1. I thought of keeping the title of the thesis as **Analyzing the models in Diabetics Prediction**
2. As i am planning to use the **PIMA dataset** for the implementation and this dataset is the collection of medical diagnostic reports of 768 examples . The sample consist of 8 attribute values and one of the two possible outcomes is whether the patient test positive for diabetics indicated by one and test negative indicated by zero. And in the dataset there are about 268 tested positive and 500 tested negative.

**The Overall process of the thesis includes:**

1.Dataset collection

2.Data Analysis

3..Build model

4..Evaluation

* First it begins with the dataset collection as i mentioned using the PIMA dataset
* Followed by **Data preprocessing**  inorder to check for any missing values,noisy and inconsistent data.
* The next step is to choose the particular model for carrying out the training phase – As i have to train the data set using Logistic regression , SVM and decision tree and Random forest
* For the evaluation checking which model gives the best accuracy can be choosen.

**Model variable selection in Diabetics Prediction**

**1.Introduction:**

* Diabetics is the one of the metabolic disorders with the inappropriate raise of blood glucose level.
* Around 422 million people are diabetics patients and about 1.6 million deaths are related to diabetes every year.
* Over the past few decades the number of cases and the prevalence of diabetics are steadily increasing.
* Dibetics is classified as Type 1, Type2 and gestational diabetics . The condition where the pancreas will produce little or no insulin is type 1 diabetics. If the insulin is not absorbed by the cells or not produced in the enough quantity it is referred as type 2 diabetics .
* The medical experts and the data analysts collaborate continuously to make this system more accurate and useful in real life.
* To accurately predict the disorder a good model that can represent the presence of diabetics through input characteristics is required.
* Based on the model selection we select the most significant model and carry out the logistic regression approach and predict the diabetics with good accuracy.

**2.Model variable selection:**

* In order to select the best covariates for the model we proceed the selection based on p-value, AIC,BIC plotting and backward selection method
* After running the three different models we have 3 different AIC values.
* By comparing the AIC values , selecting the models having the lowest AIC values
* Selecting only those variables that would give us an efficient model.
* So, the model with the lowest AIC value will be selected for the final model building
* Lastly we do residual analysis to check if our assumptions are met.

**2.1 P-values:**

* If the selection criteria is based on the p-value then the choice for the significance level is either 0.05 or 0.10
* The optimum value of the significance level to decide which variable to be included in the model
* This suggestion assumes the absence of the few strong variables or the irrelevant variables in the dataset
* There is also the strong recommendations for using the p-value in the range of 0.15-0.20, although using the higher significance level has the disadvantages that some un-important variables may be included in the model.
* However we believe that the higher significance level for the variable selection should be considered so that the important variables related to the outcome are not missed and to avoid deleting the less significant variables that has a practical meaning .

**2.2 AIC:**

* AIC is the tool for the model selection which compares the different models.
* AIC attempts to select the model by balancing underfitting and overfitting
* Including too few variables often fails to capture the true relation and too many variables creates a problem.
* A model cannot precisely represent the true relation that exist in the data as there is some information loss in estimating the true relation through modelling .
* AIC tries to estimate the relative information loss compared with the other candidate models.
* Quality of the model is believed to be with the smaller information loss and it is important to select the model that best minimises that loss.
* Among the available models for the specific data, the model with the minimum AIC is the best.

**2.3 BIC:**

* BIC is the another variable selection criterion that is similar to AIC , but with the different penalty for the number of variable parameters included in the model.
* For the given dataset BIC is calculated for each of the candidate models , and the model corresponding to the minimum BIC value is chosen.
* Although there are similarities exist between the AIC and BIC both the criteria balance simplicity and the model fit differences exist between them.
* If our objective is to select the best model which provides the minimum predictive accuracy then AIC is superior
* However if the goal is to select the correct model that is consistent then BIC is superior , because BIC consistently selects the correct model from the candidate models that best represent the true model.
* For the large dataset the performance of both the criteria improves but with the different objectives.

**3.Application:**

**3.1 Dataset Description:**

* The PIMA dataset is from National institute of Diabetics and digestion & Kidney diseases.
* The dataset consist of 768 rows and 9 columns and there are 8 variables taken as the indicators of the dataset and the variable outcome stated that whether or not the person has the diabetics by showing the result value as 0 for NO and 1 for YES
* And the objective of the dataset is to diagnostically predict whether the patient has the diabetics based on the certain diagnostic measurements included in the dataset.

**3.2 Data Analysis:**

* We start the analysis with the data cleaning steps such as looking for the missing values, duplicate records and identifying for outliers in each co-variate.
* Checking for the missing values and the NULL-values in the given dataset.
* After checking for the missing values and NULL we can do further analysis on checking for the duplicates records in the dataset.
* Then proceed with checking for outliers and if we conclude finding the given data set is incomplete we will replace the rows contained zero values in the certain variables and then check the correlation between the variables.

**3.3 Model Variable selection in PIMA dataset:**

* we need to know that which variables highly predict the cause of diabetics from the pima dataset that influences the outcome.
* To find out which of these variables are important for predicting the relationship between the chance of diabetics and the outcome , we create several models using p-value, aic and bic
* I also have an idea of checking the how each variable performs followed by the combination of the variables
* Example code- glm(Outcome~ -BloodPressure-SkinThickness-Age,data=data,family="binomial")
* If for example model 1 and model 3 gives the lowest aic value then selecting only those variables as that would yield an efficient model
* With the significant variables we could come up with an conclusion that persons having pregnancy ,increasing glucose , Insulin,BMI, DiabeticsPedigreeFunction and skin thickness will likely have the diabetics
* And the final model will look like for ex:
* Outcome = -9.22 + 0.13\* Pregnancies + 0.03\* Glucose + 0.04\* SkinThickness + 0.005\* Insulin + 0.05\* BMI + 0.80\*DiabetesPedigreeFunction.
* Based on the final model we carry out the next step of validation by including the significant set of variables in the approach.

**3.4 Model Validation:**

* Perform the model validation by splitting the dataset into training and test the sample size then we measure the model accuracy in the each subset
* We use the logisitic regression approach and even try one or two different approaches
* ROC is used in validating the models diagnostic ability and to give an overall measure of goodness of classification
* Then we can try with different model approaches and check whether that model that gives us the best accuracy on both the training and the testing data.

**4. Conclusion:**

* In this we implemented the logistic regression and conduct the model diagnostics
* We used the PIMA dataset to study how the diabetics prediction indicator is associated to the variables provided in the dataset.
* In order to access the fit to our model we employed techniques which will led to the conclusion that the model specification is good fit for these data and the findings are meaningful.