Contents

An introduction to the project dataset

Abstract

The scope of this project is to analyse a publicly available dataset and assign the most accurate machine learning algorithm proposing evaluation methods for the model performance accordingly. Currently, Machine Learning models are the crucial method for the data prediction purpose, which is an objective of this report. Python programming language allowed to implement various functions in the aim of pre-processing, cleaning data including visualisation creation. The dataset became obtain from the American Community Survey from the data.world repository including information about the cancer cases across the Unites States of America. The scope of this report is to provide information about the effective strategies to analyse and pre-process the dataset, the machine learning algorithm proposed including the evaluation methods which are the most suitable for this type of algorithm chosen. Furthermore, the ethical concerns were taken into account including proposed steps to overcome those issues.

KEYWORDS

Machine Learning, Linear Regression, Data Prediction, Python, Ethical Concerns

Background

The dataset chosen for the Machine Learning assignment is a publicly available CSV file consisting of structured information about cancer cases occurrence, the mean of deaths caused by cancer across the United States, the American residents' age regarding different ethnicities, sex, and marital status.

The dataset was downloaded from the www.data.world dataset repository website, whereas the information was gathered from the American Community Survey from both census.gov and cancer.gov.

Dataset characteristics:

- 33 variables
- 3047 records
- 3 datatypes (integers, decimals, and strings)
- 1 nominal attribute (Geography)
- 32 numeric attributes (age, number of cases etc)

```
[6]  print(data_csv.shape)

     (3047, 33)
```

Figure 1 Number of records and fields of dataset.

The information chosen for the analysis of this project includes data about average cancer cases annually and the number of death cases for the 1228 regions of the United States, using 3 from 33 fields for the analysis.

Objectives of the project

The report goals include the implementation of the pre-processing and cleaning processes before conducting the analysis of the large dataset corresponding to the regression problem. Moreover, the report objective is to use the machine learning model to achieve the analysis objectives.

The main objective of the data analysis within the use of Machine Learning model is the prediction of the mortality for people suffering from cancer in the United States, in each region separately. If the number of cancer cases would increase in 10 years, it can be predicted how many people will pass away within the Linear Regression model and the appropriately chosen algorithm.

The outcome of the Linear Regression model implementation could be therefore used by the US government to establish new strategies towards cancer treatment. The US government is able to use the Linear Regression algorithm to predict the number of death cases and invest more money into experimental research to help people suffering from this fatal disease.

The strategies employed to analyse and prepare/pre-process the dataset

Data pre-processing forms allow data analysts to prepare the appropriately adapted dataset for further analysis. First of all, the strategy of identifying missing values was required with the aim of removing all the blank cells from the dataset. The method 'dropna()' returns a new Data Frame with no empty cells.

A column called 'geography' was transformed into two columns within the use of Pandas library, and by default split. Furthermore, as the outliers impact the statistical analysis and output, identifying outliers eliminates noisy data. Descriptive statistical methods and Boxplots were used to identify outliers.

```
#Checking descriptive statistics of the dataframe
new_columns_df.describe()
```

index	Avg_cases	Avg_deaths
count	3047.0	3047.0
mean	606.3385437820807	185.9658680669511
std	1416.3562232267052	504.13428602778606
min	6.0	3.0
25%	76.0	28.0
50%	171.0	61.0
75%	518.0	149.0
max	38150.0	14010.0

Figure 1 Descriptive statistical information about the two fields of interest

It has been discovered that the outliers appear in both variables of interest, and the removal of outliers has taken place accordingly within use of numpy.percentile() method and Numpy library. The outliers were removed within the use of IQR approach, which is an acronym for Interquartile Range. Three quartiles (Q1, Q2, Q3 for 25%, 50%, 75% accordingly.) (www.askpython.com, 2022). Those were replaced with numpy.nan as the Null values. (www.askpython.com, 2022). The dropna() function allowed to remove the outliers completely from the dataset.

Preliminary visualisations

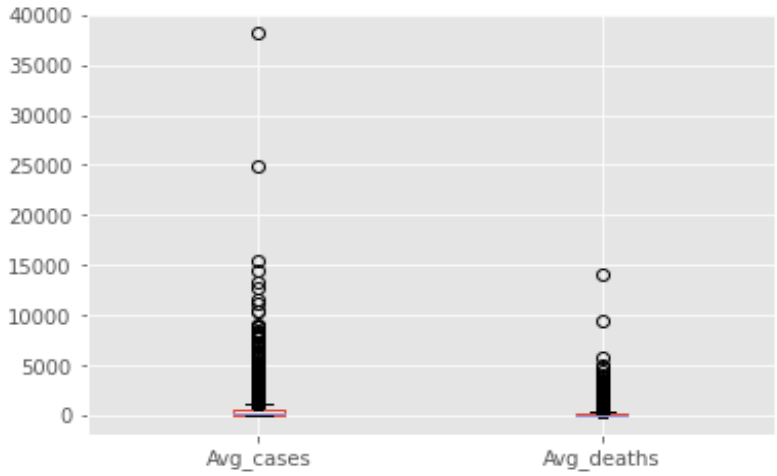

Figure 2 Boxplot presenting the outliers identification for Average cancer cases and Average cancer deaths

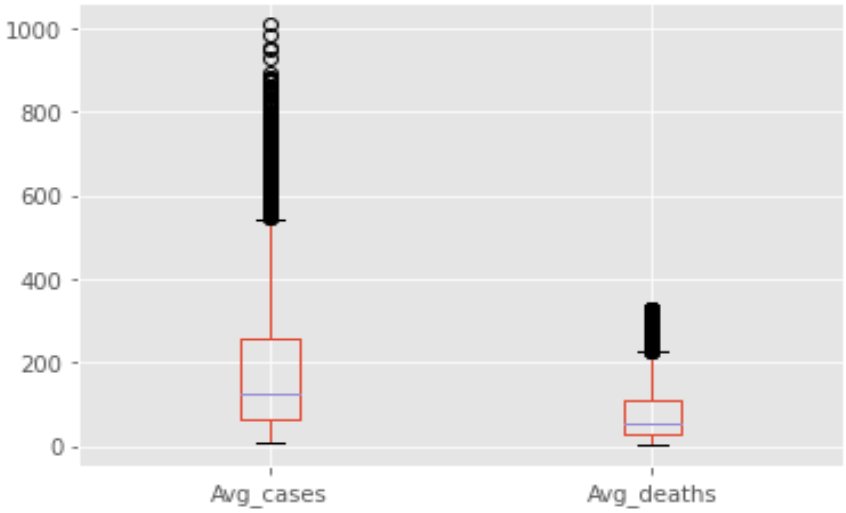

Figure 3 Boxplot confirming outliers' removal

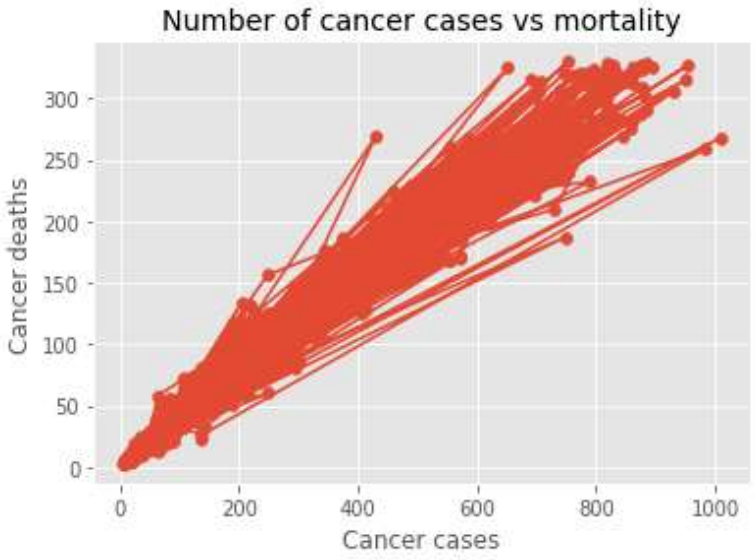

Figure 4 Scatter plot presenting relation between number of cancer vs mortality

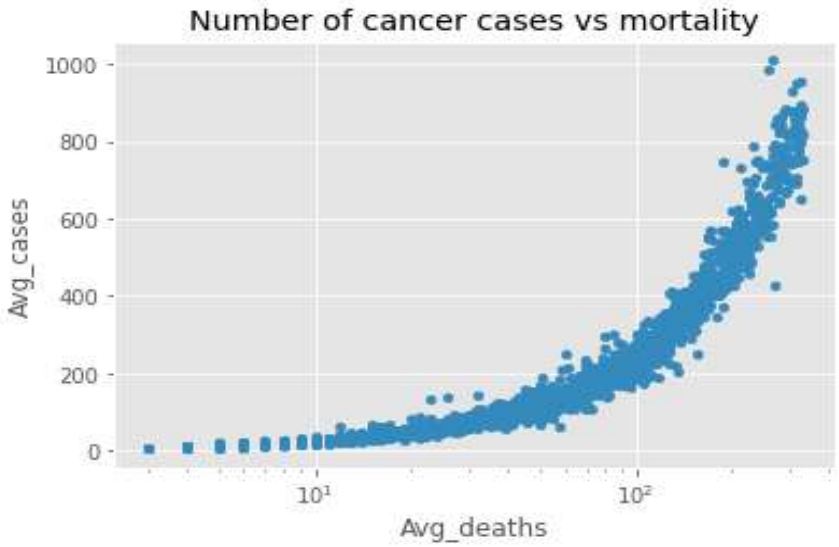

Figure 5 Scatter plot presenting relation between number of cancer vs mortality

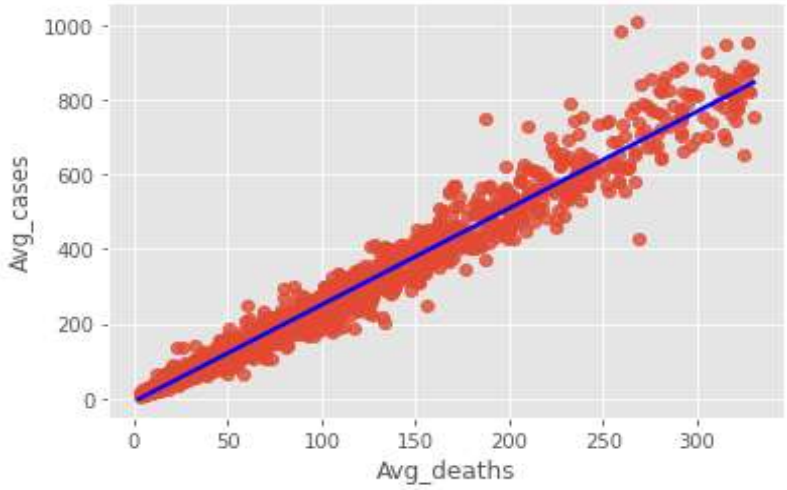

Figure 6 Scatter plot presenting relation between number of cancer vs mortality

The scatter plot presents the strength of the relationship between the mean of cancer cases and the mean of the cancer deaths in the United States. The relationship between both independent (y) and dependent variable (x) values is strong.

Machine Learning algorithm and provide a rational for the model selected

The chosen type of learning algorithm in the purpose of this report is **the supervised learning algorithm** as the objective include predicting death cases of cancer in the United States. *"Supervised learning algorithms work by inferring information or "the right answer" from labeled training data. The algorithms are given a particular attribute or set of attributes to predict."* (McClendon & Meghanathan, 2015). There are various Regression Analysis algorithms available for investigating relationships between different variables and predicting values including Linear Regression, Multiple-Linear Regression, kNN, Decision Tree, Random Forest, SVM or Gradient Boosting algorithms.

The objective stated of this report aim into the prediction of the mortality rate caused by cancer disease across the population of the United States. The observation of the dataset allowed to establish the analysis objectives simultaneously for the model selection. Regression model analyses the relationship between two variables. As explained by Maulud and Abdulazeez (2020) *"regression analysis can be used in some cases to determine causal relations between the independent and dependent variables."* The chosen algorithm allows to predict future values by comparing both independent and dependent variables. As explained by Maulud and Abdulazeez (2020) *"Regression analysis estimates dependent 'y' variable value due to the range of independent variable values 'x'"*.

Linear Regression

$$\beta = \frac{\left[\sum (x_i - mean_x)(y_i - mean_y)\right]}{\left[\sum (x_i - mean_x)^2\right]} \quad \text{and } \alpha = mean_y - \beta * mean_x$$

Figure 2 Linear Regression formula

"This method of regression is simple and provides an adequate and interpretable description of how the input affects the output." (McClendon & Meghanathan, 2015). **Linear Regression** was used with the aim of predicting the average numerical values of the response variables to a value of the predictor variable, including the assumption of the response variable being quantitative and continuous. (2a Regression presentation, NCIRL, 2022).

```
0      100
1      200
2      300
3      400
4      500
5     1000
6     2000
7     3000
8     4000
9     5000
10    6000
```

```
0      43.011168
1      80.491483
2     117.971799
3     155.452114
4     192.932429
5     380.334006
6     755.137160
7    1129.940314
8    1504.743468
9    1879.546622
10   2254.349776
dtype: float64
```

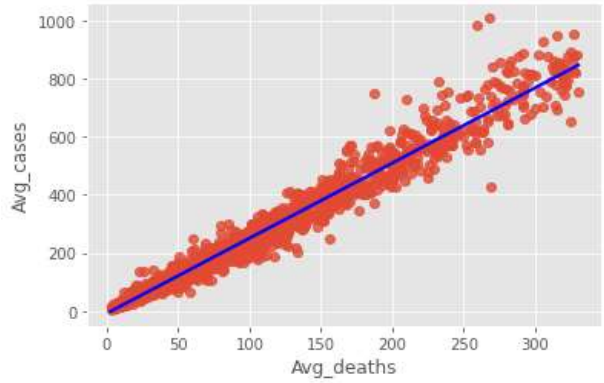

Figure 7 New added independent values (y) Generated new dependent values (x)

Evaluation for the model's performance

R^2 and AdjustedR^2

The coefficient of determination uses other metrics to check how much of the total variation of the targeted variable is explained by the variation in the regression line.

$R^2 = 0.972$

R-squared of 0.972 calculated for the Linear Regression Model has a strong relationship and very good fit of the regression line, meaning the regression was able to capture almost 100% of the variance in the target variable.

- The intercept is equal to -8.7765.
- Adjusted R^2 is equal to the R^2 meaning that the model overfits the data and the learning had no occurrence.

OLS Regression Results

Dep. Variable:	Avg_cases	R-squared:	0.972
Model:	OLS	Adj. R-squared:	0.972
Method:	Least Squares	F-statistic:	8.598e+04
Date:	Fri, 24 Jun 2022	Prob (F-statistic):	0.00
Time:	22:19:03	Log-Likelihood:	-12224.
No. Observations:	2503	AIC:	2.445e+04
Df Residuals:	2501	BIC:	2.446e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-8.7765	0.949	-9.249	0.000	-10.637	-6.916
Avg_deaths	2.5927	0.009	293.230	0.000	2.575	2.610

Omnibus:	1026.777	Durbin-Watson:	1.714
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27803.345
Skew:	1.350	Prob(JB):	0.00
Kurtosis:	19.103	Cond. No.	159.

Although, the accuracy of the regression model cannot be calculated, is it possible to report the performance of a regression model as an error of the predictions. (J.BrownLee, 2021)

MSE (Mean Squared Error)

"The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset." (J.BrownLee, 2021) In the aim of establishing weather the predicted values are to the expected values, the Mean Squared Error is the most common metric used for regression problems.

Two methods for using the MSE metric by Python programming:

1. Metrics.mean_squared_error(y_tre, y_pred, *) (ML, 3a.Model Evaluation and Selection, NCI)
2. err = (expected[i] - predicted[i])**2 (J.BrownLee, 2021)

[1]*"The squared error between each prediction and expected value is calculated and plotted to show the quadratic increase in squared error."* (J.BrownLee, 2021)

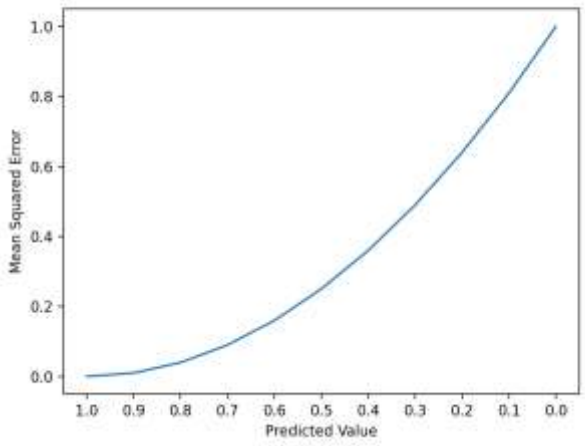

Figure 3 Line Plot of the Increase Square Error with Prediction

RMSE (Root Mean Squared Error)

"… the units of the RMSE are the same as the original units of the target value that is being predicted." (J.BrownLee, 2021) The common evaluation strategy is to use the Mean Squared Error first to train the Regression Model and to use the Root Mean Squared Error afterwards to evaluate and report its performance. (J.BrownLee, 2021)

Two methods for using the MSE metric by Python programming:

1. Sqrt(mean_absolute_error(y_true,y_pred,*)
2. RMSE = sqrt(MSE) (J.BrownLee, 2021)

[1] https://machinelearningmastery.com/wp-content/uploads/2020/12/Line-Plot-of-the-Increase-Square-Error-with-Predictions.png

MAE(Mean Absolute Error)

Third metric method called the Mean Absolute Error presenting the linear and intuitive changes. *"the MAE score is calculated as the average of the absolute error values."* (J.BrownLee, 2021) In fact the absolute error between the prediction value and expected one is calculated therefore plotted in the aim of presenting the linear increase in error. (J.BrownLee, 2021)

Two methods for using the MSE metric by Python programming:

1. Metrics.mean_absolute_erorr(y_true,y_pred,*)
2. err = abs((expected[i] - predicted[i]))

Ethical issues/concerns surrounding the dataset obtained and proposed machine learning algorithm

The privacy

Privacy is the key consideration regarding ethical concern in the Big Data current world. Data sourcing require respecting the privacy of people; therefore, it should be ethical. All the processes of collecting data, managing it, sharing, and using it afterward require following ethical behaviour. As the dataset obtained for this report include data about the disease of cancer, which has high mortality amongst population of white ethnicity, breaching of the people suffering of cancer' privacy may result in lawsuits for the company obtaining the data and sharing the information in a virtual world. Therefore, the data collected about people without their consent is a significant concern surrounding the data sourcing.

Lack of transparency

As Floridi (2021) stated, *"Lack of transparency- whether inherent due to the limits of technology or acquired by design decisions and obfuscation of the underlying data."* In fact, the algorithms may exhibit negative tendencies such as the discrimination mentioned above may result in a lack of transparency. It is required to produce trustworthy Machine Learning research.

Another concern includes the difficulty for human to interpret algorithmic models and datasets plus the code might have poor structure being impossible to read.(Tsamados, A, Aggarwal. N, Cowls.J, Morley.J, Roberts.H, Taddeo.M, Floridi, L, 2021)

Steps to be taken to address the ethical concerns identified

Companies using Machine Learning models with the aim of improving their marketing or business strategies should implement Artificial Intelligence audit trails to monitor the development of the algorithms avoiding the customers' harm or product liability cases. (M. Seeds, 2021)

Furthermore, the lack of algorithmic transparency could be treated by the use of technical tools to test and implement audits of the algorithmic systems. This step could be taken in the aim of auditing the prediction obtained from the Linear Regression model of machine learning. Moreover, the algorithmic impact assessment guidance could be implemented to address the ethical concerns and control algorithms. (Floridi, 2021). Moreover, it is recommended for the businesses to create their own ethical and fair procedures and policies regarding all the aspects of consumer data. Those should be balanced with fair usage and transparency. (ML - 1a. Introduction and Ethics, NCI)

References

Dataset source

Data.world. (2022). Available at: https://data.world/exercises/linear-regression-exercise-1/workspace/file?filename=cancer_reg.csv [Accessed on 17th June 2022].

Books

Floridi, L (2021) Ethics, Governance, and Policies in Artificial Intelligence. Springer. the UK

Articles & journals

- BrownLee, J. (2021). Regresion Metrics for Machine Learning. Available at: https://machinelearningmastery.com/regression-metrics-for-machine-learning/ [Accessed on 2nd July 2022].
- Floridi, L, Tsamados, A, Aggarwal N, Cowls, J, Morley. J, Roberts , Taddeo, M (2021). The ethics of algorithms: key problems and solutions. Available at: https://link.springer.com/article/10.1007/s00146-021-01154-8 [Accessed on 27th June 2022].
- Mahesh, B. (2019). Machine Learning Algorithms – A Review. Available at: https://www.researchgate.net/profile/Batta-Mahesh/publication/344717762_Machine_Learning_Algorithms_-A_Review/links/5f8b2365299bf1b53e2d243a/Machine-Learning-Algorithms-A-Review.pdf?eid=5082902844932096 [Accessed on: 20th June 2022].
- Maulud, D & Abdulazeez, A (2020) Journal of Applied Science and Technology Trends. A Review on Linear Regression Comprehensive in Machine Learning . Vol.01, pp.140 – 147.
- McClendon. L & Meghanathan. N (2015) Machine Learning and Applications: An International Jorunal(MLAIJ). USING MACHINE LEARNING ALGORITHMS TO ANALYZE CRIME DATA. Vol.2, No.1.
- Seeds, M. (2021) The Trouble with Algorithms: Algorithm Ethics. Available at: https://clarkstonconsulting.com/insights/algorithm-ethics/ [Accessed on 27th June 2022].

Websites

- Askpython, (2022). Detection and Removal of Outliers in Python – An Easy to Understand Guide. Available at: https://www.askpython.com/python/examples/detection-removal-outliers-in-python [Accessed on: 20th June 2022].