COMPUTERS AND INFORMATICS

2757-8259 2023, Volume 3 Issue 1

Research Article

Prediction of healthcare insurance costs

Shoroog Albalawi
University of Tabuk, Faculty of Computers and Information Technology, Saudi Arabia, 421010123@stu.ut.edu.sa
Lama Alshahrani
University of Tabuk, Faculty of Computers and Information Technology, Saudi Arabia, 421009995@stu.ut.edu.sa
Nouf Albalawi
University of Tabuk, Faculty of Computers and Information Technology, Saudi Arabia, 421009998@stu.ut.edu.sa
Rawan Alharbi
University of Tabuk, Faculty of Computers and Information Technology, Saudi Arabia, 421010012@stu.ut.edu.sa
A'aeshah Alhakamy*
University of Tabuk, Faculty of Computers and Information Technology, Artificial Intelligence and Sensing Technologies (AIST) Research Center, Saudi Arabia, aalhakami@ut.edu.sa

Submitted: *11.02.2023*
Accepted: *20.02.2023*
Published: *30.06.2023*

** Corresponding author*

Abstract:

Machine learning (ML) is one of the computational intelligence aspects that can offer diverse solutions. Medical insurance cost prediction using ML methods is still a problem that must be investigated and improved in the healthcare industry. Two approaches are presented in this study the first uses computational intelligence to predict healthcare insurance costs using ML algorithms. And the second is Spark considered a big data tool. Among the first approach, the algorithms are the well-known linear regression and polynomial regression—based on the features of the input data. Linear regression is a method that shows the relationship between two or more variables. However, in polynomial analysis, the relationship between dependent and independent variables is modeled using polynomials of the nth degree. In this work, we use the KAGGLE repository to analyze the various regression models that can predict the cost of medical insurance. These data are divided based on essential features such as age, sex, BMI, region, number of children, smokers, and charges. The results show that the performance of the polynomial regression model is much better than the linear regression model. The polynomial regression model precisely fits the data according to the target. This is because the given task is non-linear which is hard for a linear model to predict the output as desired. Through the second approach, the data was built on a Jupyter notebook by interfacing tools to get the benefits that coding is very similar to python ML. Also, the cell could be closed, and usual ML coding is resumed on the same notebook. For this method, the obtained results show that the performance of the gradient-boosted tree regression model is much better than a multi-variate and random forest with R2 = 0.9067. This is because of its sequential technique of regression.

Keywords: *Apache spark, Cost prediction, ML algorithms, Medical insurance*

Cite this paper as: Albawi, S., Alshahrani, L., Albawi, N., Alharbi, R., & Alhakamy, A., Prediction of healthcare insurance costs, *Computers and Informatics*, 2023, *3*(1), 9-18

1. INTRODUCTION

It is our right to have access to healthcare whenever we need it. Medical care is a part of life we all perceive as essential. However, this is simply not an option for many people around the world. This could be the result of poverty or being affected by war and conflict. In addition, many vulnerable individuals lack access to medical supplies. The risk of illness is much higher for people without access to healthcare.

Healthcare expenses have been recently rising in both absolute and relative terms. Those in the healthcare industry need to understand the drivers of healthcare spending, how spending varies across regions and the role that technology can play. Both patients and others play an important role in driving healthcare spending. Individual patients can have a significant impact on supply, demand, and pricing by choosing one prescription or treatment over another, opting for elective surgery, or using too much or too little care. Healthcare providers are the supply side of the equation, while patients serve as the demand side. A healthcare provider's choice of services and treatments and the cost associated with them are typically dictated by the patient's needs. The decision may also be influenced by several other factors.

A basic understanding of health insurance is very important for every member of the family today. Citizens should know how much medical expenses cost to incur, dividing the risk among many people. An insurer can then estimate the health system expenditures and the health risk pool over the years to determine the amount of money it will need to pay for the benefits specified in its insurance agreement. This can be done by developing a monthly premium or payroll tax structure.

The goal of this work is to analyze the data collected by the KAGGLE platform and predict the medical prices that will be charged by the insurance company. The data collected includes various features such as age, sex, and BMI. The proposed system can be used by government officials and patients to predict the cost of health care. It can then help them choose the most cost-effective providers. It can also help healthcare administrators plan budgets for the coming years. The proposed system is built on a combination of machine learning algorithms and linear regression. It aims to analyze the various models that can predict the cost of medical insurance. Second, The system is built by Spark(a big data analysis tool) by creating a spark session on Jupyter Notebook. This method had the advantage of coding is very similar to python ML also, the session could be closed and the usual ML code to be resumed on the same notebook. The used algorithm for spark will be mentioned later in the Mythology section of this work.

2. BACKGROUND

It can be very challenging to predict the prices of various products and services, such as electricity, stock prices, and home prices. There are various methods that can be used to analyze and predict these prices, such as neural networks, fuzzy logic, and genetic algorithms.

Several studies have been conducted on the use of machine learning and statistical techniques to predict medical costs. These studies are currently being conducted to estimate the costs of various healthcare services for the next few months. [1]. One of the most common methods used to predict the cost of healthcare is by using linear regression and random forest regression. This method can then be used to estimate the future costs of a person's health care. [2]. In addition to predicting the costs of healthcare, hierarchical regression analysis is also being used in studies to analyze price prediction problems. Multilevel linear regression is being used to study the influence of patient characteristics and physician biases on the diagnostic testing process [3]. A hierarchical decision tree is used to make classification decisions when class labels are hierarchical in nature.

This project aims to provide a unique solution to the problem of predicting the cost of healthcare by analyzing the data collected by an insurance company. The data collected by the system includes various features such as age, sex, and BMI. After analyzing the estimates from the company's system, the system will then determine the monthly premium that the customers pay. The goal of the project is to analyze the various factors that affect the cost of healthcare. In addition to determining the exact amount of insurance that a customer should pay; the system also aims to analyze the multiple factors that influence the premium.

3. RELATED WORK

Recent studies presented in this section describe the various machine learning algorithms that can be used to estimate the cost of healthcare. One of the most recent models that were developed by Taloba, et al [4] was a linear regression model. The researchers used a business analytic method to develop the proposed model. It was compared with the random forest algorithm and the naive Bayes classifier. The results of the study revealed that the linear regression model has a maximum accuracy of 97.89 percent. The healthcare data dataset was obtained from the Kaggle platform. Based on the same dataset, in [5], The researchers also evaluated the various machine learning algorithms that are used in the development of the proposed model. Some of these include the random forest algorithm, the XGBoost, Stochastic Gradient Boosting (SGB), k-Nearest Neighbors, and the support vector regression. The results of the study revealed that the SGB had a high accuracy of 86 percent with an RMSE of 0.340. The Japanese Public Health Insurance Database was used in the study by Nomura, et al. [6], the neural network model was the best way for predicting healthcare costs among machine learning technologies. The cost of healthcare was primarily determined by the previous year's medical healthcare costs.

Furthermore, the nationwide claims database in France is used in [7] to understand how well a basic neural network (NN) and a random forest (RF) are compared to a generalized linear model (GLM) in predicting medical costs at the individual level. The results of the study revealed that the use of RF in the development of the proposed model was beneficial for the prediction of medical expenditures. In addition, the GLM was well-matched with the other variables when it came to analyzing the contribution of predictors. Another example of estimating professional expenses, pharmacy prices, pharmaceutical costs, and inpatient and outpatient healthcare expenditures was provided in the work of Sushmita, et al. [8]. Both algorithms viz Naive Bayes and Decision Tree algorithms were utilized for heart disease prediction in humans [9]. Their results conclude that the Naive Bayes algorithm had better accuracy on small datasets whereas the Decision Tree algorithm executes better on large datasets. Finally, an analysis of Apache Spark is presented by Salloum, et al. [10] showed its features, key components, and abstractions. They also discussed the various features of Apache Spark, which can be used for the development of big data pipelines and machine learning algorithms.

4. METHODOLOGY

This section discusses the main software tools required to develop a healthcare insurance cost prediction system using big data and machine learning approaches. The selected healthcare charges dataset and the development environment are presented and discussed.

4.1. Problem Definition

In this study, the health care charges dataset was obtained from the "Kaggle" platform and was uploaded by Miri Choi in 2018 in Seoul, South Korea. The data are divided based on some important features such as their age, sex, BMI, region, number of children, smokers, and charges. The data that we collected has seven different variables, see Table 1.

Table 1. Variables of the collected data include age, sex, bmi, children, smoker, region, and charges

SN	Feature	Description	Value
(1)	Age	One of the most important aspects of healthcare is age	it has an integer value
(2)	Sex	Gender	(Male=1, Female=0)
(3)	Body Mass Index (BMI)	Understanding the human body: weight that is exceptionally high or low in relation to height	An object body weight index (kg/m2) based on the height-to-weight ratio, ideally 18.5-25
(4)	Children	Number of children/ dependents	it has an integer value
(5)	Smoker	Smoking state	(Smoker=1, nonsmoker=0)
(6)	Region	Area of residence	(Northeast=0, northwest=1, southeast=2, southwest=3)
(7)	Charges	Medical costs paid by healthcare insurance	It has an integer value

The basic statistics of the chosen data are given in the table below. It is worth noting that the data count is 1338, where their mean is calculated accordingly. In addition, the standard deviation (std), maximum (max), and minimum (min), of the data are listed in Table 2.

Table 2. Basic statistics of the chosen data that include count, mean, standard deviation, minimum and maximum values

Stat	Age	Gender	BMI	Children	Smoker	Region	Charge
count	1.34E+03	1338	1.34E+03	1338	1338	1338	1.34E+03
mean	4.95E-17	0.494768	3.36E-17	1.094918	0.204783	1.514948	3.07E-17
std	1.00E+00	0.50016	1.00E+00	1.205493	0.403694	1.105572	1.00E+00
min	-1.51E+00	0	-2.41E+00	0	0	0	-1.00E+00
max	1.77E+00	1	3.69E+00	5	1	3	4.17E+00
25%	-8.69E-01	0	-7.16E-01	0	0	1	-7.05E-01
50%	-1.47E-02	0	-4.32E-02	1	0	2	-3.21E-01
75%	8.40E+01	1	6.61E-01	2	0	2	2.78E-01

4.2. Research Questions

We will analyze the dataset to answer the following research questions:

> ***(R.Q.1.)*** *How much will it differ between the Machine learning and big data tool Spark in the prediction of the result of healthcare cost based on the submitted data?*
> ***(R.Q.2.)*** *For the submitted data will the Linear Regression be more accurate than Polynomial Regression in the ML model?*
> ***(R.Q.3.)*** *For the submitted data will the Multi-variate Linear Regression be more accurate than Random Forest Regression or Gradient-Boosted Tree Regression in the Spark model?*

4.3. First Method: ML

4.3.1. Pre-processing and evaluation matrix

The data is preprocessed before being implemented on machine learning algorithms. A MinMaxScaler is used to transform features by scaling each feature to a zero and one as described in Equation 1:

$$X_{Scal} = \frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

In addition, we used $pandas.get\_dummies$ to convert a categorical variable into dummy variables. Different type of visualization is presented to get a clear analysis of the used dataset. A line chart, scatter plot, histogram, and box plot are used in this study. Feature engineering is implemented to extract features from raw data to improve the performance of ML algorithms. Therefore, unimportant features such as region are removed because it does not affect the charges. The study sample was split into two groups: the training set (75%) and the test set (25%). The training set was used to develop the proposed model while the validation was carried out in the test set. The performance of the final models was evaluated by the RMSE, which is a measure of the difference between the predicted and actual costs.

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

The y_i observed values, n is the number of observations, and the predicted values $\hat{y}_i$ were used to evaluate the performance of the proposed model. The mean absolute error (MAE) was computed by considering the difference between the actual costs and the predicted costs (a smaller value indicates better performance).

$$\sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{n} \tag{3}$$

The performance of the proposed model was evaluated in the entire test set. All of the programmings was performed in Python version 3.8 using various libraries such as Matplotlib, Scikitlearn, and Pandas. To predict healthcare charges in this method, we will use two well-known algorithms - linear regression and polynomial regression - based on the features of the input data.

4.3.2. Data visualization

Getting the most out of data is very challenging when it's in tabular form. This is because it can be hard to understand the data's details and select the appropriate models for it. One of the most important steps that a company should take when it comes to analyzing its data is to visualize it in a pictorial form. This allows them to easily access the trends and patterns that can be found in the data. Matplotlib has various types of tools that can be used to visualize data. One of these is the Seaborn, which can perform complex statistics visualization. It can also be used to create simple bar graphs and line charts. Through this code, seaborn plots were used. Some of the plots are depicted in Figure 1 Figure 2 and Figure 3.

4.3.3. Regression models

While linear regression describes the relationship between two variables by using an equation as follows:

$$y = b_0 + b_1\, x_1 \tag{4}$$

where x is the explanatory variable and y is the dependent variable. Regression analysis using polynomial equations is a method for modeling the relationship between independent and dependent variables using polynomials of the *nth* degree.

$$y = b_0 + b_1\, x_1 + b_2\, x_1^2 \tag{5}$$

Polynomial regression provides a better approximation of the relationship between the dependent and independent variables than the linear regression model. It can be used for a broad range of functions. Polynomials can fit a wide range of curvatures.

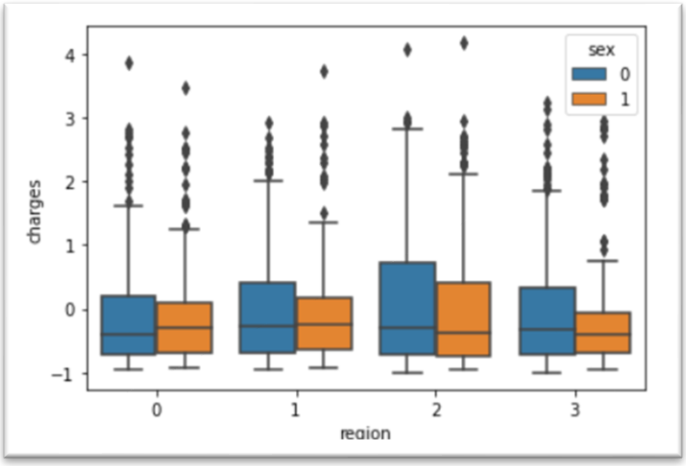


Figure 1. Comparing the charges paid among males & females at each region

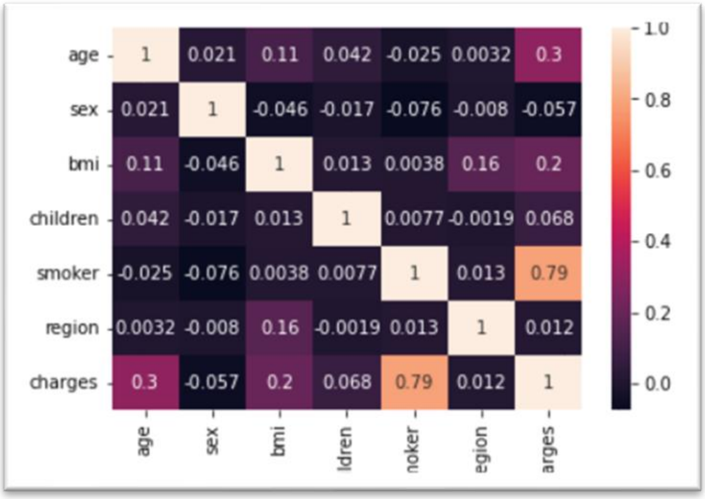


Figure 2. Correlation matrix heatmap

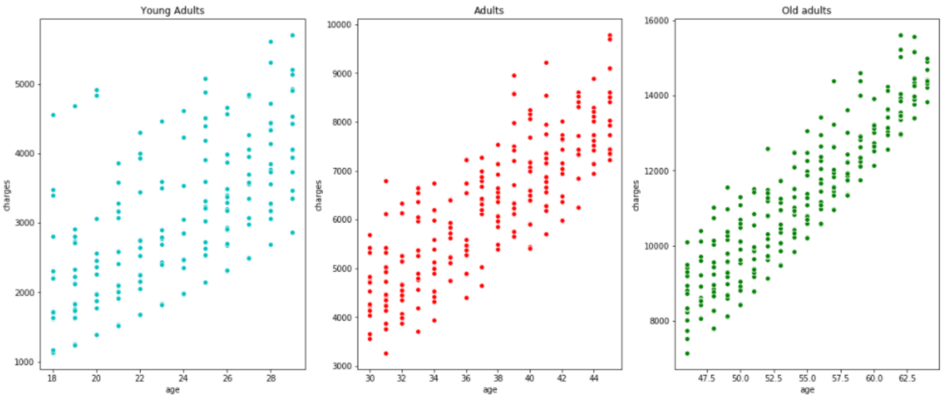
4.4. Second Method: Spark

4.4.1. Data calling and pre-processing

Similar to pandas there is a spark data frame that can be created and could help in pre-processing whit built-in functions. The data is cleaned from any null values and deleted off unnecessary columns converting the string data type into zeros and ones then applying data description. Finally, the modified data frame is saved into a new CSV file since at spark this is an important step for data analysis and modeling. The prepared data is set to a vector assembler which is a transformer that combines a given list of columns into a single vector column, see Table 3.

Table 3. Vector assembler

SN	Features	Charges
0	[19.0, 1.0, 27.899999618530273, 0.0, 1.0]	16884.92383
1	[18.0, 0.0, 33.77000045776367, 1.0, 0.0]	1725.552246
2	[28.0, 0.0, 33.0, 0.0]	4449.461914
3	[33.0, 0.0, 22.704999923706055, 0.0, 0.0]	21984.4707
4	[32.0, 0.0, 28.8799991607666, 0.0, 0.0]	3866.855225



(a)Young adults [18-30] (b) Adults [30-45] (c) Old adults [older than45]

Figure 3. Age groups and charge representation (a)Young adults [18-30] (b) Adults [30-45] (c) Old adults [older than 45].

4.4.2. Multi-variate linear regression model

It's linear regression on multiple variables like the simple linear regression model or Uni-Variate linear regression model, but with multiple independent variables contributing to the dependent variable, resulting in many coefficients to calculate and a more complicated computation due to the additional variables.

$$h\theta\ (x)\ = \theta\ 0x0\ + \theta\ 1x1\ + \theta\ 2x2 + \ldots + \theta\ n - 1xn - 1\ + \theta\ n \tag{6}$$

Where $h\theta$ = target output variable as predicted by our hypothesis, and θ 0= linear regression coefficient, and θ 1 = y-intercept. $x0, x1, x2, \ldots xn - 1$ are the independent variables or features and n is the number of independent variables.

4.4.3. Random forest regression model

One of the most important steps that a company should take when it comes to analyzing its data is to create a model that is more powerful than the original. This can be done using ensemble learning, which involves taking multiple algorithms and putting them together in a powerful model. The accuracy of the proposed model is significantly improved due to the number of predictions that are considered.

4.4.4. Gradient-boosted tree regression

Machine learning techniques are commonly used in the development of prediction models, such as classification and regression. One of the most common techniques that are used in this process is the boost method, which involves combining multiple learning algorithms in a series. This method allows a strong learner to be obtained from many weak prediction models. The goal of boosting is to minimize the errors that occurred in the previous tree. This method achieves a highly accurate and efficient model by adding many trees in a series. Unlike other techniques, boosting does not require background sampling.

5. RESULTS

5.1. Spark Model Result

Among the three models of PySpark (Multi-variate Linear, Random Forest, and Gradient-boosted tree). The obtained results show that the performance of the gradient-boosted tree regression model is much better than others with $R2 = 0.9067$. This is because of its sequential technique of regression. In the case of the multi-variate Linear regression model, the accuracy is not high due to the nature of the given dataset which is non-linear for each variable. For the random forest model, the accuracy is better than the multi-variate and less than the gradient-boosted, see Table 4.

Table 4. Prediction result of PySpark models.

Charges	Gradient-boosted	Random forest	Multi-variate
1121.8739	1995.587627	3434.738322	42.66622343
1131.5066	1489.448142	3698.6897	2190.227374
1163.4626	2141.911792	6111.8343	9314.676475
1241.565	1980.428479	3785.166242	-757.754578

6.2. ML Model Result

As a comparison between both machine learning models (linear & polynomial). The obtained results show that the performance of the polynomial regression model is much better than the linear regression model. This is because it precisely fits the data according to the target. In the case of the linear regression model, the accuracy is not high due to the nature of the given dataset which is non-linear. Therefore, it is hard for a linear model to predict the output as desired, see Figure 4 and Table 5.

Table 5. Prediction result of tested models in our work.

SN	True value	Linear Regression Predictions	Polynomial Regression Predictions
108	2867.11960	2840.936464	2685.405616
986	8410.04685	8554.397424	8702.717332
132	11163.56800	12151.669318	11302.431742
596	7640.30920	7302.061404	7900.293628
869	4391.65200	3926.076473	4368.090689
…	…	…	…
763	3070.80870	2959.403098	2991.403814
771	11150.78000	12136.929803	2991.403814
527	9861.02500	9778.623501	9991.640986
385	1261.85900	1845.014497	1409.337840
135	2155.68150	2314.454494	2147.161122

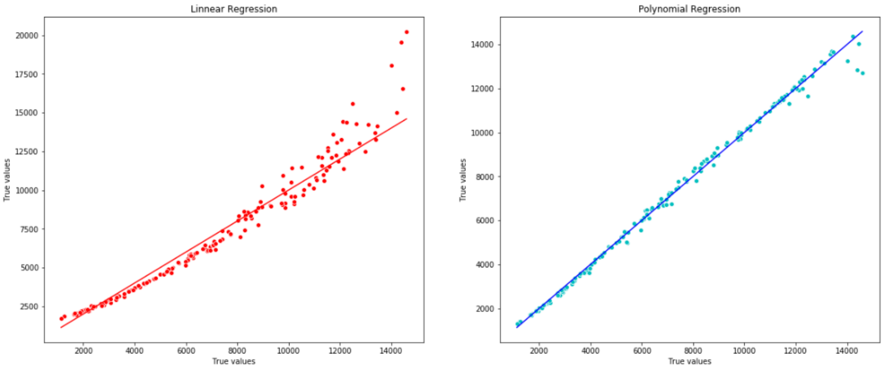
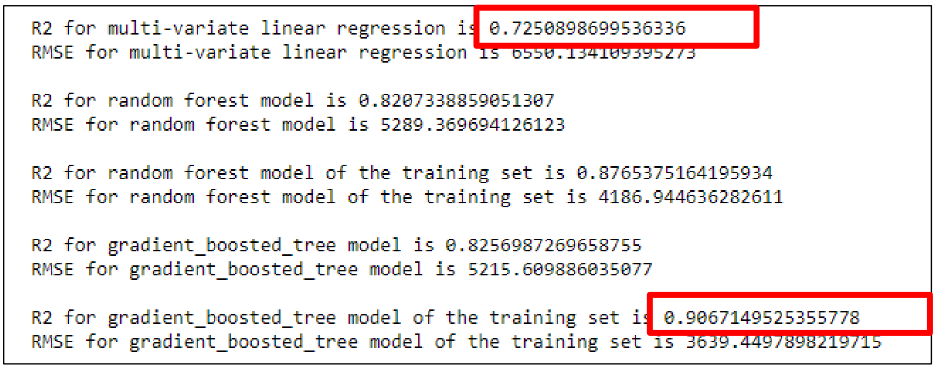


Figure 4. Scatter plot of the predicted values between the two models.

6. EVALUATION

6.1. Spark Model

The root square is the most important parameter to evaluate any model the closer to 1.0 the better. At the spark three models the gradient boosted tree has the greatest value which is 0.9067 and as noticed above its prediction values are the closest to the real charge's values. Also, it has the smallest value of root mean square error RMSE, see Figure 5.

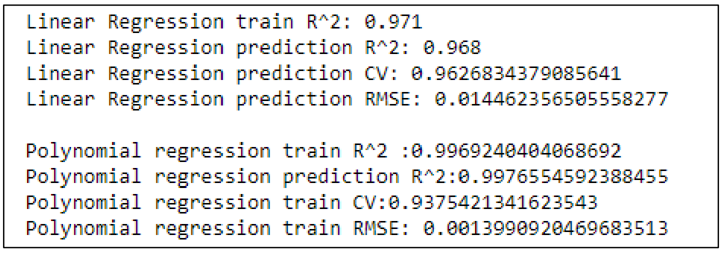


```
R2 for multi-variate linear regression is 0.7250898699536336
RMSE for multi-variate linear regression is 6550.134109395273

R2 for random forest model is 0.8207338859051307
RMSE for random forest model is 5289.369694126123

R2 for random forest model of the training set is 0.8765375164195934
RMSE for random forest model of the training set is 4186.944636282611

R2 for gradient_boosted_tree model is 0.8256987269658755
RMSE for gradient_boosted_tree model is 5215.609886035077

R2 for gradient_boosted_tree model of the training set is 0.9067149525355778
RMSE for gradient_boosted_tree model of the training set is 3639.4497898219715
```

Figure 8. The evaluation parameters between pyspark models MLV, RF and GBT.

6.2. ML Model

The root square value for the polynomial regression is higher which is 0.9969. Also, the RMSE value is less than the linear regression value. That matches the results shown above for polynomial prediction closer to real charges than linear predictions, see Figure 6.



```
Linear Regression train R^2: 0.971
Linear Regression prediction R^2: 0.968
Linear Regression prediction CV: 0.9626834379085641
Linear Regression prediction RMSE: 0.014462356505558277

Polynomial regression train R^2 :0.9969240404068692
Polynomial regression prediction R^2:0.9976554592388455
Polynomial regression train CV:0.9375421341623543
Polynomial regression train RMSE: 0.0013990920469683513
```

Figure 9. the evaluation parameters between the ML models LP, and PNR

7. CONCLUSION

This work discussed two types of Machine Learning methods with a different experiment dealing with a Big Data tool Apache Spark through its Jupiter notebook interface Pyspark. The data processing was faster in the PySpark model than in the ordinary model of ML at Jupiter. Also, the training sets finish in a few seconds compared with the ordinary sets. About the accuracy for PySpark the bigger data the more accurate results. That's maybe obvious from its category (big data tool). The two methods were implemented in one notebook since pyspark.ml was used. It has the ability to start a season (a set of cells programmed with spark commands) then stop it and complete the usual commands in the same notebook.

REFERENCES

[1] J. L. Moran, P. J. Solomon, A. R. Peisach, and J. Martin, "New models for old questions: generalized linear models for cost prediction," *Journal of evaluation in clinical practice*, vol. 13, no. 3, pp. 381–389, 2007.

[2] B. Lahiri and N. Agarwal, "Predicting healthcare expenditure increase for an individual from medicare data," in *Proceedings of the ACM SIGKDD workshop on health informatics*, 2014, pp. 73–79.

[3] P. C. Austin, V. Goel, and C. van Walraven, "An introduction to multilevel regression models," *Canadian journal of public health*, vol. 92, no. 2, pp. 150–154, 2001.

[4] A. I. Taloba, A. El-Aziz, M. Rasha, H. M. Alshanbari, and A.-A. H. El-Bagoury, "Estimation and prediction of hospitalization and medical care costs using regression in machine learning," *Journal of Healthcare Engineering*, vol. 2022, 2022.

[5] J. Iqbal, S. Hussain, H. AlSalman, M. A. Mosleh, S. Sajid Ullah *et al.*, "A computational intelligence approach for predicting medical insurance cost," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[6] Y. Nomura, Y. Ishii, Y. Chiba, S. Suzuki, A. Suzuki, S. Suzuki, K. Morita, J. Tanabe, K. Yamakawa, Y. Ishiwata *et al.*, "Does last year's cost predict the present cost? an application of machine leaning for the japanese area-basis public health insurance database," *International journal of environmental research and public health*, vol. 18, no. 2, p. 565, 2021.

[7] A. Vimont, H. Leleu, and I. Durand-Zaleski, "Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in france," *The European Journal of Health Economics*, vol. 23, no. 2, pp. 211–223, 2022.

[8] S. Sushmita, S. Newman, J. Marquardt, P. Ram, V. Prasad, M. D. Cock, and A. Teredesai, "Population cost prediction on public healthcare datasets," in *Proceedings of the 5th International Conference on Digital Health 2015*, 2015, pp. 87–94.

[9] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," in *2018 second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 2018, pp. 1275–1278.

[10] S. Salloum, R. Dautov, X. Chen, P. X. Peng, and J. Z. Huang, "Big data analytics on apache spark," *International Journal of Data Science and Analytics*, vol. 1, no. 3, pp. 145–164, 2016.