Regression Analysis on Diabetes Dataset

\*Note: Sub-titles are not captured in Xplore and should not be used

line 1: 1st Given Name Surname   
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 2nd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 3rd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*—Today’s, Diabetes has recognized as most common disease all over the world which leads to premature death and other critical health problems. Generally, physicians are gathers certain parameters from patients to diagnose diabetes and as lack of experience there are chances of error. In this project, we have adopted regression classifiers to train over given dataset and evaluate their performance. The comparison results has presented in this report. The classifiers are built in Python programming environment by using scikit-learn library.

Keywords—Diabetes, Scikit-learn, regression

# INtroduction

In recent years, Diabetes has recognized as major health challenge across the world. The prevalence of Diabetes arises the financial burden as it hit the epidemic proportions. The rapid growth in Diabetes patients across world attracted the researchers attention [1-2]. The Diabetes disease leads to early death and other problems such as heart-attack. The reports presented by World Health Organization (WHO) stated that there were above 400 million patients who had infected by diabetes in 2014 [3] and in 2045 this number has projected to around 700 million [4]. In order to diagnose Diabetes in a person, typically a physician considers several attributes. The information gathered from the patients is critical for decision makers such as physician. In this project, we have aim to apply machine learning approaches over patients attributes to diagnose diabetes.

## Contribution

* In this project, the main objective is to build and compare the regression models over diabetes observations.
* Build an interactive user interface to enable ease of use

## Structure

The rest of the report has organized as follows: Section describes the dataset, Section 3 describes the methodology, and Section 4 describes the experiment setup and libraries which are used to perform this experiment. Section 5 presents the results and discussion; Section 6 concludes the presented methodology and results.

# Dataset

The data on Diabetes patients has download from Sk-learn[[1]](#footnote-1) which is open-source.

Table 1 Dataset

|  |  |
| --- | --- |
| Total Instances | 42 |
| Dimension | 10 |
| Features | real, -.2 < x < .2 |
| Outcome | integer 25 - 346 |

# methodology

## Decision Tree

The decision tree model considers the data observations where each observation presents target attributes with nodes in a tree structure. It follows numerous path in structure to reach at leaf node to make prediction or take decision. The trees are built over a given dataset and adopts to solve regression and classification problems.

## Regression

A supervised learning approach based on statistical computation. During training phase, the regression algorithms refine error iteratively to minimize error to achieve a better score as accuracy. Further, a main objective of regression model is to reduce error between actual and forecasted value.

### Linear Regression

### Ridge

### Lasso

## Random Forest

Random Forest is an ensemble model that generates multiple models and then combine their output to achieve better accuracy.

## Support Vector Machine

A supervised learning based model widely used for regression problems. In this algorithm, each instance of data is considers as point over n-dimensional hyperplane where n denotes the number of attributes in dataset. It classifies data points over hyper place if two classes are different from each other.

Each regression model has setup with certain parameters as presented in Table 2.

Table 2 Algorithms Settings

|  |  |  |
| --- | --- | --- |
| # | Algorithm | Parameters |
| 1 | Decision Tree | max\_depth=4,splitter="best",criterion="entropy", random\_state=12345 |
| 2 | Linear Regression | copy\_X=True, fit\_intercept=True, n\_jobs=1, normalize=False |
| 3 | Ridge | alpha=0.01 |
| 4 | Lasso | alpha=0.01 |
| 5 | Random Forest | n\_estimators = 10,bootstrap=True,max\_features='sqrt' |
| 6 | Support Vector Machine | kernel='linear', C=1e3 |

# Experiment Setup

It is essential to evaluate the algorithms in terms of accuracy, there is it is crucial to setup simulation environment. In this project, we have setup Python 3.6 Programming language and corresponding libraries such as sickit-learn which provides high level interface to call regression model.

Further, a software has developed in Python using Tkinter library which allows to build a graphical user interface to allow user to interact with software in an easy way. The developed software enables the user to load, pick certain parameters, algorithm and simulate the results.

# Results

The Figure 1 and 2 depicts the User Interface of developed simulator. The application interface enables the user to load and execute regression mode and the results are displayed to user screen.

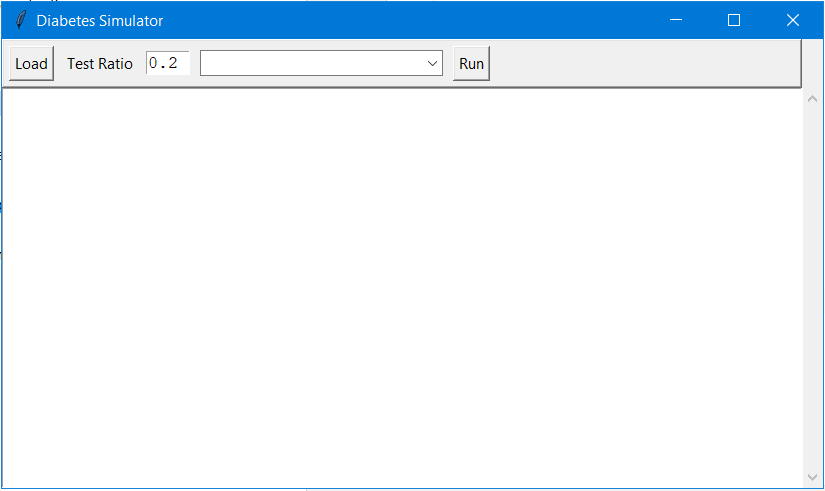


Figure 1Interface

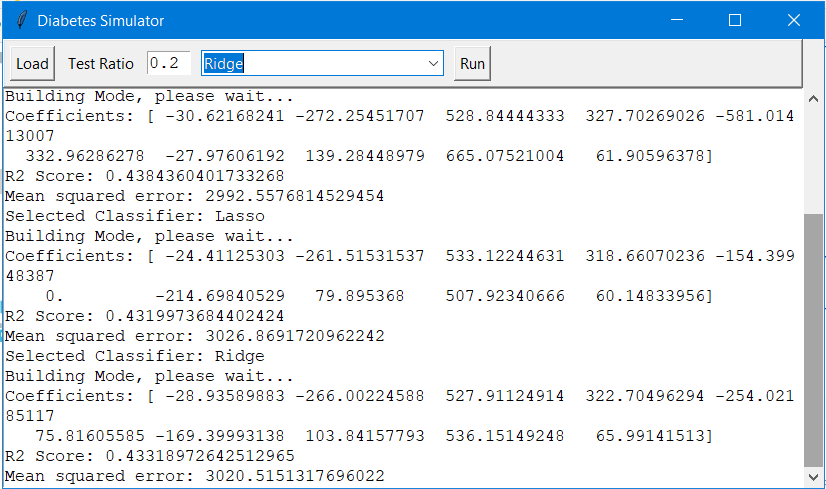


Figure 2 Interface

Results are summarized in Table 3, MSE and R2 score are used to evaluate each regression model. Linear Regression and SVM are yield .43 R2 score while Decision Tree and Random Forest yield worst performance over given diabetes dataset.

Table 3 Results

|  |  |  |  |
| --- | --- | --- | --- |
| # | Algorithm | Mean Square Error | R2 Score |
| 1 | Decision Tree | 6554.865 | -0.23 |
| 2 | Linear Regression | 2992.557 | 0.43 |
| 3 | Ridge | 3020.515 | 0.43 |
| 4 | Lasso | 3020.515 | 0.43 |
| 5 | Random Forest | 8211.539 | -0.54 |
| 6 | Support Vector Machine | 3024.902 | 0.43 |

# Conclusion

In this project, a diabetes simulator along UI has developed which allows user to select regression model and simulate classifier behavior in terms of accuracy. The Python and Scikit-learn library has used to build the solution. The result shows that Linear Regression, Ridge, Lasso and SVM are good model for diabetes dataset.

##### References

1. Gao Y, Wang Y, Zhai X, He Y, Chen R, Zhou J, et al. Publication trends of research on diabetes mellitus and T cells (1997–2016): a 20-year bibliometric study. PLoS ONE. 2017;12(9):e0184869.
2. Abutaleb MH. Diabetes mellitus: an overview. Pharm Pharmacol Int J. 2016;4(5):406–11. https:// doi.org/10.15406/ppij.2016.04.00087.
3. World Health Organization. Health topics: diabetes. https://www.who.int/diabetes/en/.
4. International Diabetes Federation. IDF diabetes atlas. 8th ed. Brussels: International Diabetes Federation 2017

1. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load\_diabetes.html [↑](#footnote-ref-1)