



*A PROJECT REPORT ON*

*breast cancer prediction using machine learning*

***Submitted by***

**Arpita gogoi**

**Mustafiz hussain**

**Nijit Krishna hazarika**

***Under the guidance of***

**MR MD FARMANUL HAQUE**

**ACKNOWLEDGEMENT**

We hereby like to express our sincere gratitude and respect to our project guide Mr. Md Farmanul Haque for his stimulating guidance and continuous supervision, monitoring and constant encouragement throughout the project completion. The blessing, help and guidance given by him/her time to time shall us a long way in the journey of life on which we are about to embark.

We are obliged to our project team member for the valuable information provided by then in their respective fields. We are grateful for everyone’s cooperation during the period of our project assignment.

**INDEX**

Abstract

Chapter-1

* 1. Introduction.
  2. Background Of Project.

Chapter-2 (Mechanism)

* 1. Importing Libraries
  2. EDA
  3. Feature Engineering
  4. Feature scaling
  5. Logistic Regression
  6. SVM
  7. Random Forest
  8. One Hot Encoder

Chapter-3(Visualization of the data)

### 3.1. Seaborn for Python Data Visualization

3.2. What is Heatmap

Chapter-4(Evaluating Model and predictions)

4.1. Confusion Matrix

Chapter-5(Improving the model)

5.1. Improvisation part 1

5.2. Improvization part 2

Conclusion

Future Scope and Work

References

**ABSTRACT**

**Aim:** Certification of breast centers helps improve the quality of care but requires additional resources, particularly for documentation. There are currently no published data on the actual staff costs and financial resources required for such documentation. The aim of this study was to determine the time and resources required to document a patient with primary breast cancer from diagnosis to the end of follow-up, to establish a database for future strategic decisions.

**Material and Methods:** All diagnostic and therapeutic procedures of patients with primary breast cancer were recorded at the University Breast Center of Franconia. All time points for documentation were evaluated using structured interviews. The times required to document a representative number of patients were determined and combined with the staff costs of the different professional groups, to calculate the financial resources required for documentation.

**Results:** A total of 494 time points for documentation were identified. The study also identified 21 departments and 20 different professional groups involved in the documentation. The majority (54 %) of documentation was done by physicians. 62 % of all documentation involved outpatients. The results of different scenarios for the diagnosis, therapy and follow-up of breast cancer patients in a certified breast center showed that the time required for documentation can be as much as 105 hours, costing € 4135.

**Conclusion:** This analysis shows the substantial staffing and financial costs required for documentation in certified centers. A multi-center study will be carried out to compare the costs for certified breast centers of varying sizes with the costs of non-certified care facilities.

* 1. INTRODUCTION

According to information from the Federal Statistical Office of Germany, breast cancer is the most common cause of death from malignant disease in women in Germany. According to figures of the Robert Koch Institute for 2010, 70 340 women develop breast cancer every year in Germany. The projection for 2014 is 75 200 new cases for that year [1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4153817/#R826-1). Allowing for an annual mortality of 17 466 women and taking the 10-year follow-up into account, more than half a million women in Germany are currently receiving care and being followed up. An estimated 60 000 women additionally have metastases. To put this into context, approximately 32 033 000 adult women were living in Germany in 2010 [2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4153817/#R826-2), meaning that one in 55 women is receiving oncological care or follow-up for breast cancer. These figures show that documentation and quality assurance for breast cancer patients are extremely relevant for healthcare policies and health economics. Documentation and quality assurance are increasingly central for health economics, both nationally and internationally, and it is important that further studies on these topics are carried out [3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4153817/#R826-3). German healthcare policies have also recognized the importance of obtaining more detailed information and have begun to focus more on this aspect. On June 16, 2008, the German Federal Ministry of Health (BMG) launched the National Cancer Plan [4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4153817/#R826-4), [5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4153817/#R826-5). The goal of the National Cancer Plan is to optimize the care of cancer patients [6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4153817/#R826-6). Key aspects include improving the provision of oncology services, ensuring guideline-based care and improving quality assurance while ensuring that documentation is both effective and cost-efficient. Cancer is a highly complex disease and it is indisputable that documentation plays an important role during the course of disease and treatment [7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4153817/#R826-7). Moreover, more than for any other disease, cancer treatment involves numerous medical specialties and professional groups from many different care services, and the complexity of the disease means that patients may have to be followed up for the rest of their lives. A reliable and neutral representation of important stations during the course of disease, starting from diagnosis and including treatment and follow-up, is therefore indispensable to optimize care [7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4153817/#R826-7). Tumor documentation is not only necessary when formulating the complex treatment process. In addition to healthcare providers and patients, healthcare insurance companies, researchers and politicians also depend on tumor documentation as a source of reliable information on the quality of oncological care in Germany. However, there are currently a number of problems associated with the quality indicators used to document care for quality assurance purposes. Numerous quality indicators are used to describe the quality of structures, processes and outcomes. But the same indicators are defined differently in different documentation systems. At present, comparisons using different documentation systems are usually not possible (for example, the number of patients with primary breast cancer in a single center: some systems count individual patients, others list each breast separately). There are also many different quality assurance systems in use. Examples for this include the documentation required by the German Cancer Society (DKG) and the German Society for Senology (DGS) from certified breast centers, the data collected by the AQUA Institute, the data required to participate in the Disease Management Program (DMP) Breast Cancer, the data collected for quality assurance purposes in breast cancer screening, etc. The existing documentation systems have very different objectives and describe different aspects of care provision, and data are processed using different systems for data acquisition and reporting.

Certified breast centers provide tertiary care to breast cancer patients and stand in particular need of additional personnel and financial resources to comply with the requirements to document quality parameters.

The tumor documentation of breast cancer patients is thus a prime example of how developments can go wrong. Quality assurance in oncology is currently very heterogeneous in Germany, and clear and uniform guidelines are lacking. Multiple data collections, different documentation systems and partial data collection should be unacceptable; data documentation should always be defined in terms of the goal of the data collection, for example, to certify the quality of services, reinforce compliance with guidelines, improve the process quality and – the most important aspect for patients – improve outcomes. Changes in documentation could reduce the number of staff required and save costs.

But in order to achieve these goals it is necessary to take stock and review the **situation as it currently stands**, including:

* At which points in the patientʼs course (from the initial diagnosis to treatment, follow-up and quality assurance) are data generally collected?
* Which parameters are documented?
* Who (professional group/medical specialty) does the documenting?
* Which financial resources are currently required for a complete documentation of a breast cancer patient?

To calculate the resources actually needed for documentation, identify which data is collected unnecessarily several times over, and establish where interfaces exist which could be used to optimize data collection across all areas, it is necessary to capture and describe the current status of data collection. The obvious gap between currently available resources for data collection and what would actually be needed has not been previously investigated. This is surprising, considering the costs and time spent on documentation. The aim of this study was to provide a detailed horizontal cross-section of the time, costs and staff involved in documenting a patient with primary breast cancer from initial diagnosis and including treatment, follow-up and quality assurance, and to determine the different time points for data collection and the collected parameters.

* 1. BACKGROUND OF PROJECT

Breast cancer is the most common non-skin cancer among American women. An estimated 271,270 new cases of breast cancer will be diagnosed in women in the United States in 2019, according to the American Cancer Society. Breast cancer accounts for 15 percent of all new cancer diagnoses and 7 percent of all cancer deaths each year.

### What causes breast cancer?

Breasts are made of a variety of different tissues, including ducts, lobes and lobules and glands that produce milk and carry it to the nipple. The breasts also contain lymph nodes and fatty tissue. Cancer develops when cells in the breast mutate and grow out of control, forming a tumor. Most breast cancers—about 80 percent—are ductal carcinomas, which begin in milk ducts. About 10 percent of all breast cancers are lobular carcinomas, which develop in the lobes or glands that produce milk.

Other factors that may increase a woman’s risk for developing breast cancer include:

* Obesity
* Breast density
* Menstrual history
* A sedentary lifestyle
* Heavy drinking
* Previous medical treatments

### Who gets breast cancer?

The risk for developing breast cancer increases with age. According to the National Cancer Institute:

* The average age of a woman diagnosed with breast cancer is 62.
* The average age of a woman who dies from breast cancer is 68.
* Breast cancer is the most common cancer diagnosed in women between age 55 and 64.
* About 10 percent of breast cancers occur in women younger than 45.

Women with a family history of breast cancer may be at a higher risk for developing the disease. For example:

* Women whose mother, sister or daughter has or had breast cancer may have double the risk.
* Women who have inherited mutations in the BRCA1 or BRCA2 gene are at higher risk.

Breast cancer also occurs in men, but is very rare. Approximately 2,670 American men will learn they have breast cancer in 2019, the American Cancer Society estimates. Male breast cancer accounts for 1 percent of all breast cancer diagnoses.

### Types of breast cancer

There are many types and subtypes of breast cancer, but most are [adenocarcinomas](https://www.cancercenter.com/adenocarcinoma) of the breast. Adenocarcinoma tumors are found in many common cancers, including prostate, lung and colorectal. These types of tumors form in glands or ducts that secrete fluid. Breast adenocarcinomas form in milk ducts or milk-producing glands called lobules. Each type of breast cancer may be determined based on where in the breast it develops, whether it is considered invasive or non-invasive and whether it is driven by hormones or proteins. Types of breast cancer include:

* Adenocystic carcinoma
* Angiosarcoma
* Ductal carcinoma
* Inflammatory breast cancer
* Lobular carcinoma
* Metaplastic carcinoma
* Phyllodes tumor

Subtypes of breast cancer include those driven by specific hormones, such as estrogen, progestogen or the protein HER2. Sixty percent of breast cancers are estrogen-positive. Twenty percent of breast cancers are HER2-positive. Another 20 percent are triple-negative breast cancers, a type of breast cancer that tests negative for estrogen, progesterone and HER2. Triple-negative breast cancer is among the more aggressive forms of the disease.

Breast cancer cells can spread into the lymph nodes in and around the breasts and, from there, travel and form tumors in distant parts of the body. When that occurs, it is called metastatic breast cancer. When it spreads, breast cancer is most often found in the brain, bones, liver and lungs. It is still considered breast cancer even if it is found on other parts of the body.

### Breast cancer symptoms

A lump, mass and change in the feel or position of the breast are among the most common symptoms of breast cancer. Other symptoms include:

* Swelling, redness or inflammation
* Changes in the nipple
* Nipple discharge
* Pain in the breast
* Itchy or irritated breasts
* Changes in color
* Peeling or flaky skin

### Diagnosing breast cancer

Tools and tests used to diagnose breast cancer include:

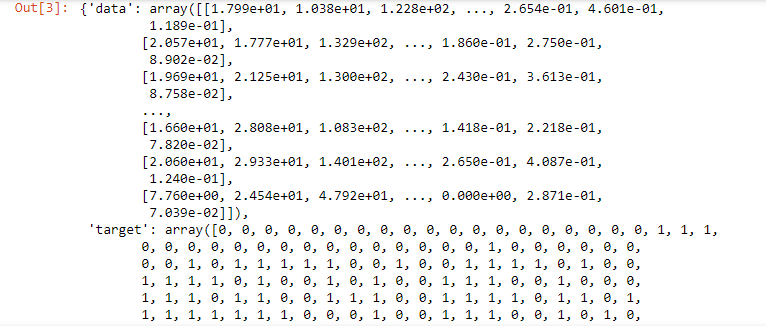
* Lab tests, including advanced genomic testing
* Biopsy
* Imaging tests, including ultrasound and mammography

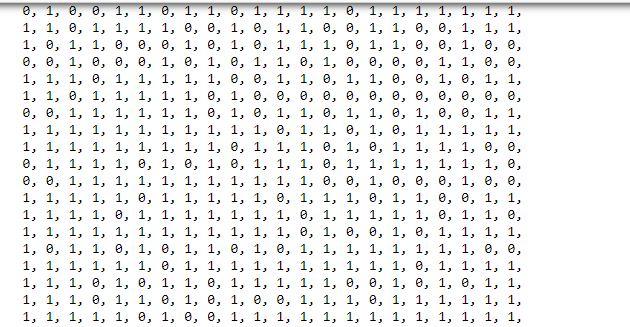
Different tests are used to determine whether the breast cancer has metastasized. These tests include:

* Radiofrequency ablation
* Endobronchial ultrasound
* Bone scan

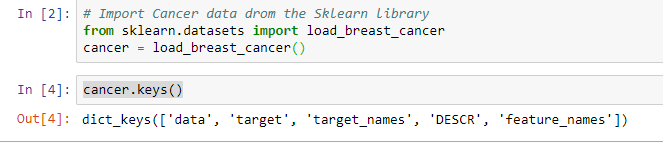
MECHANISM:

* 1. IMPORTING LIBRARIES AND FUNCTIONS:
     1. **NumPy** is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very fundamental scientific computations in machine learning. It is particularly useful for linear algebra , Fourier transform , and random number capabilities. High-end libraries like TensorFlow uses *NumPy* internally for manipulation of Tensors. *np* is an object of NumPy library.
     2. **Pandas**is a popular python library for data analysis. It is not directly related to Machine Learning. It is an opensource library that allows to perform data manipulation in python. It is used for data retrivals of csv files and datafiles. It provides an easy way to create , manipulate, and wrangle the data. It is designed to make working with ‘relational’ or ‘labeled’ data both easy and intuitive . pd is an object of pandas library .
     3. **Read.csv()**is a function belongs to pandas’s library and it takes the file name as parameters.

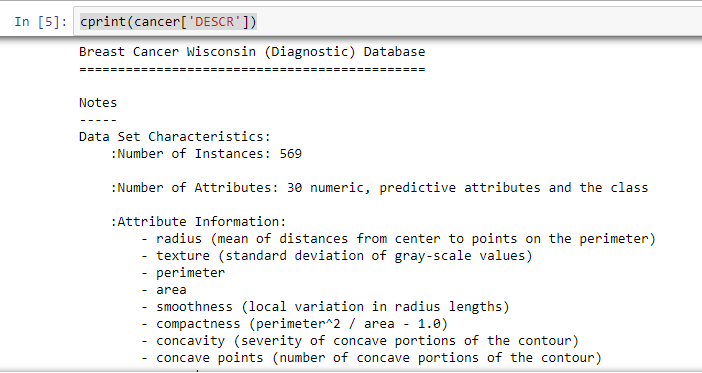


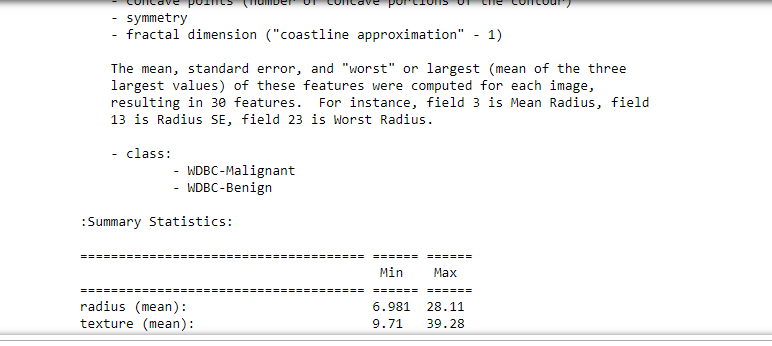


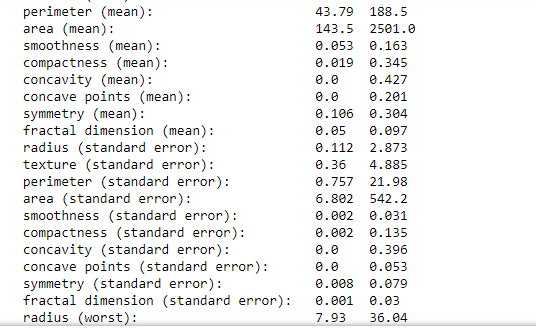
* + 1. [**Scikit-learn**](http://scikit-learn.org/stable/) is a library in Python that provides many unsupervised and supervised learning algorithms. It’s built upon some of the technology you might already be familiar with, like NumPy, pandas, and Matplotlib!
    2. **Seaborn library** is a data visualization library based on matplotlib. It provides a highlevel interfaces for drawing attractive statistical graphics. Plots are: DISTRIBUTION PLOT, REGRESSION PLOT, CATEGORICAL PLOT, MATRIX PLOT, TIME SERIES PLOT .

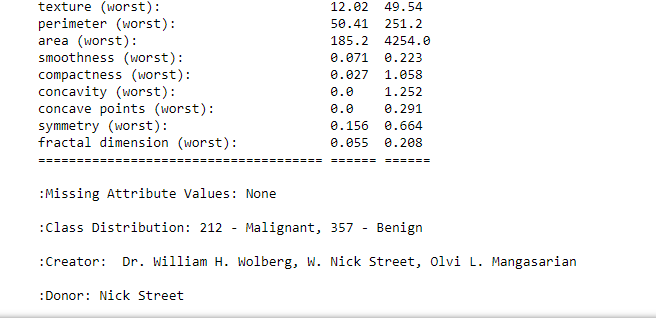


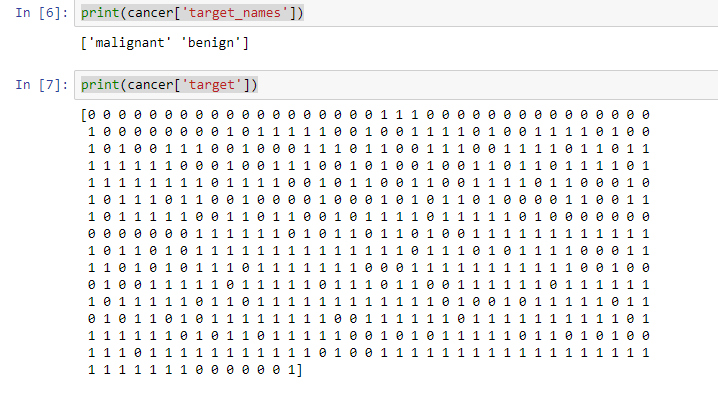
* + 1. **cprint** is a minimalist python library which gives you the possibility to print in color.











* + 1. Predicting if the cancer diagnosis is benign or malignant based on several observations/features

30 features are used, examples:

- radius (mean of distances from center to points on the perimeter)

- texture (standard deviation of gray-scale values)

- perimeter

- area

- smoothness (local variation in radius lengths)

- compactness (perimeter^2 / area - 1.0)

- concavity (severity of concave portions of the contour)

- concave points (number of concave portions of the contour)

- symmetry

fractal dimension ("coastline approximation" - 1)

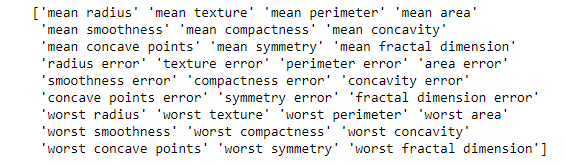
Datasets are linearly separable using all 30 input features

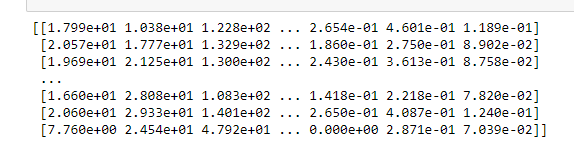
Number of Instances: 569

Class Distribution: 212 Malignant, 357 Benign

Target class:

Malignant and Benign





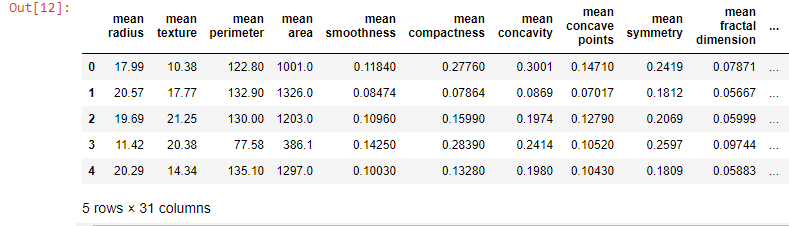
* + 1. **.Shape**

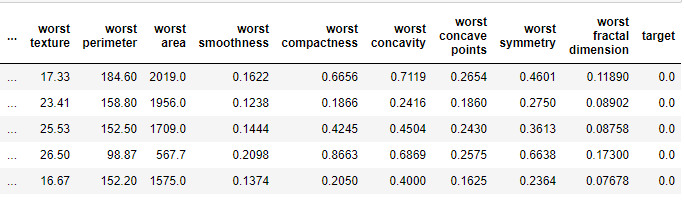
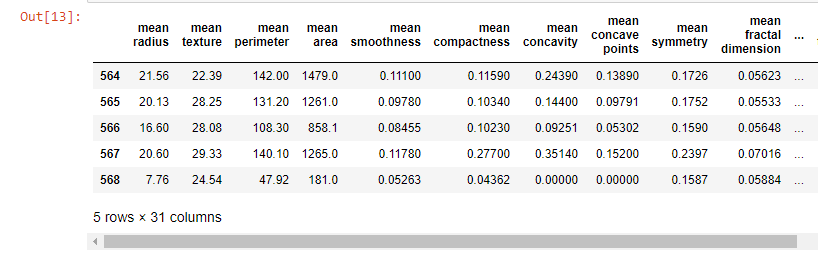
The shape command gives information on the data set size – ‘shape’ returns a tuple with the number of rows, and the number of columns for the data in the DataFrame.

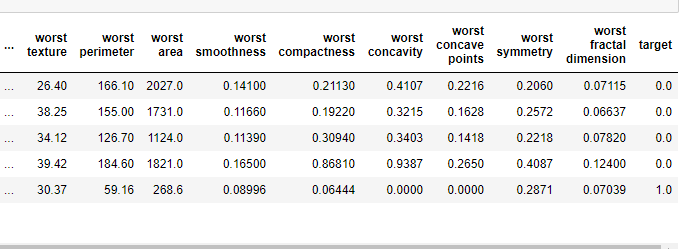
18.PNG

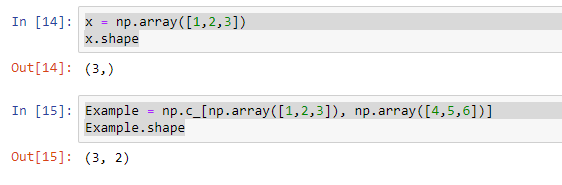
* + 1. **head() and tail()**

The DataFrame.head() function in Pandas, by default, shows you the top 5 rows of data in the DataFrame. The opposite is DataFrame.tail(), which gives you the last 5 rows. The first 5 rows of a DataFrame are shown by head(), the final 5 rows by tail(). For other numbers of rows – simply specify how many you want.







* 1. WHAT IS EDA ?

EDA(Exploratory Data Analysis ) is essentially a type of storytelling for statisticians . It allows to uncover patterns and insights , often with visual methods , within data . EDA is often the first step of the data modelling process. It is used for gaining a better understanding of data aspects like: - main features of data, -variables and relationships that holds between them, - identifying which variables are important . Various exploratory data analysis methods like:

Descriptive Statistics which is a way of a giving a brief overview of the dataset which is dealing with, including some measure and features of the sample .Pandas provide describe() applies basic statistical computations on the dataset like extreme values , count of data , point standard deviation etc. Any missing value or NaN value is automatically skipped. Describe() function gives a group picture of distribution of data.

Grouping data an interesting measure available in pandas which figure out effect of different categorical attributes on other data variables.

Anova stands for Analysis of Variance. It is performed to figure out the relation between the different group of categorical data.

Under Anova have two measures as result:

-F-testscore: which shows the variation of groups mean over variation.

-p-value: it shows the importance of the result. This can be performed using Python module scipy methodname*f-oneway().*

Correlation and Correlation computation is a simple relationship between two variables in a context such that one variable affects the other. Correlation is different from act of causing. One way to calculate correlation among variables is to find Pearson correlation . Here two parameters namely, Pearson coefficient and p-value . There is a strong correlation between two variables when Pearson coefficient is close to either 1 or -1 and the p-value is less than 0.0001. Scipy module also provides a method to perform Pearson correlation analysis.

* 1. FEATURE ENGINEERING:

Feature Engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work . Feature engineering is fundamental to the machine learning and is both difficult and expensive . The feature engineering process is

1) Brainstorming or testing feature ;

2) Deciding what feature to create;

3) Creating features;

4) Checking how the features work with the model;

5) Improving features if needed;

6) Go back to brainstorming / creating more features until the work is done .

7) Preparing the proper input dataset , compatible with machine learning algorithm requirements;

8) Improving the performance of machine learning .

This metric is very impressive to show the importance of feature engineering in datascience .

Techniques are listed as: IMPUTATION , HANDLING OUTLIERS , BINNING, LOG TRANSFORM, ONE HOT ENCODER, GROUPING OPERATIONS, FEATURE SPLIT, SCALING, EXTRACTING DATE.

In machine learning and pattern recognition , a feature is an individual measurable property or characteristic of a phenomenon being observed . Choosing informative , discriminating and independent features is a crucial step for effective algorithms in pattern recognition , classification and regression.While feature engineering requires label times, in [our general-purpose framework](https://towardsdatascience.com/how-to-create-value-with-machine-learning-eb09585b332e), it is not hard-coded for specific labels corresponding to only one prediction problem. If we wrote our feature engineering code for a single problem — as feature engineering is traditionally approached — then we would have to redo this laborious step every time the parameters change.

Instead, we use APIs like [Featuretools](https://github.com/Featuretools/featuretools) that can build features for any set of labels without requiring changes to the code. This means for the customer churn dataset, we can solve multiple prediction problems — predicting churn every month, every other week, or with a lead time of two rather than one month — using the exact same feature engineering code.

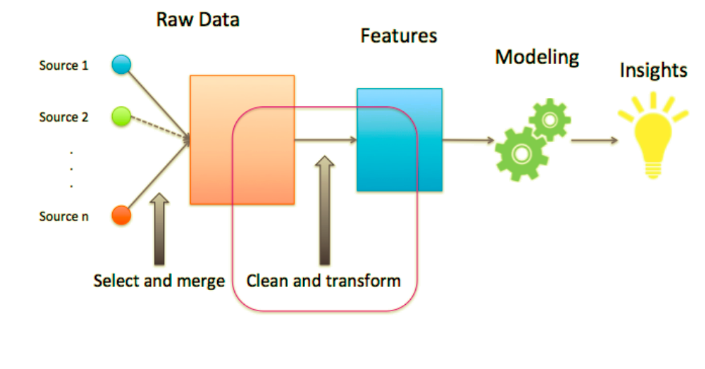


Fig: the workflow of feature engineering

Feature engineering, the second step in the [machine learning pipeline](https://towardsdatascience.com/how-to-create-value-with-machine-learning-eb09585b332e), takes in the [label times from the first step](https://towardsdatascience.com/prediction-engineering-how-to-set-up-your-machine-learning-problem-b3b8f622683b) — prediction engineering — and a raw dataset that needs to be refined. Feature engineering means building features for each label while filtering the data used for the feature based on the label’s cutoff time to make valid features. These features and labels are then passed to modeling where they will be used for training a machine learning algorithm.

* 1. FEATURE SCALING:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range . It is present in the data in a fixed range . It is performed during the data pre-processing to handle highly varying magnitudes or values or units . If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values , regardless of the unit of the values.

Techniques are:

1) MIN-MAX NORMALIZATION: This techniques re-scales a feature or observation Value with distribution value between 0 and 1.

2) STANDARDIZATION: It is a very effective technique which re-scales a feature value so that it has distribution with 0 and mean value and variance equals to 1 . Prediction of the class of new data point : The model calculates the distance of this datapoint from centroid of each class group. Finally this data point will belong to that class , which will have a minimum centroid distance from it .

The distance can be calculated between centroid and data point using these methods:

3) EUCLIDEAN DISTANCE: It is the square -root of the sum of squares of differences between the coordinates of datapoint and centroid of each class . The **Euclidean distance** between points **p** and **q** is the length of the [line segment](https://en.wikipedia.org/wiki/Line_segment) connecting them (p q ¯ {\displaystyle {\overline {\mathbf {p} \mathbf {q} }}} ).

In [Cartesian coordinates](https://en.wikipedia.org/wiki/Cartesian_coordinates), if **p** = (*p*1, *p*2,..., *pn*) and **q** = (*q*1, *q*2,..., *qn*) are two points in [Euclidean *n*-space](https://en.wikipedia.org/wiki/Euclidean_space), then the distance (d) from **p** to **q**, or from **q**to **p** is given by the [Pythagorean formula](https://en.wikipedia.org/wiki/Pythagorean_theorem):[[1]](https://en.wikipedia.org/wiki/Euclidean_distance#cite_note-Anton-1)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| d ( p , q ) = d ( q , p ) = ( q 1 − p 1 ) 2 + ( q 2 − p 2 ) 2 + ⋯ + ( q n − p n ) 2 = ∑ i = 1 n ( q i − p i ) 2 . {\displaystyle {\begin{aligned}d(\mathbf {p} ,\mathbf {q} )=d(\mathbf {q} ,\mathbf {p} )&={\sqrt {(q\_{1}-p\_{1})^{2}+(q\_{2}-p\_{2})^{2}+\cdots +(q\_{n}-p\_{n})^{2}}}\\[8pt]&={\sqrt {\sum \_{i=1}^{n}(q\_{i}-p\_{i})^{2}}}.\end{aligned}}} | |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  | |  |  | |  |

The position of a point in a Euclidean *n*-space is a [Euclidean vector](https://en.wikipedia.org/wiki/Euclidean_vector). So, **p** and **q** may be represented as Euclidean vectors, starting from the origin of the space (initial point) with their tips (terminal points) ending at the two points. The [**Euclidean norm**](https://en.wikipedia.org/wiki/Euclidean_norm), or **Euclidean length**, or **magnitude** of a vector measures the length of the vector:[[1]](https://en.wikipedia.org/wiki/Euclidean_distance#cite_note-Anton-1)

‖ p ‖ = p 1 2 + p 2 2 + ⋯ + p n 2 = p ⋅ p , {\displaystyle \left\|\mathbf {p} \right\|={\sqrt {p\_{1}^{2}+p\_{2}^{2}+\cdots +p\_{n}^{2}}}={\sqrt {\mathbf {p} \cdot \mathbf {p} }},} 

where the last expression involves the [dot product](https://en.wikipedia.org/wiki/Dot_product).

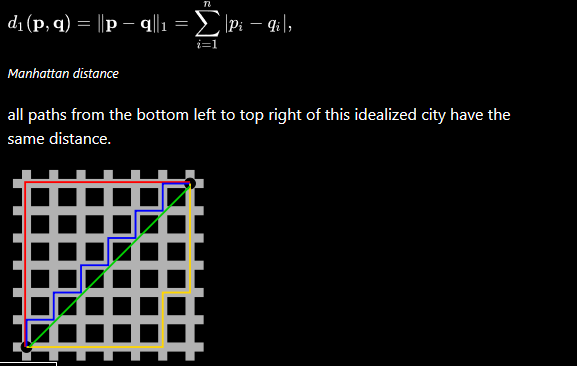
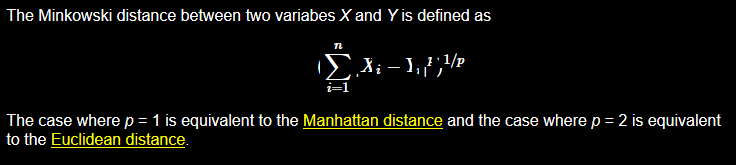
4) MANHATTAN DISTANCE: It is calculated as the sum of absolute differences between the coordinates of data point and centroid of each class. The distance between two points measured along axes at right angles.The Manhattan distance between two vectors (or points) a and b is defined as ∑i|ai−bi| over the dimensions of the vectors.

Fig: showing the Manhattan distance between the two points.

5) MINKOWSKI DISTANCE: It is a generalization of two above methods .

Although *p* can be any real value, it is typically set to a value between 1 and 2. For values of *p* less than 1, the formula above does not define a valid distance metric since the triange inequality is not satisfied.

* 1. LOGISTIC REGRESSION:

Logistic Regression is a classification algorithm used to assign observations to a discrete set of classes . Logistic Regression transforms its output using the logistic sigmoid function to return a probability value . There are two types are: BINARY AND MUILTI – LINEAR FUNCTION FAILSCLASS LOGISTIC REGRESSION . It is used for the classification problems , it is a predictive analysis algorithm and based on the concept of probability .The hypothesis of logistic regression tends it to limit the cost function between o and 1 . Therefore linear function fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression . We can call a Logistic Regression a linear Regression model but the logistic regression uses a more complex cost function can be defined as the ‘*SIGMOID FUNCTION’* or also known as the ‘logistic function’ instead of a linear function .It is used to map predictions to probabilities . The func. maps any real value into another value between 0 and 1 .t is a technique to analyse a data-set which has a dependent variable and one or more independent variables to predict the outcome in a binary variable, meaning it will have only two outcomes.

The dependent variable is **categorical** in nature. Dependent variable is also referred as **target variable** and the independent variables are called the **predictors**.

Logistic regression is a special case of linear regression where we only predict the outcome in a categorical variable. It predicts the probability of the event using the log function.

We use the **Sigmoid function/curve** to predict the categorical value. The threshold value decides the outcome(win/lose).

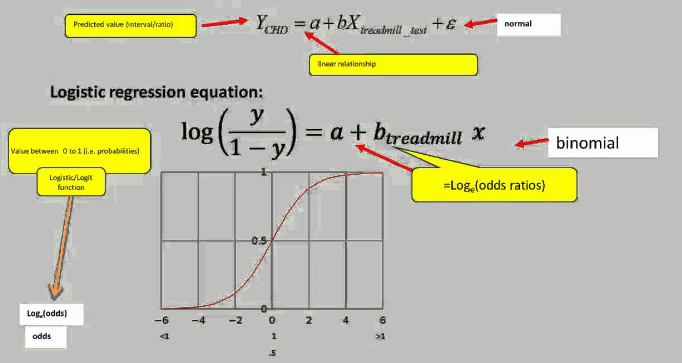
Linear regression equation:    **y = β0 + β1X1 + β2X2 …. + βnXn**

* Y stands for the dependent variable that needs to be predicted.
* β0 is the Y-intercept, which is basically the point on the line which touches the y-axis.
* β1 is the slope of the line (the slope can be negative or positive depending on the relationship between the dependent variable and the independent variable.)
* X here represents the independent variable that is used to predict our resultant dependent value.

Sigmoid function:***p =*** **1 / 1 + e-y**

Logistic Regression equation:  ***p =*** **1 / 1 + e-(β0 + β1X1 + β2X2 …. + βnXn)**

**Logisticregression** measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a **logistic** function. It uses a black box function to understand the relation between the categorical dependent variable and the independent variables.Assumptions of Linear Regression. Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions: **Linear relationship. Multivariate normality. No or little multicollinearity**. No auto-correlation. Homoscedasticity.

**Logisticregression** is **considered** a generalized **linearmodel** because the outcome always depends on the sum of the inputs and parameters. Or in other words, the output cannot depend on the product (or quotient, etc.) ... "A statistician calls a **model** "**linear**" if the mean of the response is a **linear** function of the parameter, and this is clearly violated for **logisticregression**. Consider a model with two predictors, x 1 {\displaystyle x\_{1}} and x 2 {\displaystyle x\_{2}} , and one binary (Bernoulli) response variable Y {\displaystyle Y} , which we denote p = P ( Y = 1 ) {\displaystyle p=P(Y=1)} . We assume a [linear relationship](https://en.wikipedia.org/wiki/Linear_model) between the predictor variables, and the [log-odds](https://en.wikipedia.org/wiki/Log-odds) of the event that Y = 1 {\displaystyle Y=1} . This linear relationship can be written in the following mathematical form (where *ℓ* is the log-odds, b {\displaystyle b} is the base of the logarithm, and β i {\displaystyle \beta \_{i}} are parameters of the model):

ℓ = log b ⁡ p 1 − p = β 0 + β 1 x 1 + β 2 x 2 {\displaystyle \ell =\log \_{b}{\frac {p}{1-p}}=\beta \_{0}+\beta \_{1}x\_{1}+\beta \_{2}x\_{2}}

We can recover the [odds](https://en.wikipedia.org/wiki/Odds) by exponentiating the log-odds:

p 1 − p = b β 0 + β 1 x 1 + β 2 x 2 {\displaystyle {\frac {p}{1-p}}=b^{\beta \_{0}+\beta \_{1}x\_{1}+\beta \_{2}x\_{2}}} .

By simple algebraic manipulation, the probability that Y = 1 {\displaystyle Y=1} is

p = b β 0 + β 1 x 1 + β 2 x 2 b β 0 + β 1 x 1 + β 2 x 2 + 1 = 1 1 + b − ( β 0 + β 1 x 1 + β 2 x 2 ) {\displaystyle p={\frac {b^{\beta \_{0}+\beta \_{1}x\_{1}+\beta \_{2}x\_{2}}}{b^{\beta \_{0}+\beta \_{1}x\_{1}+\beta \_{2}x\_{2}}+1}}={\frac {1}{1+b^{-(\beta \_{0}+\beta \_{1}x\_{1}+\beta \_{2}x\_{2})}}}} .

The above formula shows that once β i {\displaystyle \beta \_{i}} are fixed, we can easily compute either the log-odds that Y = 1 {\displaystyle Y=1} for a given observation, or the probability that Y = 1 {\displaystyle Y=1} for a given observation. The main use-case of a logistic model is to be given an observation ( x 1 , x 2 ) {\displaystyle (x\_{1},x\_{2})} , and estimate the probability p {\displaystyle p} that Y = 1 {\displaystyle Y=1} . In most applications, the base b {\displaystyle b} of the logarithm is usually taken to be [*e*](https://en.wikipedia.org/wiki/E_(mathematical_constant)). However in some cases it can be easier to communicate results by working in base 2, or base 10.

We consider an example with b = 10 {\displaystyle b=10} , and coefficients β 0 = − 3 {\displaystyle \beta \_{0}=-3} , β 1 = 1 {\displaystyle \beta \_{1}=1} , and β 2 = 2 {\displaystyle \beta \_{2}=2} . To be concrete, the model is

log 10 ⁡ p 1 − p = ℓ = − 3 + x 1 + 2 x 2 {\displaystyle \log \_{10}{\frac {p}{1-p}}=\ell =-3+x\_{1}+2x\_{2}}

where p {\displaystyle p} is the probability of the event that Y = 1 {\displaystyle Y=1} .Y =1 .

This can be interpreted as follows:

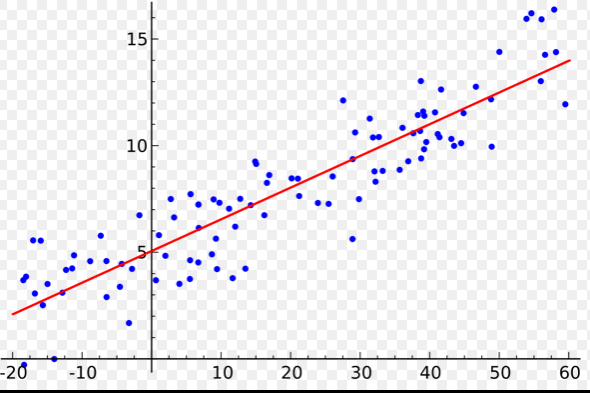
* β 0 = − 3 {\displaystyle \beta \_{0}=-3} is the [*y*-intercept](https://en.wikipedia.org/wiki/Y-intercept). It is the log-odds of the event that Y = 1 {\displaystyle Y=1} , when the predictors x 1 = x 2 = 0 {\displaystyle x\_{1}=x\_{2}=0} . By exponentiating, we can see that when x 1 = x 2 = 0 {\displaystyle x\_{1}=x\_{2}=0} the odds of the event that Y = 1 {\displaystyle Y=1} are 1-to-1000, or 10 − 3 {\displaystyle 10^{-3}} . Similarly, the probability of the event that Y = 1 {\displaystyle Y=1} when x 1 = x 2 = 0 {\displaystyle x\_{1}=x\_{2}=0} can be computed as 1 / ( 1000 + 1 ) = 1 / 1001 {\displaystyle 1/(1000+1)=1/1001} .
* β 1 = 1 {\displaystyle \beta \_{1}=1} means that increasing x 1 {\displaystyle x\_{1}} by 1 increases the log-odds by 1 {\displaystyle 1} . So if x 1 {\displaystyle x\_{1}} increases by 1, the odds that Y = 1 {\displaystyle Y=1} increase by a factor of 10 1 {\displaystyle 10^{1}} .
* β 2 = 2 {\displaystyle \beta \_{2}=2} means that increasing x 2 {\displaystyle x\_{2}} by 1 increases the log-odds by 2 {\displaystyle 2} . So if x 2 {\displaystyle x\_{2}} increases by 1, the odds that Y = 1 {\displaystyle Y=1} increase by a factor of 10 2 . {\displaystyle 10^{2}.} Note how the effect of x 2 {\displaystyle x\_{2}} on the log-odds is twice as great as the effect of x 1 {\displaystyle x\_{1}} , but the effect on the odds is 10 times greater.

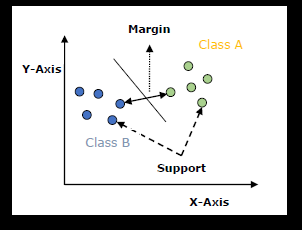
Logistic regression can be seen as a special case of **generalized linear model** and thus analogous to **linear regression**. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of **linear regression**.

**Multivariatelogisticregression** is like simple **logisticregression** but with multiple predictors. **Logisticregression** is similar to linear **regression** but you can use it when your response variable is binary. As in linear regression let’s represent our hypothesis(Prediction Of Dependent Variable) in classification. In classification our hypothesis representation which tries to predict the binary outcome of either o or 1, will look like,

**hθ(x) = g(θ T x) = 1/ 1 + e −θ T x ,**

Here g(z) = 1/( 1 + e ^−z), is called the l**ogistic function or the sigmoid function.**

 here dots are scatterplot suggest the form and strength of the relationship between the dependent variable and regressors.

* 1. SVM:
* SVM(Support Vector Machine) is a supervised machine learning algorithm which can be used for both classification or regression challenges . In algorithm we plot each data item as point in n-dimensional space ( where n is number of features) with the value of each feature being the value of a particular coordinate . Then, we perform classification by finding the *hyper-plane* that differentiate the two classes . **SVM** is a supervised machine learning algorithm which can be **used** for classification or **regression** problems. It **uses** a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH). **Support Vectors** − Datapoints that are closest to the hyperplane is called support vectors. Separating line will be defined with the help of these data points.
* **Hyperplane** − As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.
* **Margin** − It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin. 

### SVM algorithm is implemented with kernel that transforms an input data space into the required form. SVM uses a technique called the kernel trick in which kernel takes a low dimensional input space and transforms it into a higher dimensional space. Linear Kernel

It can be used as a dot product between any two observations. The formula of linear kernel is as below −

*K*(*x*,*xi*)=*sum*(*x*∗*xi*)K(x,xi)=sum(x∗xi)

From the above formula, we can see that the product between two vectors say 𝑥&𝑥𝑖 is the sum of the multiplication of each pair of input values.

### *Polynomial Kernel*

It is more generalized form of linear kernel and distinguish curved or nonlinear input space. Following is the formula for polynomial kernel −

*k*(*X*,*Xi*)=1+*sum*(*X*∗*Xi*)^*d*k(X,Xi)=1+sum(X∗Xi)^d

Here d is the degree of polynomial, which we need to specify manually in the learning algorithm.

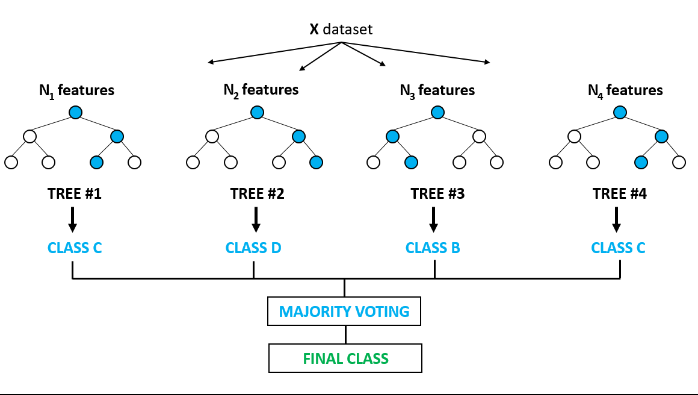
### *Radial Basis Function (RBF) Kernel*

RBF kernel, mostly used in SVM classification, maps input space in indefinite dimensional space. Following formula explains it mathematically −

*K*(*x*,*xi*)=*exp*(−*gamma*∗*sum*(*x*−*xi*^2))K(x,xi)=exp(−gamma∗sum(x−xi^2))

Here, *gamma* ranges from 0 to 1. We need to manually specify it in the learning algorithm. A good default value of *gamma* is 0.1.

* 1. RANDOM FOREST

Random forests or random decision forests are an **ensemble learning method for classification, regression** and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Random Forests are an improvement over bagged decision trees.

A problem with decision trees like CART is that they are greedy. They choose which variable to split on using a greedy algorithm that minimizes error. As such, even with Bagging, the decision trees can have a lot of structural similarities and in turn have [high correlation](https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/) in their predictions.

Combining predictions from multiple models in ensembles works better if the predictions from the sub-models are uncorrelated or at best weakly correlated.

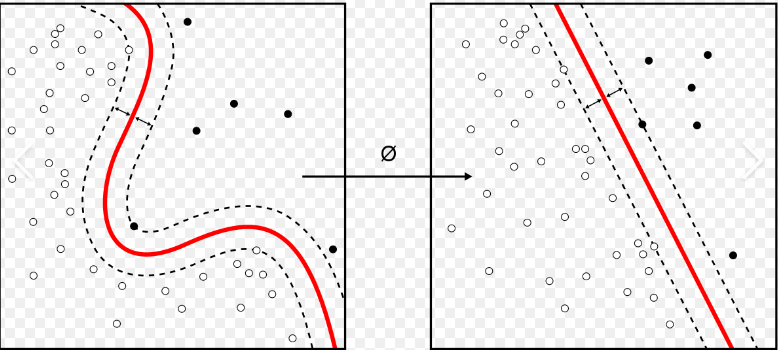
Random forest changes the algorithm for the way that the sub-trees are learned so that the resulting predictions from all of the subtrees have less correlation.

It is a simple tweak. In CART, when selecting a split point, the learning algorithm is allowed to look through all variables and all variable values in order to select the most optimal split-point. The random forest algorithm changes this procedure so that the learning algorithm is limited to a random sample of features of which to search.

* The number of features that can be searched at each split point (m) must be specified as a parameter to the algorithm. You can try different values and tune it using cross validation.For classification a good default is: m = sqrt(p)
* For regression a good default is: m = p/3

Where m is the number of randomly selected features that can be searched at a split point and p is the number of input variables. For example, if a dataset had 25 input variables for a classification problem, then:

* m = sqrt(25)
* m = 5
* simple mathematical representation.
* Random forest works by building decision trees & then aggregating them & hence the Beta values have no counterpart in random forest. Though you do get the 'Variable Importance /Gini Index' values for the forest, which can be used for making sense of the model but not as a multiplication factor.
* Now to answer your question, from your code snippet it seems that you have stored you model as rf. To use this model for prediction, you can simply call the predict method in python associated with the random forest class.
* use:
* prediction = rf.predict(test)
* This will give you the predictions for you new data (test here) based on the model rf. The predict method won't build a new model, it'll use the model rf to use for prediction on new data.



**classification** is the problem of identifying to which of a set of [categories](https://en.wikipedia.org/wiki/Categorical_data) (sub-populations) a new [observation](https://en.wikipedia.org/wiki/Observation) belongs, on the basis of a [training set](https://en.wikipedia.org/wiki/Training_set) of data containing observations (or instances) whose category membership is known. Examples are assigning a given email to the ["spam" or "non-spam"](https://en.wikipedia.org/wiki/Spam_filtering) class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition).

In the terminology of machine learning,[[1]](https://en.wikipedia.org/wiki/Statistical_classification#cite_note-1) classification is considered an instance of [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning), i.e., learning where a training set of correctly identified observations is available. The corresponding [unsupervised](https://en.wikipedia.org/wiki/Unsupervised_learning) procedure is known as [clustering](https://en.wikipedia.org/wiki/Cluster_analysis), and involves grouping data into categories based on some measure of inherent similarity or [distance](https://en.wikipedia.org/wiki/Distance).

Often, the individual observations are analyzed into a set of quantifiable properties, known variously as [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variables) or *features*. These properties may variously be [categorical](https://en.wikipedia.org/wiki/Categorical_data) (e.g. "A", "B", "AB" or "O", for [blood type](https://en.wikipedia.org/wiki/Blood_type)), [ordinal](https://en.wikipedia.org/wiki/Ordinal_data) (e.g. "large", "medium" or "small"), [integer-valued](https://en.wikipedia.org/wiki/Integer) (e.g. the number of occurrences of a particular word in an [email](https://en.wikipedia.org/wiki/Email)) or [real-valued](https://en.wikipedia.org/wiki/Real_number) (e.g. a measurement of [blood pressure](https://en.wikipedia.org/wiki/Blood_pressure)). Other classifiers work by comparing observations to previous observations by means of a [similarity](https://en.wikipedia.org/wiki/Similarity_function) or [distance](https://en.wikipedia.org/wiki/Metric_(mathematics)) function.

An [algorithm](https://en.wikipedia.org/wiki/Algorithm) that implements classification, especially in a concrete implementation, is known as a **classifier**. The term "classifier" sometimes also refers to the mathematical [function](https://en.wikipedia.org/wiki/Function_(mathematics)), implemented by a classification algorithm, that maps input data to a category.

Terminology across fields is quite varied. In [statistics](https://en.wikipedia.org/wiki/Statistics), where classification is often done with [logistic regression](https://en.wikipedia.org/wiki/Logistic_regression) or a similar procedure, the properties of observations are termed [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable) (or [independent variables](https://en.wikipedia.org/wiki/Independent_variable), regressors, etc.), and the categories to be predicted are known as outcomes, which are considered to be possible values of the [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable). In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), the observations are often known as *instances*, the explanatory variables are termed *features* (grouped into a [feature vector](https://en.wikipedia.org/wiki/Feature_vector)), and the possible categories to be predicted are *classes*. Other fields may use different terminology: e.g. in [community ecology](https://en.wikipedia.org/wiki/Community_ecology), the term "classification" normally refers to [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis), i.e., a type of [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning), rather than the supervised learning.

* 1. ONE HOT ENCODER:

A one hot encoding is a **representation of categorical variables as binary vectors**. Each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

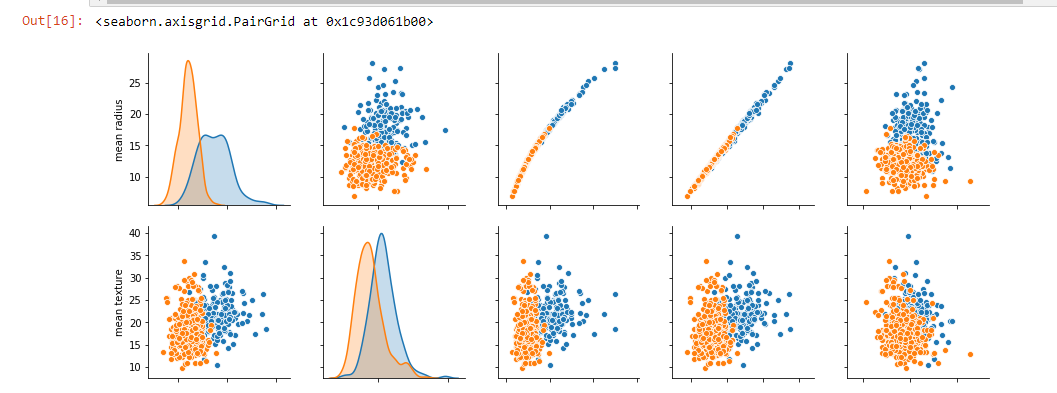
**Onehotencoding** is a powerful technique to transform categorical data into a numerical representation that machine learning algorithms can utilize to perform optimally without falling into the misrepresentation issue previously mentioned. You should now be able to easily perform **one**-**hotencodingusing** the **Pandas** built-in functionality.Using categorical data in Multiple Regression Models is a powerful method to include non-numeric data types into a regression model. Categorical data refers to data values which represent categories - data values with a fixed and unordered number of values, for instance gender (male/female) or season (summer/winder/spring/fall). In a regression model, these values can be represented by [dummy variables](http://en.wikipedia.org/wiki/Dummy_variable_%28statistics%29) - variables containing values such as 1 or 0 representing the presence or absence of the categorical value. The Dummy Variable trap is a scenario in which the independent variables are [multicollinear](http://en.wikipedia.org/wiki/Multicollinearity) - a scenario in which two or more variables are highly correlated; in simple terms one variable can be predicted from the others. The ***DUMMY VARIABLE TRAP***  is a scenario in which the independent variables are multicollinear – a scenario in which two or more variables are highly correlated .In statistics, especially in regression models, we deal with various kind of data. The data may be quantitative (numerical) or qualitative (categorical). The numerical data can be easily handled in regression models but we can’t use categorical data directly, it needs to be transformed in some way.

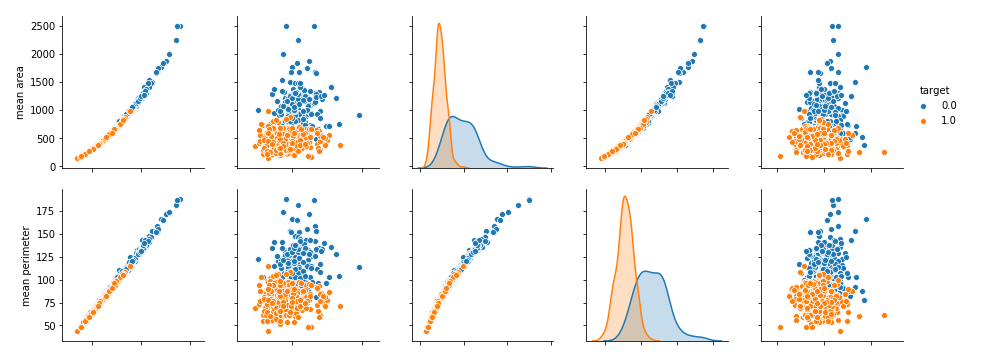
For transforming categorical attribute to numerical attribute, we can use label encoding procedure (label encoding assigns a unique integer to each category of data). But this procedure is not alone that much suitable, hence, ***One hot encoding*** is used in regression models following label encoding. This enables us to create new attributes according to the number of classes present in the categorical attribute i.e if there are *n* number of categories in categorical attribute, *n* new attributes will be created. These attributes created are called ***Dummy Variables***. Hence, dummy variables are “proxy” variables for categorical data in regression models.  
These dummy variables will be created with *one hot encoding* and each attribute will have value either 0 or 1, representing presence or absence of that attribute.

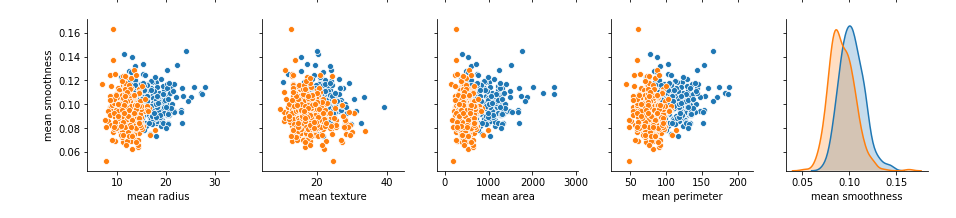
# VISUALIZING THE DATA:

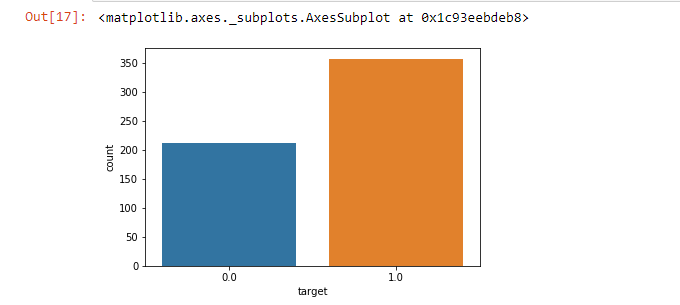
### Seaborn for Python Data Visualization

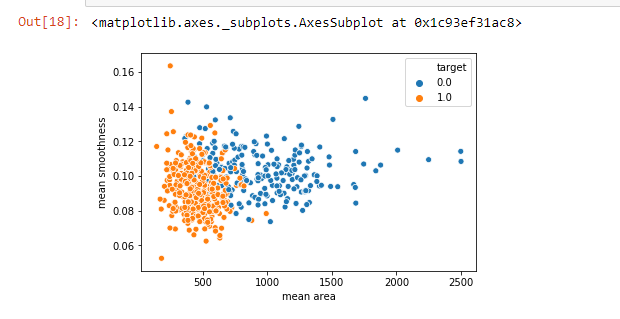
Seaborn Python is a data visualization library based on [Matplotlib](https://blog.quantinsti.com/python-matplotlib-tutorial/). It provides a high-level interface for drawing attractive statistical graphics. Because seaborn python is built on top of Matplotlib, the graphics can be further tweaked using Matplotlib tools and rendered with any of the Matplotlib backends to generate publication-quality figures.





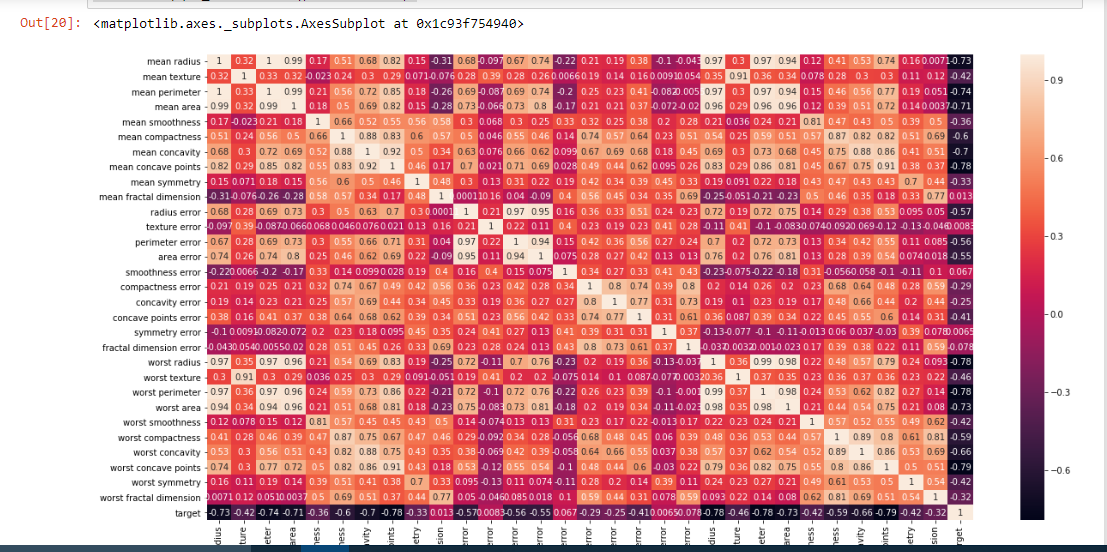


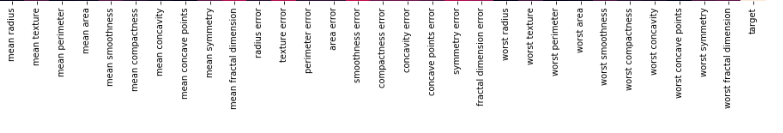


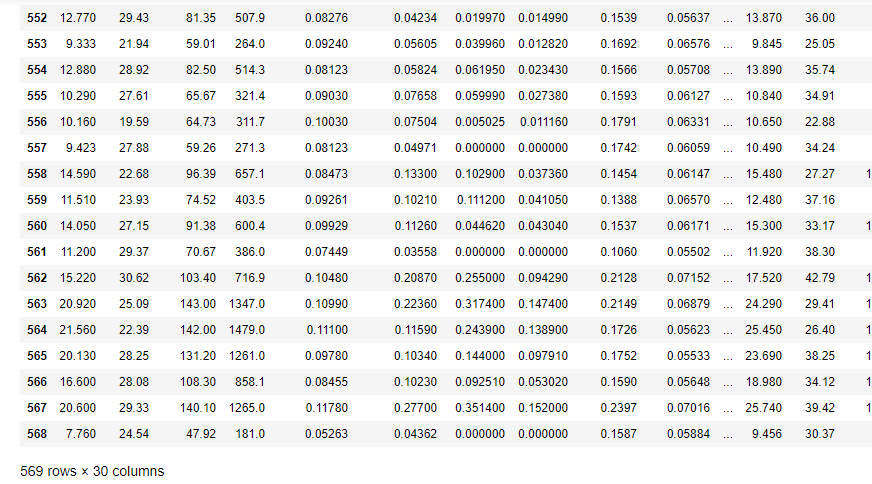
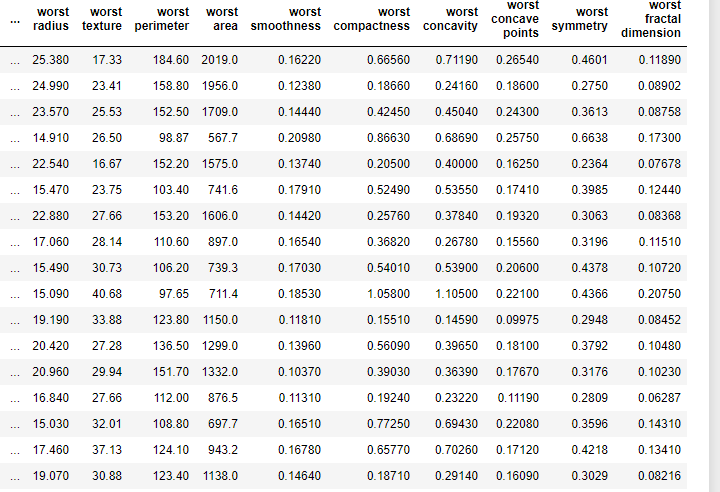
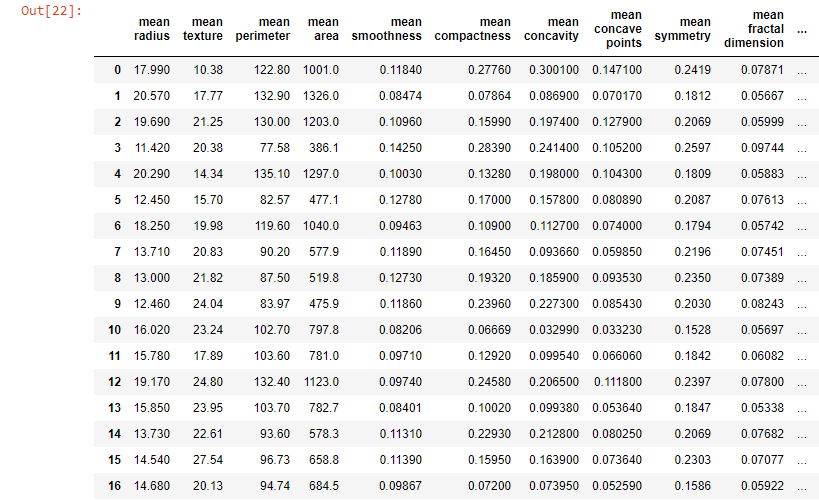


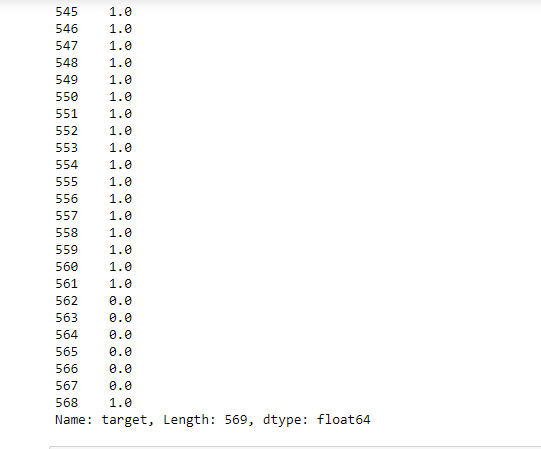
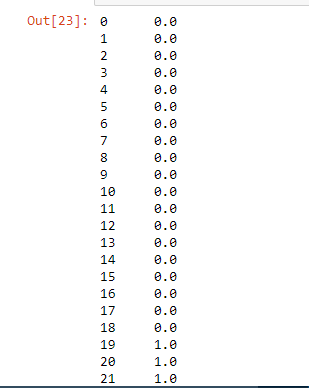
### What is a heatmap?

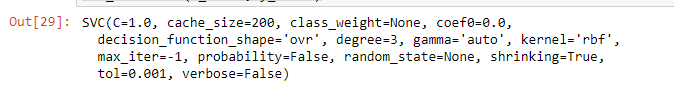
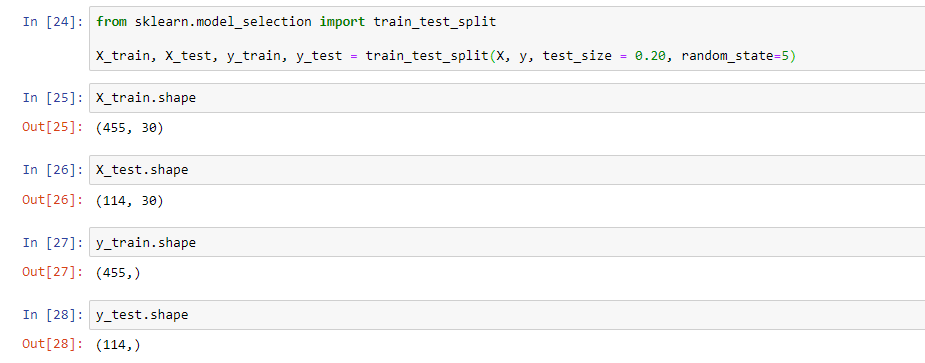
A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors. The seaborn python package allows the creation of annotated heatmaps which can be tweaked using Matplotlib tools as per the creator’s requirement.











1. EVALUATING THE MODEL:
   1. CONFUSION MATRIX:

A confusion matrix is a summary of prediction results on a classification problem.The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.**The confusion matrix shows the ways in which your classification modelis confused when it makes predictions.** There could be four possible outcomes. Let us look at all four. -----**True Positives (TP)** - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

**True Negatives (TN)** - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

False positives and false negatives, these values occur when your actual class contradicts with the predicted class.

**False Positives (FP)** – When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

**False Negatives (FN)** – When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

Once you understand these four parameters then we can calculate Accuracy, Precision, Recall and F1 score.

**Accuracy** - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.803 which means our model is approx. 80% accurate.

Accuracy = TP+TN/TP+FP+FN+TN

**Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Precision = TP/TP+FP

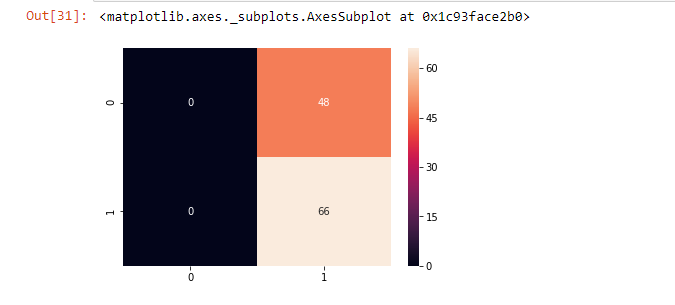
**Recall**(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: Of all the passengers that truly survived, how many did we label? We have got recall of 0.631 which is good for this model as it’s above 0.5.

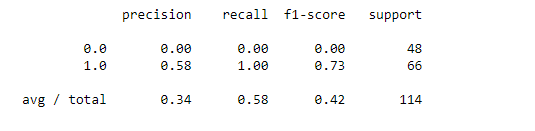
Recall = TP/TP+FN

**F1 score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall. In our case, F1 score is 0.701.

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

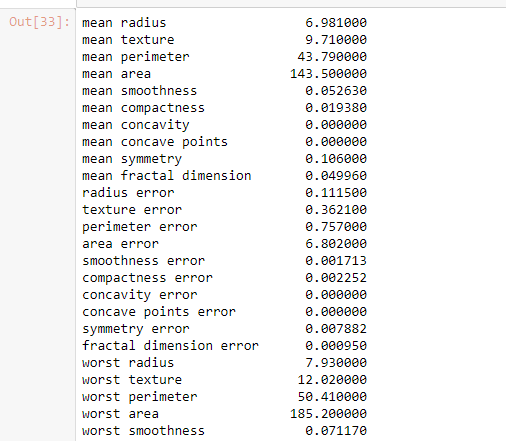
**The support** is the number of samples of the true response that lie in that class.

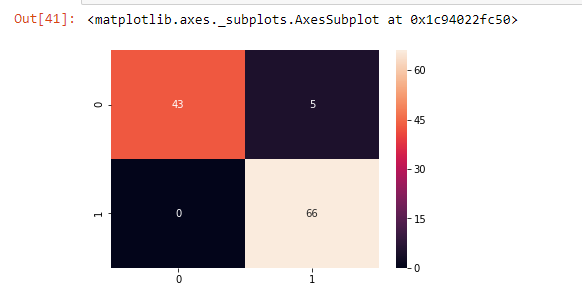
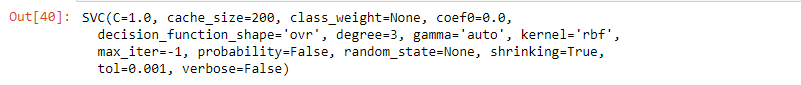
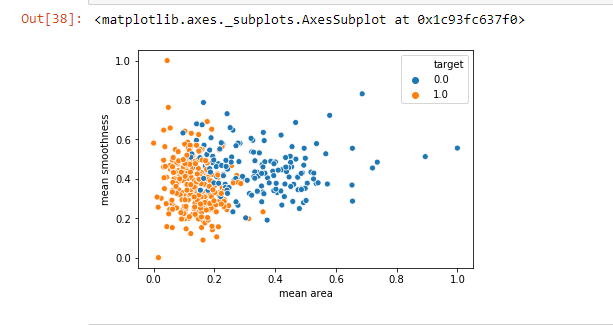
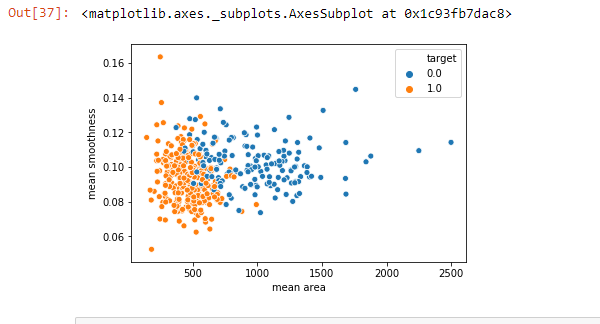
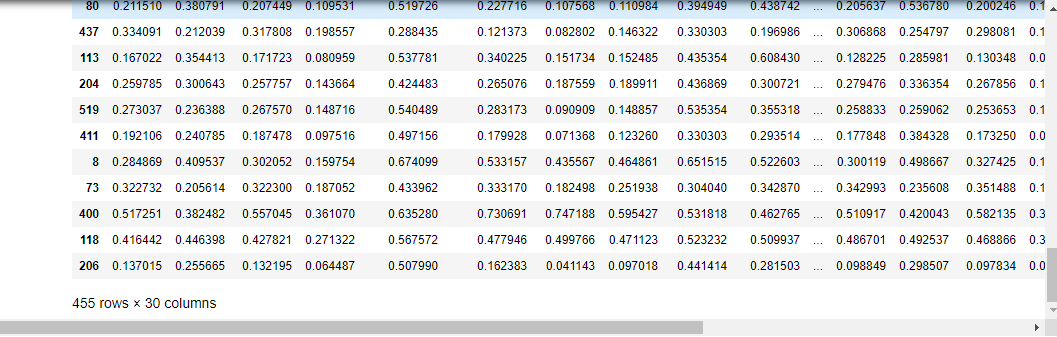
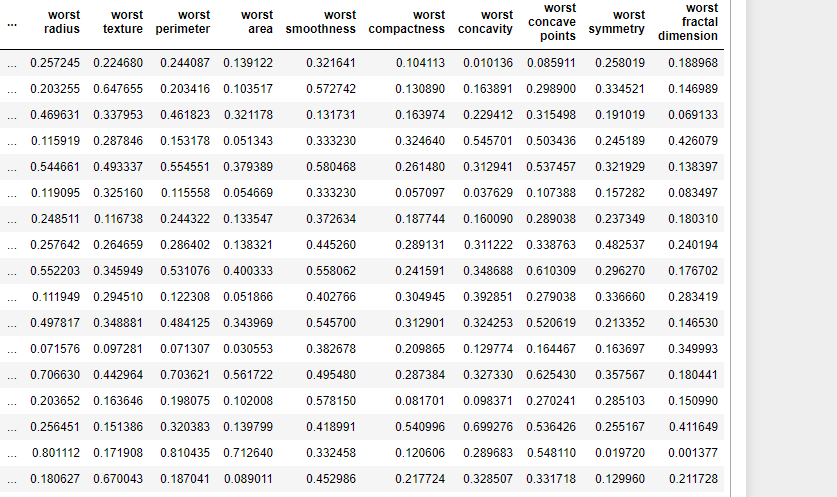
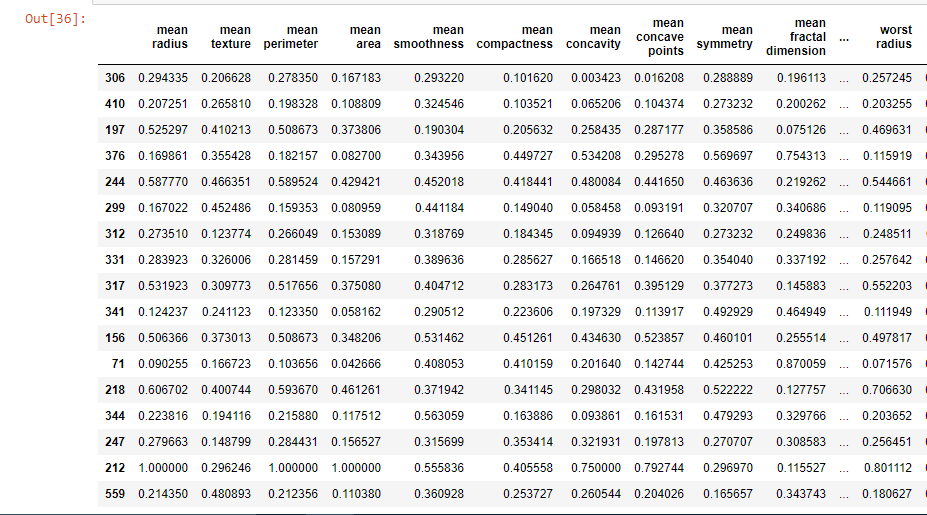
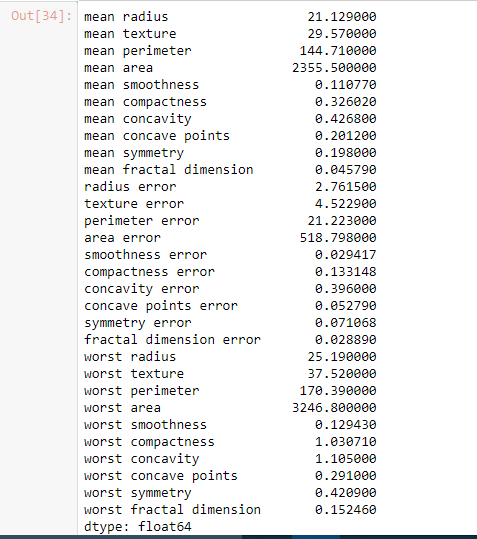
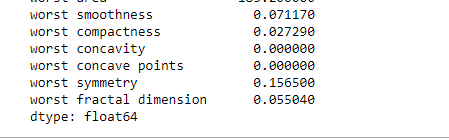




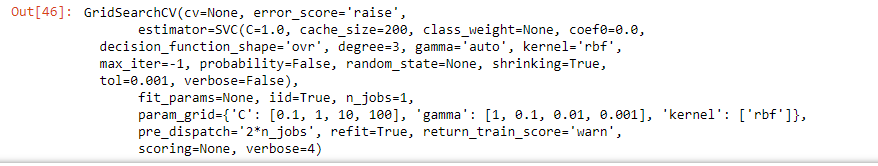
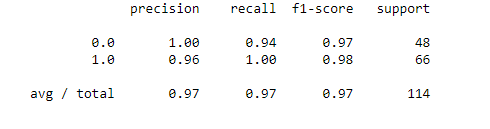
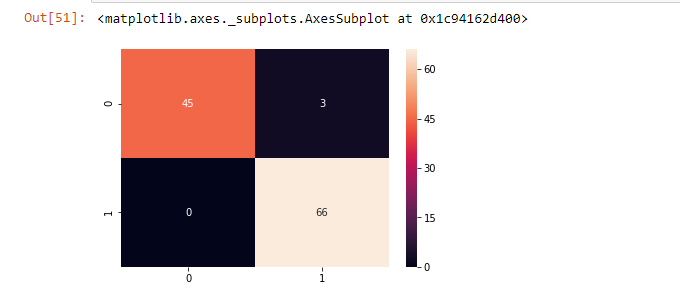
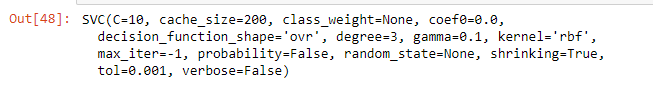
# IMPROVING THE MODEL

* 1. IMPROVISATION PART 1





* 1. IMPROVISATION PART 2

52.PNG

CONCLUSION

In Germany the quality of oncological care is already very high. This is also due, in no small measure, to the establishment of certified facilities for oncological care. In addition to the positive effect of improving the quality indicators (QI) which serve as indirect parameters for the quality of outcomes certified centers have had a highly significant impact on improving the mortality of breast cancer patients  – as demonstrated by the 3 certified centers in Middle Franconia. Similar results have also been reported for a single-center study carried out in the Breast Center of the University Hospital Heidelberg. But care providers are currently burdened with additional costs for which they are not being adequately reimbursed. The financial and staffing costs required to document breast cancer patients are considerable; in addition to clinical data, considerable time and resources are spent on documentations for quality assurance and quality management required as part of certification, notwithstanding the fact that clinical documentation can also be very detailed, as these data are used secondarily for quality assurance. In the long term, the current high standard of care can only be maintained if there is adequate financial support and if facilities are relieved of some of the costs of documentation. The most important aspect of this expenditure is the financial and staffing cost currently needed for documentation. Measures to reduce the time and cost of documentation are urgently required. A more effective way of collecting and collating data is necessary. Investment in suitable documentation systems with compatible interfaces is necessary. The subsequent impact on healthcare could be considerable:

1. Staffing and financial costs currently used for documentation could be reduced and the resources could be invested in other areas of the healthcare system;
2. This could relieve some of the burden on doctors, many of whom are already working at the limits of their capacity, particularly as there is an increasing shortage of young people entering the profession – even if enough money were available to leave documentation primarily in the hands of doctors, this would not be possible due to the lack of doctors;
3. This could strengthen the position of new professional groups working in healthcare, for example medical documentation assistants specializing in tumor documentation;
4. Quality assurance could be optimized by defining fewer but more relevant quality indicators and ensuring that these data are documented in the same standard format by all professional groups and medical specialties.

Based on the findings of this single-center project, a multi-center survey will be carried out to validate the results of this study and to highlight differences in documentation costs and times in facilities offering different levels of care as well as differences between certified and non-certified facilities.

FUTURE SCOPE OF PROJECT

Breast cancer remains the most common invasive cancer among women. The primary patients of breast cancer are adult women who are approaching or have reached menopause; 90 percent of new cases in U.S. women in 2009 were diagnosed at age 45 or older. Growing knowledge of the complexity of breast cancer stimulated a transition in breast cancer research toward elucidating how external factors may influence the etiology of breast cancer.

Breast Cancer and the Environment reviews the current evidence on a selection of environmental risk factors for breast cancer, considers gene-environment interactions in breast cancer, and explores evidence-based actions that might reduce the risk of breast cancer. The book also recommends further integrative research into the elements of the biology of breast development and carcinogenesis, including the influence of exposure to a variety of environmental factors during potential windows of susceptibility during the full life course, potential interventions to reduce risk, and better tools for assessing the carcinogenicity of environmental factors. For a limited set of risk factors, evidence suggests that action can be taken in ways that may reduce risk for breast cancer for many women: avoiding unnecessary medical radiation throughout life, avoiding the use of some forms of postmenopausal hormone therapy, avoiding smoking, limiting alcohol consumption, increasing physical activity, and minimizing weight gain.

Breast Cancer and the Environment sets a direction and a focus for future research efforts. The book will be of special interest to medical researchers, patient advocacy groups, and public health professionals.

REFERENCES

1. Robert Koch-Institut ; Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V., Hrsg . Berlin: Robert Koch-Institut; 2013. Krebs in Deutschland 2009/2010. 9. Ausgabe. [[Google Scholar](https://scholar.google.com/scholar?q=Robert+Koch-Institut++Gesellschaft+der+epidemiologischen+Krebsregister+in+Deutschland+e.V.,+Hrsg++Krebs+in+Deutschland+2009/2010.+9.+Ausgabe+Berlin+Robert+Koch-Institut+2013+)]

2. Bundesministerium für Bildung und Forschung Bevölkerung in Deutschland nach Alter und Geschlecht (Tabelle 0.14). 2012Online:[http://www.datenportal.bmbf.de](http://www.datenportal.bmbf.de/)last access: 08.05.2014

3. Lüftner D, Lux M P, Maass N. et al.Advances in breast cancer – looking back over the year. Geburtsh Frauenheilk. 2012;72:1117–1129. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4651151/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/26640285)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Geburtsh+Frauenheilk&title=Advances+in+breast+cancer+%E2%80%93+looking+back+over+the+year&author=D+L%C3%BCftner&author=M+P+Lux&author=N+Maass&volume=72&publication_year=2012&pages=1117-1129&)]

4. Beckmann M W. Nationaler Krebsplan des Bundesministeriums für Gesundheit – Strategieplan der Deutschen Krebsgesellschaft. Frauenheilkunde up2date. 2009;5:323–329. [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Frauenheilkunde+up2date&title=Nationaler+Krebsplan+des+Bundesministeriums+f%C3%BCr+Gesundheit+%E2%80%93+Strategieplan+der+Deutschen+Krebsgesellschaft&author=M+W+Beckmann&volume=5&publication_year=2009&pages=323-329&)]

5. Bundesministerium für Gesundheit Nationaler KrebsplanOnline:<http://www.bmg.bund.de/cln_169/nn_1168248/SharedDocs/Standardartikel/DE/AZ/N/Glossarbegriff-Nationaler-Krebsplan.html%23doc1632822bodyText14>last access: Februar 2010

6. Kowalski C, Wesselmann S, Kreienberg R. et al.The patientsʼ view on accrediated breast cancer centers: Strengths and potential for improvement. Geburtsh Frauenheilk. 2012;72:137–143. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4168360/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/25284830)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Geburtsh+Frauenheilk&title=The+patients%CA%BC+view+on+accrediated+breast+cancer+centers:+Strengths+and+potential+for+improvement&author=C+Kowalski&author=S+Wesselmann&author=R+Kreienberg&volume=72&publication_year=2012&pages=137-143&)]

7. Querschnitts-AG Dokumentation Datensparsame einheitliche Tumordokumentation – eine Kernforderung des Nationalen Krebsplans. Endfassung 10.10.2011Online:<http://www.tumorzentren.de/Nationaler-Krebsplan.html>last access: 07.08.2014

8. Tarifgemeinschaft deutscher Länder, Hrsg. Tarifvertrag für Ärztinnen und Ärzte an Universitätskliniken (TV-Ärzte). vom 30. Oktober 2006 in der Fassung des Änderungstarifvertrages Nr. 1 vom 27. August 2009Online:<http://www.uni-erlangen.de/einrichtungen/personalabteilung/>last access: 17.04.2011

9. Bundesärztekammer Hrsg.Gehaltstarifvertrag für Medizinische Fachangestellte. In: Deutsches Ärzteblatt (106/24)Online:<http://www.aerzteblatt.de/download/files/2009/06/down137850.pdf>last access: 24.11.2011

10. Kolberg H-C, Lüftner D, Lux M P. et al.Breast cancer 2012 – New aspects. Geburtsh Frauenheilk. 2012;72:602–615. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4168404/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/25324576)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Geburtsh+Frauenheilk&title=Breast+cancer+2012+%E2%80%93+New+aspects&author=H-C+Kolberg&author=D+L%C3%BCftner&author=M+P+Lux&volume=72&publication_year=2012&pages=602-615&)]

11. Kreienberg R, Albert U-S, Follmann M. et al.Interdisciplinary GoR level III guidelines for the diagnosis, therapy and follow-up care of breast cancer. Geburtsh Frauenheilk. 2013;73:556–583. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3963234/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/24771925)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Geburtsh+Frauenheilk&title=Interdisciplinary+GoR+level+III+guidelines+for+the+diagnosis,+therapy+and+follow-up+care+of+breast+cancer&author=R+Kreienberg&author=U-S+Albert&author=M+Follmann&volume=73&publication_year=2013&pages=556-583&pmid=24771925&)]

12. Brucker S Y, Schumacher C, Sohn C. et al.Benchmarking the quality of breast cancer care in a nationwide voluntary system: the first five-year results (2003–2007) from Germany as a proof of concept. BMC Cancer. 2008;8:358. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2647938/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/19055735)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=BMC+Cancer&title=Benchmarking+the+quality+of+breast+cancer+care+in+a+nationwide+voluntary+system:+the+first+five-year+results+(2003%E2%80%932007)+from+Germany+as+a+proof+of+concept&author=S+Y+Brucker&author=C+Schumacher&author=C+Sohn&volume=8&publication_year=2008&pages=358&pmid=19055735&)]

13. Brucker S Y, Bamberg M, Jonat W. et al.Certification of breast centres in Germany: proof of concept for a prototypical example of quality assurance in multidisciplinary cancer care. BMC Cancer. 2009;9:228. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2719663/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/19602242)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=BMC+Cancer&title=Certification+of+breast+centres+in+Germany:+proof+of+concept+for+a+prototypical+example+of+quality+assurance+in+multidisciplinary+cancer+care&author=S+Y+Brucker&author=M+Bamberg&author=W+Jonat&volume=9&publication_year=2009&pages=228&pmid=19602242&)]

14. Beckmann M W, Brucker C, Hanf V. et al.Quality assured health care in certified breast centers and improvement of prognosis of breast cancer patients. Onkologie. 2011;34:362–367. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/21734422)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Onkologie&title=Quality+assured+health+care+in+certified+breast+centers+and+improvement+of+prognosis+of+breast+cancer+patients&author=M+W+Beckmann&author=C+Brucker&author=V+Hanf&volume=34&publication_year=2011&pages=362-367&pmid=21734422&)]

15. Heil J, Gondos A, Rauch G. et al.Outcome analysis of patients with primary breast cancer initially treated at a certified academic breast unit. Breast. 2012;21:303–308. [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/22310244)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Breast&title=Outcome+analysis+of+patients+with+primary+breast+cancer+initially+treated+at+a+certified+academic+breast+unit&author=J+Heil&author=A+Gondos&author=G+Rauch&volume=21&publication_year=2012&pages=303-308&pmid=22310244&)]

16. Beckmann M W, Bani M R, Loehberg C R. et al.Are certified breast centers cost-effective? Breast Care. 2009;4:245–250. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2941653/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/20877662)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Breast+Care&title=Are+certified+breast+centers+cost-effective?&author=M+W+Beckmann&author=M+R+Bani&author=C+R+Loehberg&volume=4&publication_year=2009&pages=245-250&pmid=20877662&)]

17. Lux M P, Hildebrandt T, Beyer-Finkler E. et al.Relevance of health economics in breast cancer treatment – the view of certified breast centres and their patients. Breast Care. 2013;8:15–21. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3971810/)] [[PubMed](https://www.ncbi.nlm.nih.gov/pubmed/24715838)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Breast+Care&title=Relevance+of+health+economics+in+breast+cancer+treatment+%E2%80%93+the+view+of+certified+breast+centres+and+their+patients&author=M+P+Lux&author=T+Hildebrandt&author=E+Beyer-Finkler&volume=8&publication_year=2013&pages=15-21&pmid=24715838&)]