COBEM 2023 27TH INTERNATIONAL CONGRESS OF MECHANICAL ENGINEERING

ABCM Brazilian Society of Mechanical Sciences and Engineering
27th ABCM International Congress of Mechanical Engineering
December 04-08, 2023, Florianópolis-SC, Brazil

COB-2023-0823
EXPLORATORY DATA ANALYSIS APPLIED TO BEARING MANUFACTURING PROCESS IN THE AUTOMOTIVE FIELD

Alan Lopes
Isabelle Therezinha Simão
Luiz Eduardo Thomaz
Mechanical Engineering Graduate Program (PPGEM), Pontifical Catholic University of Parana (PUCPR), Curitiba, PR, Brazil
lopes.alan@pucpr.edu.br, isabelle.therezinha@pucpr.edu.br, luiz.thomaz@pucpr.edu.br

Viviana Cocco Mariani
Mechanical Engineering Graduate Program (PPGEM), Pontifical Catholic University of Parana (PUCPR), Curitiba, PR, Brazil
viviana.mariani@pucpr.br

Leandro dos Santos Coelho
Industrial and Systems Engineering Graduate Program (PPGEPS), Pontifical Catholic University of Parana (PUCPR), Curitiba, PR, Brazil
Department of Electrical Engineering, Federal University of Parana (UFPR), Curitiba, PR, Brazil
leandro.coelho@pucpr.br

Abstract. *Bearings are mechanical transmission elements used to reduce friction between two surfaces in relative motion. It is used in a variety of engineering applications, and are widely used in transportation equipment such as cars, trucks, trains and airplanes to help ensure that the movement of parts is smooth and accurate and reduce vibration and noise. Bearings are composed of series of balls, rollers, or needles arranged in a mechanical separator and an inner and outer ring. However, over time, these elements can exhibit defects that impair their performance and can even lead to equipment failure. Therefore, research involving the control of the constructive characteristics of bearings is indispensable. In bearing manufacturing characteristics studies, correlation analysis can be used to evaluate the relationship between different variables that affect bearing quality and performance. Therefore, this research aims to evaluate two case studies derived from experimental data of geometric characteristics generated through the manufacturing process of automotive bearings, containing data of bearings without defects and with defects, respectively. Through correlation analysis, the goal is to verify how the variables behave and if there are any relationship between the variables. In both case studies, values above 0.5 were obtained, therefore positive correlations. From Phik's correlation the positive correlations were obtained with values close to one. A difference was observed in the values obtained for the correlations using bearings with defects when compared to the data for bearings without defects. It is worth noting that this is a preliminary study applied to the automotive industry. Thus, it was possible to identify the existence of the relationship between the constructive variables, such as vibration, roughness, and shape errors, derived from the manufacturing process of the bearings.*

Keywords: *Correlation analysis, bearing manufacturing, feature engeneering, regression, random forest*

1. INTRODUCTION

Bearings are mechanical transmission elements widely used in engineering to reduce friction between two surfaces in relative motion. They play an important role in the efficient operation of a wide range of machines and equipment such as engines, turbines, machine tools, transport and processing equipment. The ability of bearings to reduce friction and withstand loads is essential to ensure smooth and precise movement of moving parts, as well as to reduce unwanted vibrations and noise. These mechanical elements also play a key role in optimising the performance and lifespan of machinery and equipment, contributing to the operational efficiency, reliability and safety of industrial processes (Liu, Tan and Huang, 2022). In addition to their operational importance, bearings also have a significant economic impact on industry. The proper use of reliable, high-quality bearings helps to reduce maintenance and repair costs, increasing equipment availability and productivity. In addition, the continuous evolution of bearing technology, with advances in materials, geometries and manufacturing techniques, drives innovation and the development of new products and technical solutions in engineering (Chen et al., 2023).

Automotive bearings are designed with the specific requirements of the automotive sector in mind. They must be able to withstand high loads, high speeds and operate in adverse environmental conditions such as extreme temperatures, humidity and the presence of contaminants. In addition, the durability and reliability of the bearings are critical, as any failure will result in serious consequences for the safety of the vehicle and its occupants (Srivania, Arunkumarb and Ashok, 2018).

The bearing consists of different essential parts that work together, as shown in Figure 1. The inner ring is one of the main parts of the bearing. It is usually mounted directly on the shaft or component that is in motion. The inner ring is responsible for transmitting the loads from the bearing to the shaft or component and also provides the contact surface for the balls or rollers. The outer ring surrounds the inner ring and is usually fixed in a bracket or housing. The outer ring also carries the loads of the bearing and provides the contact surface for the balls or rollers. The cage keeps the balls evenly spaced around the inner ring and prevents them from touching, allowing smooth and efficient movement. The cage also helps distribute loads evenly and minimises friction between balls. Balls are the most common rolling elements in ball bearings. They are balls made of steel or other high-strength material that are inserted between the inner and outer rings. The balls roll between the rings, allowing friction reduction and load transmission. The sealing component allows the bearing not to be contaminated with fine particles or moisture, for example (Srivania, Arunkumarb and Ashok, 2018).

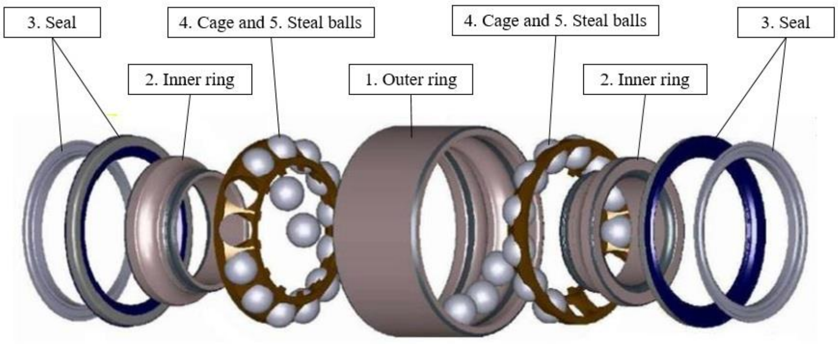


Figure 1. Bearing parts.

The control of the constructive characteristics of bearings is extremely important to ensure their adequate performance and avoid premature failures. Several factors influence the constructive characteristics of bearings, such as dimensional tolerances, raceway geometry, surface roughness, internal clearance and materials used. Various bearing fault detection methods, including the wavelet packet transform method (Lau and Ngan, 2010), the acoustic emission method (Morhain and Mba, 2003, Elforjani and Mba, 2010), the vibration method (Ghafari, Golnaraghi and Ismail, 2006) the principal component analysis method (Safizadeh and Latifi, 2014) the vibration and acoustic monitoring method (Liu, Wu and Liu, 2011), the current and temperature monitoring method (Seo *et al*, 2011) and the wear debris detection method (Peng, and Kessissoglou, 2003) have all been developed by researchers in the recent years. The majority of the time, diagnostic techniques based on the evaluation of mechanical vibration signals have shown to be successful (Samanta and Al-Balushi, 2003). Some researchers implemented various soft computing techniques, such as support vector machines, artificial neural networks, genetic algorithms, adaptive neuro-fuzzy, and hybrid partial swarm optimization and support vector machines, to develop multisensory data fusion based fault diagnostic methods (Safizadeh and Latifi, 2014).

There are several research works focused on extracting statistical characteristics from vibration data, but the analysis of the harmonic components of vibration signals is considered very limited, hence the importance of controlling the constructive characteristics of bearings. Moreover, over time, bearings may present defects that compromise their performance and may even lead to equipment failure. Therefore, research that approaches the control of the constructive characteristics of bearings becomes indispensable. The analysis of correlations between variables is a statistical technique that can be used to evaluate the relationships between important bearing variables. In studies of bearing manufacturing characteristics, correlation analysis can be employed to evaluate the relationship between different variables that affect bearing quality and performance.

The main contributions of this study can be summarized as follows: the first contribution is related to evaluation of the constructive characteristics of bearings, allowing understanding how variables such as vibration, roughness, and shape errors, affect bearing performance. Secondly, analysis of the use of statistical correlations and, finally, application of the Random Forest (RF) model in order to regression exploring the relationships among the variables and identifying possible patterns or complex interactions that may influence the bearing performance.

The rest of this paper is structured as follows. Section 2 describes the dataset employed for analysis. Section 3 defines the adopted correlations and the regression model applied at the data sets. Section 4 presents the results obtained and the discussions. Finally, Section 5 concludes with the final considerations and future works.

2. DATASET DESCRIPTION

The high-precision component manufacturer, based in the Curitiba metropolitan area, provided samples for this study's usage of and analysis of data on bearing properties. Axial clearance control, vibration, roughness, and ovalization are the parameters for which data were retrieved, with 25 input and 3 output variables described in Table 1.

Table 1. Description of the variables present in dataset.

Variable name	Type	Group	Unit	Variable description
LB1	Output	Quality control	dB	Level of bearing vibration, measured at raceway position 50 Hz and 300 Hz
LB2				Level of bearing vibration, measured at raceway position 300 Hz and 1800 Hz
LB3				Level of bearing vibration, measured at raceway position 1800 Hz and 10000 Hz
CJA	Input	Dimensional	µm	Bearing axial clearance
BE - PP	Input	Geometric	µm	Outer ring raceway deformation between two peaks
BE - PV	Input	Geometric	µm	Outer ring raceway deformation between peaks and valleys
BE – PP A	Input	Geometric	µm	Outer ring raceway deformation between two peaks on zone A
BE – Radius A	Input	Dimensional	µm	Outer ring raceway radius on zone A
BE – PV A	Input	Geometric	µm	Outer ring raceway deformation between peaks and valleys on zone A
BE - Ra	Input	Surface state	µm	Outer ring raceway roughness
BE - Circularity	Input	Geometric	µm	Outer ring raceway roundness - overall deviation
BE - Ovality	Input	Geometric	µm	Outer ring raceway roundness - 2 points deformation
BE - Triangulation	Input	Geometric	µm	Outer ring raceway roundness - 3 points deformation
BE – Profile-Height	Input	Dimensional	µm	Outer ring raceway shoulder
BE - Concentricity	Input	Dimensional	µm	Outer ring concentricity between outer diameter and raceway
BE – Perpendicularity Face/Øext	Input	Dimensional	µm	Outer ring perpendicularity between face and outer diameter
BI - PP	Input	Geometric	µm	Inner ring raceway deformation between two peaks
BI - PV	Input	Geometric	µm	Inner ring raceway deformation between peaks and valleys
BI – PP A	Input	Geometric	µm	Inner ring raceway deformation between two peaks on zone A
BI – Radius A	Input	Dimensional	µm	Inner ring raceway radius on zone A
BI – PV A	Input	Geometric	µm	Inner ring raceway deformation between peaks and valleys on zone A
BI - Ra	Input	Surface state	µm	Inner ring raceway roughness
BI - Circularity	Input	Geometric	µm	Inner ring raceway roundness - overall deviation
BI - Ovality	Input	Geometric	µm	Inner ring raceway roundness - 2 points deformation
BI - Triangulation	Input	Geometric	µm	Inner ring raceway roundness - 3 points deformation
BI – Profile-Height	Input	Dimensional	µm	Inner ring raceway shoulder
BI – Perpendicularity Face/Øint	Input	Dimensional	µm	Inner ring perpendicularity between face and bore diameter
Ball class	Input	Dimensional	µm	Steel ball class - Deviation between nominal diameter

3. METHODS

This section presents the main aspects of the methods used in this study.

3.1 Pearson Correlation

In characterizing the correlation profile, Pearson's basic correlation is commonly used to describe the relationship between time series. Huang *et al*., (2019) used Pearson correlation analysis to study the correlation between national stock indices before and after the establishment of the European Union (EU) and found a significant increase in intercorrelation between stock indices across Europe after the establishment of the EU. Relationships between financial markets are influenced by multiple factors and are typically nonlinear; thus, linear correlation methods do not fully exploit the potential information they can provide.

The association between continuous quantitative variables that represents the simplest form is the linear type. In this case, this measure is intended to assess how close a straight line can be to a point cloud formed by the ordered pairs across two attributes. The degree of association between two or more variables is also known as linear relationship and the most popular coefficient used to quantify it is Pearson's coefficient. Karl Pearson was the one who first described the standard method for calculating it.

Any value within the interval [-1,1] can be assumed by the correlation coefficient. All values that are equal to 0 denote that there is no linear association between the two variables. On the other hand, values above zero indicate that there is a positive linear correlation, and that therefore there is a tendency that an increase or decrease can occur jointly. If the value of one variable increases and the value of the other decreases, it is said that there was a negative linear association, and in this case, the values are below zero. In addition, the coefficient value indicates the intensity of the existing association. The sign, on the other hand, represents the direction of the relationship between the pair of attributes. Therefore, the stronger the association between the variables, the closer to 1 or -1 the Pearson coefficient will be. Otherwise, if the value of the coefficient is closer to 0, then the variation around the straight line that best fits the data will be greater Baak *et al.*, (2020).

3.2 Sperman Correlation

Spearman's correlation coefficient is tied to measuring the strength and direction of the association between two continuous or ordinal variables. As Hauke and Kossowski (2011) described, Spearman's coefficient assesses how well an arbitrary monotonic function can represent the relationship that exists between two variables. Therefore, in the monotonic relationship there is a tendency for the variables to change together, not necessarily meaning at a constant rate, as happens in linear relationships. Furthermore, motonicity is not the rule in Spearman correlation, just as linearity is not mandatory in Pearson correlation either.

According to Baak *et al.*, (2020) Spearman's coefficient describes the strength of the monotonic relationship and also varies in the interval [-1,1]. Therefore, the closer the absolute value of rs is to 0, the monotonic relationship between two variables will be weaker, consequently, the closer to -1 or 1, the stronger the monotonic relationship will be. If they have values equal to -1 or 1 it means a good association between the variables classified according to the order of their values. Furthermore, there is the possibility of Spearman's coefficient being equal to 0, in which case the variables are related in a non-monotonic way.

3.3 Phik Correlation

Phik (ϕk) is a new and handy correlation coefficient that works between categorical, ordinal and interval variables, captures non-linear dependence and reverts back to Pearson's correlation coefficient in the case of a bivariate input normal distribution. In many fields (not just data science), Pearson's correlation coefficient is a standard approach to measuring the correlation between two variables. However, there are some negatives such as: it only works with continuous variables, it only represents a linear relationship between variables, it is sensitive to outliers.

As for the positive aspects, the main differentiators of Phik can be described as: it is based on several refinements of Pearson's χ2 (chi-squared) contingency test - a hypothesis test of independence between two (or more) variables, it works consistently across categorical, ordinal and interval (continuous) variables, it captures non-linear dependencies, it reverts to Pearson's correlation coefficient in the case of a bivariate normal distribution of the input, the algorithm contains an integrated noise reduction technique against statistical fluctuations (Baak *et al.*, 2020).

Baak *et al.* (2020) described that the metric most similar to Phik is Cramer's ϕ, which is a correlation coefficient intended for two categorical variables and is also based on Pearson's χ2 test statistic. What is important to note is that while it is a measure used for categorical variables, it can also be used for ordinal variables and binned interval variables.

However, the value of the coefficient is highly dependent on the binning chosen per variable and can therefore be difficult to interpret and compare. This is not the case for Phik. Furthermore, Cramer's ϕ is sensitive to outliers, especially for smaller samples. As for the downsides of this new method, one can exemplify how: the calculation of ϕk is computationally expensive (due to the calculation of some integrals under the hood), no closed-form formula, no indication of direction, when working with only numerical variables, other correlation coefficients will be more accurate, especially for small samples.

3.4 Feature Engineering

The process of choosing a pertinent and instructive subset of features from the original data set to be utilized in creating machine learning models is referred to as feature selection. The objective is to eliminate unnecessary or redundant characteristics and select the most crucial features that contribute significantly to the model's accurate classification or prediction. In machine learning, feature selection is crucial, especially when dimensionality reduction is involved. By choosing an appropriate subset of features, you may minimize the dimensionality of the data. This is particularly helpful when there are several characteristics because the curse of dimensionality might cause the model to use excessive computational resources and lead to overfitting issues. Additionally, by removing pointless or redundant features, the model tends to focus on the most informative features, which tends to increase model accuracy and performance. This may lead to increased model interpretability, less training time, and enhanced prediction performance (Wang *et al*., 2023).

3.5 Random Forest Model

Random Forest (RF) is a widely used machine learning algorithm for classification and regression problems. It is based on the concept of ensemble learning, where multiple decision trees are built and combined to produce a final prediction. Each decision tree in the forest is trained on a random sample of training data, called bagging, and combining the predictions from all the trees results in the final prediction. Randomness is introduced in the construction of the decision trees by selecting a random subset of features for each node split. Furthermore, during classification, an average or vote of the predictions from all trees is used to determine the final class. This approach of combining multiple trees and introducing randomness helps to reduce overfitting and improve the generalizability of the model. RF is known for its ability to handle large data sets, handle both continuous and categorical input variables, and provide variable importance estimates that aid in the interpretation of results (Garg and Goel, 2023).

The construction of the RF involves several steps. First, random selection of a sample of the training data for each tree is performed. Then, the trees are built recursively by splitting the nodes based on a criterion such as node purity or impurity reduction. During splitting, a random subset of features is selected and the best split is chosen based on criteria such as information gain or Gini index. These steps are repeated until all trees are built (Garg and Goel, 2023). For classification, predictions are obtained by combining the predictions of all trees using an average or voting. According to Eq. (1), for regression problems, the final prediction is usually the average of the tree predictions:

$$\overline{h}(x) = \frac{1}{T}\sum_{t=1}^{T}\{h(x, \theta_t\,)\} \tag{1}$$

where $\overline{h}$ is the predicted outcome of the model, x is the independent variable, T is the number of regression decision trees, θ_t is the vector of independent and equally distributed random features and $h(x, \theta_t)$ is the outcome based on x and θ_t.

3.6 Performance metrics

In order to run Python code directly from the browser without first setting up a local environment, Google Collaboratory, a free cloud-based program, was used to perform the calculations. It has capabilities including real-time collaboration, code editing, and result visualization and is built on the Jupyter Notebook platform. Mean absolute error (MAE), root mean square error (RMSE), mean square error (MSE), and the coefficient of determination (R^2), an assessment metric used in machine learning models to gauge how well the model fits the training and test data, were used to compare the performances of the different techniques.

To assess how well a prediction model performs in comparison to the actual values, regression analysis frequently uses the statistical metric known as the MAE, given by Eq. (2). It computes the mean of the absolute differences between the predicted values of the dependent variable and their actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{x}_i| \tag{2}$$

A statistical indicator called MSE is frequently used in regression analysis to assess how well a predicted model performs in comparison to real values. The MSE, given by Eq. (3) calculates the mean of the squares of the discrepancies between the dependent variable's predicted values and its actual values. Additionally, because the mistakes are squared, it is a metric that penalizes larger errors more severely. This indicates that the overall MSE value is more significantly affected by larger errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{x}_i)^2 \tag{3}$$

The RMSE, is a statistical metric used to assess how well a predictive model performs when compared to real values. The RMSE, given by Eq. (4) is a variation of the MSE in which the final result is square-rooted to put the metric on the same scale as the initial dependent variable. Because the errors are squared before obtaining the square root, the RMSE, like the MSE, penalizes greater errors more severely. This metric also takes into account the variance between the predicted values of the model and the actual values of the dependent variable and is sensitive to outliers.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{x}_i)^2} \tag{4}$$

A statistical metric used in regression analysis to evaluate the model's fit to the observed data is the coefficient of determination, or R^2, given by Eq. (5). This statistic gives an indication of the percentage of the dependent variable's overall variability that the model developed by Huang *et al*., (2011) and Liu *et al*., (2018) accounts for. This performance metrics can all be expressed as follows:

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{x}_i)^2}{\Sigma(y_i - \bar{x}_i)^2} \tag{5}$$

where y_i and x_i are the desired output and estimated output, respectively, and n represents each sample in the data set. The flowchart of the proposed methodology as described in this section is illustrated in Figure. 2.

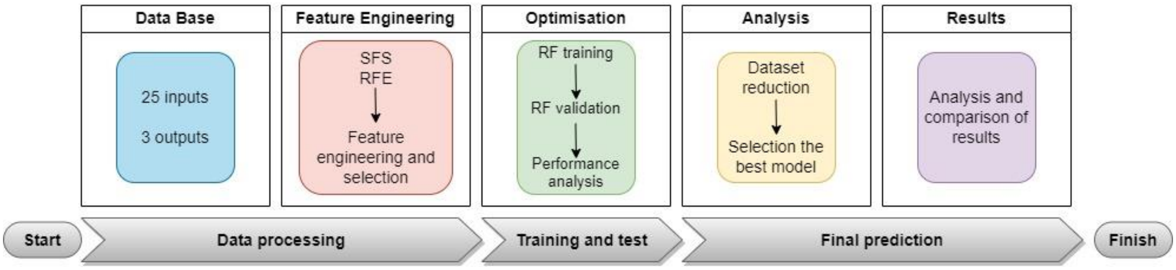


Figure 2. Flowchart of the proposed methodology.

4. RESULTS AND DISCUSSION

4.1 Correlation Analysis

This section presents the results obtained using the data set referring to the bearing characteristics, thus being 25 input variables and 3 output variables. First, correlation studies using Pearson, Spearman, and Phik were carried out.

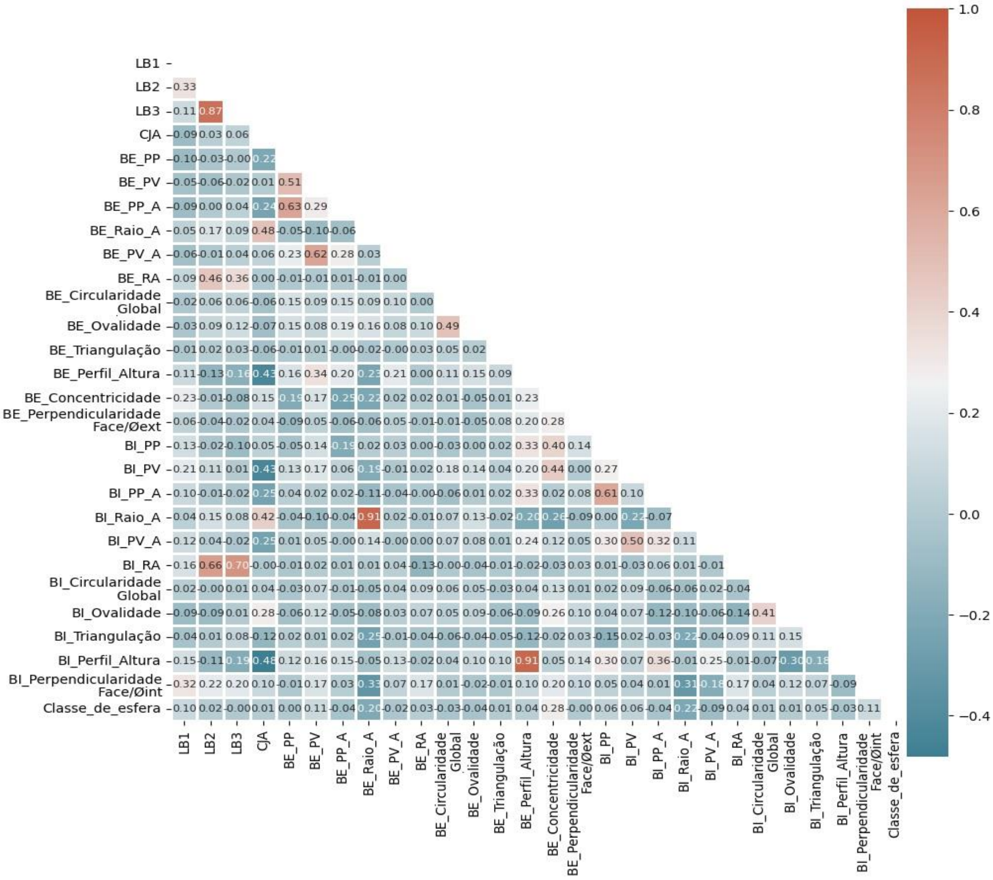
The range of Pearson's correlation coefficient is -1 to +1. A value of +1 denotes a perfect positive correlation. Values of -1, on the other hand, denote a complete negative correlation. The output variables for this Pearson correlation study are LB1, LB2 and LB3, which are variables representing the level of vibration, see Figure. 3(a). LB1, LB2 and LB3 initially represent the vibration of the measurement system, the conditions of the outer ring and the conditions of the inner ring, respectively; LB1 will be ignored for the analyses in this work. The second and third columns of this analysis are the most important.

The relationship between the vibration level and the RA of the inner and outer rings can be easily understood by doing a straightforward analysis. The measurement of the raceway of the bearing rings yields a number called RA. The importance of these characteristics was demonstrated using a significance analysis; the P-value for each correlation value is less than 0.05. It is possible to analyze the monotonic relationship between the variables, i.e., whether the variables tend to increase or decrease together, regardless of the exact form of the relationship, when performing Spearman's analysis, which is typical in non-parametric data applications, such as ordinal or interval data, as well as continuous data. Figure 3(b). Spearman correlation shows a complete lack of correlation, leading one to believe that the data indicate a trend toward linearity.

The Phik is a brand-new, useful correlation coefficient that captures non-linear dependence, operates consistently across categorical, ordinal, and interval variables, and reverts to Pearson's correlation coefficient in the event of input data with a bivariate normal distribution. In this analysis, it is simple to detect the significant correlation between the RA of the inner and outer rings with the level of bearing vibration due to the Phik correlation results, which are shown in Figure 4 for this case's second and third lines from bottom to top. The radius of the outer ring raceway, a new property that has good significance, has emerged.

Pearson's correlation, which in this instance indicates the correlation of a variable with itself, discards correlations lower than 0.3 or very high correlations, as seen in Figure 3(a). As a result, only the connection with the variable BI-Perpendicularity Face/int was taken into consideration for the output variable LB1. In order to consider the two correlations with the variables BI-Ra and BE-Ra for the output variables LB2 and LB3, 26 correlations were first eliminated. As indicated in Figure 3(b), a value of 0.3 was used as an exclusion criterion for the correlations that were too low for the Spearman's correlation analysis. In order to acknowledge the correlations with the variables BI-Perpendicularity Face/int, BI-Ra, and BE-Ra, respectively, for the output variables LB1, LB2, and LB3, 27 correlations were first eliminated. The same criterion was taken into consideration during Phik's examination of the correlations, taking into account a value of 0.3. For the outcome variable LB1, four correlations—BI- Profile-Height, BE-Concentricity, BE- Profile-Height, and BE-Radius A were absorbed whereas 24 correlations were eliminated. 19 out of the 28 outputs for LB2 and LB3 were eliminated.

(a)



(b)

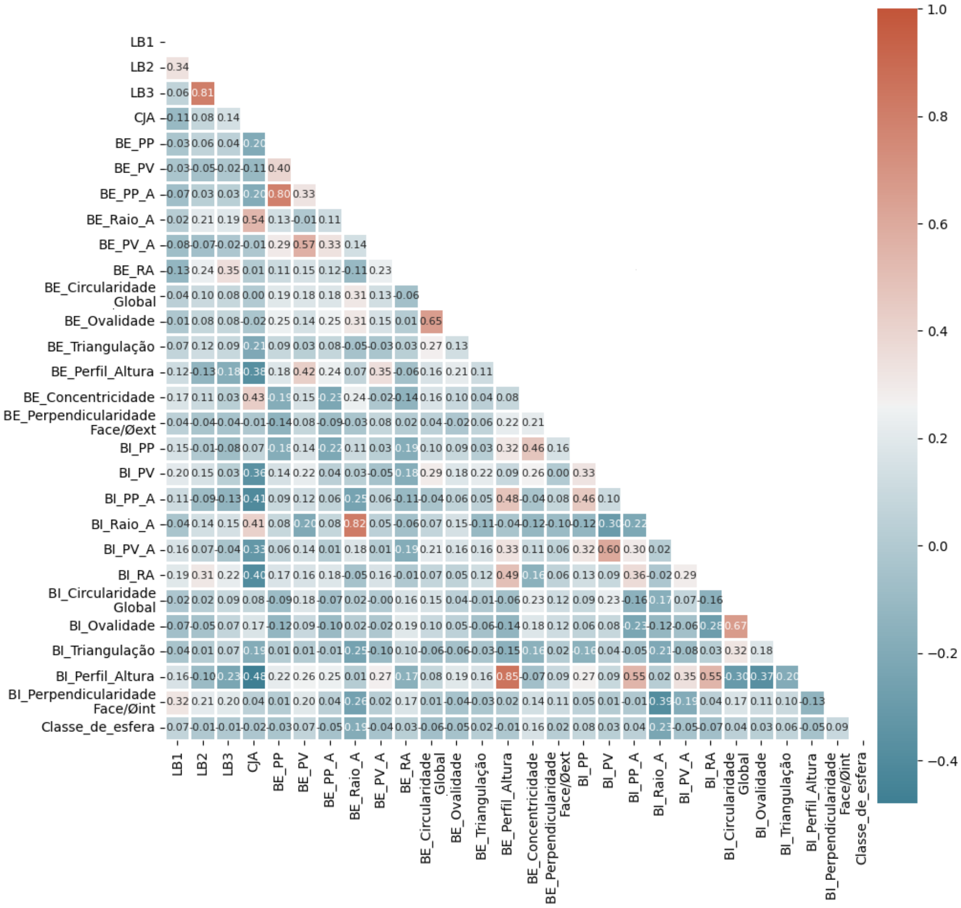


Figure 3. (a) Pearson and (b) Spearman correlation matrix.

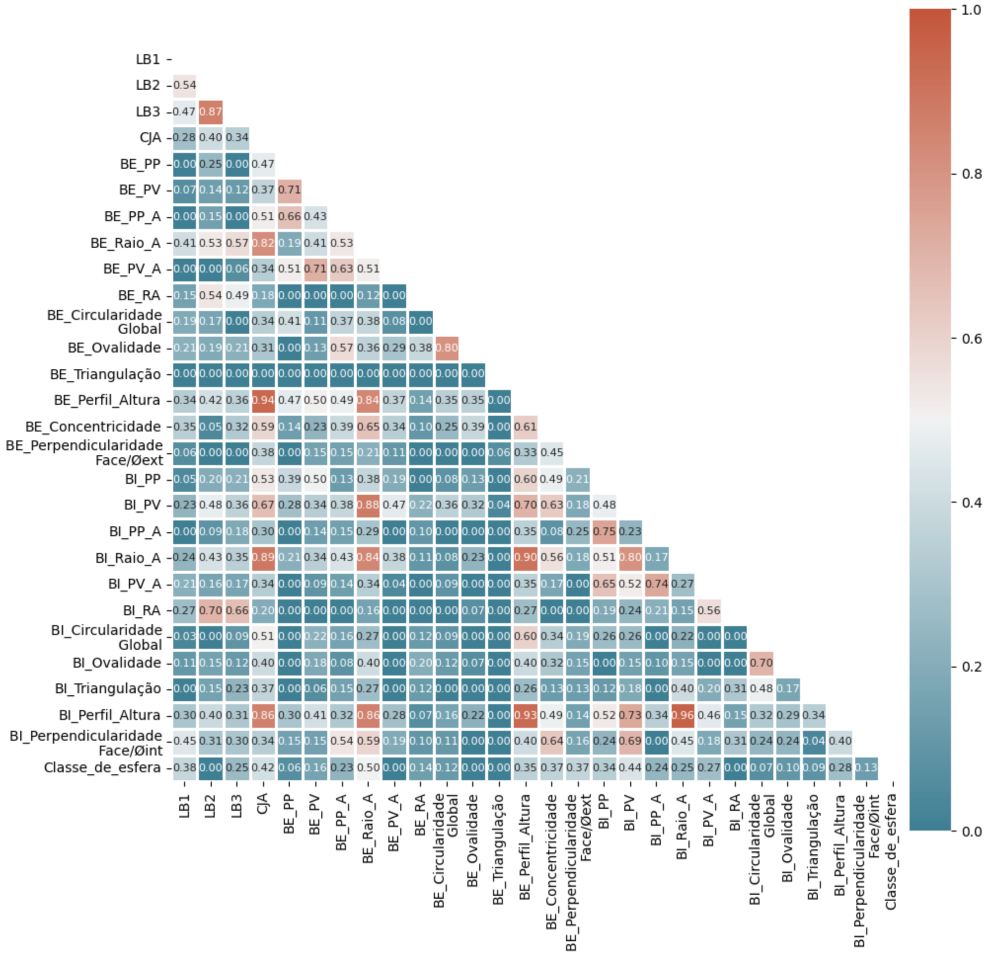


Figure 4. Phik correlation matrix.

4.2 Feature Selection

Choosing the most pertinent and useful variables to include in a predictive model is known as feature selection. In order to pick a subset of features from the complete set of accessible features, was employed Sequential Feature Selection (SFS), a feature selection technique, for this study. It was also used to rank the most significant features in a dataset using the recursive feature elimination (RFE) feature selection technique. The LGBM Regressor (Light Gradient Boosting Machine), Ridge, and LASSO models were employed for the training and subsequent provision of the feature importance metrics, see Table 2.

Table 2. Recursive feature elimination

	Characteristic of variable LB2			**Characteristic of variable LB3**		
	LGBM	**Ridge**	**LASSO**	**LGBM**	**Ridge**	**LASSO**
BI-Ra	1	1	1	1	1	1
BI- Perpendicularity Face/Øint	2	1	2	5	1	1
BE- Concentricity	1	9	3	1	3	2
BI-PV	1	1	4	2	1	3
CJA	1	10	5	1	1	4
BI- Profile-Height	1	1	6	1	1	5
BI- Triangulation	1	8	7	1	1	6
BI- Ovality	1	1	8	1	1	7

Because distinct traits were discovered for each LB, as shown in Table 2, conflicting findings were discovered when using recursive feature elimination (RFE). However, this scenario is different when analysis is done using Sequential Feature Selection (SFS). Similar to the Phik Correlation's findings, this approach identified three shared characteristics:

the radius of the outer ring raceway and the Ra of the inner and outer rings. With the aid of these findings, it is achievable to comprehend a solid and substantial connection between the raceway ring roughness, outer ring raceway radius, and the vibration level of LB2 and LB3.

LGBM Regressor, Ridge and LASSO are important tools in feature engineering. LGBM Regressor is commonly used to identify and create relevant features, while Ridge and LASSO are useful for selecting and reducing the dimensionality of the feature set, increasing the generalisation ability and improving the predictive performance of machine learning models. The combined use of these approaches results in more efficient and accurate models, thus justifying the choice to use them in this study.

4.3 Random Forest Model

In this investigation, the collected dataset was randomly divided into two homogenous subsets: a training subset and a test subset, which referred to 70% and 30% of the total data, respectively. This approach allows to efficiently capture the complexity and non-linear interactions of the data. Several tests were made changing the number of trees and the maximum depth of trees in the RF and the results obtained are described in Table 3.

Analyzing the values found, one can observe that in the case study referring to the output variable LB3, the model presented the lowest MAE, MSE and RMSE, as highlighted in bold in Table 4. Furthermore, the highest R2 value for the LB3 case study also indicates that the model has a good capacity to explain the variability of the data in this specific case. Thus, it was possible to determine the optimal configuration of the RF Regression model for the data and the problem in question, aiming to obtain the most accurate and reliable results possible. Other studies have also used RF Regression to make predictions, such as Takoutsing and Heuvelink (2022), Palomino et al. (2022), Onyelowe et al. (2022) and Coelho et al. (2024), but they have focused on case studies in different areas.

Table 3. The results obtained to each case study using RF model.

Case study	**MAE**	**MSE**	**RMSE**	**R^2**
LB1	5,78576	48,2327	6,94497	0,00884
LB2	3,88075	26,9996	5,19612	0,62022
LB3	**2,68448**	**12,2509**	**3,50013**	**0,70386**

5. CONCLUSION

The analysis was conducted through correlation analysis using experimental data obtained from the manufacturing process of automotive bearings, including both defect-free bearings and bearings with defects. The results obtained through correlation analysis provided valuable insights into the relationships between the variables. Pearson's correlation analysis revealed a linear relationship between the variables, indicating how they change together. Spearman's correlation analysis, on the other hand, demonstrated an increasing trend between the variables, suggesting a monotonic relationship. Notably, both correlation analyses yielded positive correlation coefficients above 0.5, indicating significant associations between the variables. Additionally, Phik's correlation analysis, which accounts for asymmetry and missing values in the data, revealed strong positive correlations with values close to 1. This finding further supported the existence of robust relationships among the constructive variables. Importantly, discrepancies were observed in the correlation values obtained when comparing bearings with defects to those without defects. This indicates that the presence of defects can significantly impact the relationships between the constructive variables, potentially leading to compromised bearing performance. Furthermore, for data analysis and predictive modelling, RF was used, and through a systematic evaluation of the different combinations of hyperparameters, the optimal model configuration for each output variable was determined. This determination was based on different benchmarking metrics. For future work, it is suggested to investigate other machine learning algorithms or comparing the RF model with other machine learning algorithms; incorporating more variables and resources and evaluation of the model's stability and robustness. Furthermore, it would be interesting to perform sensitivity analyses to evaluate the impact of different hyperparameter configurations of the model and identify possible sources of variability in the results.

6. ACKNOWLEDGEMENTS

This paper was carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Financial Code 001. The authors thank NTN Rolamentos do Brasil for making available the database used in this study. The authors Mariani and Coelho thank the National Council for Scientific and Technological Development - CNPq (Grants number: 307958/2019-1-PQ, 307966/2019-4-PQ and 408164/2021-2-Universal) and the Araucária Foundation PRONEX 042/2018 for the financial support to this work.

7. REFERENCES

Baak, M., Koopman, R., Snoek, H., Klous, S., 2020. “A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics”. *Computational Statistics & Data Analysis*, Vol. 152. doi: 10.1016/j.csda.2020.107043.

Chen, J., Huang, R., Chen, Z., Mao, W., Li, W., 2023. “Transfer learning algorithms for bearing remaining useful life prediction: A comprehensive review from an industrial application perspective”. *Mechanical Systems and Signal Processing,* Vol. 193. pp 110239. doi: 10.1016/j.ymssp.2023.110239.

Coelho, L.S., Ayala, H. V. H., Mariani, V.C., 2024. “CO and NOx emissions prediction in gas turbine using a novel modeling pipeline based on the combination of deep forest regressor and feature engineering”. *Fuel,*Vol. 355, No 129366. doi: 10.1016/j.fuel.2023.129366.

Das, O., Das, D. B., Birant, D., 2023. Machine learning for fault analysis in rotating machinery: A comprehensive review. *Heliyon*, Vol. 9. doi: 10.1016/j.heliyon.2023.e17584.

Elforjani, M., Mba, D., 2010. “Accelerated natural fault diagnosis in slow speed bearings with acoustic emission”. *Engineering Fracture Mechanics*. Vol. 77 pp112-127. doi: 10.1016/j.engfracmech.2009.09.016.

Garg, M., Goel, A., 2023. “Preserving integrity in online assessment using feature engineering and machine learning” *Expert Systems With Applications*. Vol. 225. doi: 10.1016/j.eswa.2023.120111.

Ghafari, S.H., Golnaraghi, F., Ismail, F., 2006. “Fault diagnosis based on chaotic vibration of rotor systems supported by ball bearings”. *Proceeding of COMADEM,* pp 819-826. doi: 10.1016/j.eswa.2010.07.119.

Huang, L., Qiu, T., Chen, G., Zhong, L., 2019. “European Union effect on financial correlation dynamics”. *Physica A: Statistical Mechanics and its Applications*. Vol. 528. doi: 10.1016/j.physa.2019.121457.

Huang, G.-B., Zhou, H., Ding, X., and Zhang, R., 2011. “Extreme learning machine for regression and multiclass classification”. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* Vol. 42. pp 513–529. doi: 10.1109/TSMCB.2011.2168604.

Jan, H., and Tomasz, K., 2011. “Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data”. *Quaestiones Geographicae*. Vol. 30. Doi: 10.2478/v10117-011-0021-1.

Lau, E.C., and Ngan, H.W., 2010. “Detection of motor bearing outer raceway defect by wavelet packet transformed motor current signature analysis”. *IEEE Transactions on Instrumentation and measurement,* Vol. 59. pp2683-2690.

Liu, C., Tan, J., and Huang, Z., 2022. “Fault Diagnosis of Rolling Element Bearings Based on Adaptive Mode Extraction”. *Machines*, Vol. 10. pp260. doi: 10.3390/machines10040260.

Liu, H., Mo, Z., Zhang, H., Zeng, X., Wang, J., and Miao, Q., 2018. “Investigation on rolling bearing remaining useful life prediction: A review”. *Prognostics and System Health Management Conference (PHMChongqing),* pp 979–984. doi: 10.1109/PHM-Chongqing.2018.00175.

Onyelowe, K.C., Gnananandarao, T., Ebid, A.M., 2022. “Estimation of the erodibility of treated unsaturated lateritic soil using support vector machine-polynomial and -radial basis function and random forest regression techniques”. *Cleaner Materials*, Vol. 3, No 100039. doi: 10.1016/j.clema.2021.100039.

Palomino, A.F., Espino, P.S., Reyes, C.B., Rojas, J.A.J., Silva, F.R., 2022. “Estimation of moisture in live fuels in the mediterranean: Linear regressions and random forests”. *Journal of Environmental Management,* Vol. 322, No 116069. doi: 10.1016/j.jenvman.2022.116069.

Peng, Z., and Kessissoglou, N., 2003 “An integrated approach to fault diagnosis of machinery using wear debris and vibration analysis”. *Wear.* Vol. 255. pp1221-1232. doi: 10.1016/S0043-1648(03)00098-X.

Samanta, B., Al-Balushi, K.R., Al-Araimi, S.A., 2003. “Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection”. *Engineering Applications of Artificial Intelligence*. Vol. 16. doi: 10.1016/j.engappai.2003.09.006.

Safizadeh, M.S., and Latifi, S.K., 2014. “Using multi-sensor data fusion for vibration fault diagnosis of rolling element bearings by accelerometer and load cell”. *Information Fusion*, Vol. 18, pp 1-8. doi: 10.1016/j.inffus.2013.10.002.

Seo, J.J., Yoon, H., Ha, H., Hong, D.P., Kim, W., 2011. “Infrared thermographic diagnosis mechanism for fault detection of ball bearing under dynamic loading conditions”. *In Advanced materials research. Trans Tech Publications*. Vol. 295. pp1544-1547.

Srivani, A., Arunkumarb, T., Ashok, S. D., 2018. “Fourier Harmonic Regression Method for Bearing Condition Monitoring using Vibration Measurements”. *Materials Today: Proceedings,* Vol 5, pp 12151–12160. doi: 10.1016/j.matpr.2018.02.193.

Takoutsing, B., and Heuvelink, G.B.M., 2022. “Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors”. *Geoderma,* Vol. 428, No 116192. doi:10.1016/j.geoderma.2022.116192.

Wang, J., Wang, H., Nie, F., Li, X., 2023. “Feature selection with multi-class logistic regression”. *Neurocomputing*. Vol. 543. No 126268 doi: 10.1016/j.neucom.2023.126268.

8. RESPONSIBILITY NOTICE

The authors are solely responsible for the printed material included in this paper.