


Predicting the Success of a Startup in Information Technology Through Machine Learning

Edilberto Vasquez, AI Group, Universidad Nacional Mayor de San Marcos, Peru

José Santisteban, AI Group, Universidad Nacional Mayor de San Marcos, Peru*

David Mauricio, AI Group, Universidad Nacional Mayor de San Marcos, Peru

 <https://orcid.org/0000-0001-9262-626X>

ABSTRACT

Predicting the success of a startup in information technology (SIT) is a very complex problem due to the diverse factors and uncertainty that affects it. The focus of automatic learning (ML) is promising because it presents good results for prediction issues; however, it presents a diversity of parameters, factors, and data that require consideration to improve prediction results. In this study, a systematic method is proposed to build a predictive model for SIT success, based on factors. The method consists of four processes, a hybrid model, and an inventory of 79 success factors. The method was applied to a database of 265 SITs from Australia with seven ML algorithms and three hybrid models based on the Voting strategy and the GreedyStepwise algorithm to reduce the factors. On average, precision increments in 11.69%, specificity in 3.25%, and accuracy in 21.75%; the prediction has precision of 82% and accuracy of 88%.

KEYWORDS

Critical Success Factors, Forecast, Machine Learning, Startups

INTRODUCTION

Predicting the Success of a Startup in Information Technology Through Machine Learning

A technology-based startup is defined as the grouping of people around an innovative technology-based idea with a replicable and scalable business model (Nadežda et al., 2019); it is an innovative venture that provides solutions to emerging problems or creates new demands by developing new forms of business (OECD, 2005). It is widely established that entrepreneurship is important for the wealth and economic growth of countries (Cabrera & Mauricio, 2017). In this regard, the importance of startups in information technology (SITs) lies in the revitalization of economies, directly impacting the creation of jobs, products and/or services with high added value. Moreover, various World Bank studies show that emerging technological companies, in 2017, contributed more than 5% of the gross

DOI: 10.4018/IJITWE.323657

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

domestic product in developed countries, moving a business of 42,300 million euros in turnover, compared to 3.1%, 34,900 million, from the previous year (World Bank, 2018); likewise, financing for these SITs, by 2021, surpassed 600,000 million dollars (Jurgens, 2022). Despite the importance of startups, still eight out of 10 ventures fail in less than five years, that is, they do not reach success (Bernard & Tariskova, 2017; Honorine & Emmanuelle, 2019). To improve this alarming situation, various efforts are being developed, including management models and indicators (Gbadegeshin et al., 2022; Satyanarayana, et al., 2021), critical factors of success (Santisteban & Mauricio, 2017), promotion policies (Horne & Fichter, 2022), and the extension of financing. In general, both the government and the private sector need to estimate the future success of a venture to direct resources and minimize risks.

The methods to predict the success of a SIT can be classified into statistical methods and machine learning (ML) methods. The studies that use statistical methods are based on logistic regression and a heuristic solution approach, with 74.8% precision as their best result (Asmoro et al., 2018). On the other hand, studies based on ML, in general, obtain better results, reaching their best precision of 89% through Extreme Gradient Boosting (XGBost) and k-nearest neighbors (KNN) (Ross et al., 2021). These results show that there are still efforts to be made in improving precision, but this depends on several elements, such as factors, data set, preprocessing method, and ML method (Krishna et al., 2016; Ross et al. 2021; Tomy & Pardede, 2018); all this deserves the building of a method to obtain a predictive model and thus achieve the best results in the prediction.

In this study, the authors propose a method to build an ML-based predictive model to predict the success of a SIT, which considers the processes of data extraction, preprocessing, and prediction. The main contributions of this paper are:

- Providing a systematic method to build an ML model to predict the success of a SIT that is applicable to any scenario.
- Showing the usability of the proposed model through its application to build nine ML models, two of them hybrid for a data set of 256 SITs.

RELATED WORK

Many definitions exist for the success of a SIT. Martens et al. (2011) defined success as the growth in sales and good profitability, while Elhedhli et al. (2014) defined it as good financial performance. Santisteban, Mauricio et al. (2021) compiled nine definitions for success:

Successful startups satisfy the demands of clients and organizations, have greater benefits than other companies in the same industrial sector, are acquired by another company for a price higher than their value, and have a market value higher than its base value. (p. 401)

The success of a SIT depends, mainly, on the elements that condition its success or failure, which are termed critical factors (Abou-Moghli & Al-Kasasbeh, 2012). Since 1984, scholars have conducted studies to identify these factors, analyze their incidence, and understand their relationships between each other. Some of these factors are entrepreneurs' motivation (Van de ven et al., 1984), technology (Alvarez & Barney, 2001), and customer satisfaction (Santisteban, Mauricio et al., 2021). Santisteban and Mauricio (2017) carried out a literature review and identified 21 factors, Santisteban, Mauricio et al. (2021) proposed 10 factors, and Santisteban, Inche et al. (2021) studied the influence of 27 factors on the lifecycle of a SIT.

The first studies on predicting business success date back to over two decades ago and are based on a wide variety of techniques; among them, the Markov chain method (Back et al., 1996), a multiagent system based on case based reasoning for Spanish companies (Borrajó et al., 2011), and

regression for small companies in Chile (Helabi & Lussier, 2014). On the other hand, Martens et al. (2011) were the first to apply ML; they used support vector machine (SVM) and achieved a precision of 65% for Spanish SITs; Table 1 shows this and another five studies based on ML.

Moreover, there are other approaches to predict SIT success, such as the Benders decomposition method and a heuristic based on Tabu search that reaches 82% precision with data from Canada (Elhedhli et al., 2014), and the method based on structural equations of 74.8% precision with data from Indonesia (Asmoro et al., 2018).

METHOD AND MODEL

Method

The authors propose a systematic method to generate a predictive ML model for SIT success, which considers four related sequential processes: Factor selection, data extraction, preprocessing, and learning. The method starts the process with factor selection, from which data are extracted from various sources (e.g., incubators, tax agencies, financial organizations, investment funds, and surveys); next, the data are preprocessed, and, with this, the learning process that has the ML predictive model as its output is conducted (Figure 1).

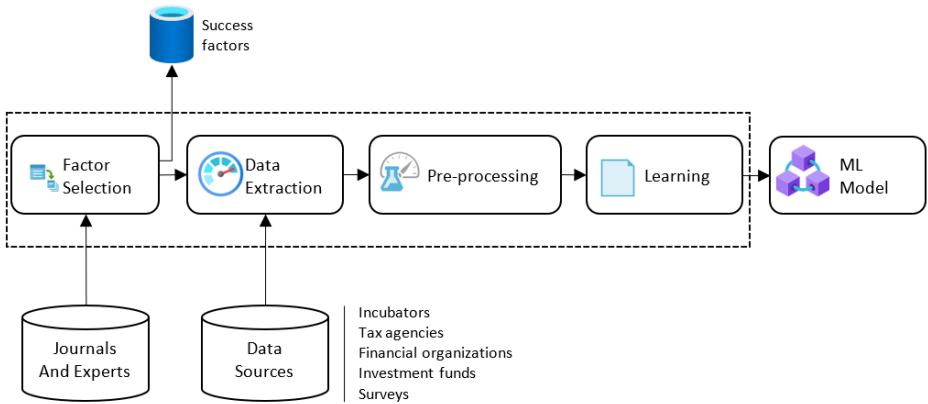
Factor Selection

Factors are characteristics that determine the success or failure of a startup (Ko & An, 2019). Factors that influence success are selected considering the study's context (since the factors may correspond to different realities and periods, they may not be valid) and the availability of significant data. Some techniques for factor selection are principal component analysis (Shlens, 2014), fast Fourier transform (Chandrashekar & Sahin, 2014), GreedyStepwise algorithm (Pourhashemi, & Mashalizadeh, 2013), forward algorithm, and backward algorithm.

Table 1. Studies on predicting SIT success based on ML

Author	Data set	Algorithm	Precision (%)
Martens et al. (2011)	218 records, Spain	SVM	65
Krishna et al. (2016)	11,000 records, India	Lazy lb1 Random forest (RF) Naïve Bayes (NB) Decision tree (DT) Simple logistics BayesNet	69.0 96.3 89.0 94.0 94.0 91.0
Antretter et al. (2018)	542 records	Text Datamining	91
Tomy & Pardede (2018)	265 records, Australia	NB KNN SVM	87.87 82.35 87.09
Ross et al. (2021)	942,605 records, USA	MLP RF XGBoost KNN	80.00 88.00 89.00 89.00
Akhavan et al. (2021)	266 records, Iran	Bayes network	61.13

Figure 1. Method to generate a predictive ML model for a SIT's success



Data Extraction

This process consists in obtaining reliable data on the selected factors of many SITs (with or without success) from various sources, such as company incubators, investment funds, government programs, tax agencies, and surveys. Data sets of the literature can also be used if they are close to the study's context.

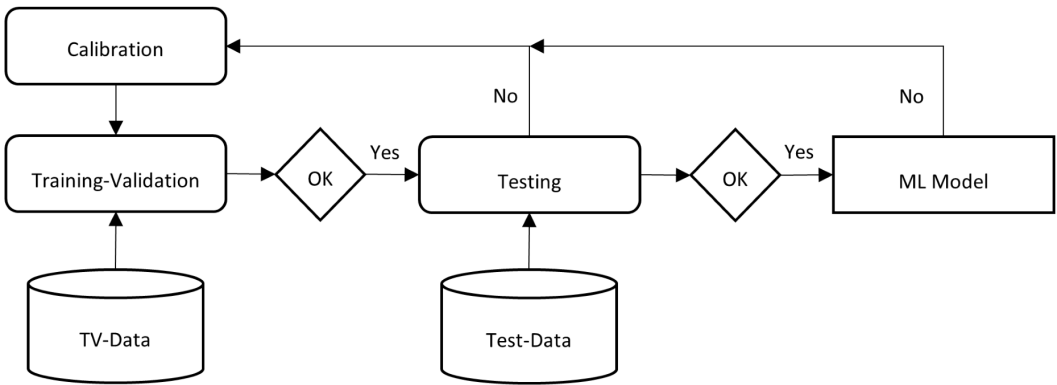
Preprocessing

This process consists in transforming the data obtained during the data extraction process into data that can be used by ML models and algorithms. For this purpose, there are various subprocesses, such as data cleansing (Corrales et al., 2020), value imputation (Useche & Mesa, 2006), categorization (Kampen, 2019), normalization (Singh, 2019), and data balancing (Kamiran & Calders, 2012).

Learning

This process receives the preprocessed data as input and, through three subprocesses (i.e., training-validation, calibration, and testing), generates a predictive ML model for SIT success (Figure 2). The preprocessed data are divided into two: TV-data and test-data. TV-data are used by the training-validation subprocess, in which an ML algorithm (Table 1) is trained with much of these data and is later validated (e.g., cross-validation) with the remaining data. If the obtained result is satisfactory,

Figure 2. Learning process



it goes to the testing subprocess; otherwise, the calibration subprocess is conducted and the process is repeated. During testing, the model obtained in the previous process is tested with test-data; if the result is satisfactory, then a predictive model is obtained; otherwise, the calibration subprocess is conducted and returns to the training-validation subprocess. Calibration consists in adjusting the hyperparameters of the ML algorithm, a process that is automated in many libraries for ML.

The SSFM Model

Based on the ML obtained by the proposed method, the authors propose a model to predict the success of a startup (SSFM), which considers three processes: Data extraction, preprocessing, and prediction. The processes are interrelated and follow an execution sequence (Figure 3).

The model receives a SIT to be evaluated; then, through a data extraction process, the data associated with the success factor obtained by the method are extracted from various sources. These data go through a preprocessing phase to obtain consistent and adequate data that are forwarded to the prediction process, which uses the ML model obtained by the method and predicts the success or failure of the SIT. This result is important for the analysis and decision-making by the agents, such as the incubator, investment fund, investors, and government programs for the venture and the entrepreneur.

The data extraction and preprocessing processes are the same given in the method, with the difference that the data corresponds to the SIT to be evaluated. For the prediction process, the authors considered a hybrid model (Figure 4), which consists in an odd number of ML models (ML_1, ML_2, \dots, ML_k) obtained by the method of applying “k” times to different ML algorithms, and a decision strategy, such as voting, that is, the result that most models present. This strategy generally presents better results than ML models by separate; this is explained because the error probability of a hybrid model is reduced as the results of most ML models coincide.

VALIDATION

To validate the method, the authors applied it to a public data set to generate a predictive ML model, and, with this, they implemented the SSFM model through a Web application.

Application of the Method

The authors considered a public data set used in some studies, such as in Tomy and Pardede’s (2018) work, for which the application of the method is reduced to the preprocessing and learning processes

Figure 3. The SSFM model

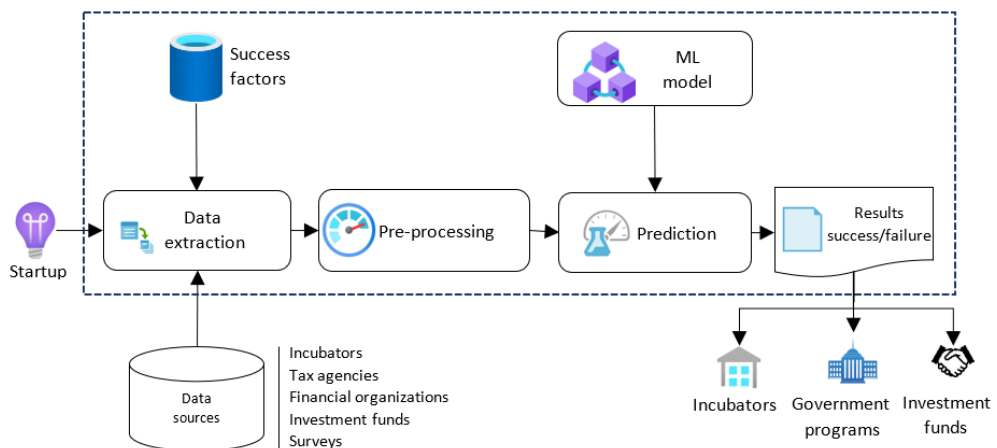
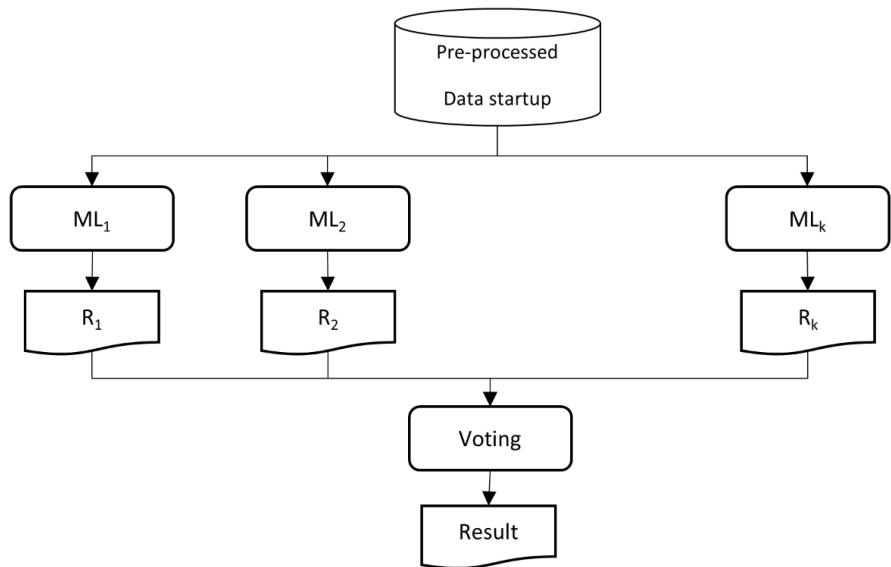


Figure 4. Hybrid ML Model



since the factor selection and data extraction processes are given by the factors and the data that the data set contemplates, respectively.

The authors considered the data set from the ICT Industry Statistics Survey of Australia (Tomy & Pardede, 2018), which has 2013 data on 265 SITs from Victoria, Australia, with one record per company. Each record has 63 data on 23 factors from the literature and other attributes, and 182 SITs are labelled as success, 80 as failure, and 3 have no information. Software development and installation, systems analysis and computer programming, and computer software consulting companies stand out in the data set, with these being 53% of the total and their success considered as profitable.

The authors conducted a preprocessing of the data set considering the following: Data cleansing, value imputation, categorization, normalization, and balancing done manually. In the data cleansing, the authors identified 27 columns with more than 50% of blank data, which they discarded; moreover, the researchers discarded 16 columns due to attribute similarity, that is, they reduced the number of columns to 20, where each corresponds to a factor. Furthermore, the authors identified three records without success or failure labels and 14 with more than 50% of blank data, which they eliminated, leaving 248 valid records. Additionally, due to the diversity of attributes that the selected factors present, the researchers classified them according to is the elements in Table 2.

In 10 records, the authors completed the data with the average integer value of its corresponding factor (value imputation). Then, the researchers balanced the data through oversampling to obtain the same number of records for success or failure, leaving the data set with 342 records (whose values are in the note of Table 2 and whose characteristics are in Table 3); they separated 10% (34 records) of the total records for testing and used 90% (308 records) of them for training-validation.

The authors applied the learning process to ML algorithms, SVM, RF, multilayer perceptron (MLP), NB, DT, KNN, and Gradient Boosting (GB), with 10-fold cross-validation and with three metrics (Table 4).

The authors conducted experiments with each of the algorithms to obtain the values of the parameters that produce the best results for the prediction. For SVM the researchers made, tests with PUK, Linear, Polynomial, and RBF kernels; for PUK, they considered the values $\omega = 0.2, 0.4, 0.6, 0.8, 1.00$, $\sigma = 0.2, 0.4, 0.6, 0.8, 1.00$. For MLP, they varied the solver function with Adam, sgd, and

Table 2. Factors and classes of selected attributes

ID	Factor	Class	ID	Factor	Class
F1	Location	A	F11	Financial capital	I
F2	Age	B	F12	R&D	I
F3	Startup size	C	F13	Availability of infrastructure	I
F4	Amount employee skills	D	F14	Innovation environment	I
F5	Company revenue	E	F15	Government regulation	I
F6	Export products	F	F16	Access to target market	I
F7	Innovation of product/ service	G	F17	Global economic environment	I
F8	Size of investment	H	F18	Exchange rates	I
F9	Environment	I	F19	Competition	I
F10	Availability of skilled employees	I	F20	Access to export market	I

Note. Class and percentage:

A: 1 = Metropolitan (77%); 2 = Regional (10%); 3 = Interstate (9%); 4 = Overseas (4%).

B: 1 = Less than 2 years (11%); 2 = 2–4 years (10%); 3 = 5–9 years (28%); 4 = 10–19 years (30%); 5 = 20+ years (22%).

C (employees): 1 = 1–5 (34%); 2 = 6–20 (34%); 3 = 21–100 (21%); 4 = 101–1000 (7%); 5 = 1000+ (4%).

D (number of skills): 1 (61%), 2 (20%), 3 (10%), 4 (5%), and 5 (5%).

E: 1 = \$0–199,999 (24%); 2 = \$200,000–499,999 (13%); 3 = \$500,000–999,999 (11%); 4 = \$1,000,000–4,999,999 (28%); 5 = \$5,000,000+ (23%).

F: 1 = Never (32%); 2 = Irregularly (33%); 3 = Regularly (32%).

G: 1 = Developing new products and services (80%); 2 = Implementing new or significantly improved operational processes/services (15%); 3 = Implementing new or significantly improved marketing method (3%); 4 = Company not involved in innovation (2%).

H: 1 = 0–10% (59%); 2 = 11–25% (21%); 3 = 26–50% (4%); 4 = 51–75% (6%); 5 = 75%+ (10%).

I: 1 = Major enabler (17%); 2 = Enabler (37%); 3 = Neither a barrier nor an enabler (29%); 4 = Barrier (12%); 5 = Major barrier (5%).

Table 3. Characteristics of original and preprocessed data set

Data set ICT - Australia	Preprocessed data set ICT - Australia
Number of records: 265	Number of records: 342
Number of factors: 23	Number of factors: 20
Class of success: Profitability	Class of success: Profitability
Type of data: Nominal and numeric	Type of data: Numeric
Number of SIT with success: 182	Number of SIT with success: 171
Number of SIT without success: 80	Number of SIT without success: 171

lbfgs. For RF, DT, NB, KNN, and GB, the authors considered the values of their default parameters. Finally, Table 5 provides the calibrated parameters for the ML models.

Implementation of the SSFM Model

The authors implemented the SSFM model, considering for the backend: Python v.3.7 in the Anaconda Navigator version v.4.8.3 environment, and the use of scipy v.1.4.1, numpy v.1.18.1, pandas v.1.0.3, statsmodels v.0.11.0, and sklearn v.0.22.1 libraries. Moreover, for the frontend, the researchers used the Flask framework. Figure 5 shows part of the code concerning the hybrid model.

Table 4. Metrics used in the prediction

Metric	Description	Formula
Acu	Rate of correctly classified predictions.	$\frac{TP + TN}{TP + FP + TN + FN}$
Pre	Rate of correctly classified positives.	$\frac{TP}{TP + FP}$
Esp	Rate of correctly classified negatives.	$\frac{TN}{TN + FP}$

Note. TP = True positives, which means the SIT is successful and the algorithm predicts success; TN = True negatives, which means the SIT is not successful and the algorithm predicts failure; FP = False positives, which means the SIT is not successful and the algorithm predicts success; FN = False negatives, which means the SIT is successful and the algorithm predicts failure.

Table 5. Parameters of ML models

SVM	RF	MLP	NB	DT	KNN	GB
C = 1 Kernel = PUK ($\sigma=0.6$, $\omega=0.6$)	Bach size= 100 split=2 Max_ samples=300	Hidden layers = 100 Learning rate=0.001 Alpha=0.0001 Bach-size=200 Max-iter=10000 Solver=adam Activation=relu	Bach size = 100 Number places = 2 var_ smoothing=1e-9	random_state=1 samples_split=10	neighbors= 10	Estimators=1000 Random state=3

Figure 5. Part of the SSFM Hybrid Model's Code Written in Python

```

ensemble_model_voting = VotingClassifier(
    estimators=[('mp', MLP), ('rf', RF), ('nb', NB), ('svm', SMV), ('dt', DT), ('knn', KNN), ('gb', GB)], voting='hard')

for clf, label in zip([MLP, RF, NB, SMV, DT, KNN, GB, ensemble_model_voting], ['Multilayer', 'Random Forest', 'Naive Bayes',
    'SVM', 'D tree', 'K-NN', 'GB', 'Voting7']):
    scores = cross_val_score(clf, X_train, y_train, scoring='accuracy', cv=10)
    ensemble_model_voting = ensemble_model_voting.fit(X_train, y_train)
    predictions = ensemble_model_voting.predict(X_test)
    print('Accuracy:', accuracy_score(y_test, predictions))
    print('Precision:', precision_score(y_test, predictions))
    print('Recall: ', recall_score(y_test, predictions))

```

The authors conducted the deployment of the implementation on a Ngnix Web and application server and a MySql database server (Figure 6). The system contemplates many functionalities, such as the recording of data on the factors of the SIT to be evaluated and the results of the success prediction (Figure 7).

RESULTS AND DISCUSSION

To understand the efficacy of the proposed method and the SSFM model, the authors considered three experiment scenarios:

- **First Scenario:** Seven ML algorithms with default parameters, with an initial data set (23 factors) with partial preprocessing.

Figure 6. Application Deployment Diagram From ML to Unified Modeling Language

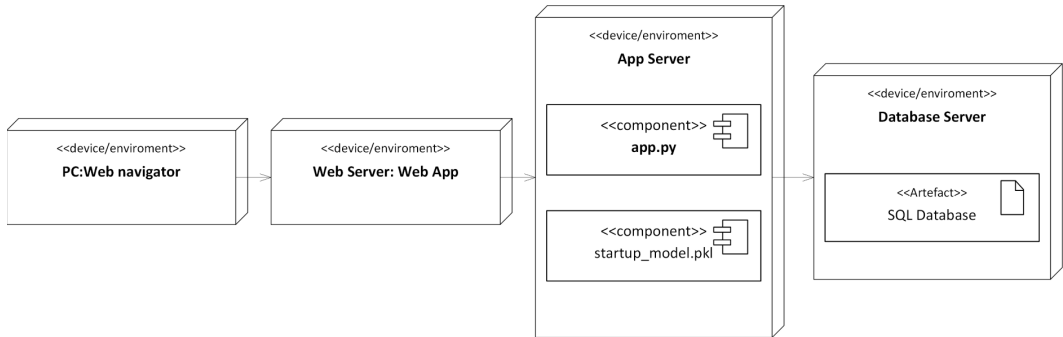
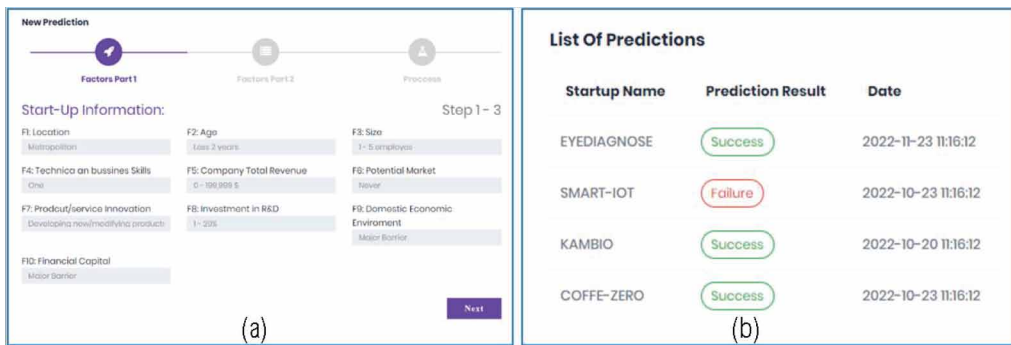


Figure 7. System Interfaces: a) Data Record; b) Prediction Results



- **Second Scenario:** Seven algorithms and three hybrid ML models using the proposed method (20 factors).
- **Third Scenario:** Seven algorithms and three hybrid ML models using the proposed method and the *GreedyStepwise* algorithm (five factors).

The initial data set corresponds to 23 factors (first scenario), the preprocessed data to 20 factors (second scenario) removing the following factors: Structure company (public or private), access skills (grouped with other related factors) and profitability (taken as class of success), and the data in the third scenario to five factors. In the third scenario, the researchers used the *GreedyStepwise* routine of Weka 3.8 on the preprocessed data to reduce the number of factors from 20 to five, these being the following: Startup size, company revenue, R&D, financial capital, and global economic environment. On the other hand, the loss function for the training and validation of the MLP shows stability in no more than 1750 epochs for the three scenarios (Figure 8).

Table 6 shows the testing results after the training-validation for the seven models and the three hybrid models in the three scenarios. The hybrids are given by the voting strategy applied to the seven ML models, the best five ML models, and the best three ML models, which the authors designated as Voting7, Voting5, and Voting3, respectively.

The results in Table 6 evidence the following:

- The proposed method (first and second scenarios) allows the attainment of ML models with better results for seven ML algorithms. In the second scenario, on average, accuracy increases by

Figure 8. Loss Variation Per Epoch of the MLP Model for SIT Success Prediction: a) First Scenario; b) Second Scenario; c) Third Scenario

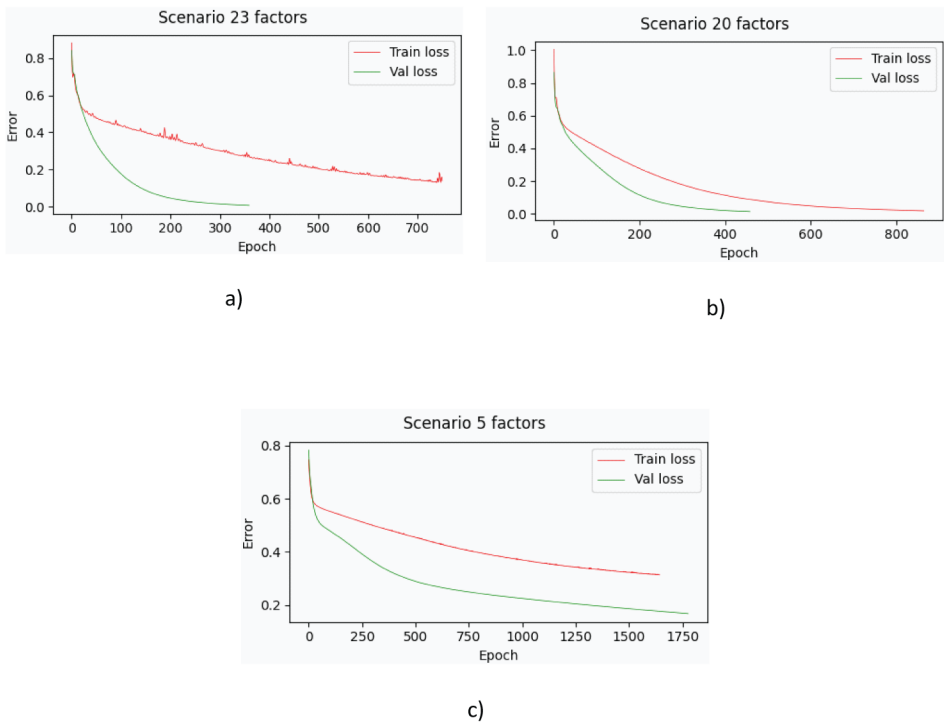


Table 6. Results of SIT success prediction for seven ML models and three hybrid models in three scenarios

Models	First scenario (23 factors)			Second scenario (20 factors)			Third scenario (5 factors)		
	Acu	Pre	Esp	Acu	Pre	Esp	Acu	Pre	Esp
MLP	71.68	77.90	85.50	100.0	100.0	100.0	82.35	76.47	88.24
GB	76.00	83.33	83.33	97.06	100.0	95.99	85.29	81.25	88.89
SVM	84.00	84.00	100.0	91.18	83.33	100.0	88.24	82.35	94.12
RF	72.00	77.77	82.35	91.18	92.86	90.00	85.29	81.25	88.89
KNN	64.00	73.68	77.77	82.86	88.24	77.78	58.82	52.94	64.71
NB	68.00	75.00	83.33	73.57	68.65	77.78	67.65	60.00	78.57
DT	60.00	63.15	80.00	67.65	64.29	70.00	64.71	61.54	66.67
Average*	70.81	76.40	84.61	86.21	85.34	87.36	76.05	70.83	81.44
Voting3	60.00	60.86	93.33	100.0	100.0	100.0	88.24	82.35	93.33
Voting5	64.00	62.50	100.0	100.0	100.0	100.0	88.24	82.35	94.12
Voting7	64.0	62.50	100.0	100.0	100.0	100.0	88.24	82.35	94.12

Note. * corresponds to the average of the seven ML models.

21.75%, precision by 11.69%, and specificity by 3.25%. These results show that preprocessing activities, including balancing and calibration in the learning process, positively impact results.

- The results with the five factors obtained by applying the GreedyStepwise routine (third scenario) show, in respect to the second scenario, an average decrease of 12% in accuracy, 17% in precision, and 7% in specificity for the seven models. This could be explained due to the loss of information produced by the routine's heuristic; however, the filtered factors have much influence on the prediction, allowing the attainment of SVM with an accuracy of 88%, precision of 82%, and specificity of 94%, which shows that it is very useful in situations where there are few studied factors on SITs.
- The individual models that present the best results in all scenarios are MLP, GB, and SVM, obtaining an accuracy of 100%, 97%, and 91%, respectively. On the other hand, the three hybrid models present better results than the ML models by separate, reaching an ideal value (100%) in the three metrics in the second scenario, even in the scenario of five factors where an accuracy of 88%, precision over 82%, and specificity over 93% are achieved.

CONCLUSION

In this study, the authors proposed a systematic method based on an ML algorithm to build a predictive model of SIT success with high precision, which consists in four processes (i.e., selection of critical factors of success, data extraction, preprocessing, and learning). Unlike other studies, which generally focus on a city or region in a country, the proposed method is systematic and applicable to any city or region; moreover, this study contemplates a hybrid model that generally presents better results and an inventory of 79 critical factors of success.

To test its efficiency, the authors applied the method to a database of 265 SITs in Australia with seven learning algorithms (i.e., SVM, MLP, DT, NB, KNN, RF, and GB); then, they implemented the predictive model using Python, considering these obtained models and three hybrid models based on the voting strategy. Moreover, the researchers considered three testing scenarios: The first, without applying the method, the second, by using the method, and the third, by using the method and the GreedyStepwise algorithm to reduce the factors.

The results show that the proposed method (first and second scenarios) allows the attainment of predictive models with better results for the seven learning algorithms that the authors used. In the second scenario, on average, accuracy increased by 21.75%, precision by 11.69%, and specificity by 3.25%. These results show that the method's processes positively impact the results. Moreover, the authors obtained the best results with MLP, GB, and SVM, with an accuracy of 100%, 97%, and 91%, respectively. Besides, the hybrid model generally provides better results than the models by separate, reaching an ideal accuracy of 100%.

The proposed method with the GreedyStepwise algorithm permits the reduction of 20 factors to five very significant factors (i.e., startup size, company revenue, R&D, financial capital, and global economic environment) and obtains, through SVM and hybrid models, a prediction with an accuracy of 88%, precision of 82%, and specificity of 94%, which shows that it is very useful in situations where there are few studied factors on SITs.

The results of the model generated by the method present high precision, accuracy, and specificity despite the high uncertainty of this type of company, demonstrating that the proposed method is systematic and applicable to other realities. However, the results depend on the quality and quantity of the data, the contemplated factors of success, the preprocessing activities, and the considered ML models. In other words, the results cannot be extrapolated to other realities, but, by following the method, good results can be guaranteed.

A future study to develop is predicting SIT success in each stage of its lifecycle, since SIT success depends on the success of each of its development stages; for example, success in the early stage can only be achieved if success has been achieved in the seed stage.

ACKNOWLEDGMENT

The authors thank the Universidad Nacional Mayor de San Marcos for the partial financing of this investigation.

REFERENCES

- Abou-Moghli, A., & Al-Kasasbeh, M. (2012). Social network and the success of business startup. *International Journal of Business and Management*, 7(9), 134–140.
- Akhavan, M., Sebt, M. V., & Ameli, M. (2021). Risk assessment modeling for knowledge based and startup projects based on feasibility studies: A Bayesian network approach. *Knowledge-Based Systems*, 222, 106992.
- Al-Fraihat, D., Joy, M., & Sinclair, J. (2020). Evaluating e-learning systems success: An empirical study. *Computers in Human Behavior*, 102, 67–86.
- Alvarez, S., & Barney, J. (2001). How entrepreneurial firms can benefit from alliances with large partners. *Acadademy Management Exececutive*, 15(1), 139–148.
- Antretter, T., Blohm, I., & Grichnik, D. (2018). Predicting startup survival from digital traces: Towards a procedure for early stage investors. In *Proceedings of the 39th International Conference on Information System, ICIS 2018*, San Francisco, United States.
- Ardito, L., Messeni Petruzelli, A., & Albino, V. (2015). From technological inventions to new products: A systematic review and research agenda of the main enabling factors. *European Management Review*, 12(3), 113–147.
- Arefin, M. S., Alam, M. S., Islam, M. R., & Rahaman, M. (2019). High-performance work systems and job engagement: The mediating role of psychological empowerment. *Cogent Business and Management*, 6(1), 1664204. doi:10.1080/23311975.2019.1664204
- Arora, S. K., Li, Y., Youtie, J., & Shapira, P. (2019). Measuring dynamic capabilities in new ventures: Exploring strategic change in U.S. green goods manufacturing using website data. *The Journal of Technology Transfer*, 45(5), 1451–1480. doi:10.1007/s10961-019- 09751-y
- Asmoro, A., Nugroho, L., & Selo, S. (2018). Prediction modeling of software startup success by PLS-SEM approach. *IACSIT International Journal of Engineering and Technology*, 7(4), 141–147.
- Back, D., Clow, K., & Box, T. (1996). *Entrepreneurial success: Test of a predictive model*. In *Proceedings of the 1996 27th Annual Meeting of the Decision Sciences Institute*, Orlando, FL, USA.
- Baum, J. A., & Silverman, B. S. (2004). Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing*, 19(3), 411–436.
- Bernard, R., & Tariskova, N. (2017). Indicators of startup failure. *Industry 4.0*, 2(5), 238 — 240.
- Bertoni, F., Colombo, M., & Grilli, L. (2011). Venture capital financing and the growth of high-tech Startups: Disentangling treatment from selection effects. *Research Policy*, 40, 1028–1043.
- Böhm, M., Weking, J., Fortunat, F., Müller, S., Welp, I., & Krcmar, H. (2017). The business model DNA: Towards an approach for predicting business model success. In in Leimeister, J.M.; Brenner, W. (Hrsg.): *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)*, St. Gallen, S. 1006-1020
- Borrajó, L., Baruque, B., Corchado, E., Bajo, J., & Corchado, J. (2011). Hybrid neural intelligent system to predict business failure in small-to-medium-size enterprises. *International Journal of Neural Systems*, 21(4), 277–296. PMID:21809475
- Cabrera, E. M., & Mauricio, D. (2017). Factors affecting the success of women's entrepreneurship: A review of literature. *International Journal of Gender and Entrepreneurship*, 9(1), 31–65. doi:10.1108/IJGE-01-2016-0001
- Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16–28.
- Corrales, D., Ledezma, A., & Corrales, J. C. (2020). A case-based reasoning system for recommendation of data cleaning. *Applied Sof Computing Journal*, 106180. doi:10.1016/j.asoc.2020.106180
- Elhedhli, S., Akdemir, C., & Astebro, T. (2014). Classification models via Tabu search: An application to early stage. *Expert Systems with Applications*, 41, 8085–8091.

- Friar, J., & Meyer, M. (2003). Entrepreneurship and startups in the Boston region: Factors differentiating high-growth ventures from micro-ventures. *Small Business Economics*, 21, 145–152.
- Ganotakis, P. (2012). Founders' human capital and the performance of UK new technology based firms. *Small Business Economics*, 39, 495–515.
- Gartner, W., & Liao, J. (2012). The effects of perceptions of risk, environmental uncertainty, and growth aspirations on new venture creation success. *Small Business Economics*, 39, 703–712.
- Gbadegeshin, S., Al Natsheh, A., Ghafel, K., Mohammed, O., Koskela, A., Rimpilainen, A., Tikkanen, J., & Kuoppala, A. (2022). Overcoming the valley of death: A new model for high technology startups. *Sustainable Futures*, 4, 100077.
- Greve, A., & Salaff, J. W. (2003). Social networks and entrepreneurship. *Entrepreneurship Theory and Practice*, 1, 1–20.
- Haltiwanger, J., Jarmin, R., & Miranda, J. (2012). *Who creates jobs? Small vs. large vs. young?* [Unpublished working paper]. University of Maryland and US Census Bureau.
- Helabi, C., & Lussier, R. (2014). A model for predicting small firm performance. *Journal of Small Business and Enterprise Development*, 21(1), 4–25.
- Honorine, A. N., & Emmanuelle, D. (2019). Stage financing and syndication in the IPO underpricing of venture-backed firms: Venture capital and IPO underpricing. *International Journal of Entrepreneurship and Innovation*, 20(4), 289–300.
- Hormiga, E., Batista-Canino, R., & Sánchez-Medina, A. (2011). The impact of relational capital on the success of new business Startups. *Journal of Small Business Management*, 49(4), 617–638.
- Horne, J., & Fichter, K. (2022). Growing for sustainability: Enablers for the growth of impact startups – A conceptual framework, taxonomy, and systematic literature review. *Journal of Cleaner Production*, 349, 131163.
- Joshi, K., & Satyanarayana, K. (2014). What ecosystem factors impact the growth of high-tech startups India? *Asian Journal of Innovation and Policy*, 3(2), 216–244.
- Jurgens, J. (2022, May 12). *How startups drive economic recovery while growing responsibly*. We Forum. <https://www.weforum.org/agenda/2022/05/how-startups-help-drive-economic-recovery-and-growth/>
- Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, 1–33.
- Kampen, K. J. (2019). Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables. *Measurement*, 137, 428–434.
- Ko, C. R., & An, J. I. (2019). Success factors of student startups in Korea: From employment measures to market success. *Asian Journal of Innovation and Policy*, 8(1), 97–121.
- Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the outcome of startups: Less failure, more success. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (pp. 798-805). IEEE.
- Lasch, F., Le Roy, F., & Yami, S. (2007). Critical growth factors of ICT Startups. *Management Decision*, 45(1), 62–75.
- Li, S., Shang, J., & Slaughter, A. (2010). Why do software companies fail? *Information Systems Research*, 21(3), 631–654.
- Luna-Perejon, F., Malwade, S., Styliadis, C., Civit, J., Cascado-Caballero, D., Konstantinidis, E., & Li, Y. C. (2019). Evaluation of user satisfaction and usability of a mobile app for smoking cessation. *Computer Methods and Programs in Biomedicine*. PMID:31473444
- Maine, E., Shapiro, D., & Vining, A. (2010). The role of clustering in the growth of new technology-based firms. *Small Business Economics*, 34, 127–146.
- Martens, D., Vanhoute, C., De Wine, S., Bsesens, B., Sels, L., & Mues, C. (2011). Identifying financial successful startup profiles with data mining. *Expert Systems with Applications*, 38, 5794–5800.

- Maxwell, A. L., Jeffrey, S. A., & Lévesque, M. (2011). Business angel early stage decision making. *Journal of Business Venturing*, 26(2), 212–225.
- Murray, A. (2019). Supporting academic entrepreneurship: A blueprint for a university based business incubator. *Journal of Higher Education Service Science and Management*, 2(2), 1–9.
- Nadežda, P., Miroslav, P., & Josef, P. (2019). Factors impacting startup sustainability in the Czech Republic. *Innovative Marketing*, 15(3), 1–15. doi:10.21511/im.15(3).2019.01
- OECD. (2005). Guidelines for collecting and interpreting innovation data. Organization for Economic Co-operation and Development (OECD)/Eurostat, Paris
- Pourhashemi, S. M., & Mashalizadeh, A. M. (2013). A novel feature selection method using CFS with Greedy-Stepwise search algorithm in e-mail spam filtering. *Advanced Modeling and Optimization*, 15(3), 867–877.
- Pugliese, R., Bortoluzzi, G., & Zupic, I. (2016). Putting process on track: Empirical research on startups' growth drivers. *Management Decision*, 54(7), 1633–1648.
- Ross, G., Das, S., Sciro, D., & Raza, H. (2021). CapitalVX: A machine learning model for startup selection and exit prediction. *The Journal of Finance as Data Science*, 7, 94–114.
- Roy, S., Modak, N., & Dan, P. K. (2020). Managerial support to control entrepreneurial culture in integrating environmental impacts for sustainable new product development. In S. K. Gosh (Ed.), *Sustainable waste management: Policies and case studies* (pp. 637–646). Springer.
- Santisteban, J., Inche, J., & Mauricio, D. (2021). Critical success factors throughout the life cycle of information technology start-ups. *Entrepreneurship and Sustainability Issues*, 8(4), 446–466.
- Santisteban, J., & Mauricio, D. (2017). Systematic literature review of critical success factors of information technology startups. *Academy of Entrepreneurship Journal*, 23(2), 1–23.
- Santisteban, J., Mauricio, D. S., & Cachay, O. (2021). Critical factors success for technology-based startups. *International Journal of Entrepreneurship and Small Business*, 42(4), 397–421.
- Satyanarayana, K., Chandrashekar, D., & Mungila Hillemane, S. (2021). An assessment of competitiveness of technology-based startups in India. *International Journal of Global Business and Competitiveness*, 16, 28–38.
- Schneider, J., Dowling, M., & Raghuram, S. (2007). Empowerment as a success factor in startup companies. *RMS*, 1, 167–184.
- Sefiani, Y., & Bown, R. (2013). What influences the success of manufacturing SMEs? A perspective from Tangier. *International Journal of Business and Social Science*, 4(7), 297–309.
- Senivongse, C., Bennet, A., & Mariano, S. (2019). Clarifying absorptive capacity and dynamic capabilities dilemma in high dynamic market IT SMEs. *VINE Journal of Information and Knowledge Management Systems*, 49(3), 372–396.
- Shepherd, D. A., & Zacharakis, A. (1999). Conjoint analysis: A new methodological approach for researching the decision policies of venture capitalists. *Venture Capital*, 1(3), 197–217.
- Shlens, J. (2014). *A tutorial on principal component analysis*. Google Research. <https://arxiv.org/pdf/1404.1100v1.pdf>
- Silver, N. (2012). *The signal and noise: Why so many predictions fail—but some don't*. Penguin Press.
- Singh, D. (2019). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97, 105524.
- Song, M., Podoyntsyna, K., Van der Bij, H., & Halman, J. (2008). Success factors in new ventures: A meta-analysis. *Journal of Product Innovation Management*, 25(1), 7–27.
- Thiranagama, R., & Edirisinghe, K. (2015). Factors affecting small business Startup of Engineers and Accountants in Sri Lanka. *NSBM Business & Management Journal*, 6(1), 84–107.
- Tomy, S., & Pardede, E. (2018). From uncertainties to successful start ups: A data analytic approach to predict in technological entrepreneurship. *Sustainability*, 10(3), 602. doi:10.5281/zenodo.4724045

Useche, L. M., & Mesa, D. M. (2006). Una introducción a la imputación de valores perdidos. *Terra Nueva Etapa*, 22(31), 127–151.

Van de ven, H., Hudson, R., & Schroeder, M. (1984). Designing new business startups. *Journal of Management*, 10(1), 87—104.

World Bank. (2018). *Poverty and shared prosperity 2018: Piecing together the poverty puzzle*. World Bank. <https://elibrary.worldbank.org/doi/epdf/10.1596/978-1-4648-1330-6>

Zahra, S. A., Matherne, B. P., & Carleton, J. M. (2003). Technological resource leveraging and the internationalization of new ventures. *Journal of International Entrepreneurship*, 1(2), 163–186.

APPENDIX

Table 7. Factors that influence SIT Success

Factor	Source	Factor	Source
1. Access to export market	Tomy & Pardede (2018)	17. Exchange rates	Tomy & Pardede (2018)
2. Access to target market	Tomy & Pardede (2018)	18. Business management experience	Back et al. (1996)
3. Age	Zahra et al. (2003)	19. Expert agent	Borrajó et al. (2011)
4. Incubator support	Santisteban et al. (2020)	20. Financial and accounting information	Helabi & Lussier (2014)
5. Government support	Lasch et al. (2007)	21. Financial capital	Martens et al. (2011) ; Tomy & Pardede (2018)
6. Availability of infrastructure	Tomy & Pardede (2018)	22. Phased financing	Santisteban et al. (2020)
7. Availability of skilled employees	Li et al. (2010)	23. Founder	Schneider et al. (2007)
8. Business agent	Borrajó et al. (2011)	24. Functional performance	Elhedhli et al. (2014)
9. Business model	Böhm et al. (2017)	25. Gender of the entrepreneur	Friar & Meyer (2003)
10. Dynamic capacity	Santisteban et al. (2020)	26. Government regulation	Tomy & Pardede (2018), Pugliese et al. (2016)
11. Risk capital	Bertoni et al. (2011)	27. Human capital	Martens et al. (2011)
12. Capital raised	Baum & Silverman (2004)	28. Industry	Thiranagama & Edirisinghe (2015)
13. Clustering	Maine et al. (2010)	29. Product / service innovation	Ardito et al. (2015)
14. Competition	Song et al. (2008) ; Tomy & Pardede (2018)	30. Innovation environment	Tomy & Pardede (2018)
15. Competitive strategy	Asmoro et al. (2018)	31. Internet	Helabi & Lussier (2014)
16. Cost of production	Elhedhli et al. (2014)	32. Knowledge support	Maxwell et al. (2011)
33. Innovative culture	Santisteban et al. (2020)	47. Marketing	Helabi & Lussier (2014)
34. Ecosystem of innovation and entrepreneurship	Santisteban et al. (2020)	48. Motivation	Ganotakis (2012); Greve & Salaff (2003)
35. Business age	Haltiwanger et al. (2012)	49. Need	Elhedhli et al. (2014)
36. Education	Back et al. (1996); Helabi & Lussier (2014)	50. Organization	Asmoro et al. (2018); Martens et al. (2011)
37. Entrepreneurial education	Baum & Silverman (2004), Maxwell et al. (2011)	51. Partners	Helabi & Lussier (2014); Sefiani & Bown (2013)
38. Entrepreneurial experience	Gartner & Liao (2012)	52. Planning	Helabi & Lussier (2014)
39. Environment	Asmoro et al. (2018); Martens et al. (2011) ; Tomy & Pardede (2018)	53. Potential market	Elhedhli et al. (2014)
40. Evaluator agent	Borrajó et al. (2011)	54. Price	Elhedhli et al. (2014)
41. Profitability	Elhedhli et al. (2014)	55. Value creation process	Asmoro et al. (2018)
42. Proof of concept	Maxwell et al. (2011)	56. Web analytics	Silver (2012)
43. Proof of value	Shepherd & Zacharakis (1999)	57. Working capital	Helabi & Lussier (2014)
44. Research and development	Baum & Silverman (2004) ; Elhedhli et al. (2014); Tomy & Pardede (2018)	58. Technological / business skills	Li et al. (2010)
45. Resource	Asmoro et al. (2018)	59. Global economic environment	Tomy & Pardede (2018)
46. Company revenue	Böhm et al. (2017)	60. Technological surveillance	Ko & An (2019)
61. Client satisfaction	Santisteban et al. (2020)	71. Knowledge absorptive capacity	Senivongse et al. (2019)
62. Service	Elhedhli et al. (2014)	72. Perceived performance	Arefin et al. (2019)
63. Size of investment	Elhedhli et al. (2014)	73. Quality of a product and/or service	Al-Fraihat et al. (2020)
64. Social capital	Martens et al. (2011)	74. Customer satisfaction	Luna-Perejon et al. (2019)
65. Store agent	Borrajó et al. (2011)	75. Staged financing	Honorine & Emmanuelle (2019)
66. Startup size	Joshi & Satyanarayana (2014)	76 Support of a business incubator	Murray (2019)
67. Technical feasibility	Elhedhli et al. (2014)	77. Innovation and entrepreneurship ecosystem	Corrales-Estrada (2019)
68. Technological hype	Maxwell et al. (2011)	78. Dynamic capability of entrepreneurs	Arora et al. (2019)
69. Technology significance	Elhedhli et al. (2014)	79. Innovative and entrepreneurial culture	Roy et al. (2020)
70. Location	Hormiga et al. (2011)		

Edilberto Vasquez has a bachelor's degree in Scientific Computing, and currently is a master candidate at the Faculty of System Engineering and Informatics, Universidad Nacional Mayor de San Marcos, Peru. He has more than seven years of experience as a consultant in innovation projects in different business sectors. His main research areas are artificial intelligence, big data, and technological entrepreneurship.

José Santisteban obtained his Ph.D in Systems Engineering at the Faculty of System Engineering and Informatics, Universidad Nacional Mayor de San Marcos, Peru. He has more than ten years of experience in implementing information systems in different business sectors. His main research areas are artificial intelligence and technological entrepreneurship.

David Mauricio obtained his Ph.D. in Systems Engineering and Computer Science and his master's degree of Science in Applied Mathematics from the Federal University of Rio de Janeiro, Brazil. He was a professor at the State University of North Fluminense Brazil from 1994 to 1998. Since 1998, he has been a Professor at the National University of San Marcos. His research interests include mathematical programming, artificial intelligence, software engineering, and entrepreneurship.