. In an effort to counteract these barriers, public institutions often define and implement public programs that seek to boost the attractiveness and appeal of investments in energy efficiency measures [55]. Specifically, public investment subsidies are one of the most prominent instruments to counter these economic-financial barriers and so encourage companies to make such investments [12,14,21,26]. Bearing this in mind, this paper seeks to identify the profile of which SMEs display the greatest potential to invest in energy efficiency measures so that governments can direct their efforts towards such SMEs and thereby enhance the effective allocation and execution of these public policies. Based on a sample of 1992 SMEs from the Region of Murcia (Spain) over the period 2013–2018, this paper uses a random forest approach and techniques for unbalanced samples to identify SME profile. Results show that the most useful predictors for predicting which SMEs in the industrial sector are potential energy beneficiaries are those related to liquidity and indebtedness.

## Funding information

This work was supported by the Fundación Cajamurcia.

## CRediT authorship contribution statement

**Susana Álvarez-Diez:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **J. Samuel Baixauli-Soler:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Gabriel Lozano-Reina:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Diego Rodríguez-Linares Rey:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

## Data availability

The data used in this study are available upon request from the corresponding author.

## Appendix A. Energy consumption based on the NACE Rev. 2 classification (2019)

| NACE Rev 2 Codes | | Electricity | Gas | Diesel | Fuel oil | Coal and coke | Biofuels | Heat and other energy | Total energy consumption |
|---|---|---|---|---|---|---|---|---|---|
| 07 | Mining of metal ores | 44,697 | 639 | 5504 | 674 | 4920 | 0 | 447 | 56,881 |
| 08 | Other mining and quarrying | 73,906 | 28,482 | 65,699 | 2989 | 0 | 0 | 436 | 175,998 |
| 10 | Manufacture of food products | 1,073,429 | 509,278 | 188,791 | 9344 | 781 | 13,569 | 15,631 | 1,828,614 |
| 11 | Manufacture of beverages | 133,368 | 57,609 | 28,180 | 25,025 | 0 | 2203 | 25 | 249,320 |
| 13 | Manufacture of textiles | 101,661 | 52,398 | 5556 | 48 | 0 | 282 | 149 | 160,623 |
| 14 | Manufacture of wearing apparel | 13,978 | 1656 | 2825 | 18 | 0 | 0 | 40 | 18,594 |
| 15 | Manufacture of leather and related products | 21,160 | 6391 | 3087 | 272 | 0 | 0 | 0 | 31,296 |
| 16 | Manufacture of wood and of products of wood and cork | 133,321 | 10,535 | 27,533 | 1013 | 0 | 4101 | 559 | 177,762 |
| 17 | Manufacture of paper and paper products | 399,019 | 231,197 | 13,576 | 14,389 | 0 | 4854 | 16,216 | 681,805 |
| 18 | Printing and reproduction of recorded media | 66,381 | 11,069 | 6120 | 0 | 0 | 0 | 46 | 84,221 |
| 19 | Manufacture of coke and refined petroleum products | 280,493 | 452,498 | 305 | 0 | 0 | 0 | 113,275 | 846,631 |
| 20 | Manufacture of chemicals and chemical products | 720,192 | 589,242 | 31,076 | 4280 | 7 | 3108 | 194,772 | 1,581,572 |
| 21 | Manufacture of basic pharmaceutical products and pharmaceutical preparations | 115,367 | 41,979 | 3760 | 5604 | 0 | 0 | 1541 | 169,302 |
| 22 | Manufacture of rubber and plastic products | 424,717 | 48,921 | 22,289 | 3798 | 0 | 214 | 18,666 | 519,741 |
| 23 | Manufacture of other non-metallic mineral products | 566,444 | 626,376 | 89,815 | 14,525 | 87,924 | 15,581 | 11,387 | 1,416,290 |
| 24 | Manufacture of basic metals | 1,113,460 | 441,515 | 14,311 | 11,061 | 10,974 | 7 | 78,333 | 1,672,215 |
| 25 | Manufacture of fabricated metal products, except machinery and equipment | 307,052 | 98,754 | 50,213 | 2836 | 24 | 393 | 5428 | 472,333 |
| 26 | Manufacture of computer, electronic and optical products | 24,258 | 2043 | 2122 | 131 | 0 | 4 | 173 | 29,101 |
| 27 | Manufacture of electrical equipment | 120,849 | 17,188 | 7625 | 365 | 0 | 0 | 1501 | 148,699 |
| 28 | Manufacture of machinery and equipment | 90,919 | 17,785 | 23,241 | 515 | 0 | 51 | 463 | 135,998 |
| 29 | Manufacture of motor vehicles, trailers, and semi-trailers | 372,113 | 100,795 | 23,025 | 1692 | 0 | 246 | 714 | 501,278 |
| 30 | Manufacture of other transport equipment | 69,192 | 14,739 | 12,639 | 0 | 0 | 1 | 966 | 100,504 |
| 31 | Manufacture of furniture | 38,619 | 3442 | 17,640 | 50 | 0 | 111 | 26 | 60,048 |
| 32 | Other manufacturing | 22,495 | 2510 | 2312 | 0 | 0 | 0 | 101 | 27,761 |
| 33 | Repair and installation of machinery and equipment | 23,719 | 4273 | 24,560 | 109 | 0 | 0 | 91 | 55,426 |
| **Total** | | 6,350,809 | 3371,314 | 671,804 | 98,738 | 104,630 | 44,725 | 460,986 | 11,202,013 |

Note: The figures are expressed in thousands of euros. Some NACE Rev. 2 codes that are part of this study are not included in this table because they do not report data.
Source: Spanish National Statistics Institute.

## References

[1] Aditsania A, Adiwijaya, Saonard AL. Handling imbalanced data in churn prediction using ADASYN and backpropagation algorithm. In: Riza LS, Pranolo A, Wibawa AP, Junaeti E, Wihardi Y, Hashim UR, et al., editors. 2017 3rd International Conference on Science in Information Technology (ICSITECH); 2017. p. 533–6.

[2] Allcott H, Greenstone M. Is there an energy efficiency gap? J Econ Perspect 2012;26(1):3–28. https://doi.org/10.1257/jep.26.1.3.

[3] Backlund S, Thollander P, Palm J, Ottosson M. Extending the energy efficiency gap. Energy Policy 2012;51:392–6. https://doi.org/10.1016/j.enpol.2012.08.042.

[4] Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16(5):412–24. https://doi.org/10.1093/bioinformatics/16.5.412.

[5] Barandela R, Sanchez JS, Garcia V, Rangel E. Strategies for learning in class imbalance problems. Patt Recognit 2003;36(3):849–51. https://doi.org/10.1016/S0031-3203(02)00257-1.

[6] Bertoldi P, Rezessy S, Oikonomou V. Rewarding energy savings rather than energy efficiency: exploring the concept of a feed-in tariff for energy savings. Energy Policy 2013;56:526–35. https://doi.org/10.1016/j.enpol.2013.01.019.

[7] Boinee P, De Angelis A, Foresti GL. Meta random forests. Int J Comp Intellig 2005;2(3):138–47.

[8] Breiman L. Random Forests. Machine Learn 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.

[9] Breiman L. Consistency for a simple model of random forests. Technical report. University of California at Berkeley; 2004.

[10] Brown MA. Market failures and barriers as a basis for clean energy policies. Energy Policy 2001;29(14):1197–207. https://doi.org/10.1016/S0301-4215(01)00067-2.

[11] Brownlee J. Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery; 2020.

[12] Brunke J-C, Johansson M, Thollander P. Empirical investigation of barriers and drivers to the adoption of energy conservation measures, energy management practices and energy services in the Swedish iron and steel industry. J Clean Prod 2014;84:509–25. https://doi.org/10.1016/j.jclepro.2014.04.078.

[13] Cagno E, Trianni A. Exploring drivers for energy efficiency within small- and medium-sized enterprises: first evidences from Italian manufacturing enterprises. Appl Energy 2013;104:276–85. https://doi.org/10.1016/j.apenergy.2012.10.053.

[14] Cagno E, Trianni A, Abeelen C, Worrell E, Miggiano F. Barriers and drivers for energy efficiency: different perspectives from an exploratory study in the Netherlands. Energ Conver Manage 2015;102:26–38. https://doi.org/10.1016/j.enconman.2015.04.018.

[15] Cagno E, Trianni A, Spallina G, Marchesani F. Drivers for energy efficiency and their effect on barriers: empirical evidence from Italian manufacturing enterprises. Energ Effic 2017;10(4):855–69. https://doi.org/10.1007/s12053-016-9488-x.

[16] Cagno E, Trucco P, Trianni A, Sala G. Quick-E-scan: A methodology for the energy scan of SMEs. Energy 2010;35(5):1916–26. https://doi.org/10.1016/j.energy.2010.01.003.

[17] Cagno E, Worrell E, Trianni A, Pugliese G. A novel approach for barriers to industrial energy efficiency. Renew Sustain Energy Rev 2013;19:290–308. https://doi.org/10.1016/j.rser.2012.11.007.

[18] Carlander J, Thollander P. Barriers to implementation of energy-efficient technologies in building construction projects — results from a Swedish case study. Resourc Environ Sustain 2023;11:100097. https://doi.org/10.1016/j.resenv.2022.100097.

[19] CARM. Orden de la Consejería de Empleo, Universidades, Empresa y Medio Ambiente, por la que se establecen las bases reguladoras del Programa de ayudas para actuaciones de eficiencia energética en PYME y gran empresa del sector industrial. https://www.borm.es/services/anuncio/ano/2019/numero/5090/pdf?id=779188; 2019.

[20] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.

[21] de Groot HLF, Verhoef ET, Nijkamp P. Energy saving by firms: decision-making, barriers and policies. Energy Econ 2001;23(6):717–40. https://doi.org/10.1016/S0140-9883(01)00083-4.

[22] Dwivedi C. Influence of production and investment tax credit on renewable energy growth and power grid. In: 2018 IEEE Green Technologies Conference (GreenTech); 2018. p. 149–54. https://doi.org/10.1109/GreenTech.2018.00035.

[23] Estabrooks A, Jo TH, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Comp Intellig 2004;20(1):18–36. https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x.

[24] European Commission. Commission Regulation (EU) N$^{o}$ 651/2014, of 17 June 2014, declaring certain categories of aid compatible with the internal market in application of Articles 107 and 108 of the Treaty. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32014R0651&from=es; 2014.

[25] Fernández A, del Río S, Chawla NV, Herrera F. An insight into imbalanced Big Data classification: outcomes and challenges. Complex Intellig Syst 2017;3(2):105–20. https://doi.org/10.1007/s40747-017-0037-9.

[26] Fleiter T, Schleich J, Ravivanpong P. Adoption of energy-efficiency measures in SMEs—an empirical analysis based on energy audit data from Germany. Energy Policy 2012;51:863–75. https://doi.org/10.1016/j.enpol.2012.09.041.

[27] Fleiter T, Worrell E, Eichhammer W. Barriers to energy efficiency in industrial bottom-up energy demand models—a review. Renew Sustain Energy Rev 2011;15(6):3099–111. https://doi.org/10.1016/j.rser.2011.03.025.

[28] García S, Luengo J, Herrera F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge-Based Syst 2016;98:1–29. https://doi.org/10.1016/j.knosys.2015.12.006.

[29] Garcia V, Sanchez JS, Mollineda RA. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowledge-Based Syst 2012;25(1, SI):13–21. https://doi.org/10.1016/j.knosys.2011.06.013.

[30] Gillingham K, Newell RG, Palmer K. Energy Efficiency Economics and Policy. Ann Rev Resource Econom 2009;1(1):597–620. https://doi.org/10.1146/annurev.resource.102308.124234.

[31] González J. Europa pacta el fin de las calefacciones de carbón y gas para el año 2040. La Verdad 2023. https://www.laverdad.es/sociedad/europa-pacta-fin-calefacciones-carbon-gas-ano-20231209072831-nt.html.

[32] Gosain A, Sardana S. Handling class imbalance problem using oversampling techniques: a review. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2017. p. 79–85.

[33] Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: review of methods and applications. Exp Syst Appl 2017;73:220–39. https://doi.org/10.1016/j.eswa.2016.12.035.

[34] Harris J, Anderson J, Shafron W. Investment in energy efficiency: a survey of Australian firms. Energy Policy 2000;28(12):867–76. https://doi.org/10.1016/S0301-4215(00)00075-6.

[35] He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks. 1-8; 2008. p. 1322–8. https://doi.org/10.1109/IJCNN.2008.4633969.

[36] Hirst E, Brown M. Closing the efficiency gap: barriers to the efficient use of energy. Resources Conserv Recycl 1990;3(4):267–81. https://doi.org/10.1016/0921-3449(90)90023-W.

[37] Holidu. The Sunniest Cities in Europe. Last access: 09-Feb-2023, https://www.holidu.co.uk/magazine/the-sunniest-cities-in-europe; 2022.

[38] Hrovatin N, Cagno E, Dolšak J, Zorić J. How important are perceived barriers and drivers versus other contextual factors for the adoption of energy efficiency measures: an empirical investigation in manufacturing SMEs. J Clean Prod 2021;323:129123. https://doi.org/10.1016/j.jclepro.2021.129123.

[39] Jaffe AB, Stavins RN. The energy-efficiency gap: What does it mean? Energy Policy 1994;22(10):804–10. https://doi.org/10.1016/0301-4215(94)90138-4.

[40] Jaffe AB, Stavins RN. The energy paradox and the diffusion of conservation technology. Resource Energy Econom 1994;16(2):91–122. https://doi.org/10.1016/0928-7655(94)90001-9.

[41] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intellig Data Analys 2002;6(5):429–49.

[42] Kostka G, Moslener U, Andreas J. Barriers to increasing energy efficiency: evidence from small-and medium-sized enterprises in China. J Clean Prod 2013;57:59–68. https://doi.org/10.1016/j.jclepro.2013.06.025.

[43] Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progr Artific Intellig 2016;5(4):221–32. https://doi.org/10.1007/s13748-016-0094-0.

[44] Kuhn M, Johnson K. Data Pre-processing BT - Applied Predictive Modeling, M. Kuhn & K. Johnson (eds.). New York: Springer; 2013. p. 27–59. https://doi.org/10.1007/978-1-4614-6849-3_3.

[45] Li K, Ma M, Xiang X, Feng W, Ma Z, Cai W, et al. Carbon reduction in commercial building operations: A provincial retrospection in China. Appl Energy 2022;306:118098. https://doi.org/10.1016/j.apenergy.2021.118098.

[46] Liaw A, Wiener M. Classification and regression by random forest. R News 2002;2(3):18–22.

[47] Lin W-J, Chen JJ. Class-imbalanced classifiers for high-dimensional data. Brief Bioinform 2013;14(1):13–26. https://doi.org/10.1093/bib/bbs006.

[48] Lopez V, Fernandez A, Garcia S, Palade V, Herrera F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inform Sci 2013;250:113–41. https://doi.org/10.1016/j.ins.2013.07.007.

[49] Ma L, Fan S. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC Bioinform 2017;18(1):169. https://doi.org/10.1186/s12859-017-1578-z.

[50] Moniz N, Branco P, Torgo L. Evaluation of ensemble methods in imbalanced regression tasks. In: Proceedings of Machine Learning Research. 74; 2017. p. 129–40.

[51] Nehler T, Parra R, Thollander P. Implementation of energy efficiency measures in compressed air systems: barriers, drivers and non-energy benefits. Energ Effic 2018;11(5):1281–302. https://doi.org/10.1007/s12053-018-9647-3.

[52] Nie P-Y, Wang C, Yang Y-C. Comparison of energy efficiency subsidies under market power. Energy Policy 2017;110:144–9. https://doi.org/10.1016/j.enpol.2017.07.053.

[53] Nimankar SS, Vora D. Designing a Model to Handle Imbalance Data Classification Using SMOTE and Optimized Classifier. In: Sharma N, Chakrabarti A, Balas VE, Martinovic J, editors. Data Management, Analytics and Innovation. Singapore: Springer; 2021. p. 323–34.

[54] Powers D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Int J Mach Learn Technol 2011;2(1):37–63.

[55] Prasad Painuly J. Financing energy efficiency: lessons from experiences in India and China. Int J Energy Sector Manag 2009;3(3):293–307. https://doi.org/10.1108/17506220910986815.

[56] Red Eléctrica. España es el segundo país europeo que más energía eléctrica generó con eólica y solar en 2021. https://www.ree.es/es/sala-de-prensa/actualidad/nota-de-prensa/2022/06/espana-es-el-segundo-pais-europeo-que-mas-energia-electrica-genero-con-eolica-y-solar-en-2021; 2022.

[57] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS One 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432.

[58] Sholihah NN, Hermawan A. Implementation of random forest and smote methods for economic status classification in Cirebon City. Jurnal Teknik Informatika (Jutif) 2023;4(6 SE-Articles):1387–97. https://doi.org/10.52436/1.jutif.2023.4.6.1135.

[59] Sorrell S, Schleich J, O'Malley E, Scott S. The Economics of Energy Efficiency: Barriers to Cost-Effective Investment. Edward Elgar; 2004.

[60] Soto S. ¿En qué zona del mapa de radiación de España está mi provincia? Roams Energía 2023. https://energia.roams.es/energia-renovable/energia-solar/radiacion-solar-espana/.

[61] Spanish Statistics National Institute. Energy consumption survey. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736146240&menu=ultiDatos&idp=1254735576715; 2019.

[62] Sudhakara Reddy B. Barriers and drivers to energy efficiency – a new taxonomical approach. Energ Conver Manage 2013;74:403–16. https://doi.org/10.1016/j.enconman.2013.06.040.

[63] Sun Y, Kamel MS, Wong AKC, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Patt Recognit 2007;40(12):3358–78. https://doi.org/10.1016/j.patcog.2007.04.009.

[64] Tallo TE, Musdholifah A. The implementation of genetic algorithm in smote (synthetic minority oversampling technique) for handling imbalanced dataset problem. In: 2018 4th International Conference on Science and Technology (ICST); 2018. p. 1–4. https://doi.org/10.1109/ICSTC.2018.8528591.

[65] Thollander P, Backlund S, Trianni A, Cagno E. Beyond barriers: A case study on driving forces for improved energy efficiency in the foundry industries in Finland, France, Germany, Italy, Poland, Spain, and Sweden. Appl Energy 2013;111: 636–43. https://doi.org/10.1016/j.apenergy.2013.05.036.

[66] Thollander P, Danestig M, Rohdin P. Energy policies for increased industrial energy efficiency: Evaluation of a local energy programme for manufacturing SMEs. Energy Policy 2007;35(11):5774–83. https://doi.org/10.1016/j.enpol.2007.06.013.

[67] Trianni A, Cagno E. Dealing with barriers to energy efficiency and SMEs: Some empirical evidences. Energy 2012;37(1):494–504. https://doi.org/10.1016/j.energy.2011.11.005.

[68] Trianni A, Cagno E, Farné S. Barriers, drivers and decision-making process for industrial energy efficiency: A broad study among manufacturing small and medium-sized enterprises. Appl Energy 2016;162:1537–51. https://doi.org/10.1016/j.apenergy.2015.02.078.

[69] Trianni A, Cagno E, Marchesani F, Spallina G. Drivers for industrial energy efficiency: an innovative approach. In: 5th International Conference on Applied Energy (ICAE); 2013.

[70] Trianni A, Cagno E, Marchesani F, Spallina G. Classification of drivers for industrial energy efficiency and their effect on the barriers affecting the investment decision-making process. Energ Effic 2017;10(1):199–215. https://doi.org/10.1007/s12053-016-9455-6.

[71] Vora D, Iyer K. Evaluating the effectiveness of machine learning algorithms in predictive modelling. J Eng Technol 2018;7(3.4):197–9. https://doi.org/10.14419/ijet.v7i3.4.16773.

[72] Wang S, Yao X. Diversity analysis on imbalanced data sets by using ensemble models. In: 2009 IEEE Symposium on Computational Intelligence and Data Mining; 2009. p. 324–31. https://doi.org/10.1109/CIDM.2009.4938667.

[73] Wu T, Fan H, Zhu H, You C, Zhou H, Huang X. Intrusion detection system combined enhanced random forest with SMOTE algorithm. EURASIP J Adv Sign Proc 2022;2022(1):39. https://doi.org/10.1186/s13634-022-00871-6.

[74] Xu Z, Shen D, Nie T, Kou Y. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. J Biomed Inform 2020; 107:103465. https://doi.org/10.1016/j.jbi.2020.103465.

[75] Yang X, He L, Xia Y, Chen Y. Effect of government subsidies on renewable energy investments: The threshold effect. Energy Policy 2019;132:156–66. https://doi.org/10.1016/j.enpol.2019.05.039.

[76] Zhao L, Chau KY, Tran TK, Sadiq M, Xuyen NTM, Phan TTH. Enhancing green economic recovery through green bonds financing and energy efficiency investments. Econom Analys Pol 2022;76:488–501. https://doi.org/10.1016/j.eap.2022.08.019.

[77] Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white-matter lesions in MR-images: Method and validation. IEEE Trans Med Imaging 1994;13(4):716–24. https://doi.org/10.1109/42.363096.

[78] Zong W, Huang G-B, Chen Y. Weighted extreme learning machine for imbalance learning. Neurocomputing 2013;101:229–42. https://doi.org/10.1016/j.neucom.2012.08.010.