Building Models based on Artificial Neural Networks to Predict Entrepreneurial Intentions among Undergraduate Students

Magaly Sandoval, Carlos Hernández
Departamento de Ingeniería Industrial y Sistemas
Universidad de La Frontera
Temuco, Chile
magaly.sandoval@ufrontera.cl, carlosalberto.hernandez@ufrontera.cl

Abstract

This research compares models based on artificial neural networks (ANN) to predict entrepreneurial intentions among undergraduate students according to the results of the University Entrepreneurial Spirit Students' Survey (GUESSS) of 2016. The research is carried out following a classic 4-stage methodology (analysis, design, development, and validation). During the analysis, surveys were thoroughly reviewed and preprocessed. During the design, the survey's questions are combined according to certain criteria to build 10 classification models. Construction and validation are carried out entirely using the software WEKA. For the purposes is this investigation 627 surveys are considered. The dataset is split up in two subsets: 80% for training and test, and the remaining 20% for validation. The approach to predict entrepreneurial intentions considers building and comparing 10 ANNs. The results reveal that, with a heavily imbalanced dataset, the proposed models classify correctly between 77% and 80%. However, the area under the curve ROC present low values. In conclusion, the investigation results show that predictive models based on ANN can help predict the entrepreneurial intention of undergraduate students by means of knowing some information about their family background, social environment, and university. However, these results might not be conclusive since the dataset is significantly imbalanced.

Keywords

Predictive Model, Machine Learning, Artificial Neural Network, GUESSS Survey, Entrepreneurial Intention

1. Introduction

Nowadays, entrepreneurship is believed to be one the phenomena behind the economic growing of countries worldwide. Mainly due to the number of new jobs that it generates (Andersen and Nielsen, 2012). Since, in many cases university programs and lectures might lack of appropriate contents on entrepreneurship or might be not concrete enough, it would be useful to discover the entrepreneurial intention among undergraduate students.

There is abundance of literature focused on nascent or active entrepreneurs, but not so much is written on how to identify future entrepreneurs among college students. It is particularly important to understand the drivers of business creation, and to define the role and influence that universities, social environment, and society have not only in the entrepreneurial intention but also in the development of the skills necessary to achieve business sustainability.

There are two important concepts in the literature: ability and skill. Ability is usually referred as the initial quality of an individual. Skill, instead, is referred as the set of requirements that are required to perform specific tasks (Autor and Handel, 2013; Guvenen et al., 2015).

Universities, seen either as research centers or as educational institutions, have the conditions to facilitate social interaction and to promote entrepreneurship. Several investigations about the influence of the academia on the entrepreneurship show that education help students discover their abilities and knowledge to create a company and to commit to their intentions (von Graveвenitz et al., 2010). However, the initiatives to analyze the entrepreneurial intention among university students have been developed from a psychological perspective, where the intention conceived as a predictor of a long term planned and goal-oriented behavior (Ajzen & Fishbein, 1980; Azjen, 1991, Azjen, 2002). Sieger, et al., (2011) identify the family and social environment as the most influential factor in the entrepreneurial intention and attitude among undergraduate students.

The data used in this investigation corresponds the results of the survey GUESSS 2016 conducted at Universidad de La Frontera, Temuco, Chile. The Global University Entrepreneurial Spirit Students' Survey (GUESSS) is a research on entrepreneurship carried out by University of Saint Gallen, Switzerland, whose focus is the entrepreneurial intentions and activities among students and their family background. It is based on the conceptual model proposed by Sieger, et al. (2011), which is in turn is based on Ajzen's Theory of Planned Behavior (1991) (Figure 1). The survey is conducted every 2 or 3 years in 50 countries approximately.

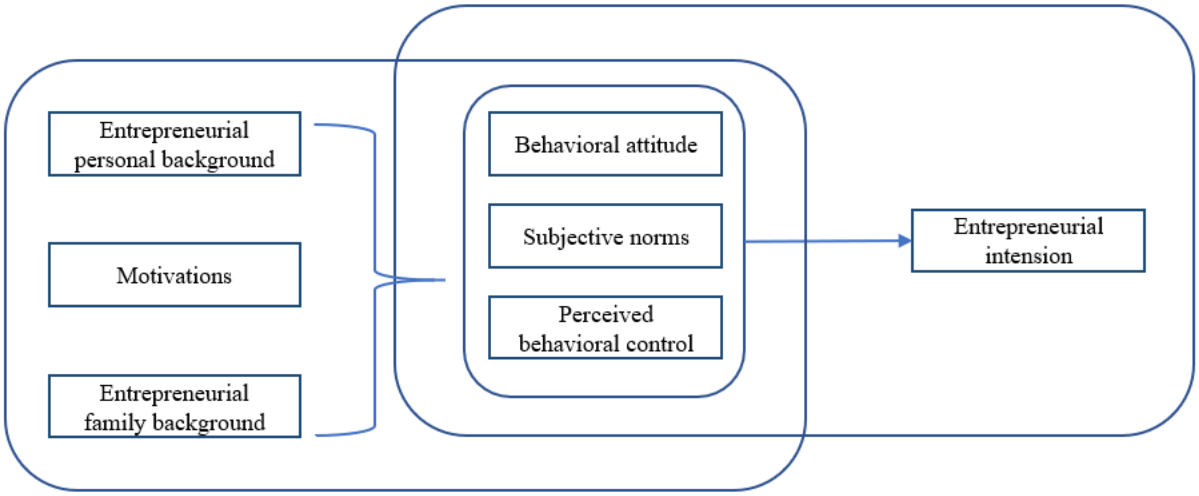
Entrepreneurial personal background
Motivations
Entrepreneurial family background
Behavioral attitude
Subjective norms
Perceived behavioral control
Entrepreneurial intension

Figure 1. GUESSS conceptual model (Sieger et al., 2011)

1.1 Objective

To apply machine learning techniques to predict entrepreneurial intentions right after graduation among undergraduate students by means of building classification models based on artificial neural networks.

2. Literature Review

2.1 Machine learning

Machine learning is a branch of artificial intelligence (AI). It uses methodologies, techniques, and algorithms to find patterns and to learn from datasets through experience.

There exist supervised, unsupervised, and reinforcement learning algorithms. In supervised learning, the training is carried out using datasets that contain the class or value to be predicted. In unsupervised learning, instead, the desired class is not known. In the reinforcement learning, on the other hand, predefined actions, parameters, and final values are used. Machine learning algorithms can be grouped in several types. Some of the most common are regression algorithms, Bayesian algorithms, decision trees, and artificial neural networks

2.2 Classification, prediction, and forecasting

Although sometimes used indistinctively, there are several important concepts in machine learning. The classification is the determination of the class with a nominal value in an unseen dataset using a model previously trained, while the prediction is estimation of a numeric value for the desired dependent variable. On the other hand, forecasting is the prediction of future values using time series.

2.3 Hold out

In machine learning, holding out refers to the split up of a dataset into a dataset for training and a test dataset. The underlying idea is to use the test dataset to assess the performance of the predictive model on unseen data. Usually, the preferred split proportion is 80% for training and 20% for testing.

2.4 Cross-validation

Cross-validation refers to the random split up of a dataset into k folds. During the model building, k-1 folds are used for training while the left one is used for testing to assess the model’s performance. Training and test are repeated iteratively k times until all folds have been used for training and for testing (Figure 2). The objective of such implementation is to minimize the risk of the overfitting that may happened when using a simple hold out. In the case of cross-validation, each iteration produces different results because the folds for training and for testing are different. Finally, the result is a weighted average.

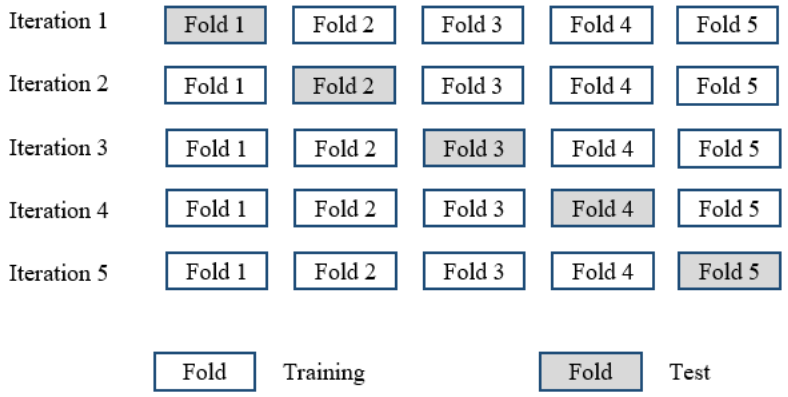

Figure 2. Cross-validation (k=5)

2.5 Overfitting

Overfitting occurs when the model learns from the training dataset so well that it is not able to generalize on unseen data. This may happen because the model incorporates details that will probably not found in new data (Figure 3).

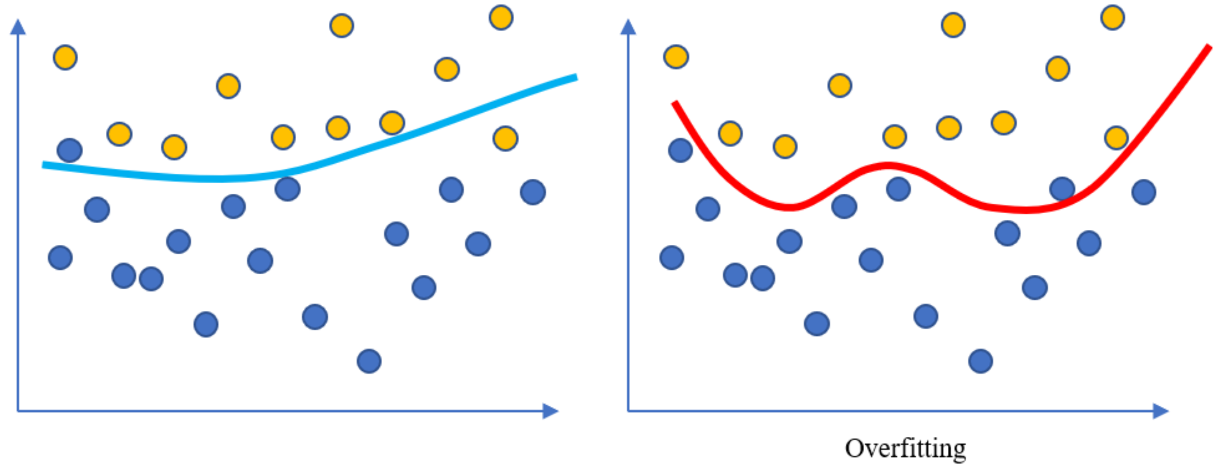

Figure 3. Overfitting

2.6 Generalization

In machine learning, generalization refers to the ability of a trained model to classify or predict unseen data. The objective of building classification models is to achieve a good performance on new data. The usefulness of the model depends on that.

2.7 Replication

In statistics, replication is the repetition of an experiment under similar conditions to estimate the variability of phenomenon under study. When applying cross-validation, the k-folds are the result of a partitioning that depends on a specific seed number. Different seeds can give rise to different folds. Since for each replication a specific set of folds are created, the results are also specific to that replication.

2.8 Artificial neural networks (ANN)

An ANN consist of nodes or neurons that are combined in an interconnected layered structure (Figure 4). The first level is the input layer, which contains the nodes that receive the externa data. In the second level are the hidden layers that transform the input data for the output layer, whose neurons are responsible for delivering the results generated by the network (Morano and Tajani 2013). The topology of an ANN is determined by the number of layers, the number of nodes in each layer and the transfer function. The main reason for using ANN is possibility of learning from highly correlated, incomplete, or previously unknown data (Ge et al. 2003).

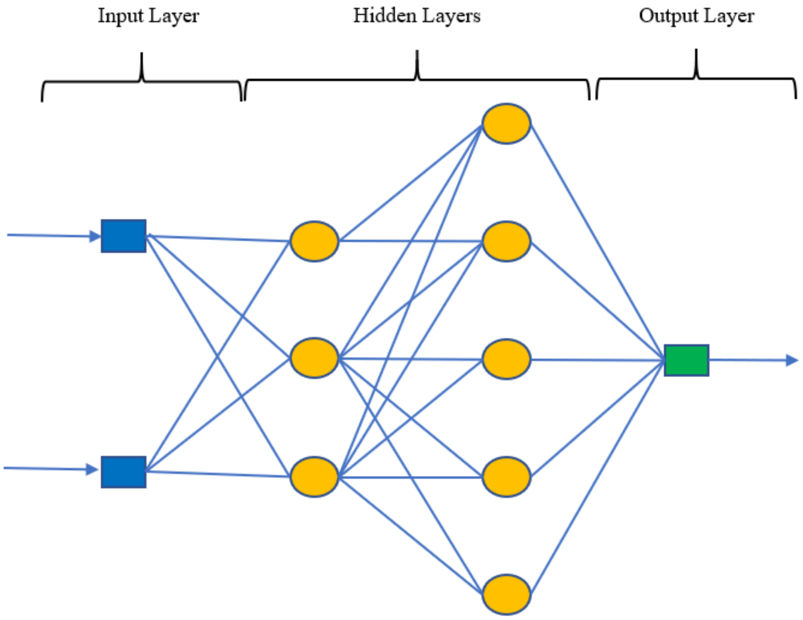

Figure 4. ANN's input, hidden, and output layers

3. Methods

The present research is carried out following a classic 4-stage model: analysis, design, construction, and validation (Figure 5).

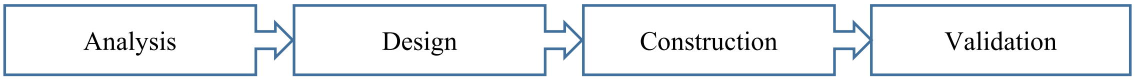

Figure 5. Four-stage model

3.1 Analysis

In the stage of analysis, a complete review and study of GUESSS 2016 survey conducted on undergraduate students at Universidad de La Frontera (Temuco, Chile) is carried out. At this point, the scope of the investigation is defined, and several software packages are compared to select the most suitable for the research's requirements.

For the purposes of this investigation, a total of 627 surveys are considered. All of which are properly pre-processed for the subsequent stages.

The GUESSS 2016 questionnaire is organized in several domains or sections. The answers for all questions are expressed in a simplified scale. Some of these questions are selected to be the attributes of the classification models, leaving the question related to the entrepreneurial intention right after finishing college as the attribute or class to be predicted. A total of 37 questions grouped in 6 domains are included in this research (Table 1).

Table 1. Selected domains and questions from GUESSS 2016

	Domain	Questions
D.1	Personal information	2
D.2	Subject	4
D.3	Career interests	2
D.4	University environment	13
D.5	Family background	2
D.6	Social environment	13

3.2 Design

The complete dataset, a matrix of 627 rows (instances) by 37 columns (attributes), is split up to create 2 subsets. The first one for training and test, contains 80% of the data. The remaining 20% of data is left in a separate dataset to be used during the validation stage.

To identity the influence of each domain, different combinations of questions defined to build and compare 10 classification models (Table 2).

Table 2. Classification models' design

	M0	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
D.1 Personal information	√	√	√	√	√	√	√	√	√	√	√
D.2 Subject	√		√	√	√	√		√			
D.3 Career interests	√	√	√	√	√	√	√	√	√	√	√
D.4 University environment	√	√		√	√	√			√		
D.5 Family background	√	√	√		√		√			√	
D.6 Social environment	√	√	√	√			√				√

The resulting ANN-based models are compared by means of the percentage of correct predictions made on the validation dataset. Additionally, curves Precision-Recall and the area under the curve ROC (ROC AUC) are considered for comparison too (Davis and Goadrich, 2006).

When dealing with classification problem it is important to consider the class balance. In the case of a heavily class imbalanced dataset, the metrics for the model performance should consider the Precision-Recall curves along with ROC AUC (Saito and Rehmsmeier, 2016).

3.3 Construction

The goal is to predict whether the students' intention right after graduation is to start their own business or to find a job in a company. Thus, predictions can take the two values: entrepreneur or employee. There is also a third value for those students who do not know yet.

All classification models presented in this work are built using the well-known data mining software WEKA (Witten et all, 2017).

For practical reasons, considering their size, all the proposed models have only 1 hidden layer, However, the number of nodes in the hidden layer depends on the number of attributes of each model (Table 3).

Table 3. Models' attributes, hidden layers, and nodes

Model	Attributes	Hidden layers	Nodes
M0	36	1	121
M1	32	1	108
M2	23	1	82
M3	34	1	117
M4	23	1	69
M5	21	1	65
M6	17	1	70

M7	8	1	26
M8	17	1	54
M9	6	1	18
M10	17	1	66

Initially, all models are trained and tested applying a cross-validation scheme of k=10 folds with a dataset of 500 instances (Table 4). An instance can be understood as a row containing all the answers of a single survey.

Table 4. Models' performance with training and test data and cross-validation k=10 (500 instances)

		Weighted Average		
Model	**%**	**ROC AUC**	**Precision**	**Recall**
M0	79.6	0.583	0.744	0.796
M1	78.0	0.588	0.739	0.780
M2	78.2	0.601	0.737	0.782
M3	78.4	0.570	0.741	0.784
M4	80.0	0.600	0.748	0.800
M5	79.6	0.622	0.759	0.796
M6	78.0	0.607	0.754	0.780
M7	80.6	0.634	0.754	0.806
M8	79.6	0.610	0.760	0.795
M9	81.4	0.781	0.755	0.814
M10	79.0	0.617	0.763	0.790

Although cross-validation helps reduce the risk of overfitting, the influence of the fold partitioning remains. A simple way to minimize this effect is by means of replicating the experiment with a different partitioning each time.
For the purposes of this research 10 replications are run, i.e., each model is trained and tested 100 times.
The results show that systematically the ratio of prediction correctness is close to 79% with an average standard deviation close to 4% (Table 5).

Table 5. Models' performance with training and test data, k=10 cross-validation, and 10 replications

Model	**Average Correct Predictions (%)**	**Standard Deviation (10 replications)**
M0	78.66	4.25
M1	78.54	3.69
M2	78.16	3.81
M3	78.18	4.26
M4	79.74	4.10
M5	80.34	4.58
M6	78.26	4.91
M7	81.52	3.60
M8	79.46	3.93
M9	79.12	3.88
M10	78.36	4.20

3.4 Validation

The validation of the trained and tested ANN-based models is carried out on the validation dataset (20%) that was held out during the stage of analysis. This dataset contains 127 unseen instances (rows), which are totally unknown to the classification models.

4. Data Collection

The results of the validation reveal that all models can generalize relatively well with unseen data since the percentage of correct classification is close to 77% (Table 6), which is not far from the results obtained on the training dataset.

Table 6. Models' performance with validation data (127 unseen instances)

Model	Correct Predictions (%)	ROC AUC	Precision	Recall
M0	77.17	0.488	0.702	0.772
M1	77.16	0.466	0.727	0.772
M2	78.74	0.450	0.725	0.787
M3	74.80	0.433	0.666	0.748
M4	74.80	0.514	0.666	0.748
M5	73.22	0.571	0.644	0.732
M6	78.74	0.469	0.704	0.787
M7	77.17	0.496	0.645	0.772
M8	76.38	0.518	0.687	0.764
M9	80.32	0.550	0.806	0.803
M10	77.95	0.491	0.682	0.780

5. Results and Discussion

Although classification percentages and resulting metrics seem to be similar, there are still significant differences in the complexity and in the number of attributes of each model.

5.1 Numerical Results

Confusion matrices are useful to summarize the classification results in tables. The diagonal of the matrix contains the number of instances correctly classified. The other cells present the incorrect classifications (Table 7).

Table 7. Confusion matrix

Class 1	Class 2	Class 3
Instance class 1 classified **correctly** as class 1	Instance class 1 classified **incorrectly** as class 2	Instance class 1 classified incorrectly as class 3
Instance class 2 classified **incorrectly** as class 1	Instance class 2 classified **correctly** as class 2	Instance class 2 classified **incorrectly** as class 3
Instance class 3 classified **incorrectly** as class 1	Instance class 3 classified **incorrectly** as class 2	Instance class 3 classified **correctly** as class 3

The results for both datasets, training-test (500 instances) and validation (127 instances), show that the proposed ANN-based models systematically classify correctly the class corresponding to students whose intention is to find a job in a company right after graduation. Instead, the class of those who have entrepreneurial intentions is usually wrongly classify as "Employee" (Table 8).

Table 8. Resulting confusion matrices for all ANN-based models

Training-test dataset (500 instances)			
M0	Employee	Entrepreneur	Other
	393	14	17
	34	5	1
	31	5	0
M1	Employee	Entrepreneur	Other
	385	22	17

Validation dataset (127 instances)			
M0	Employee	Entrepreneur	Other
	97	3	2
	5	0	0
	18	1	1
M1	Employee	Entrepreneur	Other
	96	4	2

	35	5	0
	29	7	0
M2	**Employee**	**Entrepreneur**	**Other**
	387	24	13
	35	3	2
	31	4	1
M3	**Employee**	**Entrepreneur**	**Other**
	387	20	17
	33	5	2
	31	5	0
M4	**Employee**	**Entrepreneur**	**Other**
	394	15	15
	37	3	0
	30	3	3
M5	**Employee**	**Entrepreneur**	**Other**
	389	14	21
	33	6	1
	31	2	3
M6	**Employee**	**Entrepreneur**	**Other**
	382	26	16
	33	5	2
	27	6	3
M7	**Employee**	**Entrepreneur**	**Other**
	396	13	15
	33	6	1
	34	1	1
M8	**Employee**	**Entrepreneur**	**Other**
	388	15	21
	34	6	8
	30	2	4
M9	**Employee**	**Entrepreneur**	**Other**
	401	12	10
	33	5	2
	31	4	1
M10	**Employee**	**Entrepreneur**	**Other**
	386	25	13
	30	5	5
	27	5	4

	5	0	0
	18	0	2
M2	**Employee**	**Entrepreneur**	**Other**
	99	2	1
	5	0	0
	19	0	1
M3	**Employee**	**Entrepreneur**	**Other**
	94	3	5
	5	0	0
	19	0	1
M4	**Employee**	**Entrepreneur**	**Other**
	94	3	5
	5	0	0
	10	0	1
M5	**Employee**	**Entrepreneur**	**Other**
	93	5	4
	5	0	0
	18	2	0
M6	**Employee**	**Entrepreneur**	**Other**
	99	1	2
	5	0	0
	18	1	1
M7	**Employee**	**Entrepreneur**	**Other**
	98	0	4
	5	0	0
	19	1	0
M8	**Employee**	**Entrepreneur**	**Other**
	96	3	3
	5	0	0
	18	1	1
M9	**Employee**	**Entrepreneur**	**Other**
	101	1	0
	5	0	0
	19	0	1
M10	**Employee**	**Entrepreneur**	**Other**
	98	0	4
	5	0	0
	18	1	1

6. Conclusion

GUESSS 2016 surveys contain valuable information that can be used to modify and adapt the offer of lectures and contents to help students develop the skills needed either to accomplish their entrepreneurial intentions or to be recruited by a company right after graduation.

Since GUESSS 2016 questionnaire is organized in domains, it is easy to groups questions and to define different subsets to build classification models and thus, by means of experimenting to identify those domains that are more relevant than others.

When training ANN-based models, instead of implementing a simple hold-out it is preferable to apply a cross-validation (k=10) scheme to get rid of the influence of partitioning randomness. Averaging ten (k=10) results will be always better than having only one number. Furthermore, running replications help reduce the bias caused by the partitioning of the k=10 folds. At last, having a validation dataset with unseen data is crucial to determine whether the model can generalize properly or not.

The results reveal that all proposed models can help predict correctly almost 80% of the instances with training and test data, and with validation data too. However, the complexity of the models differs significantly due to the difference in the number of attributes included.

Besides the percentage of correct predictions, it is important to consider the area under the curve ROC. And when dealing with an imbalanced dataset it is advisable to consider the Precision-Recall curves too. In the case of the proposed models, only one class (employee) presents high prediction percentages. The success rate is clearly lower for the other 2 classes.

Finally, the confusion matrices and the ROC AUC-Precision-Recall curves resulting from the validation dataset suggest that the proposed classification ANN-based models do help predict the entrepreneurial intention among university students using objective criteria based on multiple attributes. However, given that the dataset is heavily unbalanced, the real contribution of the ANN-based models is not conclusive.

Acknowledgements

The authors express their deep gratitude to The Global University Entrepreneurial Spirit Students' Survey (GUESSS) and to those who are involved in this important initiative.

References

Ajzen, I., Fishbein, M., Understanding Attitudes and Predicting Social Behavior, Englewood Cliffs, NS: Prentice Hall, 1980.

Ajzen, I., The theory of planned behavior. Organizational and Human Decision Processes, vol. 50, no. 2, pp. 179-211, 1991.

Ajzen, I., Perceived Behavioral Control, Self- Efficacy, Locus of Control, and the Theory of Planned Behavior, Journal of Applied Social Psychology, vol. 32, no. 1, pp. 1-20, 2002.

Andersen, S. & Nielsen, K. (2012). Ability or finances as constraints on entrepreneurship? Evidence from survival rates in natural experiment. The Review of Financial Studies. 25 (12): 3684-3710.

Autor, D., Handel, M., Putting Tasks to the Test: Human Capital, Job Tasks, and Wages, Journal of Labor Economics. The Princeton Data Improvement Initiative, vol. 31, no. 2, pp. 59-96, 2013.

Davis, J., Goadrich, M., The Relationship Between Precision-Recall and ROC Curves, *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Ge, J., Runeson, G., & Lim, K. (2003). Forecasting Hong Kong Housing Prices: An Artificial Neural Network Approach. International Conference on Methodologies in Housing Research, Stockholm, Sweden.

Guvenen, F., Kuruscu, B., Tanaka, S., and Wiczer, D., Multidimensional Skill Mismatch, National Bureau of Economic Research, 2015.

Morano, P., & Tajani, F. (2013). Bare ownership evaluation. Hedonic price model vs artificial neuroal network. International Journal of Business Intelligence and Data Mining. 8(4): 340-360.

Saito, T., Rehmsmeier, M., The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets, *PLoS ONE 10(3): e0118432*, 2015

Sieger, P., Fueglistaller, U., and Zellweger, T. Entrepreneurial Intentions and Activities of Students across the world. International report of GUESSS 2011, 2011.

Von Gravevenitz, G., Harhoff, D., and Weber, R., The Effects of Entrepreneurship Education, Journal of Economic Behavior and Organization, vol. 76, no. 1, pp. 90-112, 2010

Witten, I., Frank, E., Hall, M., and Pal, C., *Data Mining: Practical Machine Learning Tools and Techniques*, 4th Edition, Morgan Kaufmann, Cambridge, 2017.

Biographies

Magaly Sandoval is an industrial engineer, consultant, and university professor. She earned Licentiate Degree in Engineering from Universidad de La Frontera, Temuco, Chile, and MBA from Universidad Austral de Chile, Valdivia, Chile. She is a Babson SEE certified mentor. She has taught lectures in Project Planning & Management, Project Evaluation, Engineering Economy, Business Management, International Business Management, and Entrepreneurship for engineering students. During her academic tenure she has been appointed in different management positions and has mentored over a hundred students. Her research interests include business models, project management, and entrepreneurship & intrapreneurship.

Carlos Hernández is an industrial engineer, consultant, and university professor. He earned Licentiate Degree in Engineering from Universidad de La Frontera, Temuco, Chile, Master of Sciences in Computational Engineering and Doctor of Engineering from Technische Universität Braunschweig, Brunswick, Germany. He is the author of several scientific and engineering articles. He has taught lectures in Discrete Event Simulation, Supply Chain Management, Engineering Economics, Corporate Finances, Financial Engineering, Business Analytics, Data Mining and Machine Learning for engineering students. He has developed a professional career working for large multinational companies (PricewaterhouseCoopers, BHP Billiton, and Merck Sharp & Dohme). He also worked as a scientific researcher in the Institut für Produktionsmesstechnick at TU Braunschweig, Germany. His research interests include manufacturing process simulation, transportation systems simulation, supply chain design and simulation, and machine learning for finances. He is a member of IEOM.