JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH
ISSN: 2576-0017
2024, VOL 7, NO S8

Labor perception of the tourism sector using machine learning in the Puno region, 2024.

Cayo-Velasquez[1], Noemí Emperatriz[1] [0000-0002-9690-3006

1. National University of the Altiplano of Puno, Peru
noemicave@unap.edu.pe

Abstract:
The tourism sector is a dynamic and diverse industry that faces constant challenges in human resource management and job satisfaction. In this article, we present an innovative approach to understand job perception in the tourism sector using machine learning techniques. Using data collected from surveys applied to workers in the tourism sector, an analysis with Machine Learning algorithms to classify and predict various aspects of job perception, such as job satisfaction, employee loyalty and the probability of job turnover. Our study highlights the importance of using advanced computational approaches to better understand the complexities of the workforce in the tourism sector and provide valuable information for strategic decision making in human resource management. Gender diversification of men and women among its staff is not actively encouraged. Salary discrepancies are evident between men and women who perform identical or similar functions, strategies are not implemented to promote diversity and gender equality, nor are manifestations of prejudices avoided between the variables and indicators that contribute to the Tree model. Decisions are gender, sub-sector and the indicators are: there are no clear opportunities for positions without gender preference; gender diversification of men and women among its staff is not actively encouraged; but they do show salary discrepancies between men and women who perform identical or similar functions; Strategies to promote diversity and gender equality are not implemented, nor are manifestations of prejudice avoided; while in Random Forest the key variables are gender, position, subsector and length of service, along with the indicators that stand out are: there are no clear opportunities for positions without gender preference and the company's business culture offers a greater number of opportunities for men.
Class words: Job perception, equity and gender equality, decision tree and random forest

Introduction:
The tourism sector emerges as a fundamental pillar in the global economy, generating job opportunities for millions of people around the world. Despite its vital importance, this sector faces considerable challenges in terms of gender equality in the workplace. Job perception, which encompasses employees' subjective evaluation of their environment and working conditions, is strongly influenced by a number of factors, including gender dynamics. In this context, there is a pressing need to analyze in detail the disparities in job perception between men and women in this sector, using Machine Learning techniques as a promising tool to address these challenges more effectively.

For decades, job perception in the tourism sector has been the subject of interest, with numerous studies exploring various aspects related to employee satisfaction, commitment and retention. This previous research has identified factors such as salary, career growth opportunities, discrimination, work-life balance, and effective leadership as key determinants of job perception in this sector [1].

In addition, significant gender differences have been observed in job perception, which impacts job satisfaction and professional development opportunities [2]. Also, there are gender disparities in salaries and career advancement opportunities among employees, which consequently affects work experience, employee loyalty, and retention in jobs. This highlights the importance of organizational support in employee retention and how this can be perceived differently by men and women [3].

Therefore, it is crucial to explore the perception of the organizational climate and the importance of a favorable work environment to promote gender equality in the tourism sector.[4]

In that sense, this research adopts a comprehensive approach by addressing the phenomenon from various perspectives, despite its quantitative nature, given the relevance and persistence of gender inequalities in the tourism sector, this research seeks to answer the question: What are the factors that influence the job perception of the tourism sector using Machine Learning in the city of Puno in the year 2024?

Theoretical Framework:
This research is based on several theories and approaches that are relevant to understanding job perception in the tourism sector based on social exchange theory [5]. Employee job satisfaction is closely linked to the

perception of equity in the relationships between them and their employers [6], [7]. This theory states that employees constantly evaluate the benefits they receive from their employer in relation to the efforts and contributions they make at work. That is, employees compare what they give to the job (time, effort, skills) with what they receive in return (salary, recognition, development opportunities), and this comparison influences their overall perception of job satisfaction. The fundamental premise is that employees seek balance and fairness in these interactions, and when they perceive the relationship to be equitable, they are more satisfied with their jobs. On the contrary, if they perceive a discrepancy between what they give and what they receive, they experience job dissatisfaction and demotivation. This theory highlights the importance of equitable and fair employment relationships in determining employee job satisfaction [6].

Furthermore, job satisfaction is related to the good perception of equality and gender equity in the worker's performance, particularly in activities in the tourism sector [8].

Another relevant theory is the *Organizational Climate Model* , which focuses on how the work environment influences employees' perceptions of their work and their organization. According to this theory, the organizational climate, which encompasses aspects such as leadership style, the quality of internal communication and human resources policies, can have a significant impact on employee satisfaction and commitment in the tourism sector. In other words, organizational climate creates a context in which employees interpret their work environment and form their attitudes toward the job and the organization. A positive organizational climate, characterized by effective leadership, open communication, and fair and equitable human resources policies, can foster job satisfaction and employee engagement, which in turn can have positive effects on productivity and performance. organizational in the tourism sector [9].

Machine learning is a branch of artificial intelligence that focuses on the development of algorithms and models that allow computers to learn and perform tasks without being explicitly programmed [10].

Classification algorithms group data into categories or classes; regression algorithms predict continuous values; Clustering algorithms identify patterns and similarities in unlabeled data sets; and reinforcement learning algorithms rely on reward systems to learn how to make optimal decisions in changing situations [11].

Chi Square Test is applied for categorical data to classify into groups without a specific order whose purpose is to evaluate if there is a significant association between two variables, being a probabilistic distribution whose p-value if less than 0.05 means that they are associated [12].

Mann-Whitney U is a non-parametric statistical test for ordinal data type, it compares whether two independent numerical groups not necessarily normally distributed are significantly different in their medians, as long as their p-value is less than 0.05 [12].

Wald is a statistical method of inference that follows the chi-square distribution, used in the context of logistic regression models for the purpose of evaluating the significance of one or more parameters (multivariate) within a model. The significance of the test indicates that the coefficient is different from zero, so it has a contribution to the dependent variable [12].

Kruskal Wallis H test is an extension of the Mann-Whitney U test and represents an excellent alternative to the completely randomized one-way ANOVA.

Logistic regression is a supervised learning algorithm used to solve binary classification problems, it models the relationship between input variables and the probability that an instance belongs to a particular class, where it seeks to predict a discrete label [13].

The analysis of categorical data in the context of machine learning algorithms requires data encoding using One -Hot Encoding as a binary column with or without some inherent order, Label Encoding encodes each ordered category with a unique integer, not recommended for nominal data. Target Encoding replaces each category with a centralization measure derived from the target useful for models such as logistic regression and decision trees, but introduces overfitting if not handled properly. Frequency Encoding replaces categories with their frequency of occurrence in the dataset , appropriate when categories have a skewed distribution. Common Machine Learning Algorithms for Categorical Data are Decision Trees (Decision Trees), Random Forests set of Gradient decision trees Boosting Machines (GBM) either XGBoost , LightGBM and CatBoost , Naive Bayes good algorithm for categorical data using conditional probabilities [14].

Random Forest or Random Forests is a machine learning algorithm widely used for its ability to handle complex and high-dimensional data sets, and its flexibility to deal with different types of variables and classification problems [15].

Decision trees are machine learning algorithms that are used to classify instances based on a series of logical questions or conditions, achieving accurate and understandable classification due to their interpretability and ability to handle both categorical and continuous variables [1].

Naive Bayes is a machine learning algorithm that is based on Bayes' theorem to perform probabilistic classifications. Although it is based on a simplified and "naive" assumption, it is widely used for its efficiency and good performance in tasks such as text classification [16].
The data mining model based on the CRISP DM methodology is a predictive model to develop and carry out the collection of information to analyze the most relevant variables for the classification model, subsequently an exploration of the data is carried out to define the architecture of the model. [17].
Machine learning classification techniques, implemented in Python and the Scikit libraries Learn reveal that the overall accuracy rate of Random Forest is 94.14%, better than other algorithms for identifying students at risk of dropping out and although this research predicts students at risk of dropping out [18].
The deep learning algorithms used in deep learning, tensor flow to design the proposal of a deep learning model for making appropriate and timely decisions regarding the student dropout rate mainly in the field of virtual education in higher education institutions , being a continuous and long-term project that includes additional categorical attributes [14].
Machine learning techniques with data scaling have no effect on the final performance of linear discriminant analysis and Random Forest, on the other hand, scaling affects the performance of Support Vector Machine, but does not seem to add any benefit. The results obtained show that based on data without any pedagogical or didactic value, an attempt can be made to mitigate the problem of dropout [19].

Methodology:
To carry out this research, the methodology consisted of evaluating different machine learning algorithms for processing the information collected through the survey technique carried out with employees of various companies in the tourism sector, including hotels, travel agencies and public institutions. These surveys addressed different aspects of job perception, such as job satisfaction, work environment, and professional development opportunities. The survey included a total of 23 indicators. In addition, demographic and employment data were collected from employees, such as age, gender, educational level, and length of service in the company [19].
Once the data was collected, data preprocessing techniques were applied to clean and transform the information into a format suitable for analysis. Machine Learning algorithms, such as logistic regression, decision trees and neural networks, are used to develop predictive models of job perception. These models were trained using a portion of the data and their performance was evaluated using metrics such as accuracy, sensitivity, and specificity [1].
To evaluate the performance of the classification models, the metrics are obtained from a confusion matrix (MC) which describes the count of true positives (TP), true negatives (VN), false positives (FP) and false negative (FN). The rows represent the number of samples in the observed class and the columns the number of predictions for each class [11]. The MC diagonal corresponds to the number of samples that the algorithm correctly classifies into each class. If MC only has positive values on the diagonal, it indicates that the classifier correctly classifies all samples. The overall classification accuracy metric (PG) measures the overall proportion of well-classified samples in each class and is calculated as:

$$PG = \frac{VP + VN}{FP + FN + VP + VN}$$

The metrics to measure the performance of the classifier in each class are accuracy (P), sensitivity (S), specificity (E) and F1 score. They are defined with the following expressions:

$$P = \frac{VP}{VP + FP}$$
$$S = \frac{VP}{FN + VP}$$
$$E = \frac{VN}{VN + FP}$$
$$F1 = 2\frac{P \times S}{P + S}$$

In this case, the value of F1 summarizes P and S into a single metric, is an appropriate estimator in unbalanced classes, and varies between zero and one. The receiver operating curve characteristics (ROC) is a curve that relates values of S versus 1-E [15]. The different points on the curve correspond to the cut-off points used to determine whether the test results are positive. The AUCROC value (area under the ROC curve) is interpreted as the probability that, in two samples, one positive and one negative, the test assigns a higher probability to the positive sample, correct classification [20]. Its value ranges between zero and one;

the greater the AUCROCthe better the classification, a value close to 0.50 indicates a poor classification. The PS curve is the result of plotting P versus S. This allows us to observe from which S there is a degradation of P and vice versa. The ideal result is a curve that approaches the upper right corner (high P and S), which generates an area under the AUCP-S curve that, the closer it is to one, the better the model [21].

Table 01

Operationalization of the variables and indicators of job perception

Variable	**Indicators**	**Index**	**Guy**
Sex	Sex	1.2	nominal
Sector	Sub sector	1,2,… ,6	nominal
Prof.	Profession	1,2 and 3	nominal
Post	Position held	1, 2,3,… ,12	nominal
Tser	Service time	1,2,3, and 4	ordinal
Clab	Working Condition	1,2,1…,6	nominal
P1	There are clear opportunities for positions without gender preference.	1,2,3,4 and5	ordinal
P2	There are always opportunities to change sectors when necessary	1,2,3,4 and5	ordinal
P3	Options are provided to improve our professional qualifications	1,2,3,4 and5	ordinal
P4	Continuously finds opportunities and proposals to join work teams and projects.	1,2,3,4 and5	ordinal
P5	A clear commitment of the organization to promoting the promotion of women to leadership positions is evident.	1,2,3,4 and5	ordinal
P6	There are perceived obstacles that make it difficult for women to advance to positions of greater responsibility.	1,2,3,4 and5	ordinal
P7	There are clear opportunities for positions without gender preference.	1,2,3,4 and5	ordinal
P8	Men and women assume leadership equally	1,2,3,4 and5	ordinal
P9	Fair pay equity is detected between men and women	1,2,3,4 and5	ordinal
P10	An environment of mutual respect and trust is provided between men and women	1,2,3,4 and5	ordinal
P11	Gender diversification of men and women among its staff is actively encouraged.	1,2,3,4 and5	ordinal
P12	Women receive the same type of treatment as men in the work environment	1,2,3,4 and5	ordinal
P13	In the workplace, women are treated equally with men.	1,2,3,4 and5	ordinal
P14	In high-level positions, the tourism sector presents greater opportunities for women.	1,2,3,4 and5	ordinal
P15	At lower hierarchical levels, the tourism sector provides more opportunities for men	1,2,3,4 and5	ordinal
P16	At higher hierarchical levels, the tourism sector provides more opportunities for men	1,2,3,4 and5	ordinal
P17	The company's business culture offers a greater number of opportunities for women	1,2,3,4 and5	ordinal
P18	The company's business culture offers a greater number of opportunities for men	1,2,3,4 and5	ordinal
P19	In the organization, differentiated hiring and recruitment policies are implemented according to gender	1,2,3,4 and5	ordinal
P20	Salary discrepancies are evident between men and women who perform identical or similar functions	1,2,3,4 and5	ordinal
P21	Implement strategies to promote diversity and gender equality, avoid manifestations of prejudice	1,2,3,4 and5	ordinal
P22	Promotional job descriptions include men and women (Boss or Director).	1,2,3,4 and5	ordinal
P23	I hold an unpaid position of responsibility, that is, I carry out activities that would be those of superiors, but I do not charge for them	1,2,3,4 and5	ordinal
perceive	Tourist perception	0.1	ordinal

Adapted from:[22]

Results and discussion

The results according to the sociodemographic variables and the 23 indicators of the job perception questionnaire. For all items, the Likert scale ranged from 1: Strongly disagree to 5: Strongly agree. The percentage frequencies of the descriptive statistics in Table 02 include p-value of the chi-square test to determine the association between the characterization variables and the perception of gender equality, as well as the differences between the independent groups of said variables through U by Mann Whitney and Kruskal Wallis.

Table 02

Percentage frequencies of the characterization variables in relation to job perception

Variables	Values	No.	%	Sig , Chi-square	Next ,
Sex	Male	252	48.1		Mann Whitney U
	Female	272	51.9		
	Total	524	100.0	<0.001	<0.001
Sub sector	Public Administration	107	20.4		
	Lodging/Accommodation	282	53.8		Kruskal Wallis
	Travel agency	135	25.8		
	Total	524	100.0	0.007	0.003
Profession	Bachelor of Tourism	378	72.1		Kruskal Wallis
	Tour guide	111	21.2		
	Others	35	6.7		
	Total	524	100.0	0.008	0.547
Position held	Administrator	140	26.7		
	Store	12	23		Kruskal Wallis
	night auditor	20	3.8		
	Driver	11	2.1		
	Concierge	8	1.5		
	Consultant	29	5.5		
	Counter -office	60	11.5		
	Teaching	89	17.0		
	Guide	90	17.2		
	Reception	43	8.2		
	Restaurant	9	1.7		
	Transfer	13	2.5		
	Total	524	100.0	<0.001	0.051
Service time	less than 1 year	141	26.9		
	1 - less than 3 years	97	18.5		Kruskal Wallis
	Between 3 to 5 years	129	24.6		
	over 5 years	157	30.0		
	Total	524	100.0	0.034	0.870
Working Condition	Permanent contract	77	14.7		
	eventual contract	156	29.8		Kruskal Wallis
	Named	181	34.5		
	CAS service	27	5.2		
	others	83	15.8		
	Total	524	100.0	0.378	0.799
perceive	Bad	266	50.8		
	Good	258	49.2		
	Total	524	100.0		

Firstly , a bad perception of 50.8% is observed, compared to a good perception of 49.2%, this indicates that the perception is balanced, there is a highly significant association and difference with respect to the perception according to gender, there are more women, being 51.9 % compared to men; The activities according to sub sector, 53.8% represent lodgings/accommodations that are associated with perception and there is a highly significant difference between the subsectors; Professionally, 72.1% of the respondents had a degree in tourism, followed by guides, reflecting an association but not differentiable in perception; Depending on the position they occupy, administrators stand out more, resulting in being associated on equal

terms according to levels of perception; According to the length of service they have been performing, they are associated with the perception but without discriminating in their stay, their employment status is not associated or different in the perception.

Table 03

Percentage frequencies of the 23 job perception indicators on a Likert scale

The ordinal assessment of the Likert scale has the following scores: 1: Completely disagree; 2: Disagree; 3: Neither disagree/nor agree; 4: Agree and 5: Totally agree

Indicators/	1	2	3	4	5	Total	Next. Wald	Sig , Chi-square	Sig , Mann Whitney U
P1	1	15	53	21	10	100	0.332	0.076	0.484
P2	10	13	36	32	10	100	0.769	0.010	0.795
P3		53	30	17	1	100	0.434	0.243	0.623
P4	3	6	41	49	2	100	0.748	0.000	0.246
P5	3	41	48	8	1	100	0.932	0.784	0.917
P6		33	44	19	4	100	0.005	0.394	0.405
P7	2	38	49	11		100	---	0.702	0.431
P8	4	3	65	18	10	100	0.411	0.089	0.721
P9		11	46	43		100	0.196	0.703	0.448
P10		5	24	63	7	100	0.237	0.962	0.775
P11		1	37	55	8	100	0.514	0.012	0.276
P12		8	55	36	1	100	0.023	0.015	0.240
P13	12	31	46	10		100	0.103	0.079	0.527
P14	14	60	26			100	0.560	0.977	0.830
P15	6	36	46	12		100	0.573	0.942	0.979
P16			23	74	3	100	0.006	0.078	0.033
P17	32	49	9	10		100	0.533	0.393	0.526
P18	8	19	20	52	2	100	0.933	0.040	0.790
P19	1	8	26	64	1	100	0.093	0.085	0.542
P20		2	74	25		100	0.072	0.005	0.015
P21	5	57	34	3	0	100	0.430	0.356	0.834
P22	15	51	32	3		100	0.577	0.290	0.367
P23	9	27	61	3	1	100	0.029	0.006	0.004

In most of the indicators of job perception in equality or equity of conditions according to gender, respondents maintain an indifferent perception “Neither disagree/nor agree” with some exceptions such as case P4:She continually finds opportunities and proposals to join work teams and projects that are “in disagreement.” This position also prevails in P14: In high-level positions, the tourism sector presents greater opportunities for women; Q17: The company's business culture offers a greater number of opportunities for women; P21: Implement strategies to promote diversity and gender equality, avoid manifestations of prejudice; P22: The descriptions of promotion positions include men and women (Boss or Director) and P23: I hold an unpaid position of responsibility, that is, I carry out activities that would be those of superiors, but I do not charge for them.

Using the chi square test, the significant association of the following indicators was detected:

P2: There are always opportunities to change sectors when necessary , indifferent 36% tend to “Agree”; P4: Continuously finds opportunities and proposals to join work teams and projects in a 49% “Agree” position; P11: Gender diversification of men and women among your staff is actively encouraged, “Agree” at 55%; P12: Women receive the same type of treatment as men in the work environment, indifferent 55% followed by 36%; Q18: The company's business culture offers a greater number of opportunities for men “Agree” by 52%; P20: Salary discrepancies between men and women who perform identical or similar functions are evident at 74% indifference; P23: I carry out an unpaid position of responsibility, that is, I carry out activities that would be those of superiors, but I do not charge for them.

In the significant difference between bad and good perception according to indicators, it is then confirmed that Yes P16: At higher hierarchical levels, the tourism sector provides more opportunities for men; also if P20: Salary discrepancies are evident between men and women who perform identical or similar functions; Likewise, Q23: I carry out an unpaid position of responsibility, that is, I do carry out activities that would be those of superiors, but I do not charge for them.

Regarding the Wald Test, not all indicators have significance in the coefficients, which must be resolved to resolve the problem of multicollinearity.
Machine Learning used for categorical analysis of job perception requires prior coding of the data using with or without some inherent order using Label Encoding assigning a unique integer to each ordered category for indicators and not recommended for characterization variables.
Common Machine Learning Algorithms for Categorical Data are Logistic Regression, Decision Trees Trees), Random Forests , a set of decision trees and Naive Bayes for conditional probabilities.

Logistic regression as the machine learning method used classifies and predicts categorical variables, based on the probability of job perception according to gender equity or equality (yes/no, success/failure). Here the procedures and libraries for the survey data set applied according to the variables established above are:

```
# ======================================================================
import pandas as pd
import numpy as np

# Gráficos
# ======================================================================
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns

# Preprocesado y modelado
# ======================================================================
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import plot_confusion_matrix
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Configuración matplotlib
# ======================================================================
plt.rcParams['image.cmap'] = "bwr"
#plt.rcParams['figure.dpi'] = "100"
plt.rcParams['savefig.bbox'] = "tight"
style.use('ggplot') or plt.style.use('ggplot')
```

Loading the dataset
The class variable “Perceives” is dichotomous whose values are good and bad perception, for which there are 6 characterization variables, 23 indicators and 524 instances of labor perception related to gender equality in activities linked to the tourism sector in the Puno region.

In [9]:
```
# Datos
# ======================================================================
per = 'c:/percep3.csv'
datos = pd.read_csv(per)
datos.head(10)
```

Out[9]:

	Sexo	Ssector	Prof	Cargo	Tser	Clab	P1	P2	P3	P4	...	P15	P16	P17	P18	P19	P20	P21	P22	P23	Percibe
0	1	2	1	10	1	3	3	1	3	3	...	2	3	1	4	3	3	3	1	1	1
1	1	2	1	8	4	3	3	4	2	4	...	3	4	2	4	4	3	3	3	2	1
2	1	1	1	8	1	3	2	1	3	3	...	2	3	1	3	3	3	2	1	1	1
3	2	1	1	8	3	3	3	2	3	4	...	4	3	1	3	3	3	1	2	2	0
4	1	1	1	8	1	3	3	3	3	3	...	2	3	3	2	3	3	2	2	2	1
5	2	1	1	8	2	3	2	2	3	3	...	2	3	1	3	3	3	2	2	2	0
6	1	2	1	8	4	3	3	4	2	4	...	3	4	2	5	4	3	3	3	4	1
7	1	1	1	8	1	3	2	2	3	3	...	2	3	1	3	3	3	3	2	3	1
8	1	1	1	8	3	3	3	1	3	3	...	2	3	1	3	3	3	1	1	2	1

To obtain the classification frequencies in logistic regression, the algorithm was used:

```
In [10]: datos['Percibe'] = np.where(datos['Percibe'] == 1, 1, 0)

         print("Número de observaciones por clase")
         print(datos['Percibe'].value_counts())
         print("")

         print("Porcentaje de observaciones por clase")
         print(100 * datos['Percibe'].value_counts(normalize=True))
```

```
Número de observaciones por clase
0    266
1    258
Name: Percibe, dtype: int64

Porcentaje de observaciones por clase
0    50.763359
1    49.236641
Name: Percibe, dtype: float64
```

To fit the multiple logistic regression model, with the objective of predicting whether the perception is explained based on all the available variables and indicators, it is necessary to divide the training data of 80% and the test data of 20%.

```
In [7]: # División de los datos en train y test
        # ======================================================================
        X = datos.drop(columns = 'Percibe')
        y = datos['Percibe']

        X_train, X_test, y_train, y_test = train_test_split(
                                               X,
                                               y.values.reshape(-1,1),
                                               train_size   = 0.8,
                                               random_state = 1234,
                                               shuffle      = True
                                           )
```

Creating multiple logistic regression model

```
In [8]: # Creación del modelo utilizando matrices como en scikitlearn
        # ======================================================================
        # A la matriz de predictores se le tiene que añadir una columna de 1s para e
        X_train = sm.add_constant(X_train, prepend=True)
        modelo = sm.Logit(endog=y_train, exog=X_train,)
        modelo = modelo.fit()
        print(modelo.summary())
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: inf
         Iterations: 35

---------------------------------------------------------------------------
LinAlgError                               Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_8188\1383235006.py in <module>
      4 X_train = sm.add_constant(X_train, prepend=True)
      5 modelo = sm.Logit(endog=y_train, exog=X_train,)
----> 6 modelo = modelo.fit()
      7 print(modelo.summary())
```

When creating the model, an error " LinAlgError : Singular matrix" is obtained where the matrix to be inverted during the adjustment process is singular, that is, the matrix does not have an inverse, therefore, it is not possible to calculate the model coefficients However, this occurs when there are anomalies of perfect multicollinearity: where some predictor variables are perfectly correlated, because they are redundant indicators. In fact, it is necessary to verify multicollinearity based on very high correlations (close to 1 or -1), to eliminate these variables, using the Variance Inflation Factor (VIF) as follows:

In [ ]:
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
import statsmodels.api as sm

# Calcular matriz de correlación
correlation_matrix = X_train.corr()
print(correlation_matrix)

# Calcular VIF
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif = [variance_inflation_factor(X_train.values, i) for i in range(X_train.shape[1])]
print(vif)
```

Ord	VIF Factor	features	Ord	VIF Factor	features
0	677.434066	const	15	1.218856	P10
1	1.423236	Ssector	16	3.343566	P11
2	3.568425	Prof	17	3.950403	P12
3	1.302872	Position	18	4.752679	P13
4	2.116993	Tser	19	3.177992	P14
5	2.523289	Clab	20	3.816421	P15
6	2.949053	P1	21	3.770583	P16
7	2.759884	P2	22	2.639697	P17
8	3.413872	P3	23	1.400090	P18
9	4.510507	P4	24	1.513809	P19
10	2.330653	P5	25	2.894981	P20
11	2.833864	P6	26	3.133366	P21
12	5.187742	P7	27	5.189938	P22
13	1.891550	P8	28	2.372192	P23
14	2.657351	P9	29	1.158877	Perceive

According to the VIF, the indicators P7, P22 are eliminated, whose values are VIF 5.187742, 5.189938, respectively, coefficients greater than 5, a sufficient condition for their discarding.

The results according to the decision trees specify the dichotomous predictive models to classify job perception (good/bad).

In [1]:
```
# Tratamiento de datos
# ------------------------------------------------------------
import numpy as np
import pandas as pd

# Gráficos
# ------------------------------------------------------------
import matplotlib.pyplot as plt

# Preprocesado y modelado
# ------------------------------------------------------------
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.tree import plot_tree
from sklearn.tree import export_graphviz
from sklearn.tree import export_text
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import mean_squared_error

# Configuración warnings
# ------------------------------------------------------------
import warnings
warnings.filterwarnings('once')
```

The dataset loaded for the decision tree is:

```
data = pd.read_csv('c:/percep3.csv')
```

```
data.head(3)
```

	Sexo	Ssector	Prof	Cargo	Tser	Clab	P1	P2	P3	P4	...	P14	P15	P16	P17	P18	P19	P20	P21	P23	Percibe
0	1	2	1	10	1	3	3	1	3	3	...	2	2	3	1	4	3	3	3	1	1
1	1	2	1	8	4	3	3	4	2	4	...	2	3	4	2	4	4	3	3	2	1
2	1	1	1	8	1	3	2	1	3	3	...	1	2	3	1	3	3	3	2	1	1

Preprocessing of the dataset for the Decision Tree model is:

```
# División de los datos en train y test
# ----------------------------------------------------------------------------
X_train, X_test, y_train, y_test = train_test_split(
                                       data.drop(columns = "Percibe"),
                                       data['Percibe'],
                                       random_state = 123
                                   )
# Creación del modelo
# ----------------------------------------------------------------------------
modelo = DecisionTreeRegressor(
            max_depth         = 3,
            random_state      = 123
         )

# Entrenamiento del modelo
# ----------------------------------------------------------------------------
modelo.fit(X_train, y_train)
```

```
DecisionTreeRegressor(max_depth=3, random_state=123)
```

Decision Tree Tree Design

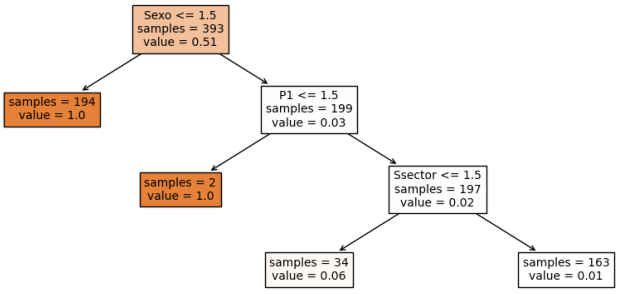
```
# Estructura del árbol creado
# ----------------------------------------------------------------------------
fig, ax = plt.subplots(figsize=(12, 5))

print(f"Profundidad del árbol: {modelo.get_depth()}")
print(f"Número de nodos terminales: {modelo.get_n_leaves()}")

plot = plot_tree(
            decision_tree = modelo,
            feature_names = data.drop(columns = "Percibe").columns,
            class_names   = 'Percibe',
            filled        = True,
            impurity      = False,
            fontsize      = 10,
            precision     = 2,
            ax            = ax
       )
```

```
Profundidad del árbol: 3
Número de nodos terminales: 4
```

Decision tree of job perception related to gender equity/equality



The variable that node rais represents is precisely gender, where 194 men consider good perception, on the other hand, 199 ladies state that there are no clear opportunities for positions without gender preference, in addition, 197 are graduates of which only 2 have good perception and 163 bad perception.

```
texto_modelo = export_text(
                    decision_tree = modelo,
                    feature_names = list(data.drop(columns = "Percibe").columns)
                )
print(texto_modelo)
```

```
|--- Sexo <= 1.50
|   |--- value: [1.00]
|--- Sexo >  1.50
|   |--- P1 <= 1.50
|   |   |--- value: [1.00]
|   |--- P1 >  1.50
|   |   |--- Ssector <= 1.50
|   |   |   |--- value: [0.06]
|   |   |--- Ssector >  1.50
|   |   |   |--- value: [0.01]
```

Importance of model predictors

```
importancia_predictores = pd.DataFrame(
                            {'predictor': data.drop(columns = "Percibe").columns,
                             'importancia': modelo.feature_importances_}
                            )
print("Importancia de los predictores en el modelo")
print("-------------------------------------------")
importancia_predictores.sort_values('importancia', ascending=False)
```

```
Importancia de los predictores en el modelo
-------------------------------------------
```

Ord	predictor	importance	Ord	predictor	importance
0	Sex	0.979046	13	P90.000000	
6	P10.020135		14	P100.000000	
1	S_sector	0.000819	12	P80.000000	
15	P110.000000		11	P60.000000	
25	P210.000000		10	P50.000000	
24	P200.000000		9	P40.000000	
23	P190.000000		8	P30.000000	
22	P180.000000		7	P20.000000	
21	P170.000000		5	Clab	0.000000
20	P160.000000		4	Tser	0.000000
19	P150.000000		3	Charge	0.000000
18	P140.000000		2	Depth	0.000000
17	P130.000000		26	P230.000000	
16	P120.000000				

In the importance of the variables and indicators of the results of the model, the gender variable stands out, where men express good perception, while women express bad perception, in second priority is the indicator "There are no clear opportunities for positions without gender preference." , followed by the sub-sector variable where lodging/lodging prevails, continue in the order of the importance of the indicators:
Gender diversification of men and women among its staff is not actively encouraged.
Salary discrepancies are evident between men and women who perform identical or similar functions
Strategies are not implemented to promote diversity and gender equality, nor are manifestations of prejudice avoided.
As for Pruning (cost complexity pruning) by cross-validation, pruning is not required since the results linked to the first process whose metrics will be the same obtained with pruning.

```
# Valores de ccp_alpha evaluados
param_grid = {'ccp_alpha':np.linspace(0, 80, 20)}

# Búsqueda por validación cruzada
grid = GridSearchCV(
        # El árbol se crece al máximo posible para luego aplicar el pruning
        estimator = DecisionTreeRegressor(
                            max_depth         = None,
                            min_samples_split = 2,
                            min_samples_leaf  = 1,
                            random_state      = 123
                        ),
        param_grid = param_grid,
        cv         = 10,
        refit      = True,
        return_train_score = True
       )

grid.fit(X_train, y_train)

fig, ax = plt.subplots(figsize=(6, 3.84))
scores = pd.DataFrame(grid.cv_results_)
scores.plot(x='param_ccp_alpha', y='mean_train_score', yerr='std_train_score', ax=ax)
scores.plot(x='param_ccp_alpha', y='mean_test_score', yerr='std_test_score', ax=ax)
ax.set_title("Error de validacion cruzada vs hiperparámetro ccp_alpha");
```

Model predictions

```
# Error de test del modelo inicial
#-----------------------------------------------------------
predicciones = modelo.predict(X = X_test)

rmse = mean_squared_error(
        y_true  = y_test,
        y_pred  = predicciones,
        squared = False
       )
print(f"El error (rmse) de test es: {rmse}")
```

The error (rmse) of test is: 0.020312157747340725

Randon Forest is a model of multiple individual decision trees, trained with a slightly different sample, using a bootstrapping technique , it predicts and combines estimates from each of all the trees that make up the model, the libraries used by the algorithm are:

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.tree import plot_tree
from sklearn.tree import export_graphviz
from sklearn.tree import export_text
from sklearn.model_selection import GridSearchCV
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
```

The dataset of job perception according to gender in tourist activities in Puno for the R andom Forest algorithm is:

```
data = pd.read_csv("c:/percep3.csv")
data.head(3)
```

	Sexo	Ssector	Prof	Cargo	Tser	Clab	P1	P2	P3	P4	...	P14	P15	P16	P17	P18	P19	P20	P21	P23	Percibe
0	1	2	1	10	1	3	3	1	3	3	...	2	2	3	1	4	3	3	3	1	1
1	1	2	1	8	4	3	3	4	2	4	...	2	3	4	2	4	4	3	3	2	1
2	1	1	1	8	1	3	2	1	3	3	...	1	2	3	1	3	3	3	2	1	1

Splitting the data for training and the corresponding Random Forest test is:

```
# División de los datos en train y test
# ------------------------------------------------------------------------
X_train, X_test, y_train, y_test = train_test_split(
                                        data.drop(columns = "Percibe"),
                                        data['Percibe'],
                                        random_state = 123
                                    )
```

Model creation

```
# Creación del modelo
# ------------------------------------------------------
modelo = RandomForestClassifier(
            n_estimators         = 100,
            random_state      = 123
         )
# Entrenamiento del modelo
# ------------------------------------------------------
modelo.fit(X_train, y_train)
```

```
RandomForestClassifier(random_state=123)
```

Model metrics:

```
# Metricas del modelo
#-----------------------------------------------------
predicciones = modelo.predict(X = X_test,)

print("Matriz de confusión")
print("-------------------")
confusion_matrix(
    y_true    = y_test,
    y_pred    = predicciones
)
```

The output result is:
confusion matrix
---
array([[72, 0],
[0, 59]], dtype =int64)
Model accuracy metric

```
accuracy = accuracy_score(
            y_true    = y_test,
            y_pred    = predicciones,
            normalize = True
           )
print(f"El accuracy de test es: {100 * accuracy} %")
```

Whose output is:
The test accuracy is: 100.0%
The accuracy metric of the model predictors is explained at 100% of the model
Importance of the model predictors

```
importancia_predictores = pd.DataFrame(
                            {'predictor': data.drop(columns = "Percibe").columns,
                             'importancia': modelo.feature_importances_}
                            )
print("Importancia de los predictores en el modelo")
print("-------------------------------------------")
importancia_predictores.sort_values('importancia', ascending=False)
```

The output of the importance of the predictors in the model is:

Ord	predictor	importance	Ord	predictor	importance
0	Sex	0.723818	18	P140.010260	
3	Charge	0.029166	21	P170.009230	
1	Sector	0.018562	11	P60.008927	
4	T_be	0.015422	10	P50.008341	
6	P10.014916		5	Clab	0.007701
22	P180.014354		15	P110.006587	
7	P20.014186		8	P30.006201	
26	P230.013620		24	P200.005666	
23	P190.013141		19	P150.005586	
9	P40.011832		2	Prof0.005139	
17	P130.011449		13	P90.004604	
14	P100.011342		16	P120.004439	
12	P80.010787		20	P160.004279	
25	P210.010445				

In this case, the order of contribution of the variables for the explanation of the Random Forest model is first gender, position held, sub sector to which it belongs and length of service, followed by indicators such as:
There are no clear opportunities for positions without gender preference
The company's business culture offers a greater number of opportunities for men

```
# Predicción de probabilidades
#-----------------------------------------------------------------
predicciones = modelo.predict_proba(X = X_test)
predicciones[:5, :]
```

array([[0.96, 0.04],

[0.87, 0.13],
[0.14, 0.86],
[0.03, 0.97],
[0. , 1.]])

```
# Clasificación empleando la clase de mayor probabilidad
# ------------------------------------------------------------------------
df_predicciones = pd.DataFrame(data=predicciones, columns=['0', '1'])
df_predicciones['clasificacion_default_0.5'] = np.where(df_predicciones['0'] > df_predicciones['1'], 0, 1)
df_predicciones.head(3)
```

	01classification_default_0.5
0	0,960,040
1	0,870,130
2	0,140,861

The application of Machine Learning techniques applied to categorical data analysis of job perception offers several significant advantages for human resource management in the tourism sector. These approaches allow a deeper and more holistic understanding of the factors that influence job perception according to gender equity/equality. The results show a good perception in men and a bad perception in women, in the subsector of licensed professionals in tourism. , who very consciously identify gaps in the exercise of their profession, requiring effective strategies to overcome these practices in the exercise of the profession. However, it is important to keep in mind that the success of these models depends largely on the quality and representativeness of the data used, as well as the ability to interpret and contextualize the results obtained.

Conclusions:

In summary, this study demonstrates the potential of using Machine Learning techniques to analyze and understand job perception in the tourism sector. By leveraging the wealth of data available in this industry, organizations can gain valuable insights to improve employee satisfaction and engagement, which in turn can lead to better organizational performance and increased talent retention. Furthermore, this approach can be adapted and applied to other economic sectors to address similar challenges in human resource management.

For the decision tree algorithm, the contribution of the variable is gender followed by sub sector, while the indicators stand out:

Gender diversification of men and women among its staff is not actively encouraged.

Salary discrepancies are evident between men and women who perform identical or similar functions

Strategies are not implemented to promote diversity and gender equality, nor are manifestations of prejudice avoided.

For the Random Forest algorithm, the contribution of the predictors are the gender variables. Position, subsector, length of service and indicators:

There are no clear opportunities for positions without gender preference

The company's business culture offers a greater number of opportunities for men

Bibliographic References:

[1] A. Moreno *et al.* , *Machine learning* . 1994. doi : 10.5821/ebook-9788483019962.

[2] Z. Qin and Y. Pan, “Design of A Smart Tourism Management System through Multisource Data Visualization-Based Knowledge Discovery,” *Electronics (Switzerland)* , vol. 12, no. 3, 2023, doi : 10.3390/electronics12030642.

[3] D. Buhalis and C. Cooper, “Tourism Management,” in *Encyclopedia of Tourism Management and Marketing* , 2023. doi : 10.4337/9781800377486.tourism.management.

[4] MO Morgan, EE Okon, WH Emu, OIE Olubomi , and HU Edodi , “Tourism management: A panacea for sustainability of hospitality industry,” 2021. doi : 10.30892/GTG.37307-709.

[5] R. Ahmad, MR Nawaz, MI Ishaq, MM Khan, and HA Ashraf, “Social exchange theory: Systematic review and future directions,” 2023. doi : 10.3389/fpsyg.2022.1015921.

[6] D. Chhabra, K. Andereck, K. Yamanoi , and D. Plunkett, “Gender equity and social marketing: An analysis of tourism advertisements,” *Journal of Travel and Tourism Marketing* , vol. 28, no. 2, 2011, doi : 10.1080/10548408.2011.545739.

[7] Y. Zhao, “STEAM Display Path for Tourism Management in Era of Industry 4.0,” *EAI Endorsed Transactions on Scalable Information Systems* , vol. 10, no. 6, 2023, doi : 10.4108/eetsis.3942.

[8] AV Barza and M. Galanakis, “The Big Five Personality Theory and Organizational Commitment,” *Psychology* , vol. 13, no. 03, 2022, doi : 10.4236/psych.2022.133027.

[9] L. Cardenas -Niño and Y. Arciniegas- Rodriguez , “Intervention model in organizational climate,” *Int J Psychol Res (Medellin)* , vol. 2, no. 2, 2009.
[10] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf Process Manag* , vol. 45, no. 4, pp. 427–437, Jul. 2009, doi : 10.1016/J.IPM.2009.03.002.
[11] B. Mahesh, “Machine Learning Algorithms – A Review,” *International Journal of Science and Research (IJSR)* , vol. 9, no. 1, 2020, doi : 10.21275/art20203995.
[12] V. Berlanga María José Rubio, V. Berlanga Silvente Professor, M. José Rubio Hurtado Professor, and V. Berlanga Silvente María José Rubio Hurtado, “How to apply them in SPSS,” *Revista d'Innovació i Recerca en Educació* , vol . 5, no. 2, pp. 101–113, 2012, doi : 10.1344/reire2012.5.2528.
[13] E. W. Ingwersen *et al.* , “Machine learning versus logistic regression for the prediction of complications after pancreatoduodenectomy,” *Surgery (United States)* , vol. 174, no. 3, 2023, doi : 10.1016/j.surg.2023.03.012.
[14] J. Díaz-Ramírez, “Machine Learning and Deep Learning,” *Ingeniare. Magazine Chilean Engineering ,* vol. 29, no. 2, 2021, doi : 10.4067/s0718-33052021000200180.
[15] MR Ali, SMA Nipu , and SA Khan, “A decision support system for classifying supplier selection criteria using machine learning and random forest approaches,” *Decision Analytics Journal* , vol. 7, 2023, doi : 10.1016/j.dajour.2023.100238.
[16] LK Foo, SL Chua, and N. Ibrahim, “Attribute weighted naïve bayes classifier,” *Computers, Materials and Continua* , vol. 71, no. 1, 2022, doi : 10.32604/cmc.2022.022011.
[17] MCSG Cornejo Sifuentes, LG Vega Pérez, MG Naranjo Cantabrana , Ing. IF Osúa Acosta, FA Ávila Santana, and M. de los Á. Sotomayor Fierro, “Predictive Model of School Dropouts in Higher Education: an Approach from Data Mining Using the CRISP-DM Methodology,” *Ciencia Latina Revista Cientifica Multidisciplinar* , vol. 7, no. 5, 2023, doi : 10.37811/cl_ rcm.v 7i5.8363.
[18] D.Y. Chen, *Pandas for Everyone: Python Data Analysis, 2nd Edition* . 2023.
[19] N. Gupta and J. Li Wen, “Evaluating machine learning algorithms to classify forest tree species through satellite imagery,” *J Emerg Investig* , 2023, doi : 10.59720/22-153.
[20] B. Hamner and M. Frasco, “Metrics: Evaluation metrics for machine learning,” 2018.
[21] L. Zúñiga Segura, “Artificial intelligence in organizations,” *Investiga TEC* , vol. 15, no. 45, 2022.
[22] R. Soares, J. Renner , L. Paola Macedo Castro Gabriel, and S. Romo, “El mercado laboral desde una perspectiva de género: percepciones en el sector turístico español,” *Revista Venezolana de Gerencia (RVG)* , vol. 25, no. 92, pp. 1478–1501, 2020, [Online]. Available : https://orcid.org/0000-0003-1490-8944

.