A

Project Report

On

**“TELECOM CHURN PREDICTION MODEL”**

Submitted in partial fulfillment of

the requirements for the final Semester Sessional Examination of

BACHELOR OF TECHNOLOGY

IN

**COMPUTER SCIENCE & ENGINEERING**

By

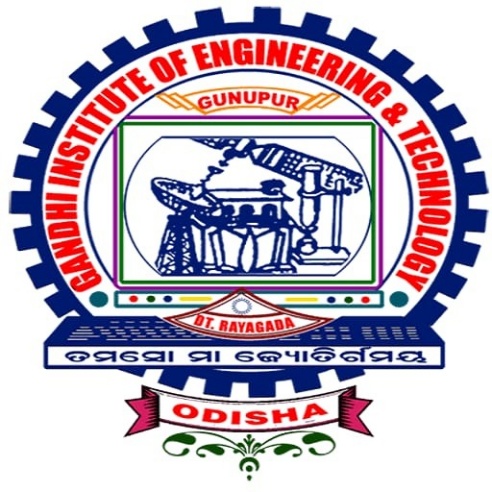
**SANSKRUTI PANDA**

**JIGYANSHU SURAJ**

**AYUSHI PRADHAN**

**Registration No.**

1801210109

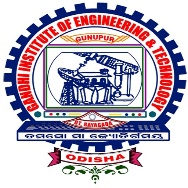


**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**GIET MAIN CAMPUS AUTONOMOUS**

**GUNUPUR – 765022**

**2021 - 22**

**GIET MAIN CAMPUS AUTONOMOUS,**

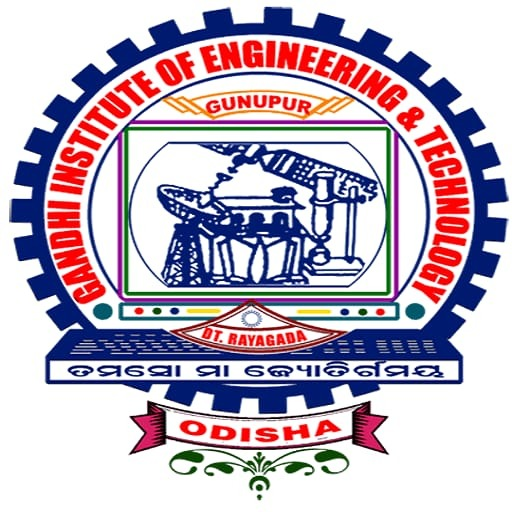
**GUNUPUR-765022, Dist.: Ravaged (Orissa), India**

**(Approved by AICTE and Govt, of Orissa)**

**Ph: 06857 – 250172(Office), 251156(Principal), 250232(Fax),**

**E-mail:** [**gandhi\_giet@yahoo.com**](mailto:gandhi_giet@yahoo.com) **visit us at WWW.giet.org**

**Department of Computer Science & Engineering**



**CERTIFICATE**

This is to certify that the project work entitled “**TELECOM CHURN PREDICTION MODEL**” is done by **Name- SANSKRUTI PANDA, AYUSHI PRADHAN, JIGYANSHU SURAJ Regd. No. –** 1801210109 in partial fulfillment of the requirements for the final Semester Sessional Examination of Bachelor of Technology in Computer Science and Engineering during the academic year 2021-22. This work is submitted to the department as a part of evaluation of final Semester Project.

Name

Class Teacher Prof. (Dr). Sanjay Kumar Kunar

Hood, CSE

i

**ACKNOWLEDGEMENT**

In the present world of competition, there is a race of existence in which those are having will to come forward succeed. Project is like a bridge between theoretical and practice working. With this willing I did this particular project. First of all, I would like to thank the supreme power the almighty God who is obviously the one has always guided me to work on the right path of the life. Without his grace this project could become a reality. Next to him are my parents, whom I am greatly indebted for me brought up with love and encouragement to this stage. I am feeling oblige in taking the opportunity to sincerely thank to *Prof. (Dr). Sanjay Kumar Kunar (Head of the Department, Computer Science and Engineering)* and special thanks to my worthy teachers of department. Moreover, I am highly obliged in taking the opportunity to sincerely thank to all the staff members of the Computer Science and Engineering Department for their generous attitude and friendly behavior. At last, but not the least, I am thankful to all my teachers and friends who have been always helping and encouraging me throughout the year. I have no valuable words to express my thanks, but my heart is still full of the favors received from every person.

Name-Jigyanshu Suraj

Ayushi Pradhan

Sanskruti Panda

Roll No.- 18CSE197

18CSE

18CSE

Regd No.- 1801210109

ii

**CONTENTS**

# Abstract

# Introduction

## Need of the project

# Requirements

## Hardware requirements

## Software requirements

# Model design

## Data flow diagram

## Structure chart

# Model description

## Phase 1: Data pre-processing

## Phase-2: Data exploration

## Phase-3: Data preparation

## Phase-4: Data visualization

## Phase-5: Categorical data

## Phase 6: Feature Scaling

## Phase 7: Model Selection

# Result

# Future work

# Conclusion

# References

iii

# ABSTRACT

Customers are the base for any business success and that is why firms become aware of

the significance of acquiring satisfaction of customers. Customer churn is an essential

issue and it is regarded as one of the most essential concerns among firms because of

increasing rivalry among firms, increased significance of marketing strategies and

customers conscious behaviour in present years. Organizations must develop different

strategies to resolve the churn issues relying on the services they offer. Customer churn

practice is essential in competitive and rapidly developing in telecom sector. The process

of migrating from one service provider to another telecom service provider occurs due to

good services or rates or due to various advantages which the rivalry firm provides

customers when signing up. Due to the greater cost related with acquiring new customers

the prediction of customer churn has developed as an indispensable part of planning

process and strategic decision making in telecom sector. The main aim of the study is to

explore the customer churn prediction in telecom using in big machine learning data

platform. Machine learning techniques have been used for estimating the customer

probability to churn. This study makes use of logistic regression and KNN with big data

for predicting consumer churn in the telecom sector. Logistic regression has been used

widely to estimate the probability of churn as a function of variables set or features of

customers. Similarly, for churn K-Nearest Neighbor is used to examine if a customer

churns or not based on their feature’s proximity to customers in every class. This study

uses Kaggle website for dataset in predicting and analyzing churn. The results of the study

show that the accuracy rate of prediction in consumer churn is found to be 0.80 percent

and area under curve is found to be 0.71 percent

1

# INTRODUCTION

## PURPOSE

The key challenge is to predict if an individual customer will churn or not. To accomplish that, machine learning models are trained based on 80% of the sample data. The remaining 20% are used to apply the trained models and assess their predictive power with regards to “churn / not churn”. A side question will be, which features actually drive customer churn. That information can be used to identify customer “pain points” and resolve them by providing goodies to make customers stay. To compare models and select the best for this task, the accuracy is measured. Based on other characteristics of the data, for example the balance between classes (number of “churners” vs. “non-churners” in data set) further metrics are considered if needed.

## NEED OF THE PROJECT

For Telco companies it is key to attract new customers and at the same time avoid contract terminations (=churn) to grow their revenue generating base. Looking at churn, different reasons trigger customers to terminate their contracts, for example better price offers, more interesting packages, bad service experiences or change of customers’ personal situations.

Churn analytics provides valuable capabilities to predict customer churn and also define the underlying reasons that drive it. The churn metric is mostly shown as the percentage of customers that cancel a product or service within a given period (mostly months). If a Telco company had 10 Mio. customers on the 1st of January and received 500K contract terminations until the 31st of January the monthly churn for January would be 5%.

Telcos apply machine learning models to predict churn on an individual customer basis and take counter measures such as discounts, special offers or other gratifications to keep their customers. A customer churn analysis is a typical classification problem within the domain of supervised learning.

2

# REQUIREMENTS

## HARDWARE REQUIREMENTS

* Processor: - Intel i5 @160 GHz
* RAM: - more than 4GB

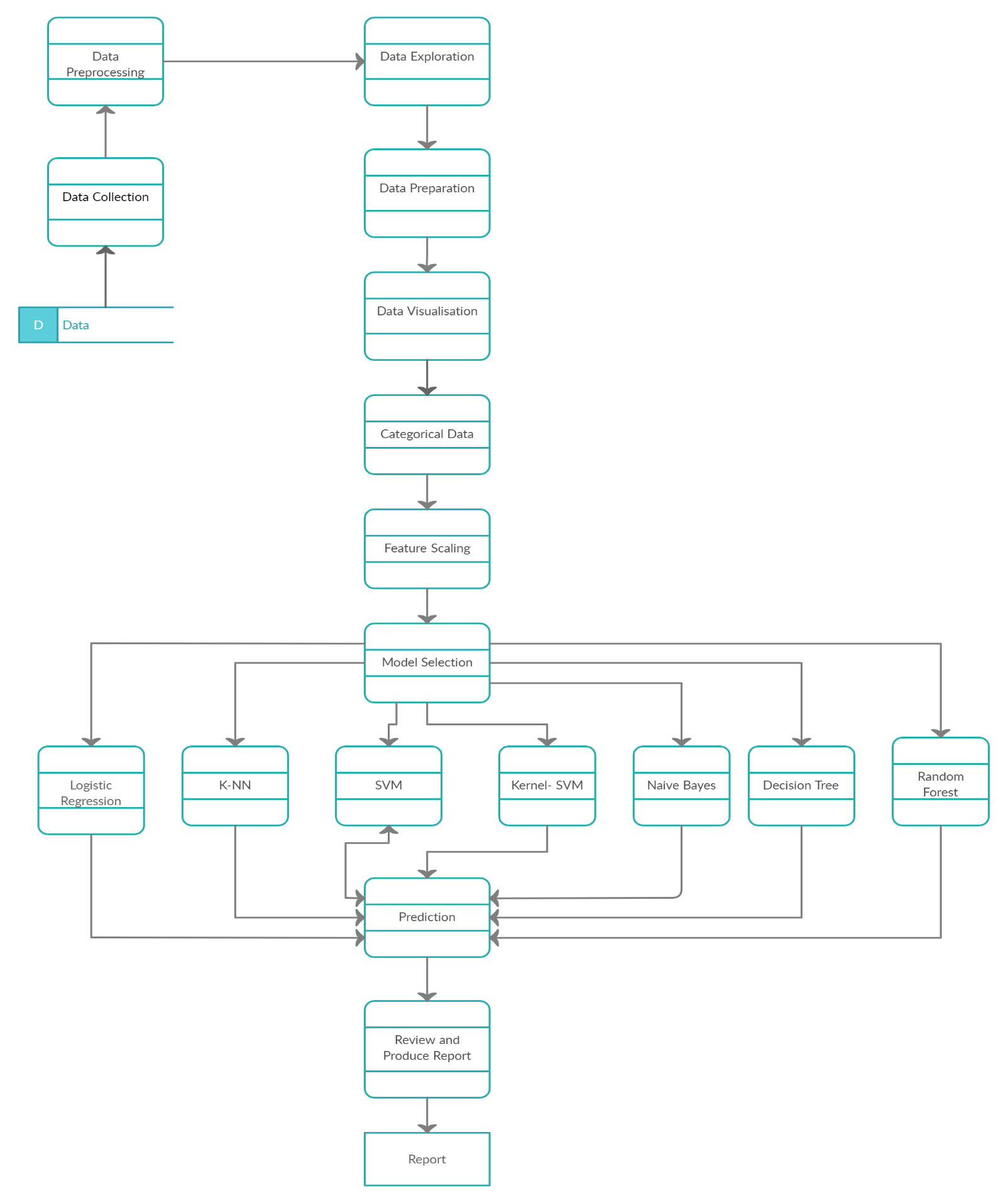
## SOFTWARE REQUIREMENTS

* Operating System-: Windows
* Language used-: Python 3.8
* Simulation environment-: Jupyter Notebook

3

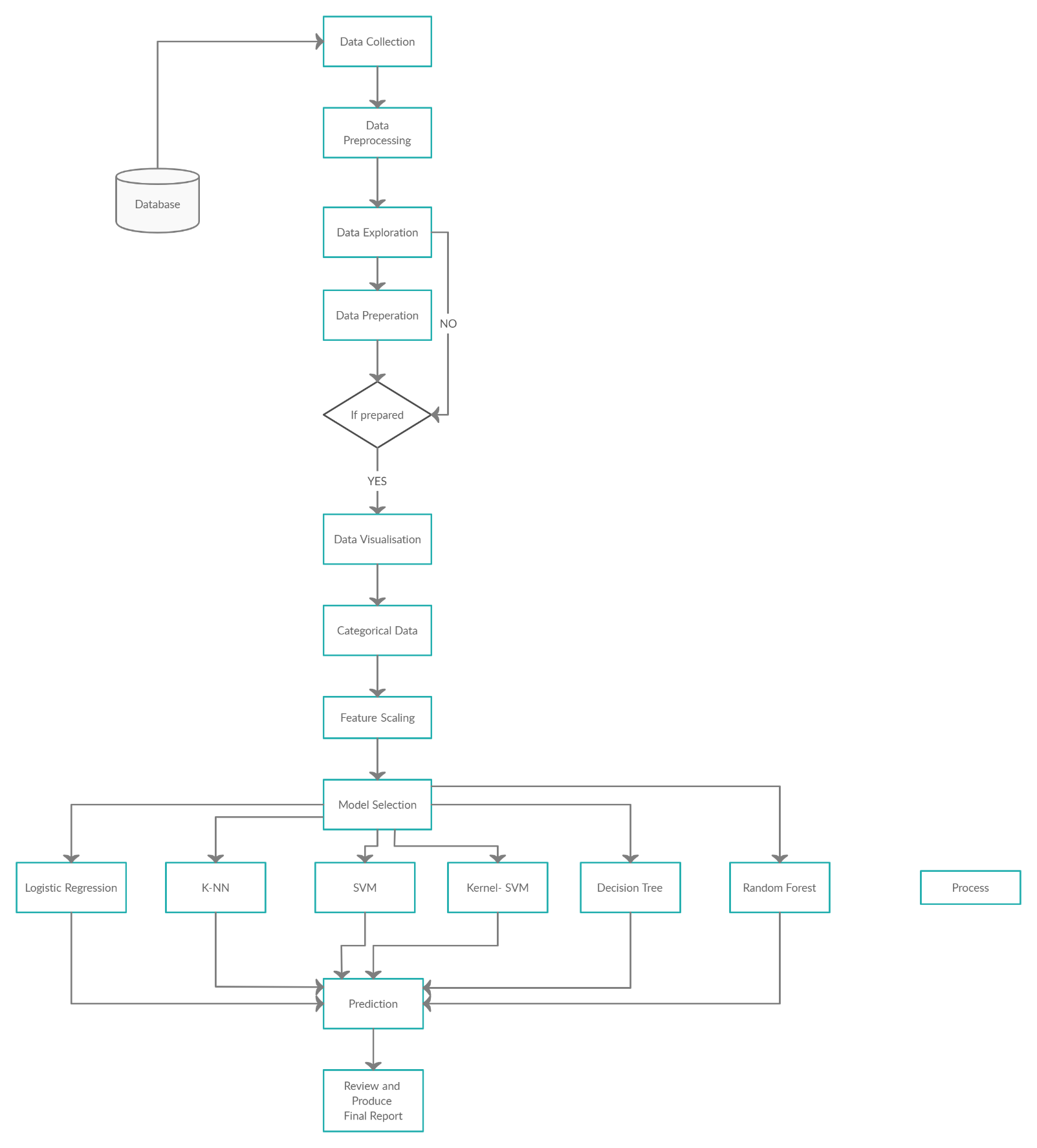
# MODEL DESIGN

## DATA FLOW DIAGRAM



4

## STRUCTURE CHART



5

# MODEL DESCRIPTION

## PHASE-1: DATA PREPROCESSING

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. The dataset used in this story is publicly available and was created by Rd. William H. Walberg. The goal is to classify whether the breast cancer is benign or malignant. To achieve this, we have used machine learning classification methods to fit a function that can predict the discrete class of new input.

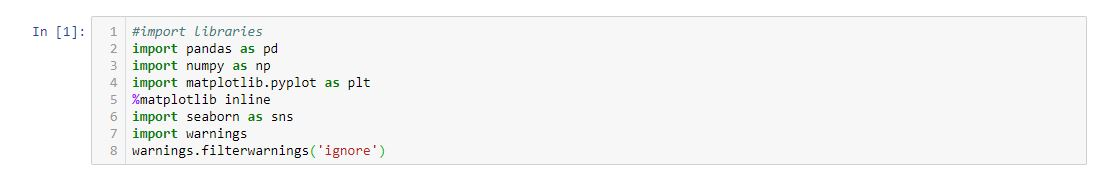
These are attributes in the dataset.

* ID number
* Diagnosis (M = malignant, B = benign)
* Ten real-valued features are computed for each cell nucleus:
* radius (mean of distances from centre to points on the perimeter)
* texture (standard deviation of Gray-scale values)
* perimeter
* area
* smoothness (local variation in radius lengths)
* compactness (perimeter² / 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)

The mean, standard error and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

## PHASE-2: DATA EXPLORATION

We will be using Jupyterto work on this dataset. We will first go with importing the necessary libraries and import our dataset to Jupyter. Then loading the dataset.



6

Fig: - Import libraries

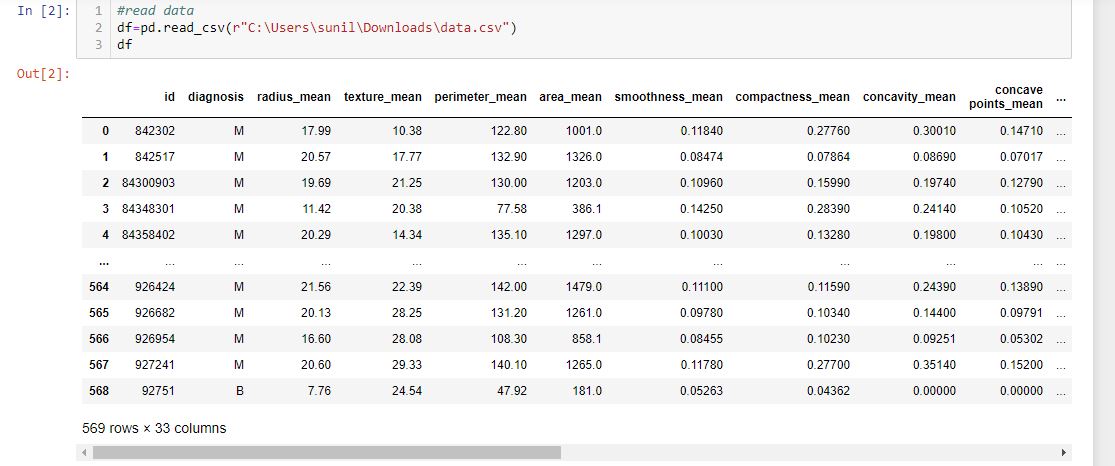


Fig: - read data

We can examine the data set using the pandas head() method. This will give the first 5 rows as output. We can find any missing or null data points of the data set (if there is any) using the following pandas function.

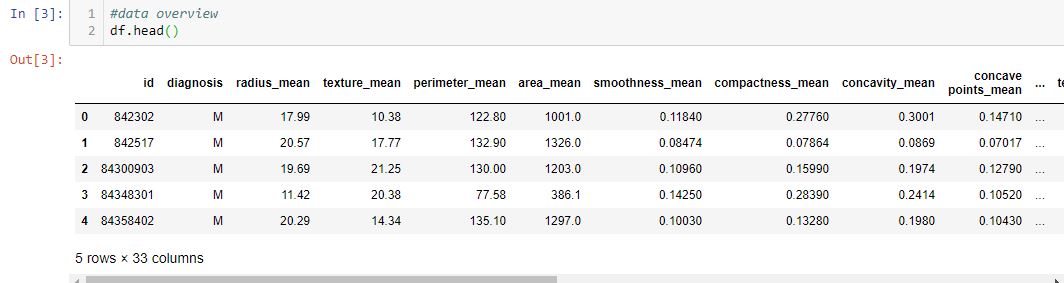


Fig: - Data overview

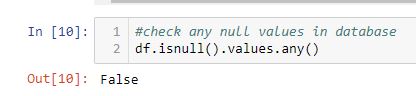


Fig: - Checks any null values in database

Fig: - Returns the number of missing values in each column

7

## PHASE-3: DATA PREPARATION

We can observe that the data set contain 569 rows and 32 columns. ‘*Diagnosis*’ is the column which we are going to predict, which says if the cancer is M = malignant or B = benign. 1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labelled as B (benign) and 212 as M (malignant). Then plot the graph using seaborn package.

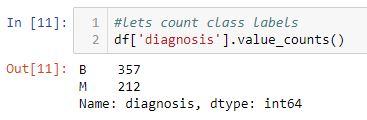
 

Fig: - bar graph representation of labels. It shows the count of benign and malignant

Fig: - counts the class labels

### Removing outliers

We use boxplot to check the outliers exists. Some of the boxplots are given below.

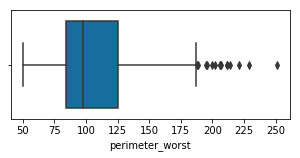
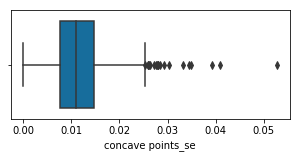


Fig: - boxplot before removing the outliers

After removing the outliers, there are 398 rows and 31 columns. Some of the boxplots are given below.

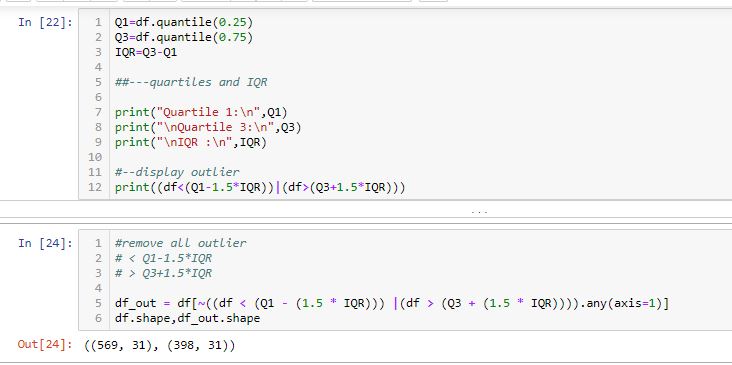


Fig: - number of columns after remove the outliers

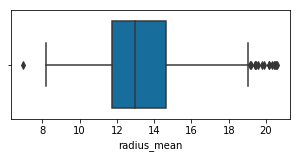
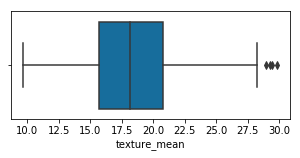
 

Fig: - boxplots after removing outliers

8

## PHASE-4: DATA VISUALIZATION

We will use pandas’ visualization which is built on top of matplotlib, to find the data distribution of the features. We use seaborn package to generate heatmap to check correlation between attributes.

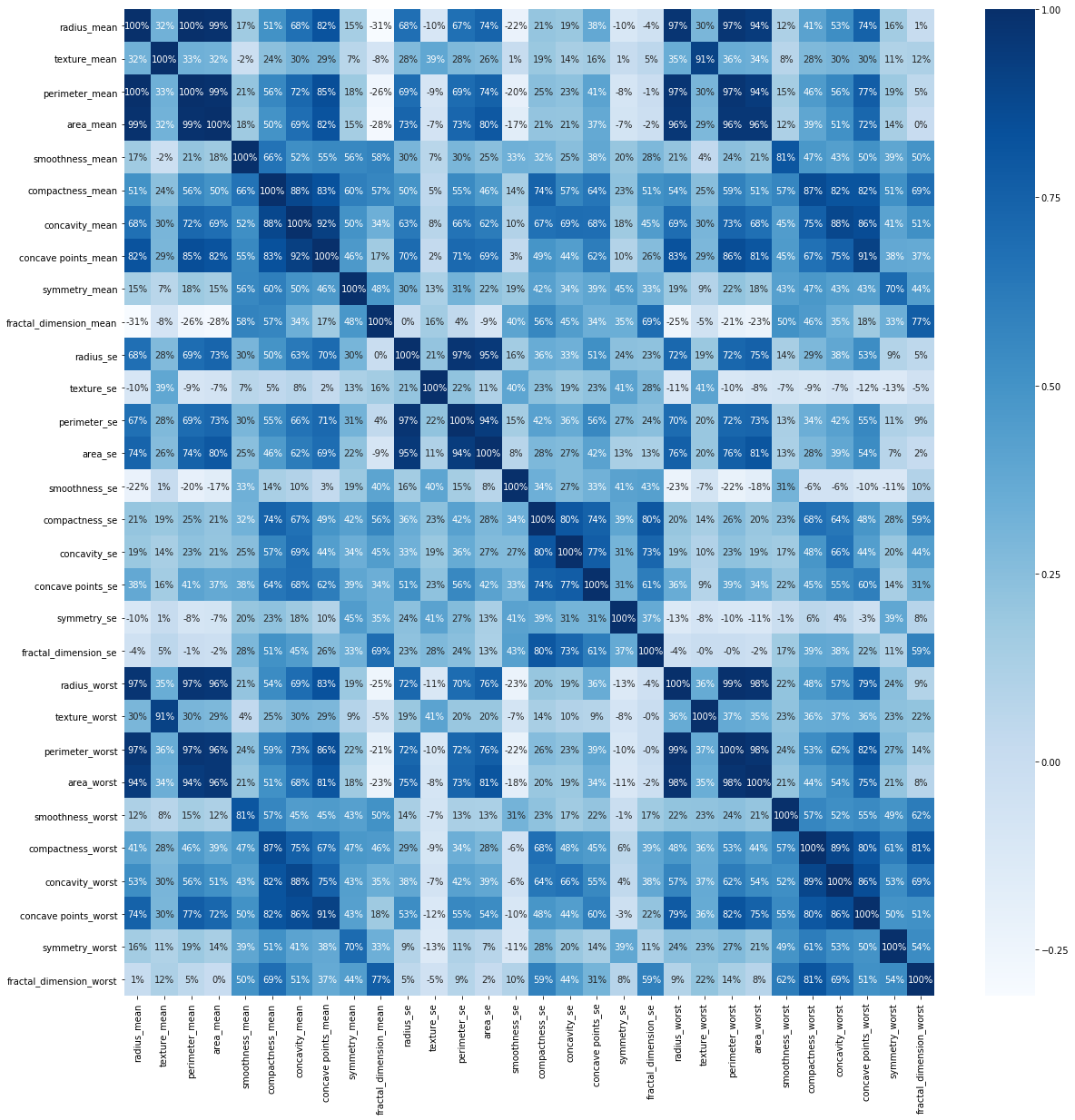


Fig: - heatmap to show correlation

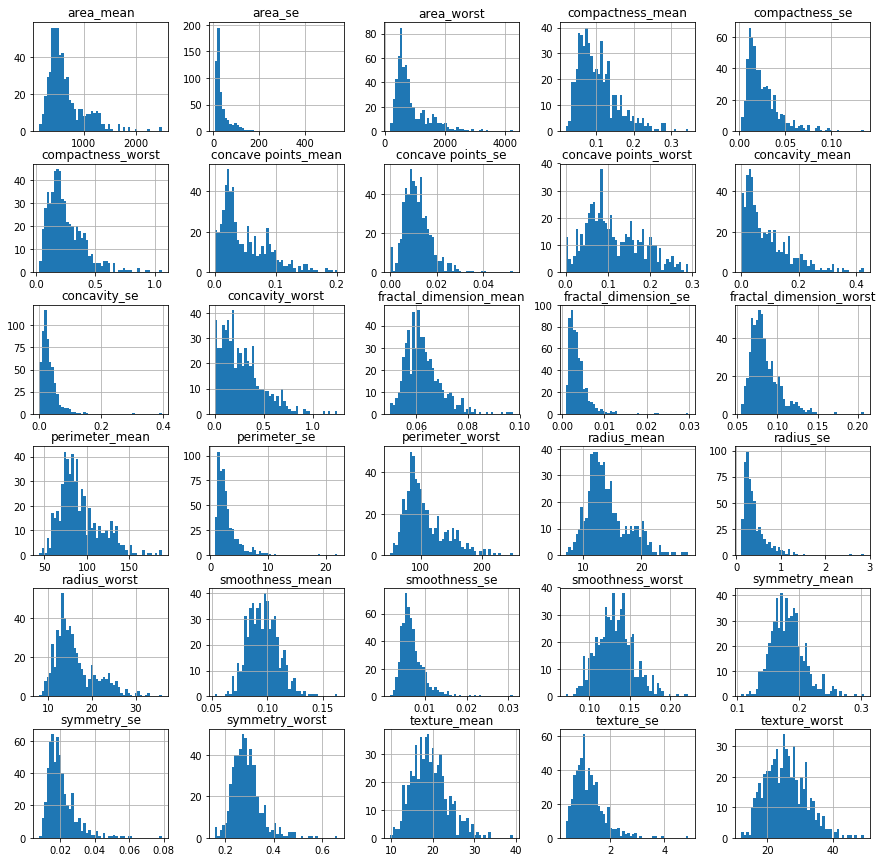


Fig: - histogram

9

## PHASE 5: CATEGORICAL DATA

Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set. We will use Label Encoder to label the categorical data and used to convert categorical data, or text data, into numbers, which our predictive models can better understand.

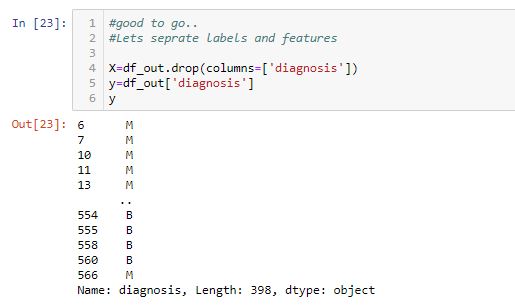


Fig: - diagnosis data before encoding

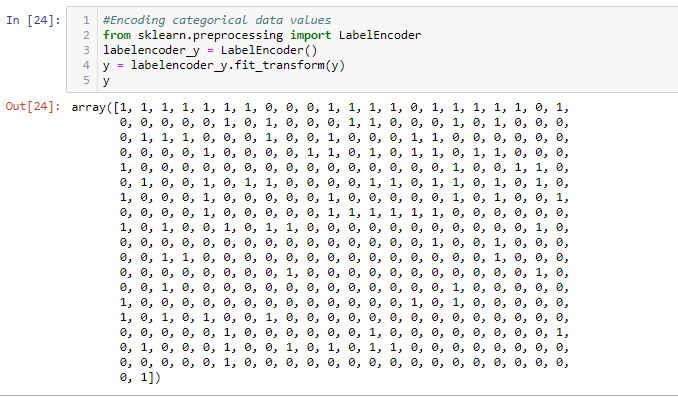


Fig: -diagnosis data after encoding

### Splitting The Dataset

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model’s prediction on this subset.

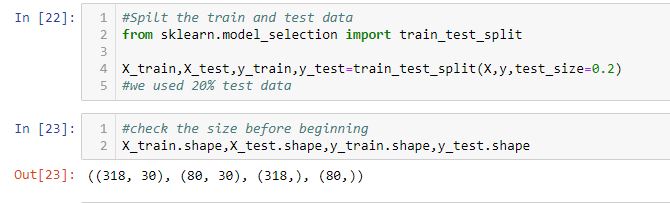


Fig: - split the data into training and test data

10

## PHASE 6: FEATURE SCALING

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. We need to bring all features to the same level of magnitudes. This can be achieved by scaling. This means that we are transforming the data so that it fits within a specific scale, like 0–100 or 0–1.

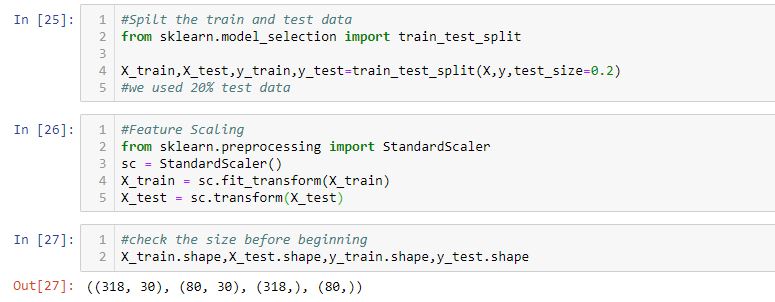


Fig: - feature scaling and checking the size of the training and test data

## PHASE 7: MODEL SELECTION

In our dataset we have the outcome variable or Dependent variable i.e., Y having only two set of values, either M (Malign) or B(Benign). So, we will use Classification algorithm of supervised learning.

We have different types of classification algorithms in Machine Learning: -

1. Logistic Regression
2. K-Nearest Neighbour
3. Support Vector Machines
4. Kernel SVM
5. Naïve Bayes
6. Decision Tree Algorithm
7. Random Forest Classification

### Logistic Regression

Logistic regression is a supervised machine learning technique, employed in classification jobs. This algorithm gives 97% accuracy.

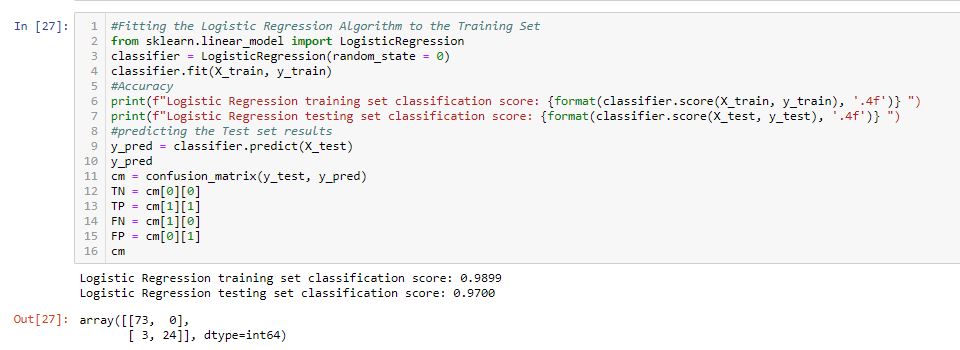


Fig: - output after applying logistic regression algorithm (Accuracy of the algorithm)

### K-Nearest Neighbour

K-Nearest Neighbor is a supervised machine learning technique. It gives 96% accuracy.

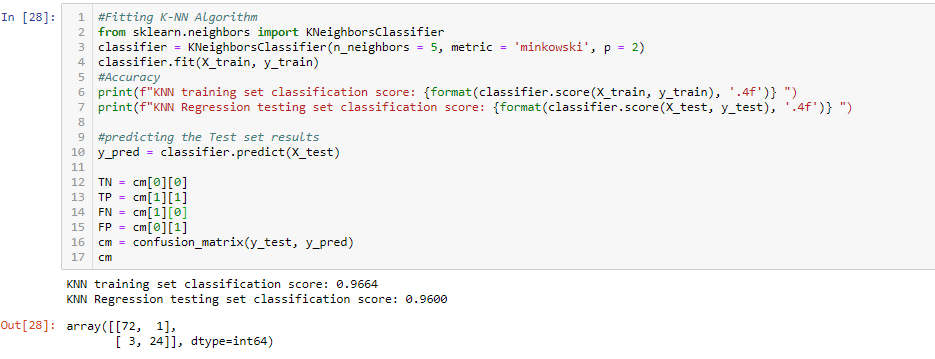


Fig: - output after applying K-NN algorithm (Accuracy of the algorithm)

11

### Support Vector Machines

Support Vector Machine is a supervised machine learning algorithm. This algorithm gives 95% accuracy.

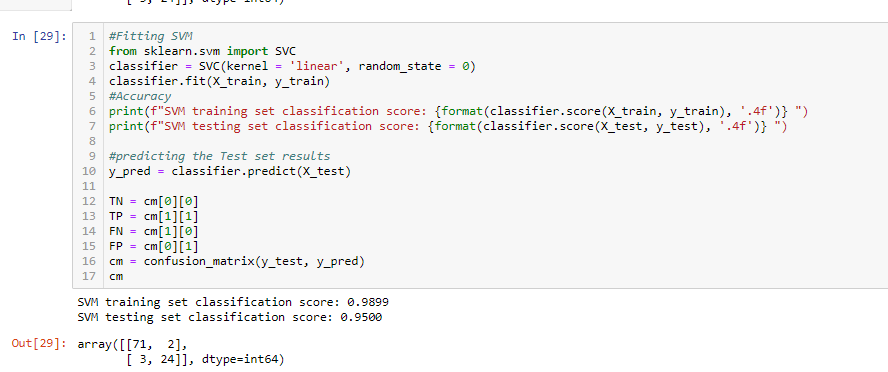


Fig: - output after applying Support Vector Machine algorithm (Accuracy of the algorithm)

### Kernel SVM

SVM algorithms use a set of mathematical functions that are defined as the kernel. This algorithm gives 96% accuracy.

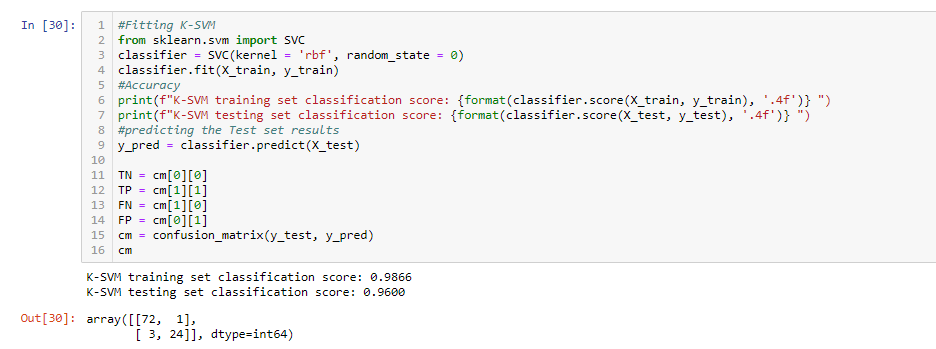


Fig: - output after applying Kernel SVM algorithm (Accuracy of the algorithm)

### Naïve Bayes

Naïve Bayes is a statistical classification technique. This technique gives 96% accuracy.

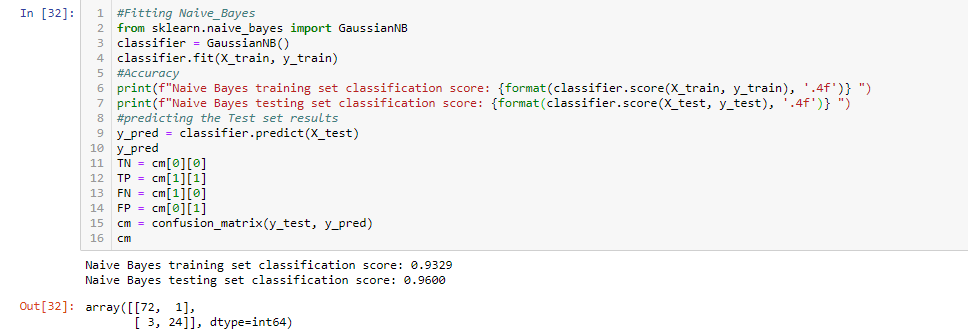


Fig: - output after applying Naïve Bayes algorithm (Accuracy of the algorithm)

### Decision Tree

A decision tree is a flowchart-like tree structure. This technique gives 92% accuracy.

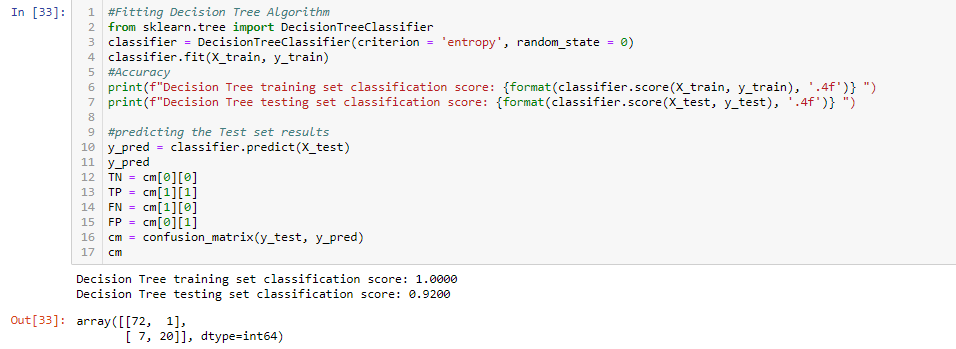


Fig: - output after applying Decision Tree algorithm (Accuracy of the algorithm)

### Random Forest

Random forest consists of many individual decision trees that operate as an ensemble. This algorithm gives 94% accuracy.

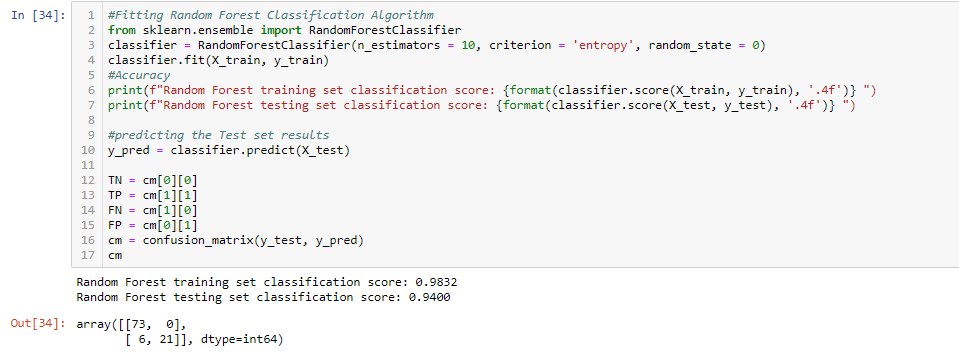
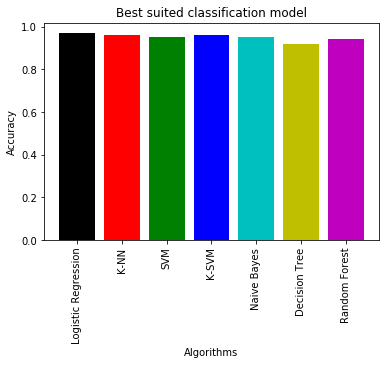


Fig: - output after applying Random Forest algorithm (Accuracy of the algorithm)

12

## PHASE 8: PREDICTION

We will use Classification Accuracy method to find the accuracy of our models. We plotted the accuracy bar graph. Here, we have got that logistic regression.



13

# RESULT

This project focus on the earlier diagnosis of breast cancer, as the detection bring about the success of about 97% accuracy by the use of ‘Logistic Regression’ algorithm for classification. By the use of python as a software platform, training and the computational time has been reduced to a greater extent than others. This project ensures the greater level of detection of breast cancer at earlier stage, by which mortality rate of cancer affected person can be reduced and earlier diagnosis would increase the life time of a patient by giving them a right treatment at a right stage.

|  |  |  |
| --- | --- | --- |
| S.NO. | ALGORITHMS | ACCURACY |
| 1 | LOGISTIC REGRESSION | 97% |
| 2 | KNN | 96% |
| 3 | SVM | 95% |
| 4 | KERNEL SVM | 96% |
| 5 | NAÏVE BAYES | 95% |
| 6 | DECISION TREE | 92% |
| 7 | RANDOM FOREST | 94% |

14

# FUTURE WORK

It is recommended to expect behavioural patterns and customer churn. Telecom service providers must spend in insight tools and powerful analytics to expect churn of customers, finds behaviour of customer and devise strategies that enhances profitability as well as retention. The customer retention strategies costs must be mapped with expected return on interest to prioritize investments effectively. Telecom firms have to realign their priorities around retention of customers. It is recommended to employ co-browsing to provide a personalized service to customers. Be in person or on phone, telecom service providers must engage with a strong welcome message to customers which makes them feel appreciated and comfortable. Co-Browsing is one of the essential ways to add a personal feeling to consumer service. Quality service to customers is useful in reducing the churn rate of customers saving their effort in convincing customers to remain when they need to cancel. Co-browsing brings the customer representative and customer together on similar page offering a visual link and helping to build trust rapidly. It is recommended that telecom service providers must increase engagement of customers. In this competitive world customers are bombarded constantly by information and choices from all around. With the appropriate strategy of marketing in place and by concentrating on customer retention and satisfaction service providers must increase engagement of customers and nurture big term relations. Telecom service providers must implement tailored programs specifically to support their customers perceive the advantages of their services and products. It is recommended that telecom service providers must delight and surprise their customers. A satisfied customer is the best strategy among all solutions to reduce the churn rate. Putting a smile on the face of customer is as easy as providing the best recognition award to customer. Telecom service provider must do something outstanding to show how much they value them. Thus, the survival of any business is based on its capability to retain customers and put huge amount of efforts in reducing the churn rate of customers

15

# CONCLUSION

Customer churn is major problems which the telecom sector is facing nowadays. It is essential to recognize possible customer churn so that the losses can be avoided. In order to maintain a loyal base of customer the service providers in telecom sector aims to retain customers with themselves. Since the costs related with obtaining a new customer is much greater than retaining older customer, the prediction of churn becomes even more essential. The big data analysis with machine learning makes the churn prediction much easier in telecom sector. Thus, it can be concluded that the big data analytics with machine learning techniques have proven to be accurate and effective to predicts customer churn in nearby future.

16

# REFERENCES

* <https://www.geeksforgeeks.org/ml-kaggle-breast-cancer-wisconsin-diagnosis-using-logistic-regression/>
* <https://pypi.org/>
* <https://pypi.org/project/scikit-learn/>
* <https://seaborn.pydata.org/generated/seaborn.heatmap.html>

17

# Step 1: Problem Definition

Based on the introduction the key challenge is to predict if an individual customer will churn or not. To accomplish that, machine learning models are trained based on 80% of the sample data. The remaining 20% are used to apply the trained models and assess their predictive power with regards to “churn / not churn”. A side question will be, which features actually drive customer churn. That information can be used to identify customer “pain points” and resolve them by providing goodies to make customers stay.

To compare models and select the best for this task, the accuracy is measured. Based on other characteristics of the data, for example the balance between classes (number of “churners” vs. “non-churners” in data set) further metrics are considered if needed.

# Step 2: Data Collection

The use case pipeline build-up is started with imports of some basic libraries that are needed throughout the case. This includes Pandas and Numpy for data handling and processing as well as Matplotlib and Seaborn for visualization.

For this exercise, the data set (.csv format) is downloaded to a local folder, read into the Jupyter notebook and stored in a Pandas DataFrame.

# Step 3: Exploratory Data Analysis

After data collection, several steps are carried out to explore the data. Goal of this step is to get an understanding of the data structure, conduct initial preprocessing, clean the data, identify patterns and inconsistencies in the data (i.e. skewness, outliers, missing values) and build and validate hypotheses.

# Understanding

In the first part of EDA the data frame is evaluated for structure, columns included and data types. The goals of this step are to get a general understanding for the data set, check domain knowledge and get first ideas on topics to investigate. In this step some standard Pandas functions are used:

The unique values for every feature are printed to the console to get a deeper understanding about the feature values.

# Meaning of Features

By inspecting the columns and their unique values, a general understanding about the features can be build. The features can also be clustered into different categories:

**Classification labels**

* Churn — Whether the customer churned or not (Yes or No)

**Customer services booked**

* PhoneService — Whether the customer has a phone service (Yes, No)
* MultipleLines — Whether the customer has multiple lines (Yes, No, No phone service)
* InternetService — Customer’s internet service provider (DSL, Fiber optic, No)
* OnlineSecurity — Whether the customer has online security (Yes, No, No internet service)
* OnlineBackup — Whether the customer has online backup (Yes, No, No internet service)
* DeviceProtection — Whether the customer has device protection (Yes, No, No internet service)
* TechSupport — Whether the customer has tech support (Yes, No, No internet service)
* StreamingTV — Whether the customer has streaming TV (Yes, No, No internet service)
* StreamingMovies — Whether the customer has streaming movies (Yes, No, No internet service)

**Customer account information**

* Tenure — Number of months the customer has stayed with the company
* Contract — The contract term of the customer (Month-to-month, One year, Two year)
* PaperlessBilling — Whether the customer has paperless billing (Yes, No)
* PaymentMethod — The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
* MonthlyCharges — The amount charged to the customer monthly
* TotalCharges — The total amount charged to the customer

**Customers demographic info**

* customerID — Customer ID
* Gender — Whether the customer is a male or a female
* SeniorCitizen — Whether the customer is a senior citizen or not (1, 0)
* Partner — Whether the customer has a partner or not (Yes, No)
* Dependents — Whether the customer has dependents or not (Yes, No)

# Data Preprocessing for EDA

The analysis shows 11 missing values for “TotalCharges”. The respective data entries (=rows) will be deleted for simplicity.

# Hypothesis Building

Looking at the features included in data and connecting them to their potential influence on customer churn, the following hypotheses can be made:

* The longer the contract duration the less likely it is that the customer will churn as he/she is less frequently confronted with the termination/prolongation decision and potentially values contracts with reduced effort.
* Customers are willing to cancel simple contracts with few associated product components quicker and more often than complexer product bundles — for bundles customers value the reduced administrative complexity. They might also be hesitant to cancel a contract, when they depend on the additional service components (e.g. security packages).
* Customers with spouses and children might churn less to keep the services running for their family.
* Tenure, contract duration terms and number of additional services are assumed to be among the most important drivers of churn.
* More expensive contracts lead to increased churn as the chances to save money by changing providers might be higher.
* Senior citizens tend to churn less due to the extended effort associated with terminating contracts.

# Data Exploration

The plot shows a class imbalance of the data between churners and non-churners. To address this, resampling would be a suitable approach. To keep this case simple, the imbalance is kept forward and specific metrics are chosen for model evaluations.

Plot insights:

* Churning customers have much lower tenure with a median of ca. 10 months compared to a median of non-churners of ca. 38 months.
* Churning customers have higher monthly charges with a median of ca. 80 USD and much lower interquartile range compared to that of non-churners (median of ca. 65 USD).
* TotalCharges are the result of tenure and MonthlyCharges, which are more insightful on an individual basis.

Plot insights:

* Senior citizens churn rate is much higher than non-senior churn rate.
* Churn rate for month-to-month contracts much higher that for other contract durations.
* Moderately higher churn rate for customers without partners.
* Much higher churn rate for customers without children.
* Payment method electronic check shows much higher churn rate than other payment methods.
* Customers with InternetService fiber optic as part of their contract have much higher churn rate.

# Check for Outliers in Numerical Features

No outliers in numerical features detected with the IQR method — no adjustments made.

# Data Cleaning

# Feature Engineering Actions

Based on the data types and the values, following actions are defined to preprocess/engineer the features for machine readibility and further analysis:

**Columns removed**

* customerID: not relevant

**No action**

* SeniorCitizen

**Label encoding** The following features are categorical and each take on 2 values (mostly yes/no) — therefore are transformed to binary integers

* gender
* Partner
* Dependents
* Churn
* PhoneService
* PaperlessBilling

**One-Hot Encoding** The following features are categorical, yet not ordinal (no ranking) but take on more than 2 values. For each value, a new variable is created with a binary integer indicating if the value occured in a data entry or not (1 or 0).

* MultipleLines
* InternetService
* OnlineSecurity
* OnlineBackup
* DeviceProtection
* TechSupport
* StreamingTV
* StreamingMovies
* Contract
* PaymentMethod

**Min-Max Scaling** Values of numerical features are rescaled between a range of 0 and 1. Min-max scaler is the standard approach for scaling. For normally distributed features standard scaler could be used, which scales values around a mean of 0 and a standard deviation of 1. For simplicity we use min-max scaler for all numerical features.

* tenure
* TotalCharges
* MonthlyCharges

# Step 4: Feature Engineering

In feature engineering, the steps identified at the end of EDA are executed. Additionally, a new feature is generated from extisting features and a correlation analysis is conducted after all features have been transformed to numerical.

# Step 5: Train-Test-Split

For conduction of model training and testing steps, the data set is split into 80% training data and 20% test data. The “Churn” column is defined as the class (the “y”), the remaining columns as the features (the “X”).

# Step 6: Model Evaluation Metrics

For performance assessment of the chosen models, various metrics are used:

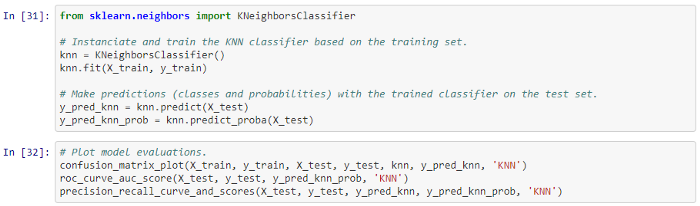
* **Feature weights:** Indicates the top features used by the model to generate the predictions
* **Confusion matrix:** Shows a grid of true and false predictions compared to the actual values
* **Accuracy score:** Shows the overall accuracy of the model for training set and test set
* **ROC Curve:** Shows the diagnostic ability of a model by bringing together true positive rate (TPR) and false positive rate (FPR) for different thresholds of class predictions (e.g. thresholds of 10%, 50% or 90% resulting to a prediction of churn)
* **AUC (for ROC):** Measures the overall separability between classes of the model related to the ROC curve
* **Precision-Recall-Curve:** Shows the diagnostic ability by comparing false positive rate (FPR) and false negative rate (FNR) for different thresholds of class predictions. It is suitable for data sets with high class imbalances (negative values overrepresented) as it focuses on precision and recall, which are not dependent on the number of true negatives and thereby excludes the imbalance
* **F1 Score:** Builds the harmonic mean of precision and recall and thereby measures the compromise between both.
* **AUC (for PRC):** Measures the overall separability between classes of the model related to the Precision-Recall curve

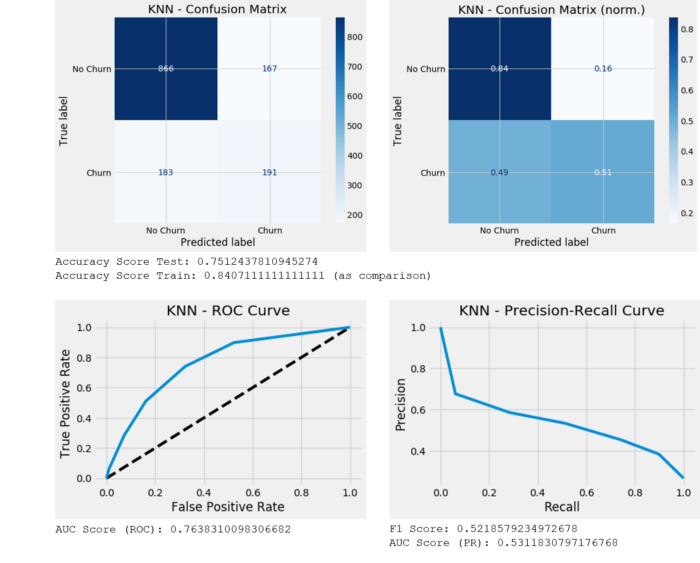
# Step 7: Model Selection, Training, Prediction and Assessment

In the beginning we will test out several models and measure their performance by several metrics. Those models will be optimized in a later step by tuning their hyperparameters. The models used include:

* **K Nearest Neighbors** — fast, simple and instance-based
* **Logistic Regression** — fast and linear model
* **Random Forest** — slower but accurate ensemble model based on decision trees
* **Support Vector Machines** — slower but accurate model used here in the non-linear form

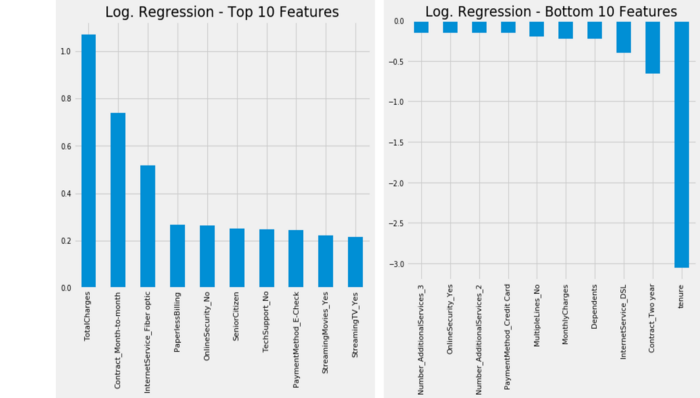
**K Nearest Neighbors**

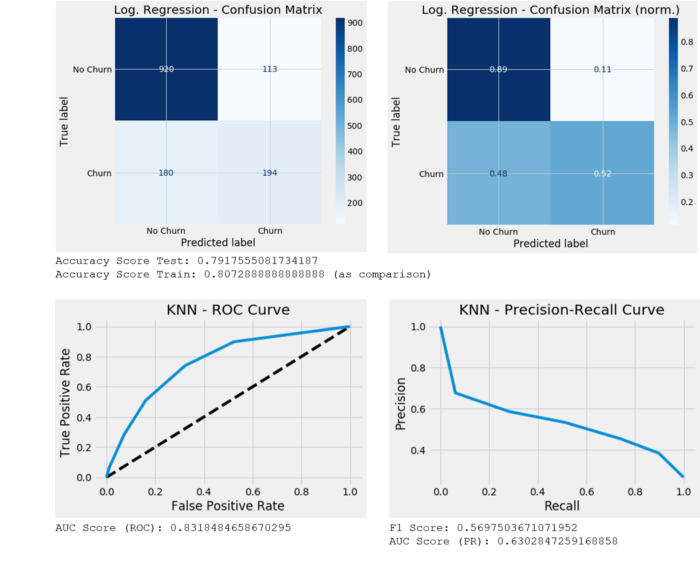




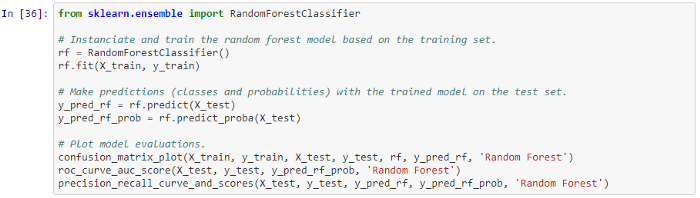
**Logistic Regression**

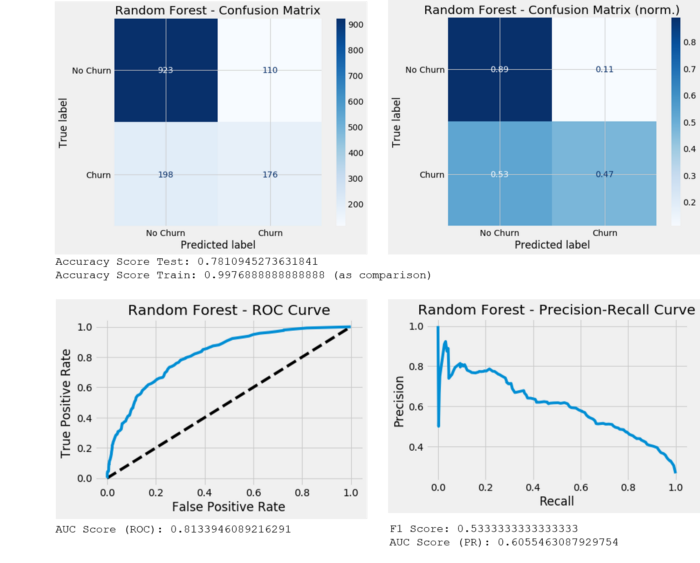




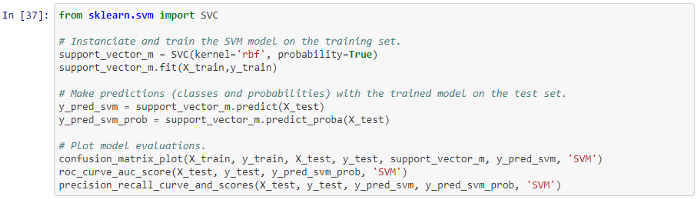


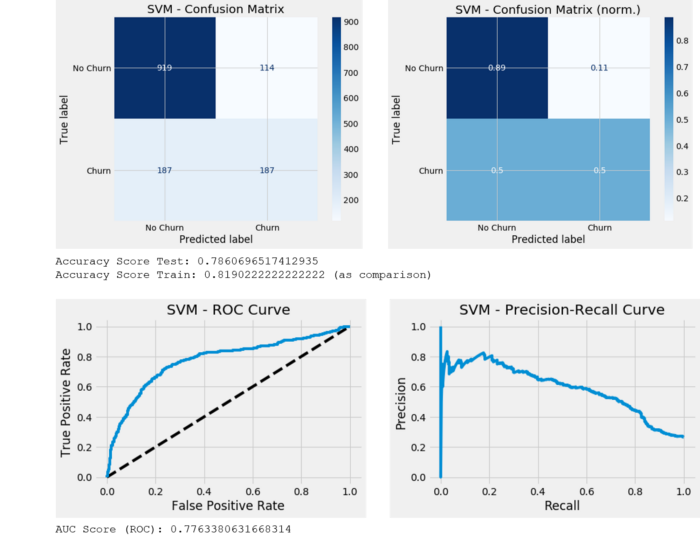
**Random Forest**





**Support Vector Machine**



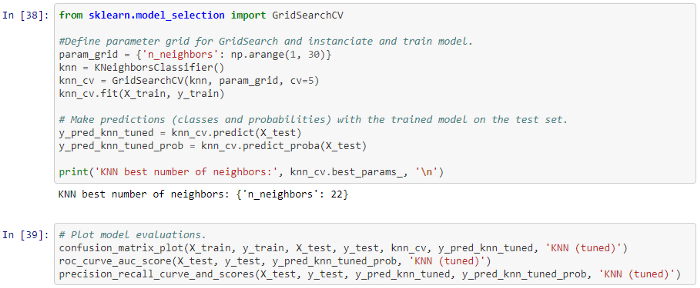


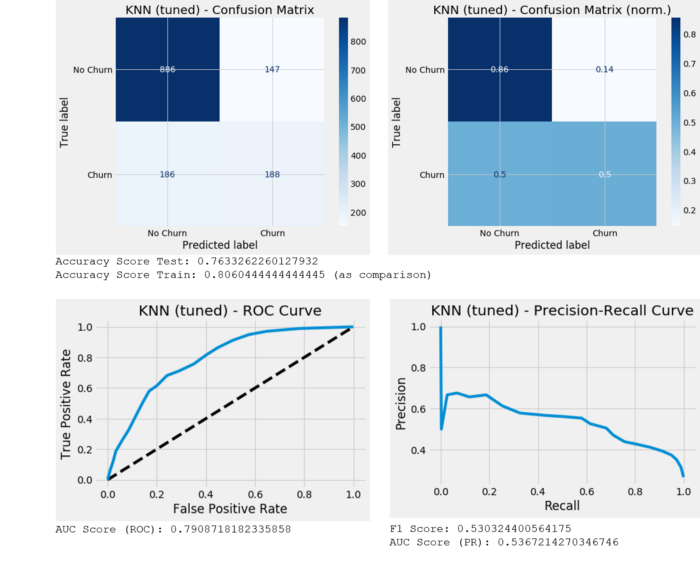
# Step 8: Hyperparameter Tuning/Model Improvement

To address a potential bias stemming from the specific split of the data in the train-test-split part, cross-validation is used during hyperparameter tuning with Grid Search and Randomized Search. Cross validations splits the training data into in a specified amount of folds. For each iteration one fold is held out as “training-dev” set and the other folds are used as training set. Result of cross-validation is k values for all metrics on the k-fold CV.

**K Nearest Neighbors (Optimized)**

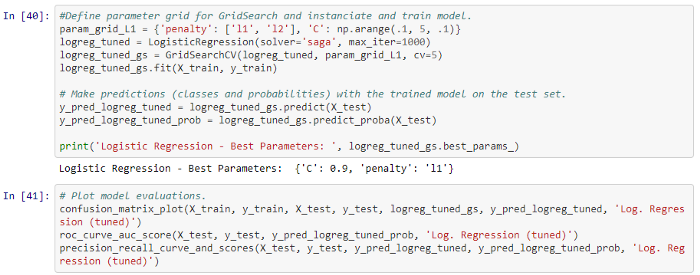
For KNN GridSearch CV is used to determine the optimal number of neighbors (k) leading to the best model performance.





**Logistic Regression (Optimized)**

For Logistic Regression GridSearchCV is used to determine the best model while applying different values of L1 or L2 regularization to turn the impact of non-meaningful feature to zero (L1) or to simplify the model by relativizing strong patterns that are picked up during training (L2).



**Random Forest (Optimized)**

For the Random Forest model RandomizedSearchCV is used to optimize for several hyperparameters including n\_estimators, max\_features, max\_depth, criterion and bootstrap.

**Support Vector Machine (optimized)**

For SVM GridSearchCV is used to determine the C value for the optimal margin around the support vector.

**Add-on: Feed Forward Neural Network**

Although the data set is relatively small and neural networks generally require lots of training data to develop meaningful prediction capabilities, a simple neural network is employed for a quick comparison to the other approaches.

# Summary

## Model Summary

Looking at model results, the best accuracy on the test set is achieved by the neural network with 0,7996. Given the high imbalance of the data towards non-churners, it makes sense to compare F1 scores to get the model with the best score on jointly precision and recall. This would also be the neural network with a F1 score of 0,5948.

Given the scores of the best performing models, it can be observed that F1 scores are not much above 50%. Further optimization efforts should be carried out to achieve a higher scores and thereby increase prediction power for more business value.

## Hypotheses Check

Looking at the evaluation results, specifically the feature weights from the logistic regression, the hypotheses can be directionally supported or refused:

* **Contract duration:** Contract duration month-to-month is the second biggest driver of churn → supported
* **Number of additional services:** This feature does not rank among the top features → refused
* **Partners and children:** Having children ranks as the fourth feature that drives not churning, but strength is relatively low → partially supported
* **Tenure:** High tenure ranks as the strongest factor for not churning and the strongest feature overall. This is also supported by the boxplot in the EDA step. → supported
* **Monthly payment:** Total payments, which is the product of tenure and monthly payment ranks as the strongest factor for churn. Indirectly, high monthly payments lead to churn. However, tenure is the highest driver of not churning → refused
* **Senior citizens:** Senior citizens does not have high feature weights. Also the ratio of senior citizens who churn is much higher than that of non-churners → refused