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-2023-0822
AUTOMOTIVE BEARINGS ANALYSIS BASED ON REGRESSION MODELS

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. *Using statistical models to predict new observations or events is one of the main goals of supervised statistical learning methods. In machine learning, one of the most important segments of Artificial Intelligence, a system or machine automatically learns to predict without being explicitly programmed to do so. Instead, it uses algorithms that allow the system to analyze data and recognize patterns based on certain prior examples, the data. The accuracy of the model depends on the data set used for training, the selection of the appropriate algorithms, and the choice of appropriate parameters. Machine learning is a constantly evolving field, with much research underway to develop new algorithms and techniques for regression, prediction, and classification. Currently, models have many applications in regression tasks in different fields of study, and the objective of this study is to apply such models to automotive bearing manufacturing processes. Regression models are widely used when it is desired to evaluate the impact of production factors on the quality of a product. The main objective of this study is to use statistical regression techniques, including linear and polynomial regression, and the neural network multi-layer perceptron (MLP), to understand and quantify the relationships between geometric variables of automotive bearings, as well as to compare the performance of the regression technique MLP through statistical metrics.*

Keywords: *Automotive bearing, Linear regression, Polynomial regression, Multi-Layer Perceptron*

1. INTRODUCTION

In supervised statistical learning approaches, the use of statistical models to predict data or events is crucial. It is possible to analyse a complex data set and extract valuable information to support the use of statistical models in decision-making. A structured and organised strategy for predicting observations or events is provided by the statistical models used in supervised statistical learning techniques. These models use methods such as linear regression, logistic regression, decision trees, neural networks, among others, and are based on sound statistical concepts (Rattan, Penrice and Simonetto, 2022).

As a central component of artificial intelligence, machine learning (ML) offers a powerful method for solving complicated questions and improving automated decision-making. By utilising machine learning techniques, systems can become adaptable and autonomous and learn and develop as they are exposed to additional data (Das, Das and Birant, 2023). Applications in fields such as healthcare, finance, data science and industrial automation are possible thanks to these ML capabilities (Dinardo, Fabbiano and Vacca, 2018).

A crucial consideration in the creation of ML-based solutions is the dependence of model correctness on the training data set, choice of relevant methods, and choice of appropriate parameters. Continuing research to enhance and create new regression, prediction, and classification algorithms and approaches has led to the continuing evolution of the ML field. The need for more precise and effective solutions for data analysis and decision-making is driving this growth. Deep neural networks, genetic algorithms, and reinforcement learning are just a few of the cutting-edge techniques that researchers are continually experimenting with to solve complicated problems and improve the prediction and classification abilities of models (Hakim et al., 2023).

A viable method for assessing the influence of production parameters on product quality is to use regression models in the manufacture of automotive bearings. Regression models can be created using statistical and mathematical analyses that relate manufacturing aspects such as temperature, pressure, speed and use of materials with bearing quality attributes

such as resistance, durability and mechanical performance. These models make it possible not only to identify the production variables that have the greatest impact on bearing quality, but also to quantify their impact and predict product performance based on various production scenarios (Lei et al., 2020) and (Liu, Tan and Huang, 2022).

Vibration analysis, thermographic analysis and oil analysis are the three main procedures for analysing bearings. Each of these methods provides information on the condition of the bearing, allowing a decision to be made regarding possible maintenance or replacement (Hakim et al., 2023). To determine the expected service life of an aircraft engine, different machine learning techniques have been used, including linear regression, decision trees, support vector machines, random forests, nearest neighbours, K-Means algorithms, gradient boosting methods, adaptive boosting, deep learning and analysis of variance in a comparative investigation carried out by Mathew et al., 2017 and Susto et al., 2013, who produced a series of publications using various regression techniques. Subsequently, Paolanti et al., 2018 reported the success of the Random Forest approach when used in cutting machine investigations.

As a result, the goal of this study is to assess how well regression models and a particular neural network can be used to forecast the quality of original bearings based on the characteristics of these elements. This entails evaluating the effectiveness of many models using the same assessment metrics. It will be feasible to make useful suggestions to enhance the bearing manufacturing process based on the comparative study of the regression models. The most pertinent traits for bearing quality may be identified, significant trends or linkages may be noted, and enhancements or corrective measures may be suggested to be used in subsequent work.

2. DATASET DESCRIPTION

The Curitiba metropolitan area-based high-precision component manufacturer contributed samples for this study's use of and analysis of data on bearing qualities. With 25 input and 3 output variables, Table 1 describes the parameters for which data were retrieved: axial clearance control, vibration, roughness, and ovalization.

Table 1. Dataset characteristics.

Type	Variable name	Variable description
Output	LB1	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 dB at raceway position 1800 Hz and 10000 Hz
Input	CJA	Bearing axial clearance
	BE - PP	Outer ring raceway deformation between two consecutive peaks
	BE - PV	Outer ring raceway deformation between consecutive peaks and valleys
	BE – PP A	Outer ring raceway deformation between two consecutive peaks on zone A
	BE – Radius A	Outer ring raceway radius on zone A
	BE – PV A	Outer ring raceway deformation between consecutive peaks and valleys on zone A
	BE - Ra	Outer ring raceway roughness
	BE - Circularity	Outer ring raceway roundness - overall deviation
	BE - Ovality	Outer ring raceway roundness - Ovality - 2 points deformation
	BE - Triangulation	Outer ring raceway roundness - Triangulation - 3 points deformation
	BE – Profile-Height	Outer ring raceway shoulder
	BE - Concentricity	Outer ring concentricity between outer diameter and raceway
	BE – Perpendicularity Face/Øext	Outer ring perpendicularity between face and outer diameter
	BI - PP	Inner ring raceway deformation between two consecutive peaks
	BI - PV	Inner ring raceway deformation between consecutive peaks and valleys
	BI – PP A	Inner ring raceway deformation between two consecutive peaks on zone A
	BI – Radius A	Inner ring raceway radius on zone A
	BI – PV A	Inner ring raceway deformation between consecutive peaks and valleys on zone A
	BI - Ra	Inner ring raceway roughness
	BI - Circularity	Inner ring raceway roundness - overall deviation
	BI - Ovality	Inner ring raceway roundness - Ovality - 2 points deformation
	BI - Triangulation	Inner ring raceway roundness - Triangulation - 3 points deformation
	BI – Profile-Height	Inner ring raceway shoulder
	BI – Perpendicularity Face/Øint	Inner ring perpendicularity between face and bore diameter
	Ball class	Steel ball class - Deviation between nominal diameter

3. METHODS

Regression analysis is a quantitative research strategy that is generally used to examine the relationships between dependent and independent variables. Data visualization makes it easier to spot trends and outliers, which yields more information (Rattan, Penrice and Simonetto, 2022). In order to determine the links between the dependent and independent variables regarding the constructional features of vehicle bearings, this study used two regression models and a neural network. Drawing a line or curve of best fit between the data is the true objective of the regression procedure. Training data made up 70% of the total data set, whereas testing data made up 30% of the total data set. The test set was employed for validation once the regression equations had been established using the training set.

3.1 Multiple Linear Regression

Multiple linear regression (MLR), also known as ordinary least squares, can be defined as an extension of linear regression. In this case, MLR adopts more than one explanatory (independent) variable to predict the outcome of a response (dependent) variable. The MLR model accurately approximates the true unknown functions using complex forms of independent variables and consequently, underlying functional relationship between the dependent and independent variables is made unknown (Ajona *et al.*, 2022). In MLR, the equation for predicting the dependent variable (y) based on multiple independent variables (x_1, x_2, ..., x_n) given by Eq. (1):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \tag{1}$$

where β_0 represents the intercept or constant term, β_1, β_2, ..., β_n are the regression coefficients associated with each independent variable, and ε is the error term. The regression coefficients represent the estimated change in the dependent variable for a one-unit change in the corresponding independent variable, assuming all other independent variables are held constant. To estimate the regression coefficients, the least squares method is commonly used. It minimizes the sum of the squared differences between the observed values of the dependent variable and the predicted values based on the regression equation. According to Eq. (2), the formula for estimating the regression coefficients is:

$$\beta_j = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ denote the mean values of the independent variables x_j and the dependent variable y, respectively. To assess the overall goodness of fit of the multiple linear regression model, the coefficient of determination, R^2, is often used. It represents the proportion of the variance in the dependent variable that can be explained by the independent variables.

3.2 Polynomial Regression

Is a regression approach that uses a polynomial equation to model the relationship between the independent (predictor) and dependent (response) variables. In other words, it makes predictions and approximations of the data using a polynomial function. The degree of flexibility of the model is determined by the polynomial's order. A polynomial regression of order 1 is comparable to basic linear regression, while those of orders 2 and 3 correspond to parabolas and cubic curves, respectively. The complexity of the model and the data fit improve with increasing polynomial order, but there is also a bigger risk of overfitting (Ajona *et al.*, 2022). Polynomial regression is a variation of linear regression where the relationship between the dependent variable and the independent variable(s) is modeled using a polynomial function. In polynomial regression, according to Eq. (3), the equation for predicting the dependent variable, y, based on an independent variable, x, can be written as:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \varepsilon \tag{3}$$

where β_0 represents the intercept or constant term, β_1, β_2, ..., β_n are the regression coefficients associated with each term in the polynomial equation, x represents the independent variables, x^2 represents the squared term of the independent variable, and x^n represents the n^{th} power term of the independent variable, ε represents the error term. To estimate the regression coefficients in polynomial regression, the least squares method is commonly used. The goal is to minimize the sum of the squared differences between the observed values of the dependent variable and the predicted values based on the polynomial equation. The equations for estimating the regression coefficients are similar to those in multiple linear regression but involve the polynomial terms, given by Eq. (4):

$$\beta_j = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^j} \tag{4}$$

where the power of the independent variable, *j*, determines the term in the polynomial equation, x^j. Polynomial regression allows for modeling non-linear relationships between variables by introducing polynomial terms. The degree of the polynomial determines the complexity of the model. For example, a polynomial regression of degree 2 will include squared terms, x^2, while a polynomial regression of degree 3 will include cubed terms, x^3, and so on. It is important to note that when using polynomial regression, it is essential to choose an appropriate degree of the polynomial that balances model complexity and overfitting. Overfitting occurs when the model fits the training data too closely but fails to generalize well to new, unseen data. Polynomial regression can be a powerful tool for capturing non-linear relationships in the data. By using higher-order polynomial terms, it can accommodate more complex curves and patterns. However, as the degree of the polynomial increases, the model can become more prone to overfitting and may require careful regularization techniques to avoid excessive complexity.

3.3 Multi-Layer Perceptron Artificial Neural Network

The neurological system of the human brain and an artificial neural network (ANN) work in a similar way, since the ANN can accurately identify the pattern of the human brain or conventional computational tools (Liu et al., 2021). For this study, it was decided to compare multiple linear regression, polynomial regression and an artificial neural network called multilayer perceptron (MLP) for the study of car bearing characteristics, due to the need to identify the model that best fits the data and provides the most accurate predictions.

A multi-layer perceptron (MLP) is a type of artificial neural network that consists of multiple layers of interconnected nodes, known as neurons. It is a feedforward neural network, meaning that information flows from the input layer through the hidden layers to the output layer without any feedback loops. MLPs are widely used for various tasks, including classification, regression, and pattern recognition.

An MLP typically consists of three types of layers: an input layer, one or more hidden layers, and an output layer. The input layer receives the initial input data, and each neuron in the input layer represents a feature or attribute of the data. The hidden layers are intermediate layers between the input and output layers, where the neurons perform computations and transform the input data. The output layer produces the final predictions or outputs of the network (Lek and Park, 2008).

The neurons in the hidden layers and the output layer are connected by weighted connections. Each connection represents the strength or importance of the connection between two neurons. During the training process, the network adjusts these weights to minimize the difference between the predicted outputs and the true outputs, using a technique called backpropagation (López, López and Crossa, 2022).

Backpropagation is a key algorithm used to train MLPs. It involves two main steps: forward propagation and backward propagation. In forward propagation, the input data is fed through the network, and the activations of the neurons are computed layer by layer until the final output is obtained. In backward propagation, the error between the predicted output and the true output is calculated, and this error is then propagated backward through the network to update the weights. This process iterates until the network's performance converges to a satisfactory level.

MLPs can have varying architectures with different numbers of hidden layers and neurons. Deeper architectures with more hidden layers allow the network to learn more complex representations and capture intricate patterns in the data. However, deeper networks can also be more challenging to train and may require more data to avoid overfitting. MLPs have been widely successful in various domains, such as image recognition, natural language processing, and financial forecasting. However, they do have limitations. MLPs may suffer from overfitting if the training data is insufficient or noisy. They also require careful hyperparameter tuning, such as selecting the appropriate number of hidden layers, neurons, and activation functions, to achieve optimal performance (Rajamanickam and Baskaran, 2022). In Figure 1 was illustrated one design about MLP.

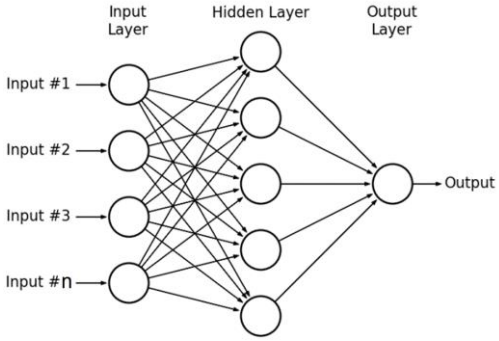

Figure 1 – MLP architecture.

3.4 Performance Metrics

Google Collaboratory, a free cloud-based software, was used to run Python code directly from the browser without first setting up a local environment. It is based on the Jupyter Notebook platform and contains features including real-time collaboration, code editing, and result display. The performances of the various methods were compared using mean absolute error (MAE) given by Eq. (5), mean square error (MSE) Eq. (6), root mean square error (RMSE) Eq. (7), and coefficient of determination (R^2) Eq. (8):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| \tag{5}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - x_i)^2 \tag{6}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - x_i)^2} \tag{7}$$

$$R^2 = 1 - \frac{\sum (y_i - x_i)^2}{\sum (y_i - x_i)^2} \tag{8}$$

where y_i and x_i are the desired output and estimated output, respectively, and n represents each sample in the data set. That statistical metrics are evaluation metric used in machine learning models to assess how well the model fits the training and test data.

4. RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis

This section summarizes the findings from the bearing characteristics data set. The first step was to conduct correlation analyses using Pearson, Spearman, and Phik. One of the main uses of correlation analysis is to discover linear relationships between variables. The Pearson correlation coefficient lies between -1 and +1. A complete positive correlation has a value of 1, which indicates that the variables are positively associated linearly that is, when one variable grows, the others climb proportionately. The opposite is true with a value of -1, which indicates a completely negative correlation. In this case, the variables are linearly unrelated and change in one variable causes a corresponding decrease in the other. A score of 0 indicates that there is no discernible linear association between the variables. The output variables for this Pearson's correlation analysis are the vibration level variables LB1, LB2, and LB3, which can be seen in Figure 2. LB1 will not be taken into consideration for the analyses in this work because it initially represents the measure system vibration, outer ring conditions, and inner ring conditions, respectively. The most important columns in this analysis are the second and third columns.

A simple analysis can be used to understand the relationship between the inner and outer rings' RA and vibration level. A number called RA is produced by measuring the bearing rings' raceways. A significance analysis can be used to show how important these characteristics are; the P-Value for each correlation value is less than 0.05. Spearman's analysis, which is common in non-parametric data applications such as ordinal or interval data as well as continuous data, allows one to examine the monotonic relationship between the variables, i.e., whether they tend to increase or decrease together, regardless of the precise form of the relationship. At Figure 3 one can assume that the data show a trend toward linearity because the Spearman correlation reveals a complete lack of association.

Regarding the Phik correlation results, which are displayed in Figure 4 for the second and third rows of this case from bottom to top, it is possible to identify the significant correlation between the AR of the inner and outer rings with the vibration level of the bearing. A new characteristic that is significant, the track radius of the outer ring, has been discovered.

As seen in Figure 2, Pearson's correlation, which in this case shows the correlation of a variable with itself, ignores correlations lower than 0.3 or extremely high correlations. Therefore, for the output variable LB1, only the relationship with the variable BI- Perpendicularity Face/int was taken into account. 26 correlations were initially eliminated before the two correlations with the variables BI-Ra and BE-Ra for the output variables LB2 and LB3 were taken into account. A value of 0.3 was utilized as an exclusion criterion for correlations that were too low for the Spearman's correlation analysis, as shown in Figure 3, 27 associations were initially disregarded before acknowledging the correlations with the variables BI-Perpendicularity Face/int, BI-Ra, and BE-Ra, respectively, for the output variables LB1, LB2, and LB3. When Phik examined the correlations, the same criterion was taken into account, accounting for a value of 0.3. Four correlations BI- Profile-Height, BE-Concentricity, BE- Profile-Height, and BE-Radius A were absorbed while 24 correlations were deleted for the outcome variable LB1. For LB2 and LB3, 19 out of the 28 outputs were removed.

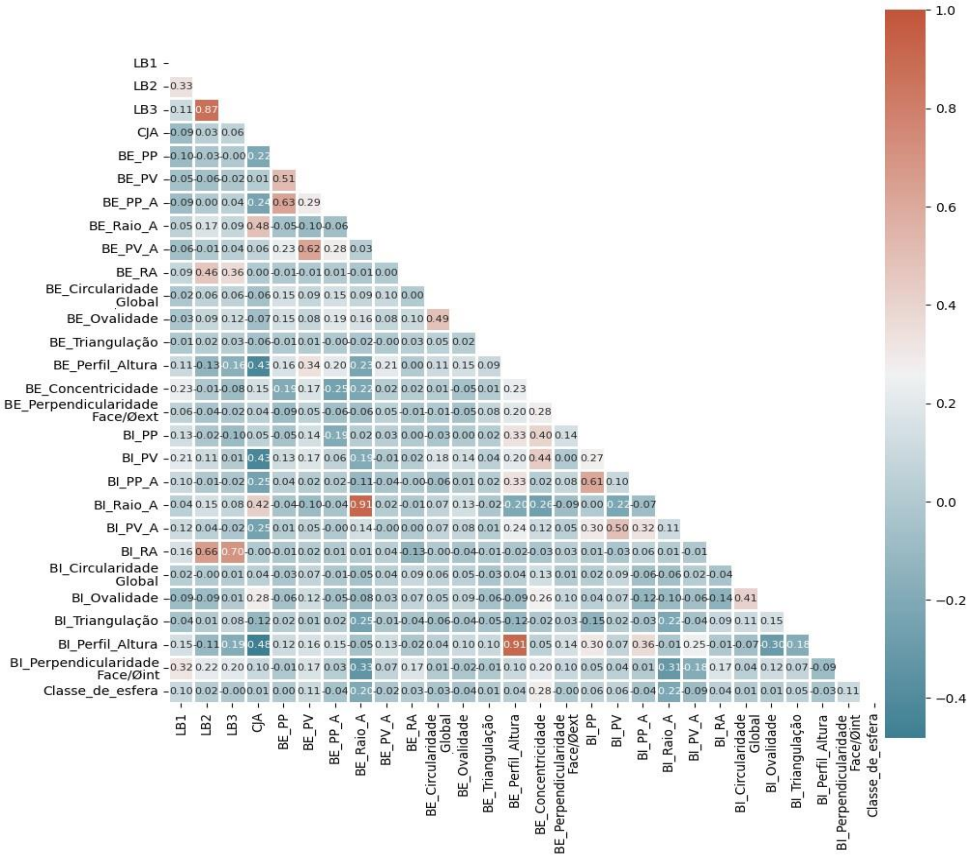

Figure 2. Pearson's correlation matrix

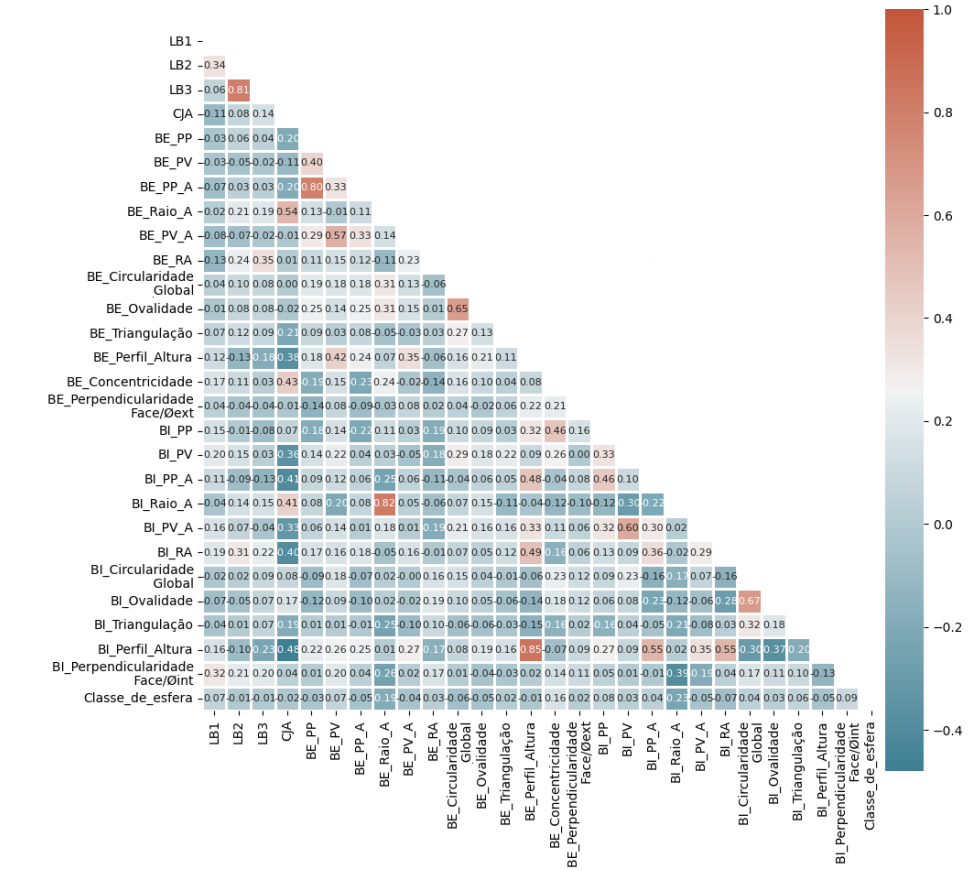

Figure 3. Spearman's correlation matrix.

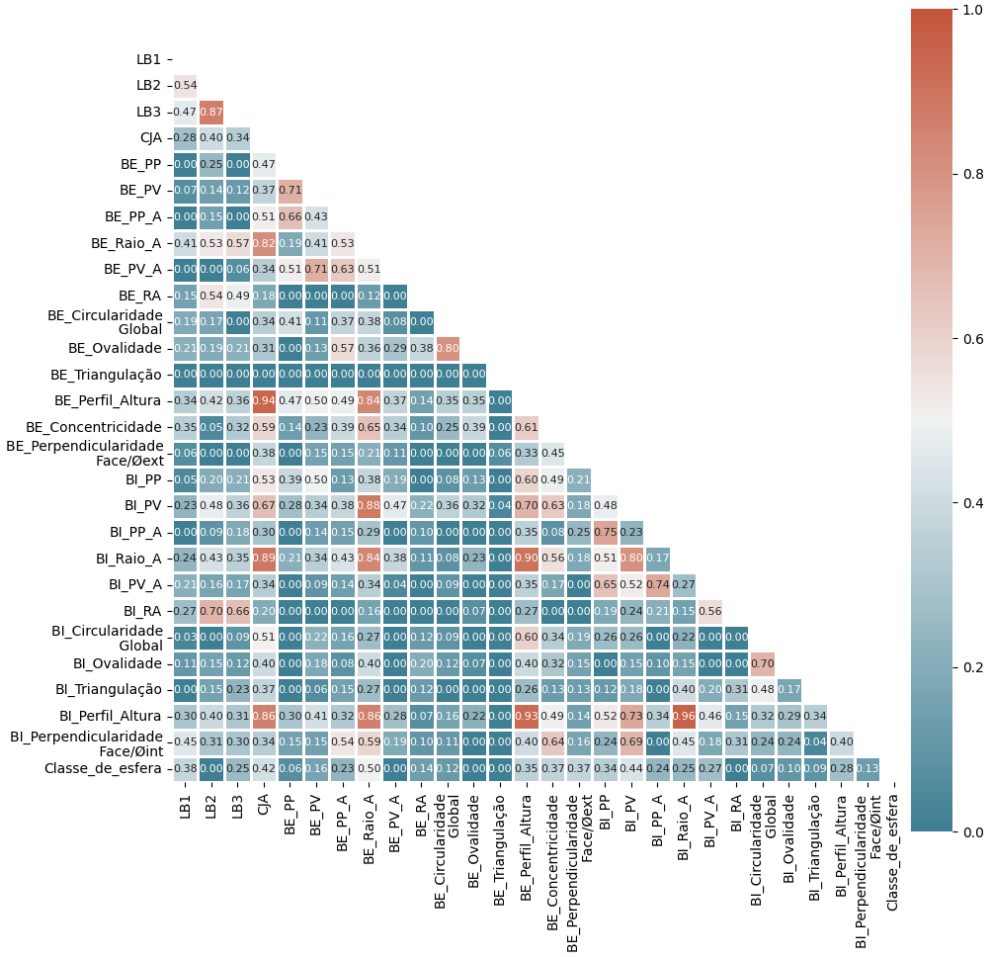

Figure 4. Phik correlation matrix.

4.2 Feature Selection

Feature selection is the process of selecting the most relevant and practical variables to include in a prediction model. For this work, we used Sequential Feature Selection (SFS), a feature selection technique, to select a subset of features from the entire set of accessible features. Recursive feature elimination (RFE) feature selection was also used to rank the most important features in a dataset. The feature importance metrics were trained on and then provided by the LGBM Regressor (Light Gradient Boosting Machine), Ridge, and Lasso models.

Recursive feature elimination (RFE) contradictory findings were found since distinct features were found for each LB. However, when analysis is conducted using Sequential Feature Selection (SFS), this situation is different. This method discovered three traits that were similar to those found by the Phik Correlation: the radius of the outer ring raceway and the Ra of the inner and outer rings. These results make it possible to understand a strong and significant relationship between the outer ring raceway radius, the vibration level of LB2 and LB3, and the raceway ring roughness.

4.3 Model Comparison

Each model has its own assumptions and representation capabilities, which may affect its suitability for the study at hand. Through the works already published, it is known that multiple linear regression is widely used when there is a linear relationship between the independent variables and the dependent variable. In the case of automotive bearings, it is possible that there are multiple independent variables (such as temperature, load, speed, dimensional and surface characteristics, among others) that influence the useful life or performance of the bearing. Polynomial regression, in turn, allows modelling non-linear relationships between variables, which may be relevant if there is evidence that bearing characteristics exhibit non-linear behavior. Neural networks are able to learn complex patterns in the data, capturing non-linear relationships and interactions between variables, being especially useful when there is high dimensionality and the exact form of the relationship between bearing characteristics and their performance is not known. This comparative approach is fundamental to inform decisions and direct future research, contributing to the advancement in the study and understanding of automotive bearing characteristics and their impact on vehicle performance.

By applying different regression methods and neural network, it was possible to find an interesting result in this application. In Table 2 are described the results obtained to LB1, LB2 and LB3, respectively. It can be seen that in case study LB1, Table 2, both Polynomial and Multiple Linear Regression showed similar results, with MAE, MSE and RMSE values close to each other. However, both models have limited performance, with low R^2 values, indicating that they failed to adequately explain the variation in the data, this can be attributed to the complexity of the relationships between variables in automotive bearings, since Liu et al. (2022) showed that multiple linear regression is widely used in the diagnosis of bearing faults, suggesting that this technique may be effective for some applications. For the case study LB2, Polynomial and Multiple Linear Regression showed similar and superior performance compared to MLP. Both techniques presented lower MAE, MSE and RMSE values, indicating a better ability to predict the data. In addition, the models obtained high R^2 values, suggesting that they were able to explain a significant percentage of the variation in the data. Liu et al. (2021) and Mathew et al. (2017), who applied non-linear models to predict bearing failures and the remaining useful life of motors and highlighted the relevance of polynomial regression when there is evidence of non-linear behaviour in bearing characteristics.

Table 2. Results based on each case study LB1, LB2 and LB3 using regression models.

	LB1				LB2				LB3			
Regression	**MAE**	**MSE**	**RMSE**	**R^2**	**MAE**	**MSE**	**RMSE**	**R^2**	**MAE**	**MSE**	**RMSE**	**R^2**
Polynomial	5.756	46.797	6.94	0.038	3.855	24.194	4.918	0.659	3.105	14.928	3.863	639
Multi Linear	5.845	47.785	6.912	0.018	3.751	22.080	4.699	0.689	2.771	12.944	3.597	687
MLP	6.072	53.356	7.304	-0.096	3.649	22.520	4.745	0.683	3.145	14.563	3.816	647

In the LB3 case study, it was possible to verify that both Multiple Linear Regression and MLP showed superior results to Polynomial Regression. Both techniques presented lower MAE, MSE and RMSE values, indicating a better fit to the data. In addition, the models achieved higher R^2 values, indicating a better ability to explain the variation in the data. Therefore, both Polynomial Regression and Multiple Linear Regression were more effective in case study LB2, Multiple Linear Regression and MLP showed better performance in case study LB3.

Through Figures 5, 6 and 7, generated using Excel software it is possible to analyse the different results obtained in the training and testing stages, during the use of the neural network for each of the case studies. In case study LB1, Figure 5, it is observed that the model underperformed during training, with an MAE of 6.17 and a negative R^2 of -0.09. This indicates that the model did not fit the data well during training. However, during testing, the model showed a significant improvement, with an MAE of 5.7 and a positive R^2 of 0.06. This suggests that the model was able to better generalise the patterns and make more accurate predictions on new data.

Figure 5. Comparison between training and testing for variable LB1.

Figure 6. Comparison between training and testing for variable LB2.

Figure 7. Comparison between training and testing for variable LB3.

For the LB2 case study, Figure 6, it was observed that the model performance during training was relatively good, with an MAE of 4.08 and an R2 of 0.6. This indicates that the model fitted the training data well. During testing, the model maintained a good performance, with an MAE of 3.7 and an R2 of 0.7, indicating that the model was able to generalise the patterns well and make accurate predictions on new data. Finally, in case study LB3, Figure 7, the model's performance during training was moderate, with an MAE of 4.5 and an R2 of 0.3. It was therefore realised that the model had difficulty in fully fitting the training data. However, during testing, the model performed better, with an MAE of 3.16 and an R2 of 0.64, indicating that the model was able to generalize the patterns well and make more accurate predictions on new data.

In general, the application of MLP in this study showed superior performance in some cases, especially when it came to high dimensionality and complex relationships between bearing characteristics and performance. Lei et al. (2020) and Susto et al. (2013) also explored the use of neural networks in fault diagnosis and predictive maintenance, highlighting the ability of neural networks to capture complex, non-linear patterns in the data. Other studies have also used the comparison of various techniques to make forecasts, such as Takoutsing and Heuvelink (2022), Palomino et al. (2022) and Coelho et al. (2024), but they have focused on case studies in different areas.

5. CONCLUSION

In summary, the analysis of experimental data from automotive bearing manufacturing, including both faulty and faultless bearings, revealed strong correlations among the variables. Pearson's and Spearman's correlation analyses indicated linear and monotonic associations, respectively, between the variables, with significant positive correlations exceeding 0.5. Phik's correlation analysis, which considers data asymmetries and missing values, further confirmed substantial positive correlations. It was also possible to observed significant differences in correlation values when comparing bearings with and without flaws. To develop predictive models, was employed MLR, Polynomial Regression, and MLP, optimized hyperparameters for each output variable through systematic comparisons. These model configurations were evaluated using various benchmarking metrics. For future research, we recommend comparing our models with other machine learning algorithms, exploring additional features and variables, and assessing model stability and robustness. Sensitivity analysis can help identify the impact of different hyperparameter settings and potential sources of result variability.

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

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, D., Cheng, W., Zhang, J., Gao, R. X., Wen, W., 2021. Integrated method of generalized demodulation and artificial neural network for robust bearing fault recognition. *Procedia Manufacturing*, Vol. 53. pp 628-635.

Susto, G. A., Schirru, A., Pampuri, S., Pagano, D., McLoone S., and Beghi, A., 2013. "A predictive maintenance system for integral type faults based on support vector machines: An application to ion implantation,". *IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 195-200. doi: 10.1109/CoASE.2013.6653952.

Dinardo, G., Fabbiano, L., Vacca, G., 2018. A smart and intuitive machine condition monitoring in the Industry 4.0 scenario”, *Measurement* Vol. 126, pp 1–12. doi:10.1016/j.measurement.2018.05.041.

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.

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.

Ajona, M., Vasanthi, P., Vijayan, D.S., 2022. Application of multiple linear and polynomial regression in the sustainable biodegradation process of crude oil. *Sustainable Energy Technologies and Assessments*, Vol. 54.

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.

Hakim, M., Omran, A. A. B., Ahmed, A. N., Al-Waily, M., Abdellatif, A., 2023. A systematic review of rolling bearing fault diagnoses based on deep learning and transfer learning: Taxonomy, overview, application, open challenges, weaknesses and recommendations. *Ain Shams Engineering Journal*, Vol. 14. doi:10.1016/j.asej.2022.101945.

Paolanti, M., Romeo, L., Felicetti, A., Mancini, A., 2018. E. Frontoni and J. Loncarski, "Machine Learning approach for Predictive Maintenance in Industry 4.0,". *14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*, pp. 1-6. doi:10.1109/MESA.2018.8449150.

López, O. A. M., López, A. M., and Crossa, J., 2022. “Fundamentals of Artificial Neural Networks and Deep Learning, Multivariate Statistical Machine Learning Methods for Genomic Prediction”. Springer, pp 379–425. doi:1007/978-3-030-89010-0.

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.

Rattan, P., Penrice, D. D., Simonetto, D. A., 2022. “Artificial Intelligence and Machine Learning: What You Always Wanted to Know but Were Afraid to Ask”. Gastro Hep Advances. *Gastro Hep Advances*, pp 70-78.

Rajamanickam, R., Baskaran, D., 2022. “In Intelligent Data-Centric Systems: Current Trends and Advances in Computer-Aided Intelligent Environmental Data Engineering”. *Academic Press*, pp 393-415.

Lek, S., Park, Y.S., 2008. “Multilayer Perceptron”. *Encyclopedia of Ecology*, pp 2455-2462. doi: 10.1016/B978-008045405-4.00162-2.

Mathew, V., Toby, T., Singh, V., Rao, B. M. and Kumar, M. G., 2017. "Prediction of Remaining Useful Lifetime (RUL) of turbofan engine using machine learning," *IEEE International Conference on Circuits and Systems (ICCS)*, pp. 306-311. doi: 10.1109/ICCS1.2017.8326010.

Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., Nandi, A.K., 2020. “Applications of machine learning to machine fault diagnosis: a review and roadmap”, *Mech. Syst. Signal Process*. Vol. 138. doi:10.1016/j.ymssp.2019.106587.

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.

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.

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.

8. RESPONSIBILITY NOTICE

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