

**Data Driven Decision Making**

*Risk of diabetes prediction*

*Sriram Ranganathan*

*Parvez Bahadur*

**School of Graduate Professional Studies**

MPS/MS in Data Analytics

DAAN 881 – Data Driven Decision Making

Fall 2019

# Document Control

## Work carried out by:

|  |  |  |
| --- | --- | --- |
| **Name** | **Email Address** | **Task description** |
| Sriram Ranganathan | Skr5613@psu.edu | Getting ideas and the dataset |
| Parvez Bahadur | Pxb99@psu.edu | Getting ideas and the dataset |
|  |  |  |

## Revision Sheet

|  |  |  |
| --- | --- | --- |
| **Release No.** | **Date** | **Revision Description** |
| 1 | 10/27/2019 | Idea generation |
| 2 | 11/3/2019 | Interpretation of data |
| 3 | 11/10/2019 | Finding problems with the dataset |
| 4 | 11/17/2019 | Data cleaning |
| 5 | 11/30/2019 | PCA |
| 6 | 12/08/2019 | Data modeling |
| 7 | 12/12/2019 | Report finalization |

**TABLE OF CONTENTS**

[Document Control 1](#_Toc1830484)

[Work carried out by: 1](#_Toc1830485)

[Revision Sheet 1](#_Toc1830486)

[Table of contents 1](#_Toc1830487)

[General guidelines 2](#_Toc1830488)

[Academic Integrity 2](#_Toc1830489)

[Week 1 Assignment 5](#_Toc1830490)

[Problem Statement 5](#_Toc1830491)

[Answer 6](#_Toc1830492)

[Week 2 Assignment 6](#_Toc1830493)

[Problem Statement 6](#_Toc1830494)

[Answer 7](#_Toc1830495)

[Week 3 Assignment 10](#_Toc1830496)

[Problem Statement 10](#_Toc1830497)

[Answer 11](#_Toc1830498)

[Week 4 Assignment 17](#_Toc1830499)

[Problem Statement 17](#_Toc1830500)

[Answer 18](#_Toc1830501)

[Week 5 Assignment 20](#_Toc1830502)

[Problem Statement 20](#_Toc1830503)

[Answer 21](#_Toc1830504)

[Week 6 Assignment 23](#_Toc1830505)

[Problem Statement 23](#_Toc1830506)

[Answer 24](#_Toc1830507)

[References 88](#_Toc1830508)

**General Guidelines**

1. To complete all the homework assignments for this course please use this template document.
2. Each assignment has to be submitting by the following Sunday 11:59 PM EST.
3. Each figure should be followed by a brief description about the figure.
4. The figures can be hand drawn and scanned in some circumstances, but the hand drawn figure should be clear and legible to obtain full credits. Unclear hand drawn figures will receive partial credits. For constructing figures and diagrams it is advised to use tools.
5. Figures and tables should have appropriate captions. For documenting and referencing styles please follow the APA or MLA writing style.
6. Please make sure that you provide a reference section.
7. Any material text or figure taken from books, journals or Internet should be referenced. If you have a sentence or a figure that does not belong (authorship) to you, they need to be clearly referenced. If you fail to do so your report will be considered as a case for plagiarism. It is your duty to make sure that your report is free from any activity related to plagiarism. In case you are suspected of attempting plagiarism then you will be responsible for the cause. The penalty for plagiarism will be a “0” awarded to your report. So, it is good to keep simple, always have the principle to acknowledge people for their contributions.

Please go through the following instructions before submitting the report

#### **Academic Integrity**

Academic integrity — scholarship free of fraud and deception — is an important educational objective of Penn State. Academic dishonesty can lead to a failing grade or referral to the [Office of Student Conduct](http://www.sa.psu.edu/ja/).

Academic dishonesty includes, but is not limited to:

* cheating
* plagiarism
* fabrication of information or citations
* facilitating acts of academic dishonesty by others
* unauthorized prior possession of examinations
* submitting the work of another person or work previously used without informing the instructor and securing written approval
* tampering with the academic work of other students

#### How Academic Integrity Violations Are Handled

In cases where academic integrity is questioned, [procedure requires an instructor to notify a student](http://www.psu.edu/oue/aappm/G-9-academic-integrity.html) of suspected dishonesty before filing a charge and recommended sanction with the college. Procedures allow a student to accept or contest a charge. If a student chooses to contest a charge, the case will then be managed by the respective college or campus Academic Integrity Committee. If a disciplinary sanction also is recommended, the case will be referred to the [Office of Student Conduct](http://www.sa.psu.edu/ja/title=).

All Penn State colleges abide by this Penn State policy, but review procedures may vary by college when academic dishonesty is suspected. Information about Penn State's academic integrity policy and college review procedures is included in the information that students receive upon enrolling in a course.

Additionally, Penn State students are expected to act with civility and personal integrity; respect other students' dignity, rights, and property; and help create and maintain an environment in which all can succeed through the fruits of their own efforts. An environment of academic integrity is requisite to respect for oneself and others, and a civil community.

#### For More Information on Academic Integrity at Penn State

Please see the [Academic Integrity Chart](http://www.campuses.psu.edu/CAO.pdf)  for specific college contact information or visit one of the following URLs:

* Penn State Senate [Policy on Academic Integrity](http://www.psu.edu/dept/oue/aappm/G-9.html)
* [iStudy for Success!](http://istudy.psu.edu/tutorials/) — learn about plagiarism, copyright, and academic integrity through an educational module
* [Turnitin](http://tlt.its.psu.edu/turnitin) a web-based plagiarism detection and prevention system

**Week 1 assignment**

Today, the top companies around the world use data to make decisions about their business. The reason they’re leading the way is that they’ve gained a strategic advantage over their rivals simply by shifting their focus to data rather than relying on business acumen alone. In the next 7 weeks your objective will be to demonstrate the potentiality of collecting, cleaning and critically analyzing the data for decision making. In summary, this course is all about presenting a case on data driven decision making. Based on the CRISP-DM process model a critical decision making process involves collecting the data from a source (Internet, data mart/warehouse, data sources, publicly available documents and audio/video etc.) and establish an in-house data mart/database/dataset, identify issues in the collected data (errors, missing values, outliers, semantic and syntactic errors etc.), clean the data, prepare the data, and use the data for critical decision making *i.e.* to address the research /business related queries. To do so you must first define your overall research or business goal. Once you have identified your research/business goal the next step would be to identify the sources for collecting data.

In week 1 your goal would be to define a research or a business goal, identify the queries to be addressed, and identify the source (s) for collecting data. Please remember that data can be collected from a single or multiple source. In this week you must just identify the source (s) from where you are planning to collect data related to your research goals. The actual data collection steps can be conducted and completed in the following week.

Sources for exploring Business problems and data sources are (This list is not exhaustive):

* <http://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html>
* <http://www.kdnuggets.com/datasets/index.html>
* Explore links to other data sources open to public at <https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public>
* <http://sherlock.ics.uci.edu/data.html>
* <http://scg.sdsu.edu/sample-data/>
* <https://www.crowdflower.com/data-for-everyone/>
* <http://www.cc.gatech.edu/gvu/user_surveys/>

It is preferable that you identify a research/business goal related to your job/interest. If you cannot find such a research/business goal a default one is given to you below:

**Default research/business goal**(use in case you cannot identify a work-related research/business goal):

**Did the Stop-and-Frisk program commissioned by the New York Police Department (NYPD) help the police in slashing down the crime rate of the New York City**?

Please remember that this is just a suggestion. You can tweak the research/business goal related to the Stop-and-Frisk program. For example:

**Did the NYPD detain, and question pedestrians of a particular ethnic group more frequently than any other ethnic groups i.e. was there any evidence of racial bias in the Stop-and-Frisk program.**

For more information you can refer to the article

**Reading paper link**: Gelman, A., Fagan, J., and Kiss, A., “An Analysis of the New York City Police Department’s “Stop-and-Frisk” Policy in the Context of Claims of Racial Bias”, retrieved from <http://www.stat.columbia.edu/~gelman/research/published/frisk9.pdf> retrieved on June 25, 2016.

Download the template document and complete the week 1 assignment. In this week you must:

1. Define your research/business goal.
2. Briefly describe the queries (research/business questions) you want to address. Limit the number of queries to a maximum of four.
3. Briefly discuss about the source (s) from which you are planning to collect the data to address your research/business queries.
4. Please update the **Reference** section which should be the last page of the template document. For referencing please follow the APA/MLA guidelines. More information about the APA/MLA guidelines can be obtained from the link [http://academictips.org/mla-format/ (Links to an external site.)Links to an external site.](http://academictips.org/mla-format/)
5. Add any Appendix section if needed.

**Research/Business Goal:**

The goal of this project is to predict the risk of diabetes for a patient, based on different diagnostic measures obtained from the patient. The objective is to tell whether a patient has high chance of diabetes if certain diagnostic conditions are met.

**Research Queries:**

• Does pregnancy have an effect on the risk of diabetes?

• what are the metrics which are used to calculate the risk of diabetes for a patient?

• Is prima Indian heritage more prone to diabetes?

• Can a machine learning model be used to predict the risk if proper diagnostic inputs are fed?

**Data Source:**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. This data is sourced from one of the Kaggle competitions and can also be found in the UCI repository for machine learning. It contains the diagnostic measures obtained from prima Indian heritage patients who are females with age at least 21 or higher.

**Week 2 assignment**

In week 1, you had defined your research/business goal (queries) and had also identified the source(s) for data collection. In this week your task will be to retrieve/collect the data from the source and maintain a local copy of the dataset/database/ data mart/data warehouse. To pursue a challenging as well as a manageable project I will recommend that your dataset/database/ data mart/data warehouse contains at least 20 +, 30 +, < 40 columns and several thousand rows (not more than 100K unless you need to have a larger dataset). Also note that you can establish a dataset/database/ data mart/data warehouse by combining one or more dataset/database/ data mart/data warehouse.

**Note**: If your work requires you to collect unstructured data (tweets from Twitter, text from blog/web sites etc.) then please consult with me to get a good idea about the requirements for data collection in this course.

**For structured data**

Once you have retrieved/collected the dataset/database/data mart/data warehouse update the template document with the following sections:

* Provide a brief description of the dataset/database/ data mart/data warehouse. The description could involve listing the following: the attributes, the data types of the attributes, number of instances, attribute values, missing values, outliers if any? etc. You can follow the format for the data description provided in the link <http://archive.ics.uci.edu/ml/machine-learning-databases/car/car.names.>
* Also describe the types of data and measurement scales (nominal, ordinal, interval, and ratio) across each attribute. Please go through this link (<http://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/>) for additional details about the types of data and measurement scale.

**For unstructured data**

Discuss with me to gather the requirements for the week 2 assignment. Please get in touch with me to set up a meeting with me during the office hours.

Submit your completed week 2 assignment to the dropbox in Canvas.

**ANSWER:**

**Data Characteristics**

Below are some of the characteristics of the obtained dataset –

* The columns are 9 attributes representing 769 Pima Female Indians.
* The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

**Column Description and Scale**

Below are the columns from the raw dataset with their description [2], measurement scale and missing values –

Attribute – exact names as provided in the dataset

Description – explanation about what actually the attributes interpret

Data Type / Scale – Data types and the measurement scale of the dataset based on data understanding

Outliers – number of outliers present in the data interpreted by plotting a graph

Missing values – counting the number of null values from the data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attribute | Description | Data Type /  Measurement Scale | Distribution | Missing Values |
| Pregnancies | Number of times pregnant | Discrete | Has Outliers | We have 55 patients for whom the Pregnancy count is zero. These are patients who did not have any pregnancies, so these are NOT missing values. |
| Glucose: | Plasma glucose concentration 2 hours in an oral glucose tolerance test | Continuous/Ratio Scale | None | We have 4 patients for whom the Blood Pressure is zero. We know that this is not possible so these are missing values. |
| Blood Pressure: | Diastolic blood pressure (mm Hg) | Continuous/Ratio Scale | Has outliers | We have 36 patients for whom the Blood Pressure is zero. We know this is not possible so these are missing values. |
| Skin Thickness: | Triceps skin fold thickness (mm) | Continuous/Ratio Scale | Has Outliers | We have 9 patients for whom the Skin Thickness is zero. We know that this is not possible so these are missing values. |
| Insulin: | 2-Hour serum insulin (mu U/ml) | Continuous/Ratio Scale | Has Outliers | We have 375 patients for whom the Insulin serum levels is zero. We know that this is possible for a person to have zero insulin level, so these are NOT missing values. |
| BMI: | Body mass index (weight in kg/(height in m) ^2) | Continuous/Ratio Scale | Has outliers | We have 10 patients for whom the BMI is zero. We know that this is not possible so these are missing values. |
| Diabetes Pedigree Function: | Diabetes pedigree function. Measurement of Risk of Diabetes based on genealogy and family history. | Continuous/Ratio Scale | Has Outliers | No missing values |
| Age: | Age (years) | Discrete | Has Outliers | No missing values |
| Outcome: | Response variable for diabetes. (0 or 1) | Categorical/Nominal | No Outliers | No missing values |

**Week 3 assignment**

This week is all about conducting a preliminary analysis and investigating for issues (errors and inconsistencies) in your in-house dataset/database/ data mart/data warehouse. You must go through your dataset/database/ data mart/data warehouse, analyze and identify all the issues. There might be several issues in your dataset including missing values, outliers, obvious inconsistencies, type conversion, string manipulation, collinearity etc.

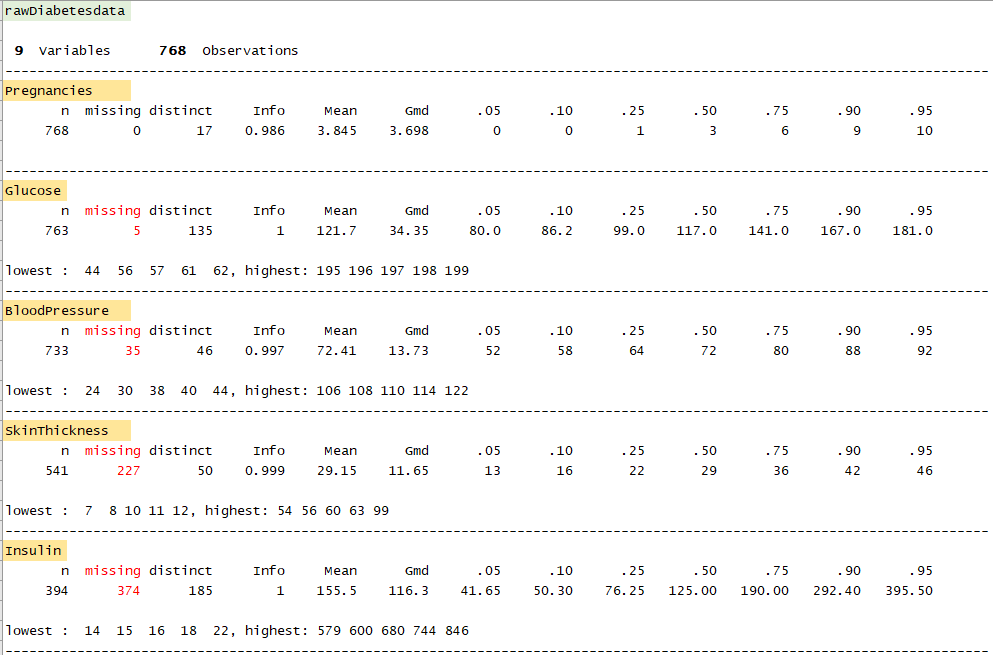
Update the template document with the following information

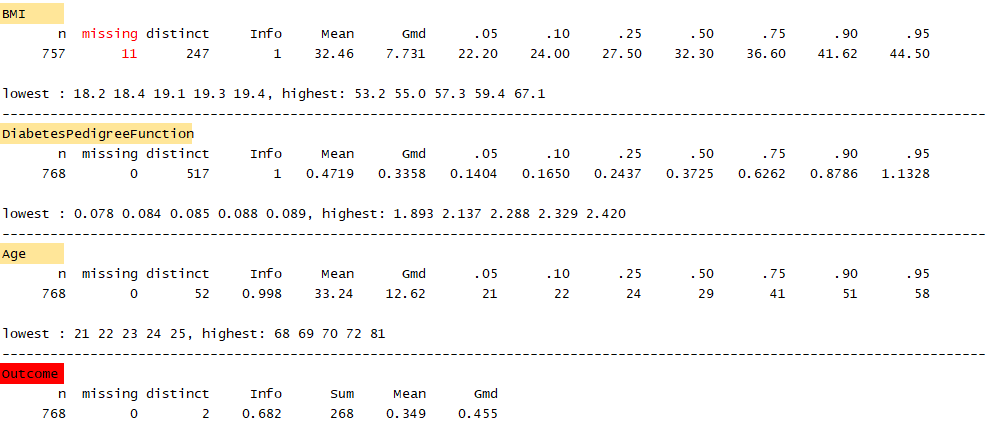
* Provide a brief analysis of your quantitative/qualitative data. Please go through this link (<http://toolkit.pellinstitute.org/evaluation-guide/analyze/analyze-qualitative-data/>) for additional details on how to analyze the quantitative/qualitative data.
* List of all the issues in your in-house dataset/database/data mart/data warehouse. Extensive cleaning (data cleaning) of your dataset/database/data mart/data warehouse will be performed in week 4. In this week you must just identify the issues and report them. In addition to that you must also list/describe all the steps that you will be taking to correct the issues in your dataset/database/data mart/data warehouse.

Submit your completed week 3 assignment to the dropbox in Canvas.

**Descriptive Statistics of the data for missing values :**

We observe that there is a large number of missing values especially for the fields skin thickness and insulin. From business knowledge we know that none of the columns can have a zero value except for the variable “Pregnancy”. For the purpose of this analysis we will be replacing the missing values with the mean. The following screenshots give details on the number of variables, sample size, missing data and few more observation to better understand the data.

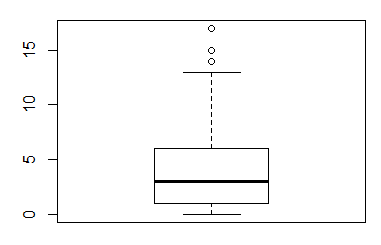




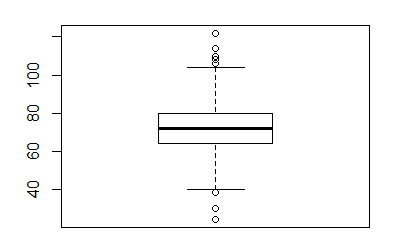
**Outliers:**

We can find outliers in the following columns:

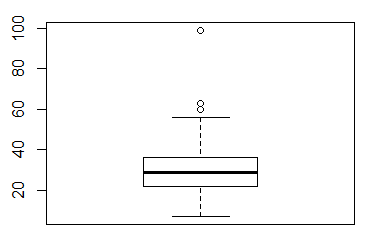
Pregnancies:



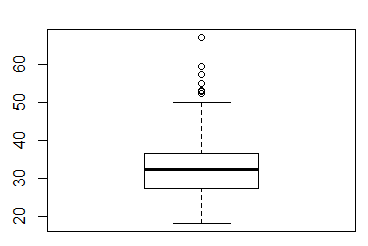
Blood Pressure:



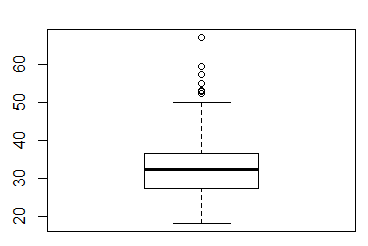
Skin Thickness :



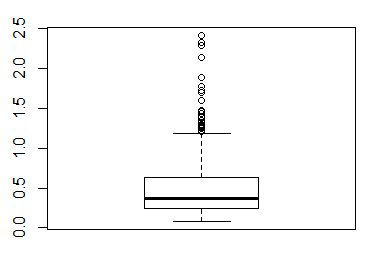
Insulin :



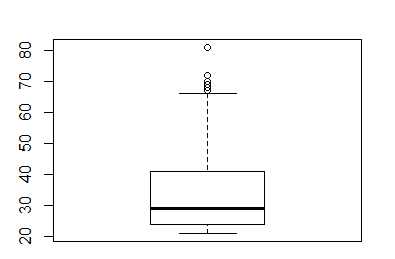
BMI:



Diabetes Pedigree function:

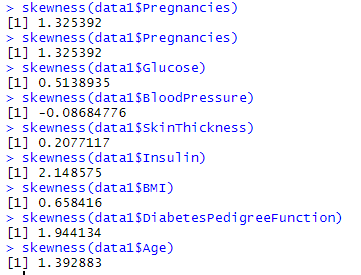


Age:



**Skewness:**

We do find some skewness in the data. The values are given below:



We can see little bit of positive as well as negative skewness. Therefore, we will be applying Box Cox transformation to remove this skewness.

**Duplicate values:**

Since our data set consists of diagnostic measures of patients, it is obvious that not all the values can be same for all the patients. The rows are not duplicated since at least one or more of the diagnostic measures vary between the patients.

**Week 4 assignment**

This week is all about cleaning your in-house dataset/database/data mart/data warehouse. Based on the issues identified in your dataset/database/data mart/data warehouse and the proposal for resolving the issues (reported in Week 3) you must take appropriate steps to clean your data. You can use any software you like to clean your data. Some of the popular software used for data cleaning are

* WEKA
* Excel
* R
* KNIME
* Python

Please visit this link <http://www.kdnuggets.com/2011/04/free-tools-data-visualization-analysis.html> . This link provides you with a list of free tools available for data cleaning, visualization and analysis.

You will have to update the template document with the following:

* Describe the entire data cleaning process and the outcome for each of the data cleaning step.
* Add appropriate tables and figures if needed.
* Feel free to add an appendix section if needed.

Submit your completed week 4 assignment to the dropbox in Canvas.

**Replacing missing values with median:**

Since there are outliers, we cannot impute the data by taking the mean. Therefore, we had chosen to impute the missing value with the column’s median value.

**Code**:

import pandas as pd

import numpy as np

data=pd.read\_csv("project.csv")

data.fillna(data.median(), inplace=True)

data

data.isna().sum()

A screenshot of a cell phone

Description automatically generated

**Skewness:**

We have found skewness in the variables. Therefore, we applied Box-Cox transformation to the variables in order to remove the skewness and see if the outlier count reduces.

**Code**:

#Skewness before box-cox

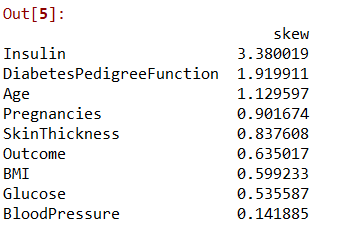
nums=data.dtypes[data.dtypes!='object'].index

nums

skws=data[nums].skew().sort\_values(ascending=False)

skewness=pd.DataFrame({'skew':skws})

skewness



#assigning the skewed values to new variables

from scipy.stats import boxcox

from scipy import stats

dft\_insulin = stats.boxcox(data['Insulin'])[0]

dft\_BMI=stats.boxcox(data['BMI'])[0]

dft\_diabetes=stats.boxcox(data['DiabetesPedigreeFunction'])[0]

data['Insulin']=dft\_insulin

data['BMI']=dft\_BMI

data['DiabetesPedigreeFunction']=dft\_diabetes

A picture containing indoor

Description automatically generated

We reduced the skewness a bit using Box-Cox transformation.

**Outliers**:

We do have some outliers even after performing Box-Cox transformation. Since all the variables represent a different diagnostic measure, we do not want to remove the outliers as we think they will be useful in interpreting some information for our research questions stated in week 1.

A screenshot of a cell phone

Description automatically generated

**Week 5 assignment**

This week the focus will be on variable selection, transformation, data reduction etc. After data cleansing in the previous week your deliverables for this week would be:

You will have to update the template document with the following:

* Describe all the steps performed for variable selection, transformation, data reduction etc.
* Discuss about your plans for data modelling *i.e.* what modelling steps will be performed in future week for addressing your research/business queries [**Example**: Perform a Multiple Linear Regression Model to determine the characteristic of the continuous response variable]

Submit your completed week 5 assignment to the dropbox in Canvas.

**PCA implementation:**

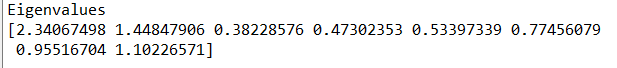
We have tried to implement PCA in order to see if we can reduce the number of independent variables.

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated



We found that there are two principal components having eigen values more than 1 and it explains about 47% variance in the data. Since it does not explain even 50% variance , We have planned to not use these components and try out implementing random forest algorithm without PCA for the final week assignment.

**Upsamling the target variable:**

**A screenshot of a cell phone

Description automatically generated**

From the plot, We can see that there is a chance of overfit for Outcome[0].

We removed this by upsampling the target variable using SMOTE.

A picture containing wall

Description automatically generated

The plot of the target variable after applying smote function:

A screenshot of a cell phone

Description automatically generated

**Future objective:**

1) In order to check query 1 stated in week 1 assignment, We will perform logistic regression with column ‘Pregnancies’ as the predictor and ‘Outcome’ as the target variable.

2) For query 2 in week 1 assignment, We will perform logistic regression with all columns as the predictor variables and ‘Outcome’ as the target variable.

3) The objective is to derive a Random-Forest Classifier model for this data set to determine the target variable (Outcome) considering all combinations of the predictor variables.

**Week 6 assignment**

This week the focus will be on data modeling. You will have to update the template document with the following:

* Describe all the steps performed for data modelling.
* Discuss about the performance measure(s) used to determine the goodness of fit for the proposed model.
* Do you suspect your initial effort (model) suffers overfitting? If so, discuss what steps you took to overcome overfitting.
* Discuss about your findings or inferences for each of the research/ business queries you identified in week 1 assignment.

Submit your completed week 6 assignment to the dropbox in Canvas.

**Logistic regression:**

In order to check query 1 stated in week 1 assignment, We tried to do logistic regression with column ‘Pregnancies’ as the predictor and ‘Outcome’ as the target variable.

A screenshot of a cell phone

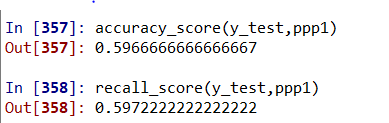
Description automatically generated

We get a score of 59%. Therefore, we can say that “pregnancy” does affect “Outcome”.

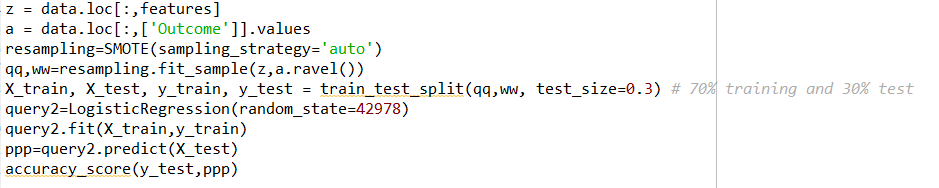
Confusion matrix:

A close up of a person

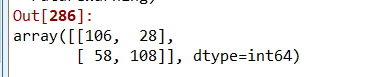
Description automatically generated

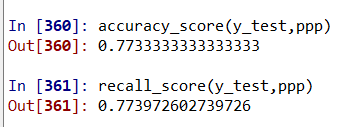


For query 2 in week 1 assignment, We tried to do logistic regression with all columns as the predictor variables and ‘Outcome’ as the target variable.



Confusion matrix:

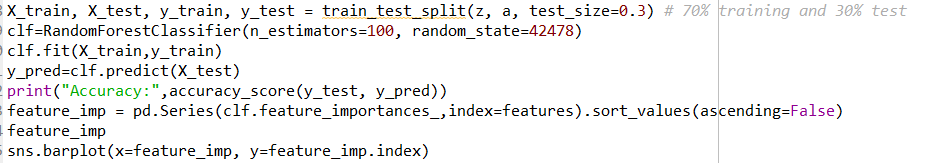




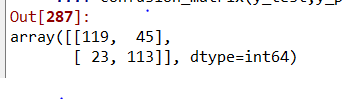
We got a score of 77% for these predictor variables.

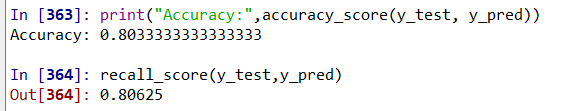
**Random Forest classifier:**

The objective is to derive a Random-Forest Classifier model for this data set to determine the target variable (Outcome) considering all combinations of the predictor variables.



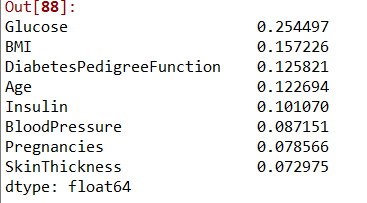
Confusion matrix:





The recall score is calculated for outcome 1 for all of the models.

The contribution of the features is shown below



A screenshot of a cell phone

Description automatically generated

We get the accuracy of 80% which is more than in logistic regression. We also get a good recall score of 81%

**Goodness of fit:**

We have calculated recall scores to validate our accuracy. Those two scores do not vary much which says us that the model is doing good.

**Overfitting removal:**

We used SMOTE function to up-sample the train data in order to remove bias for a particular class. Random forest model has seemed to improve after the resampling.

**Solution to business queries:**

1. By Logistic regression, Pregnancy do affect the outcome of diabetes and has an accuracy of 59%
2. By Logistic regression, all the features affect the outcome of diabetes and they have an accuracy of 77%
3. By Random Forest Classifier, we have built 100 random trees and estimated an accuracy of 80% with a recall score of 81%. We have also given the importance of each feature in the bar graph.

**Week 7 assignment**

**Purpose:**

To provide a demonstration of your team’s project in this course

**Tasks:**

1. Before making the final submission make sure that you have implemented all my feedbacks from week 1 to week 6.
2. Please ensure that the title page, document control section, table of contents and the reference section is updated and is up-to-date. Any appendix section should also be updated.
3. The team should prepare to demonstrate the designed data-driven Analytics System in the class (if determined by the instructor).
4. The demonstration should include a power point presentation and/or demonstration.
5. Every team will have an opportunity to go through the demonstration of the data-driven Analytics system designed by other teams.
6. The team should submit the template document, dataset(s), code/script(s) and the power point presentation deck in the box account set up by the instructor. Details about the box account will be provided in the class or by e-mail.
7. Team members should make all submissions latest by December 12 or December 13, 2019 at 11:59 PM EST.
8. More instructions (if needed) will be provided by the instructor over e-mail or in class.

Submit your completed week 7 assignment to the dropbox in Canvas.

**REFERENCES :**

<https://www.kaggle.com/uciml/pima-indians-diabetes-database/download>

<https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

<https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html>

<https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce>