**CAPSTONE PROJECT**

**PREDICTION OF INDICATORS OF CERVICAL CANCER USING BIOPSY AS THE MAIN TARGET VARIABLE**

**MENTOR:** **PRESENTED BY:**

DIPANJAN GOSWAMI AJAY SANKAR P.D

LAVANYA K

ROHIT KUMAR SINGH

SALMAN MOHAMMED

**INTRODUCTION**

**1.1 CERVICAL CANCER**

The uterine cervix is the lowest portion of a woman's uterus (womb), connecting the uterus with the vagina. Cervical cancer occurs when the cells of the cervix grow abnormally and invade other tissues and organs of the body. When it is invasive, this cancer affects the deeper tissues of the cervix and may have spread to other parts of the body (metastasis), most notably the lungs, liver, bladder, vagina, and rectum. However, cervical cancer is slow-growing, so its progression through precancerous changes provides opportunities for prevention, early detection, and treatment. Better means of detection have meant a decline in cervical cancer in the U.S. over the decades.

**1.2 AIM OF THE PROJECT**

Most women diagnosed with precancerous changes in the cervix are in their 20s and 30s, but the average age of women when they are diagnosed with cervical cancer is the mid-50s. This difference in the age at which precancerous changes are most frequently diagnosed and the age at which cancer is diagnosed highlights the slow progression of this disease and the reason why it can be prevented if adequate steps are taken.

We aim to predict these indicators of cervical cancer with a reasonable level of accuracy using collected medical data and aid in the early detection and subsequent treatment.

**PROBLEM STATEMENT:**

# Prediction of indicators of cervical cancer using Biopsy as the target variable versus using Hinselmann OR Schiller OR Citology - recommend which is the better for future predictions.

* Does combining Biopsy with any of the other tests rather than using solely Biopsy improve predictions?

**LITERATURE SUMMARY**

**1. Handle Imbalanced Classification Problems in machine learning?**

While performing the conventional machine learning on the data which is imbalanced the model will be inaccurate and biased. In this case number of observations in one class will be significantly lesser than other. This problem is predominant in scenarios where anomaly detection is crucial like electricity pilferage, fraudulent transactions in banks, identification of rare diseases, etc. Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have number of instances. This happens because Machine Learning Algorithms are usually designed to improve accuracy by reducing the error. Thus, they do not take into account the class distribution / proportion or balance of classes. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored.

Author: Upasana |Consultant of Data & Analytics in KPMG.

<https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>

**2. Algorithms for screening of Cervical Cancer: A chronological review**

There are various algorithms and methodologies used for automated screening of cervical cancer by segmenting and classifying cervical cancer cells into different categories. This study presents a critical review of different research papers published that integrated AI methods in screening cervical cancer via different approaches analyzed in terms of typical metrics like dataset size, drawbacks, accuracy etc. An attempt has been made to furnish the reader with an insight of Machine Learning algorithms like SVM (Support Vector Machines), GLCM (Gray Level Co-occurrence Matrix), k-NN (k-Nearest Neighbours), MARS (Multivariate Adaptive Regression Splines), CNNs (Convolutional Neural Networks), spatial fuzzy clustering algorithms, PNNs (Probabilistic Neural Networks), Genetic Algorithm, RFT (Random Forest Trees), C5.0, CART (Classification and Regression Trees) and Hierarchical clustering algorithm for feature extraction, cell segmentation and classification. This paper also covers the publicly available datasets related to cervical cancer. It presents a holistic review on the computational methods that have evolved over the period of time, in chronological order in detection of malignant cells.

Authors: Yasha Singh, Dhruv Srivastava, P.S. Chandranand & Dr. Surinder Singh

<https://arxiv.org/ftp/arxiv/papers/1811/1811.00849.pdf>

**3. Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches**

Cervical cancer, as the fourth most common cause of death from cancer among women, has no symptoms in the early stage. There are few methods to diagnose cervical cancer precisely at present. Support vector machine (SVM) approach is introduced in this paper for cervical cancer diagnosis. Two improved SVM methods, support vector machine-recursive feature elimination and support vector machine-principal component analysis (SVM-PCA), are further proposed to diagnose the malignant cancer samples. The cervical cancer data are represented by 32 risk factors and 4 target variables: Hinselmann, Schiller, Cytology, and Biopsy. All four targets have been diagnosed and classified by the three SVM-based approaches, respectively. Subsequently, we make the comparison among these three methods and compare our ranking result of risk factors with the ground truth. It is shown that SVM-PCA method is superior to the others.

Author: Wen Wu | Department of Blood Transfusion, Jinan Military General Hospital, Jinan, China

<https://ieeexplore.ieee.org/abstract/document/8070120>

**DATA DESCRIPTION AND EDA**

**2.1 DATA SET:**

The dataset, “Cervical Cancer Risk Factors for Biopsy” was obtained from the UCI Repository. The data was collected by Kelwin Fernandes, Jamie S. Cardoso, and Jessica Fernandes in 2017 at ‘Hospital Universitario de Caracas’ in Caracas, Venezuela. The dataset contains habits, demographic information, and medical history of 858 patients from the hospital. There are many missing values in this dataset, due to many patients not answering questions because of privacy concerns. The dataset consists of 858 instances, with 36 attributes.

([https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29#](https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29))

**1.3 VARIABLES CONSIDERED FOR ANALYSIS**

# The dataset consists of 36 variables and records of 858 women patients. Of the 36 variables 4 variables are the target variables.

# *Below is the detailed description of each of the variables.*



**TARGET VARIABLES:**

# The target variables are four tests and are characterized by ‘1’ or ‘0’ in our data set where ‘1’ represents a malignant tumor and ‘0’ indicates benign tumor.

They are detailed below:

**SCHILLER**

In this test, Schiller's iodine solution is applied to the cervix under direct vision. Normal cervical mucosa contains glycogen and stains brown, whereas abnormal areas, such as early cervical cancer, do not take up the stain. The abnormal areas can then be biopsied and examined histologically. The composition of Schiller's iodine is the same as Lugol's iodine, the latter being more concentrated. When Schiller's iodine is not available, Lugol's iodine can be used as an alternative.

Schiller's test is not specific for cervical cancer, as areas of inflammation, ulceration and keratosis may also not take up the stain.

**CYTOLOGY**

Cytology: The medical and scientific study of cells. Cytology refers to a branch of pathology, the medical specialty that deals with making diagnoses of diseases and conditions through the examination of tissue samples from the body.

Cytologic examinations may be performed on body fluids (examples are blood, urine, and cerebrospinal fluid) or on material that is aspirated (drawn out via suction into a syringe) from the body. Cytology also can involve examinations of preparations that are scraped or washed (irrigated with a sterile solution) from specific areas of the body. For example, a common example of diagnostic cytology is the evaluation of cervical smears (referred to as the Papanicolaou test or Pap smear).

Screening is performed using cervical cytology (Pap test) or a human papillomavirus (HPV) test, or a combination of the two tests.

**HIENSELMAN**

In the first experiments, colposcopic examination was almost impossible to perform because of the distance from the focus, that was no more than 80 mm. Hinselmann tried to solve this problem by pulling out the uterine cervix. The examined part is anemised by this procedure, which can prejudice the final result and a small amount of blood might leak as well. Besides that, a patient can feel pain if the portio is held by thin forceps.

Colposcopy: (Ancient Greek: κόλπος, translit. kolpos, lit. 'hollow, womb, vagina' + skopos "look at") is a medical diagnostic procedure to examine an illuminated, magnified view of the cervix and the tissues of the vagina and vulva. Many premalignant lesions and malignant lesions in these areas have discernible characteristics which can be detected through the examination. It is done using a colposcope, which provides an enlarged view of the areas, allowing the colposcopist to visually distinguish normal from abnormal appearing tissue and take directed biopsies for further pathological examination.

**BIOPSY**

Cervical biopsy is a procedure to remove tissue from the cervix to test for abnormal or precancerous conditions, or cervical cancer. Cervical biopsies can be done in several ways. The biopsy can remove a sample of tissue for testing.

Types of cervical biopsies include:

• Punch biopsy: This procedure uses a circular blade, like a paper hole puncher, to remove a tissue sample. One or more punch biopsies may be done on different areas of the cervix.

• Cone biopsy: This procedure uses a laser or scalpel to remove a large cone-shaped piece of tissue from the cervix.

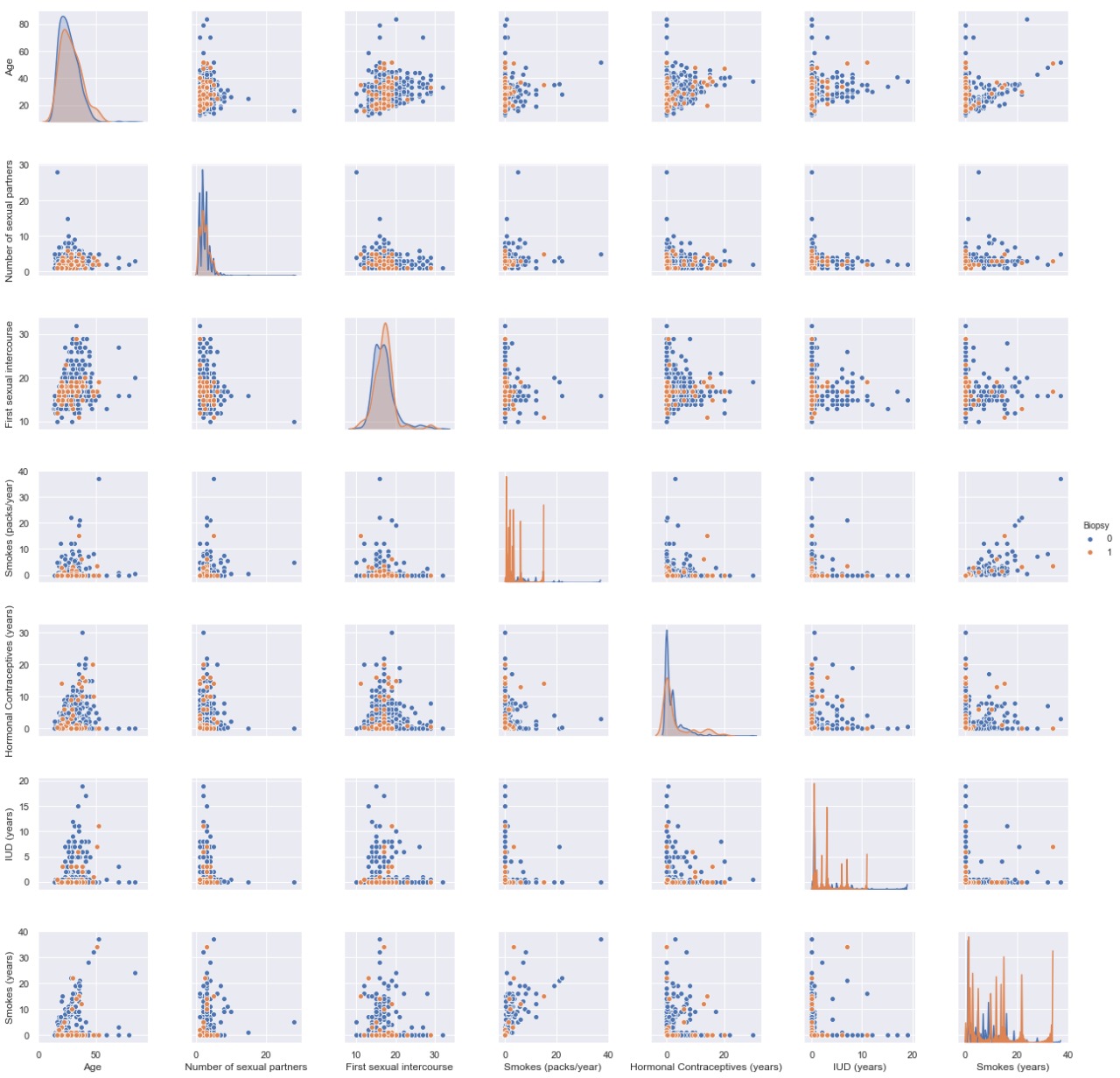
• Endocervical curettage (ECC): This procedure uses a narrow instrument called a curette to scrape the lining of the endocervical canal. This is an area that can’t be seen from the outside of the cervix.

**EXPLORATORY DATA ANALYTICS**

**INTRODUCTION:**

EDA is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of the data.

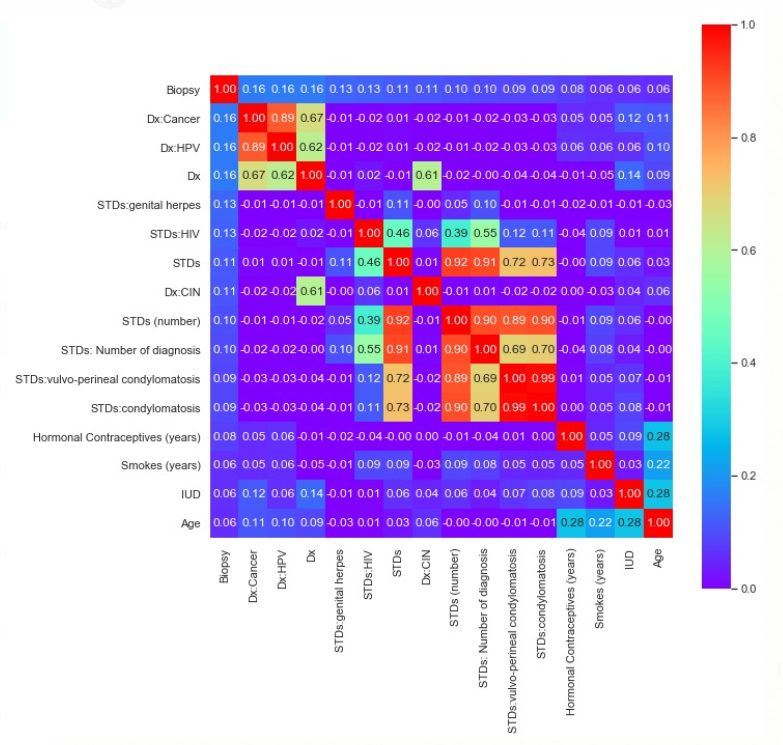
**COMPARISON OF BIOPSY RESULTS FOR ATTRIBUTES**



Human papilloma virus (HPV) is given to be the prime causer of cervical cancer [5]. In our investigation above, it seems not to be detected at all by biopsy screening method, suggesting it has no effect on cervical cancer or that screening method is very inaccurate.

For positive readings of cervical cancer by Biopsy method, there is an increase for its column-average in each feature. Except for the following features, there is a decrease. 1) number of sexual partners 2) Hormonal contraceptives 3) STDs:pelvic inflammatory disease

**HEATMAP OF ATTRIBUTE CORRELATION**



We can see that Hormonal Contraceptives and Age, IUD and Age have an effect on Biopsy.

### **3.2.1 DISTRIBUTION OF CLASS IN ALL FOUR TARGET VARIABLES**

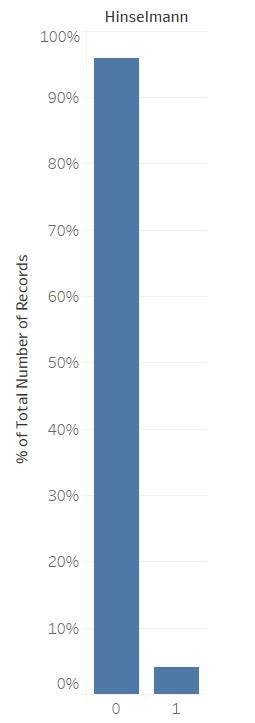
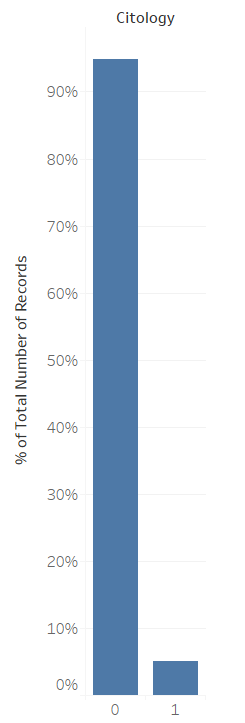
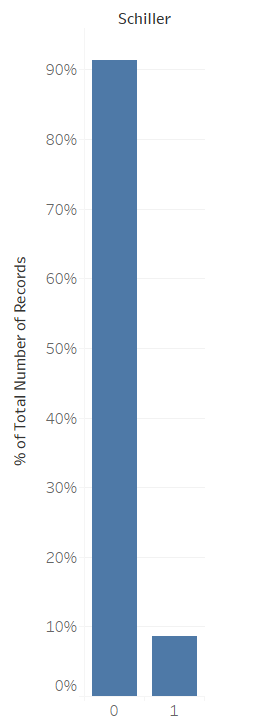
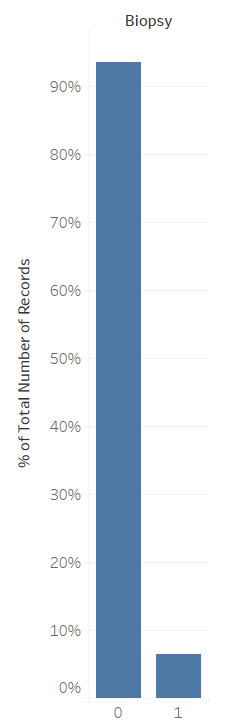
********

Figure: % Distribution in Target Variables

The classes in the Target variables of the dataset are highly imbalanced with 95.92%, 94.87%, 91.38%, 93.59% diagnosed as negative for the test of Citology, Hinselmann, Schiller and Biopsy respectively.

### **3.2.2 NUMBER OF PATIENTS WITH RESPECT TO STDs, SMOKING AND BIOPSY**

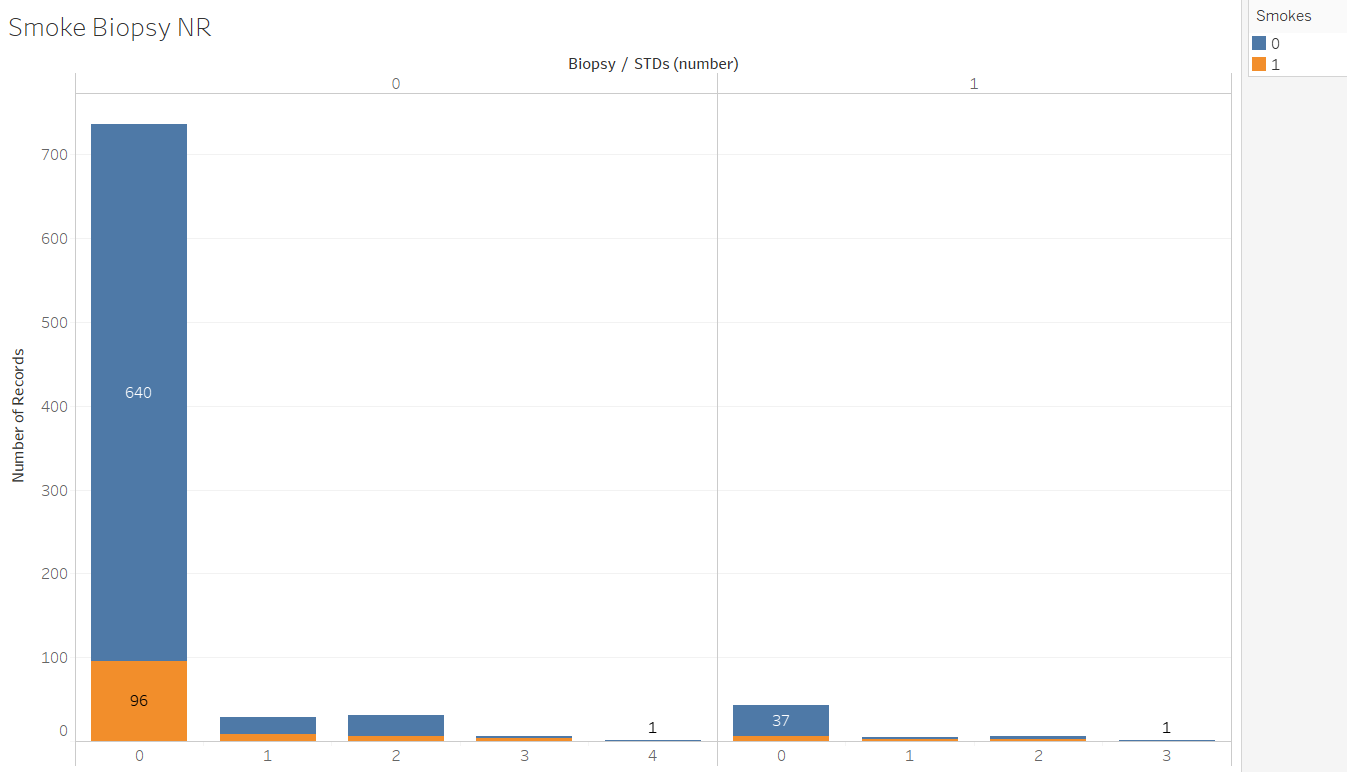


Figure: Distribution with respect to STDs, Smokers and Biopsy

We can clearly see that the patients who have been diagnosed with Biopsy are very low and smoking as a factor does not seem to have a significant impact in this.

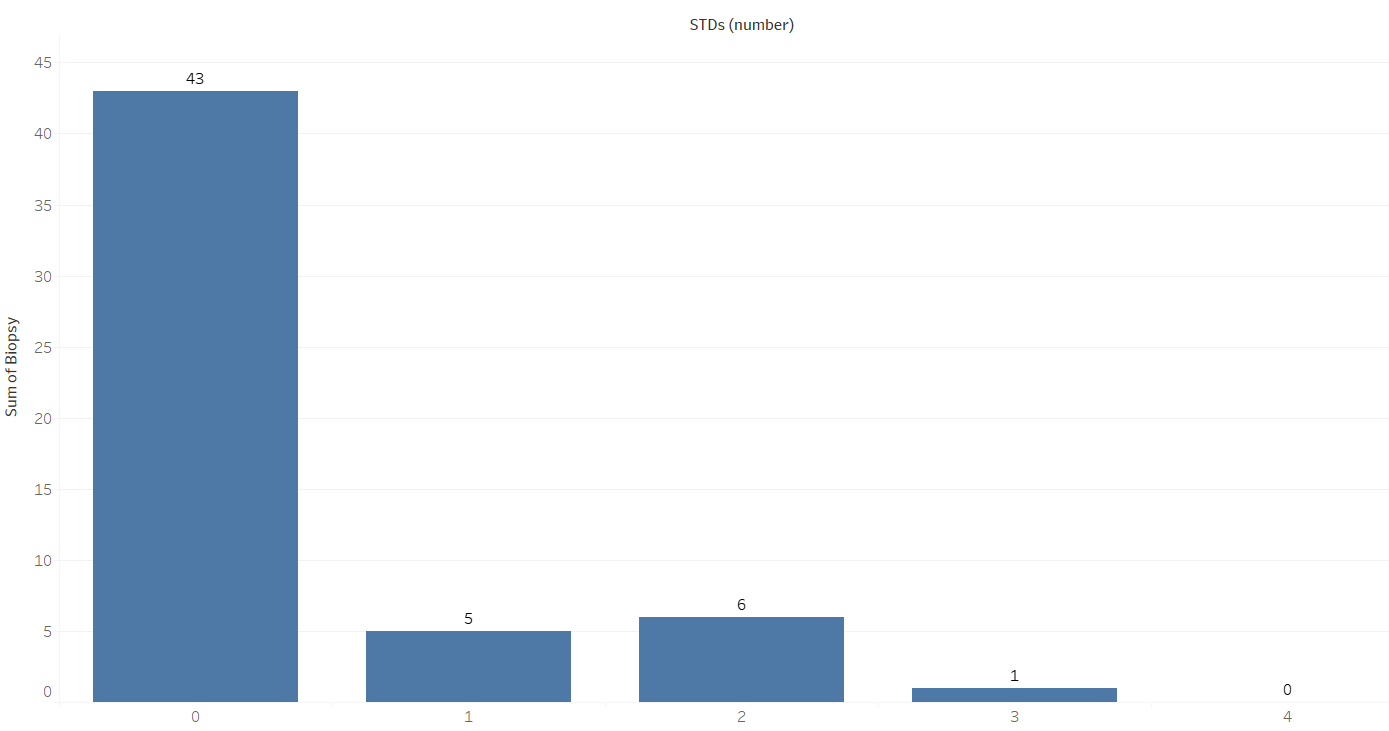


Figure: Count of positive results on Biopsy with respect to STDs

The number of people with STDs who were diagnosed with biopsy are comparatively low. We can see that only a handful have had multiple STDs from the above plot.

**3.2.3 POSITIVE BIOPSY VS OTHERS**

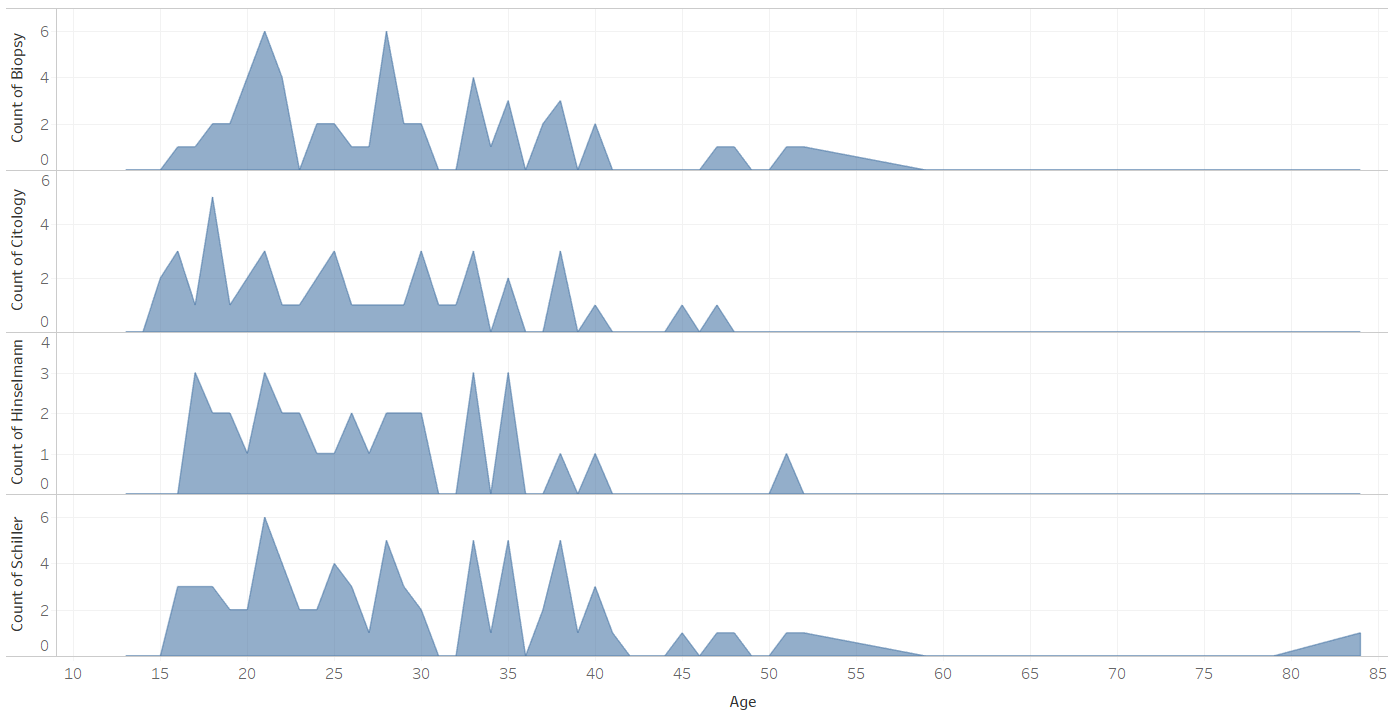


Figure: Count of all the Target Variables VS Age

The patients affected by cervical cancer is between 15 to 63 as per the majority of the tests.

From the frequency distributions, we can conclude that although there are slight variances in the diagnosis from test to test, cervical cancer is mostly affecting the young to middle aged women.

**3.2.4 POSITIVE BIOPSY VS NUMBER OF SEXUAL PARTNERS**

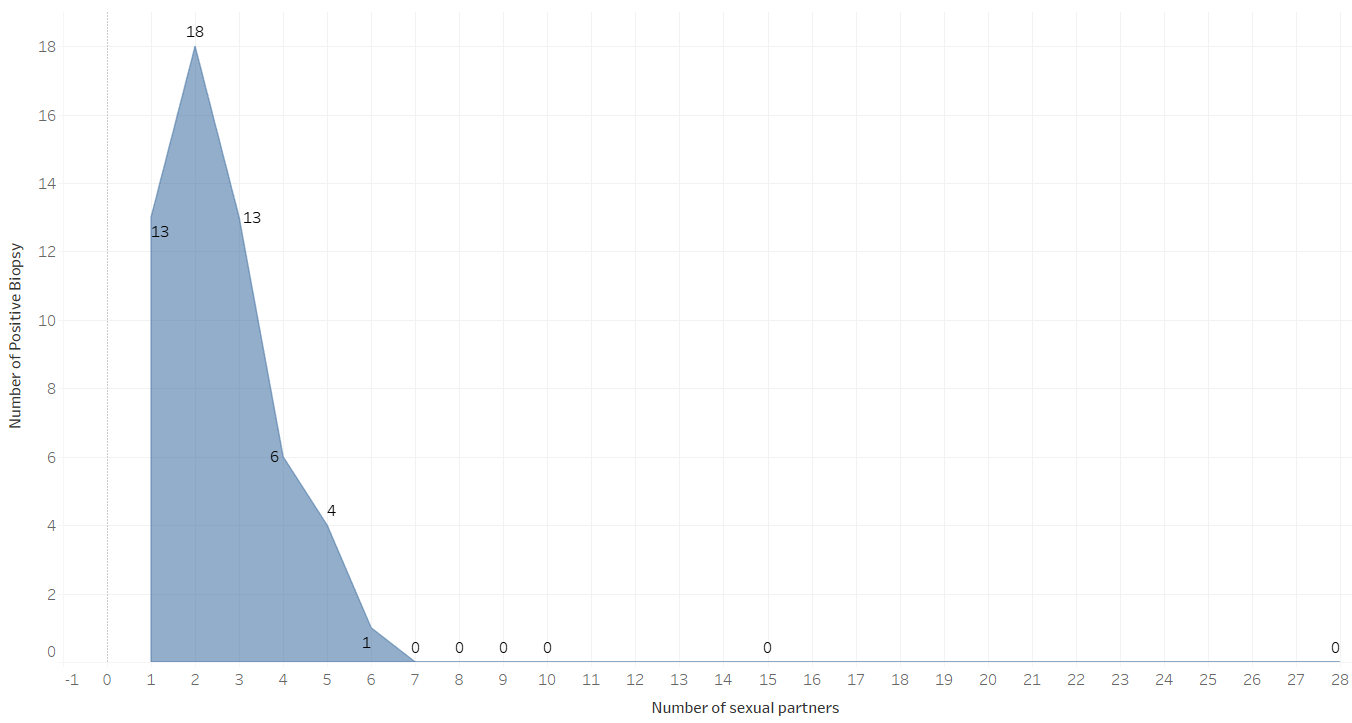


Figure: Count of Positive Biopsy VS Number of Sexual Partners

The greatest number of cervical cancers diagnosed through biopsy are for women who have had one to five sexual partners.

### **3.2.5 BIOPSY WITH NUMBER OF PREGNANCIES**

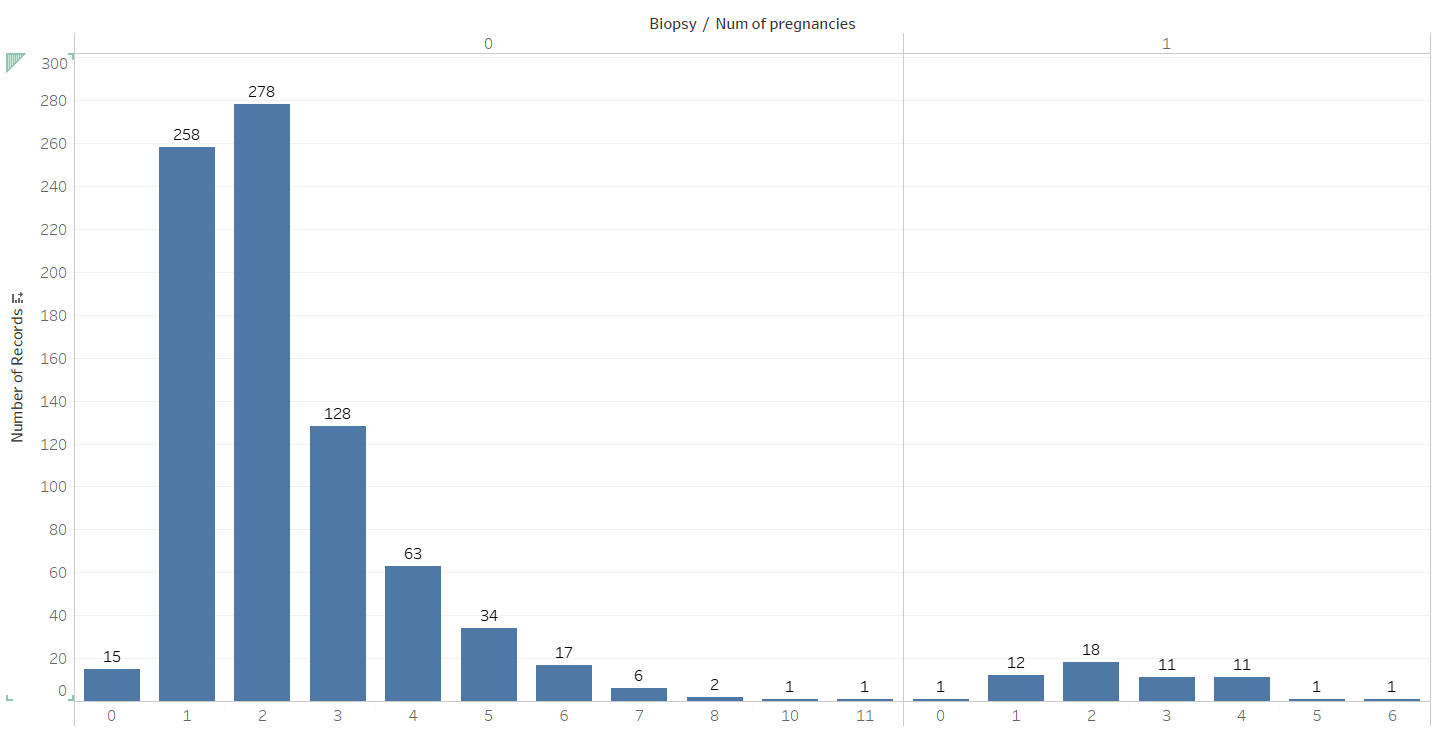


Figure: Biopsy with Number of Pregnancies

When looking at the proportion of detection through biopsy, we can conclude that women with who had one to four pregnancies are at a higher risk of cervical cancer.

**3.2.6 COUNT OF BIOPSY WITH RESPECT TO FIRST SEXUAL INTERCOURSE**

The maximum number of positive detections through Biopsy is when the first sexual intercourse age is between the range of 14 to 19. From this we can conclude that having sexual intercourse at a young age can increase the risk of cervical cancer significantly.

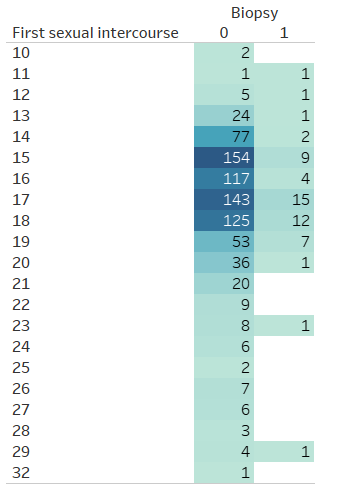


Figure: Count of Biopsy with respect to First Sexual Intercourse

# **DATA CLEANING**

**MISSING VALUE TREATMENT**

For the categorical column like Smokes, Hormonal Contraceptives, IUD, Biopsy etc., the null values were replaced by the **mode**. And null values in the numerical column like Age, Number of sexual partners etc. present were replaced by their **median** values.

**OUTLIER TREATMENTS**

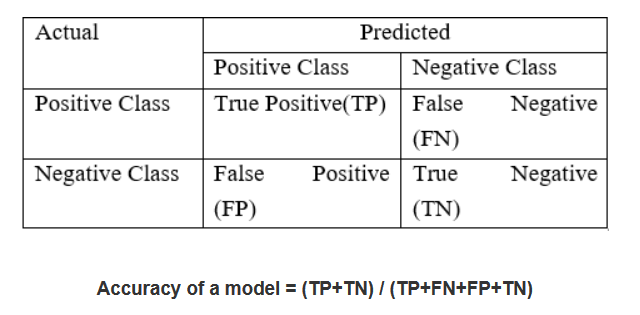
Since the data is medical in nature, outlier treatment is not advisable as each data point is significant and we need to include them as they are in our analysis.

**ARCHITECTURE**

We have to take a different approach here from a normal machine learning flow because of the nature of our data. The conventional model evaluation methods do not accurately measure model performance when faced with imbalanced datasets.

Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes which have number of instances. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored. Thus, there is a high probability of misclassification of the minority class as compared to the majority class.

Evaluation of a classification algorithm performance is measured by the Confusion Matrix which contains information about the actual and the predicted class.

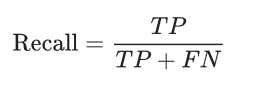
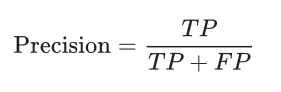


However, while working in an imbalanced domain accuracy is not an appropriate measure to evaluate model performance**.** For e.g.: A classifier which achieves an accuracy of 98 % with an event rate of 2 % is not accurate, if it classifies all instances as the majority class. And eliminates the 2 % minority class observations as noise.

To fully evaluate the effectiveness of our model, we must examine **precision** and **recall** as well. Unfortunately, precision and recall are often in tension. That is, improving precision typically reduces recall and vice versa.

**Precision :** What proportion of positive identifications was actually correct?

**Recall :** What proportion of actual positives was identified correctly?



## Handling Imbalanced Data

Dealing with imbalanced datasets entails strategies such as improving classification algorithms or balancing classes in the training data (data preprocessing) before providing the data as input to the machine learning algorithm. The later technique is preferred as it has wider application.

The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. This is done in order to obtain approximately the same number of instances for both the classes.

## Resampling Techniques

**RANDOM UNDER-SAMPLING**

Random Undersampling aims to balance class distribution by randomly eliminating majority class examples.  This is done until the majority and minority class instances are balanced out.

* **Advantages**
  + It can help improve run time and storage problems by reducing the number of training data samples when the training data set is huge.
* **Disadvantages**
  + It can discard potentially useful information which could be important for building rule classifiers.
  + The sample chosen by random under sampling may be a biased sample. And it will not be an accurate representative of the population. Thereby, resulting in inaccurate results with the actual test data set.

**RANDOM OVER-SAMPLING**

Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample.

* **Advantages**
  + Unlike under sampling this method leads to no information loss.
  + Outperforms under sampling
* **Disadvantages**
  + It increases the likelihood of overfitting since it replicates the minority class events.

**SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)**

This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models.

* **Advantages**
  + Mitigates the problem of overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances
  + No loss of useful information
* **Disadvantages**
  + While generating synthetic examples SMOTE does not take into consideration neighboring examples from other classes. This can result in increase in overlapping of classes and can introduce additional noise
  + SMOTE is not very effective for high dimensional data

**TENTATIVE LIST OF ALGORITHMS & INITIAL APPROACH**

Since our problem is a classification problem, we will be using the following algorithms in modelling:

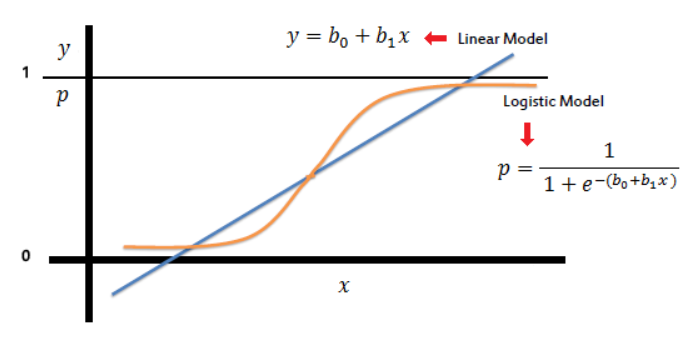
* Logistic Regression
* Tree Based Classifiers / Regressors
  + Decision Tree
  + Random Forest

**LOGISTIC REGRESSION**

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

* A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
* Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.



**Assumptions or Requirements of Logistic Regression:**

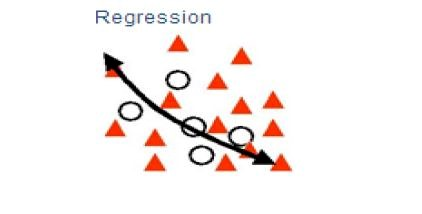
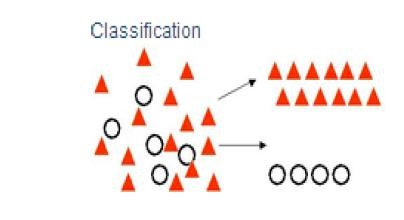
1. First, binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
2. Second, logistic regression requires the observations to be independent of each other.  In other words, the observations should not come from repeated measurements or matched data.
3. Third, logistic regression requires there to be little or no multicollinearity among the independent variables.  This means that the independent variables should not be too highly correlated with each other.
4. Fourth, logistic regression assumes linearity of independent variables and log odds.  although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.
5. Finally, logistic regression typically requires a large sample size.

**DECISION TREE (CART)**

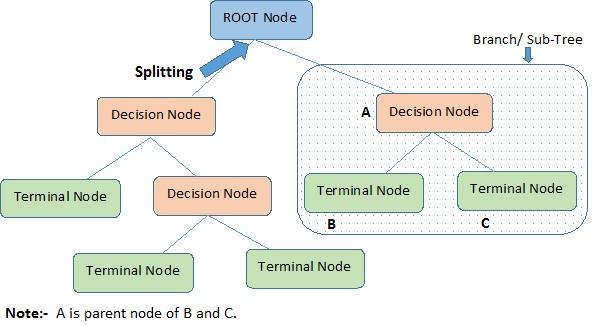
A Decision tree (CART) is a schematic, tree-shaped diagram used to determine a course of action or show a statistical probability. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

**Types of Decision Tree**

* **Classification Trees:** where the Dependent variable is categorical and the tree is used to identify the "class" within which a Dependent variable would likely fall into.
* **Regression Trees:** where the Dependent variable is continuous and tree is used to predict its value. (e.g. the price of a house, or a patient's length of stay in a hospital).



**Layout / flow of Decision Tree**



**Advantages of CART**

* Simple to understand, interpret, visualize.
* Decision trees implicitly perform variable screening or feature selection.
* Can handle both numerical and categorical data. Can also handle multi-output problems.
* Decision trees require relatively little effort from users for data preparation.
* Nonlinear relationships between parameters do not affect tree performance.

**Disadvantages of CART**

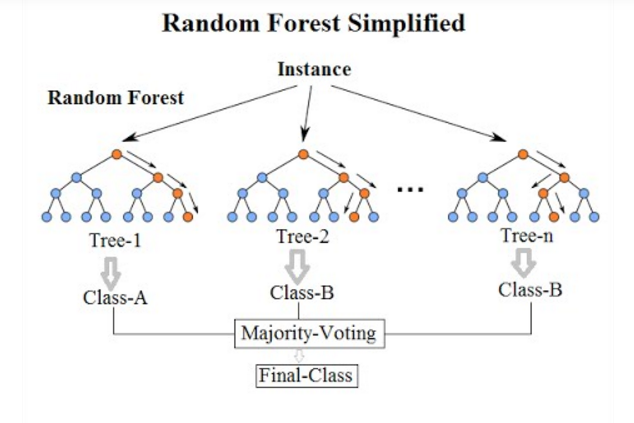
* Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting.
* Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This is called variance, which needs to be lowered by methods like bagging and boosting.
* Greedy algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees, where the features and samples are randomly sampled with replacement.
* Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the data set prior to fitting with the decision tree.

**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.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is, that it can be used for both classification and regression problems.



Random Forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, we don’t have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. Like I already said, with Random Forest, you can also deal with Regression tasks by using the Random Forest regressor.

Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

**Advantages of Random Forest**

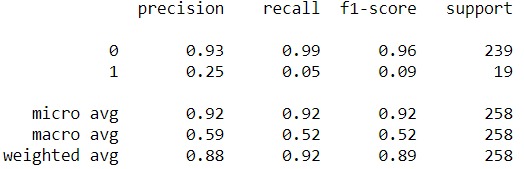
* There is no need for feature normalization
* Individual decision trees can be trained in parallel
* Reduced overfitting
* Require almost no input preparation
* Performs implicit feature selection
* It’s very quick to train

**Disadvantages of Random Forest**

* No interpretability

**INITIAL APPROACH**

1. Made a Decision Tree model on the entire data to see its overall behaviour and check entropy.

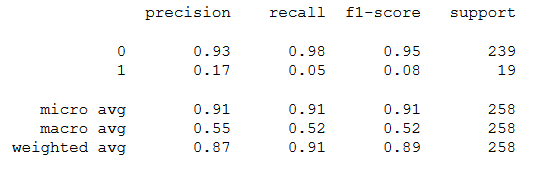


While the accuracy of this simple DT was high, it had very low recall and precision as expected because of the class imbalance.

1. Used IBM SPSS® software to get the Co-efficient of variation, coefficient of correlation, skewness and kurtosis of the predictor variables to aid in feature selection on the following guidelines employing the meta features:
2. Features which have low variance i.e. low coefficient of variation are candidates for elimination.
3. Features which are relatively unrelated with other features i.e. low average correlation can be eliminated.
4. Features which have lower entropy i.e. lesser information content can be eliminated.
5. Features which have highly asymmetric distribution measured by skewness are more suitable to be removed.
6. Features with exhibit varying peaks measure in terms of kurtosis scan be eliminated.

***This however did not provide us any significant and actionable insights on our dataset.***

1. The next step we took was to make distinct and fixed **bins** of all the continuous variables in the dataset like Age, Number of Sexual Partners, Number of Pregnancies, etc. and use these as predictors for modelling. The initial DT was run on these and the results were compared.



As we can observe, there was no significant improvement in our results.

After the above attempts, the focus was shifted towards finding the most optimal way of feature selection; And for our data set, that was by running a **chi-square test for independence** on each of the individual predictor variables on all 4 target variables and checking the resultant p-value.

**Chi-Square Test for Independence**

A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.

In our case, we need to determine whether there is indeed a relationship between a predictor variable and any of the target variables to a significant degree. We only need to consider these features for our further analysis.

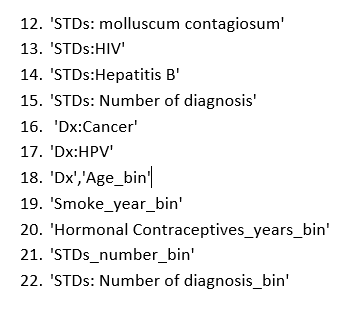
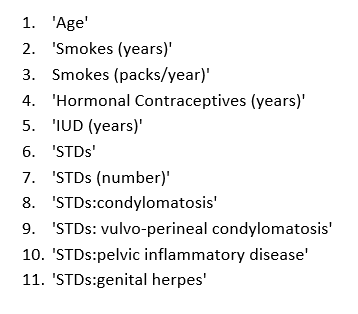
**The null hypothesis of the Chi-Square** test is that no relationship exists on the categorical variables being tested. I.e they are independent.

The p-value will tell us if our test results are significant or not. In order to perform a chi square test and get the p-value, you need two pieces of information:

* Degrees of freedom. That’s just the number of categories minus 1.
* The alpha level(α). The usual alpha or significance level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

*We reject the null hypothesis when the P-value is less than the set significance level.*

These are the attributes we got as significant from this chi-square method:



**IMPLEMENTATION**

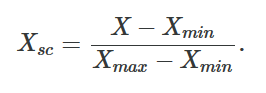
Modelling using the listed algorithms was done under multiple scenarios and results compiled for overview.

**To note:**

For Logistic Regression, the data was first normalized with the Min-Max Scaler:

*An alternative approach to Z-score normalization; In this approach, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers.*

*A Min-Max scaling is typically done via the following equation:*



**SCENARIO DESCRIPTIONS:**

**ORIGINAL DATA**

* Modelling was done on the original data after default data cleaning and scaling where necessary.

**DERIVED VARIABLES**

* New features were engineered and added from the intuitive understanding of the attributes in the data set and were added to the existing attributes.
  + YRSS: Years passed since patient had first sexual intercourse
  + NSPP: Number of sexual partners since first time as a percentage.
  + HPA: Hormonal Contraceptives usage by age
  + TPS: Total packets of cigarettes smoked
  + NPA: Number of pregnancies by age
  + NSA: Number of sexual partners by age
  + NYHC: number of years patient did not take Hormonal Contraceptives
  + APP: number of pregnancies by numbers of sexual partner
  + NHCP: number of years patient took Hormonal Contraceptives after first sexual intercourse as a percentage.

**IMPORTANT FEATURES**

* Using the “feature\_importances\_” feature of tree-based algorithms, only features that were significant to predicting the target variable Biopsy were considered for modelling.

**UNDER SAMPLING**

* Under-Sampling decreases the number of samples of the majority class on which the models are trained so that a more balanced proportion of representation of the minority class is achieved.

**OVER SAMPLING**

* Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample.

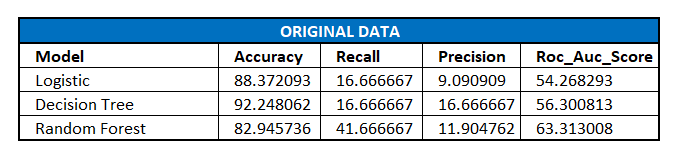
**SMOTE**

* Oversampling using SMOTE not only increases the size of the training data set, it also increases the variety. It creates new (artificial) training examples based on the original training examples and adds them as synthetic data points on which the model can be build.

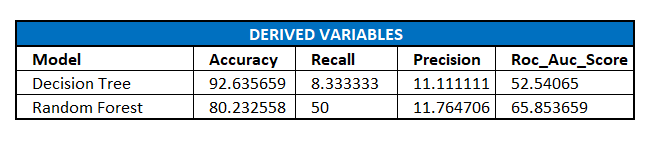
**RESULTS AND COMPARISON STUDY**

Since we are dealing with medical data, we will be focusing on **Recall** rather than Precision I.e, we will allow Type I error to creep in to our models because in our case, Type II error is far costlier.

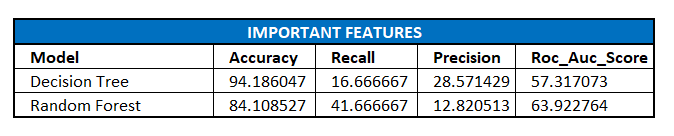
*Being diagnosed with cancer when you don’t have cancer (false positive), although not desirable, is far better than being diagnosed that you don’t have cancer when you do (false negative | Type II Error) – Even costing a life.*



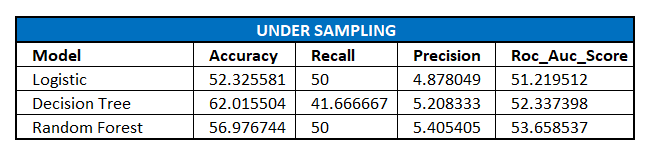
**INFERENCE:** In this case we have high accuracy; but we neither have precision nor recall. The model is not able to identify the affected people at a reasonable rate. The accuracy is high only because of the imbalance between the non-affected and affected people. We cannot rely on these.



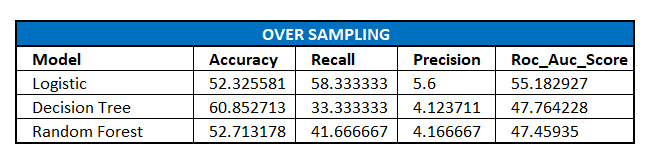
**INFERENCE:** Even after including the new derived variables, we are not able to achieve an optimal point for recall and accuracy. This case Random Forest is working the best with 50% recall and good AUC value, but Precision is still too low.



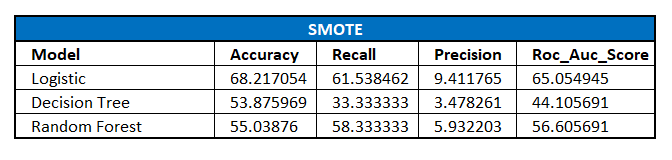
**INFERENCE:** On the derived variables, when important features were found out, these are the results we observe. It is not significantly different from the original data modelling. Even after finding imp features, we are unable to identify a satisfactory number of 1’s and the model is overfitting.



**INFERENCE:** When the data is under sampled, all the models are giving us similar results. But here we are increasing the bias error substantially as we are not considering all the data points. Even though model is working good on the recall part, the precision is almost negligible and we cannot accept this.



**INFERENCE:** Here we do not see a significant difference from when we ran the data under sampled.



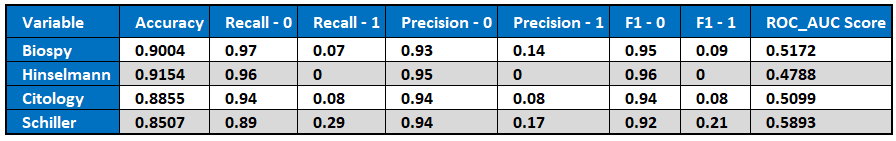
**INFERENCE:** Here we get a reasonable accuracy and recall rate with decent AUC value as well with the Logistic Regression Model.

Since we had already decided to forgo Precision for our model valuation, Logistic Regression with SMOTE is the overall best model we’ve seen so far for predicting the cancer indicator with Biopsy as the target variable**.**

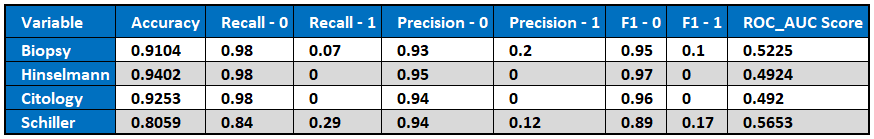
***Now that we’ve seen the model comparisons and results for target variable Biopsy, let’s check the same for the dataset using the OTHER target variables as well.***

We will be using SMOTE on Logistic Regression as well as Random Forrest on all the variables for comparison’s sake.

**LOGISTIC REGRESSION**

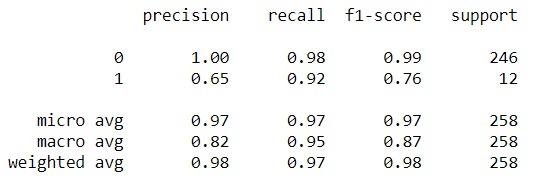


**RANDOM FOREST**



We can see that **Schiller** is the target variable that gives the most adequate results along with biopsy.

What if we use Schiller test results as an independent variable to build our model? We will run a Logistic Regression run over features selected through chi-square test with Schiller included as an independent variable.



**We see a drastic improvement across the board. Accuracy, Precision, Recall and F1 Score are almost perfect on test data!**

***Before we conclude***, we will use K-Means clustering here to check the stability of the data itself. Since we know that the data is of patients with or without cancer, we can see if that is what emerges from the data when run without specifying a target to train upon.

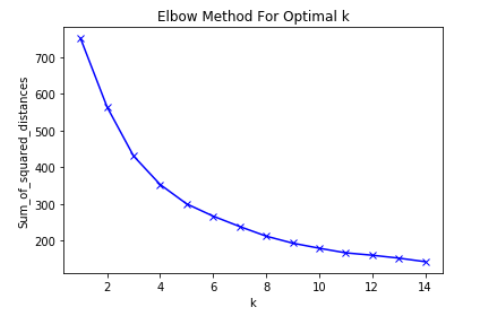
**K-Means Clustering**

K-means clustering is a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

1. The centroids of the K clusters, which can be used to label new data.
2. Labels for the training data (each data point is assigned to a single cluster)

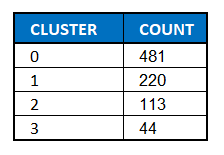
Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

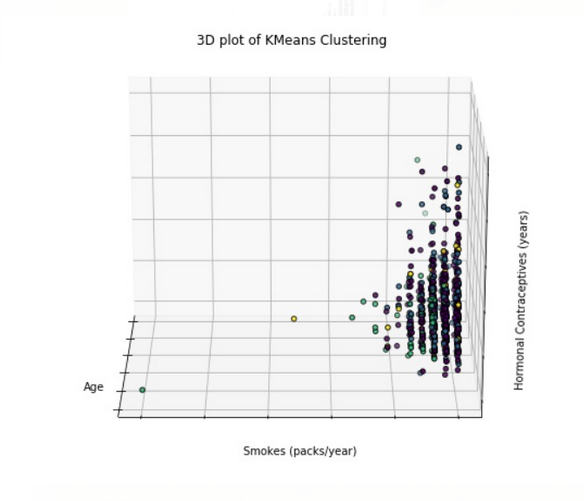
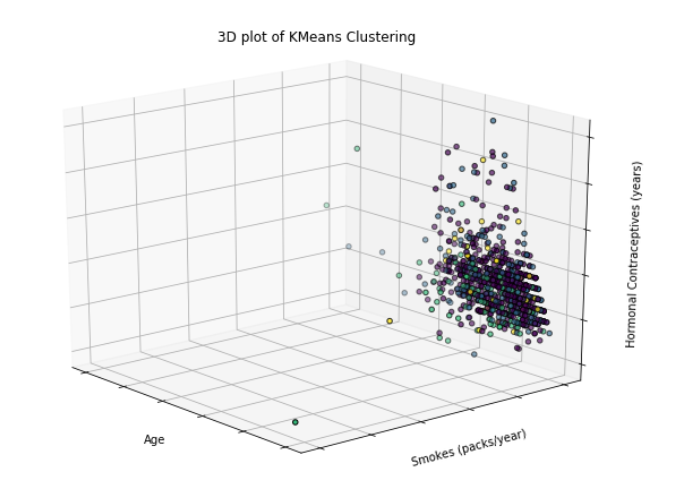
While modelling, the mean distance to the centroid as a function of K is plotted and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine K.



From the elbow plot, the most optimal K for our data here will be 4. We can already see a disparity here. Ideally, only two clusters should have formed: people with cancer and people without.

We will go ahead with the full process anyway. The result we got were:





We can clearly see that K-Means fails here as the data is unable to properly classify itself on the attributes at hand.

**CONCLUSION**

From the models built and the tests performed, Logistic Regression over data that is SMOTE oversampled is the best method to predict the indicators of cervical cancer in women from a Machine Learning perspective.

But from our study, it is clear that the original independent variables are not enough for the identification of these indicators to a reasonable accuracy just by themselves.

Due to the inherently unpredictable nature of Cervical Cancer, we cannot solely depend upon machine learning techniques here. The best we can achieve is around a 60 - 70% accuracy in stating that the patient **has** the indicators for cervical cancer, but with a high error rate of this being a false positive.

The only way to be certain is to have a full medical diagnosis with any of the stated tests, especially Biopsy; or combine the results of a Schiller’s test to our pre-built model containing the empirical data of potential the patient.

**APPENDIX**

**Python Codes**

**Tableau Link of EDA:**

* <https://public.tableau.com/profile/rohit.singh4001#!/vizhome/Cervical_Cancer_Group7/BiopsyVSNumberofSTDs>

**Reference:**

* <https://www.researchgate.net/publication/317348739_An_Univariate_Feature_Elimination_Strategy_for_Clustering_Based_on_Metafeatures>
* <https://stattrek.com/chi-square-test/independence.aspx>
* <https://www.saedsayad.com/logistic_regression.htm>
* <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>
* <https://en.wikipedia.org/wiki/Random_forest>
* <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>
* <https://www.datascience.com/blog/k-means-clustering>
* <https://en.wikipedia.org/wiki/Schiller%27s_test>
* <https://www.uptodate.com/contents/cervical-cancer-screening-tests-techniques-for-cervical-cytology-and-human-papillomavirus-testing>
* <https://www.ncbi.nlm.nih.gov/pubmed/22439022>
* <https://en.wikipedia.org/wiki/Colposcopy>
* <https://www.hopkinsmedicine.org/healthlibrary/test_procedures/gynecology/cervical_biopsy_92,p07767>
* <https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1111/tog.12279>
* <https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/>
* <https://arxiv.org/ftp/arxiv/papers/1811/1811.00849.pdf>
* <https://ieeexplore.ieee.org/abstract/document/8070120>