```
% Generated by GrindEQ Word-to-LaTeX
\documentclass{article} % use \documentstyle for old LaTeX compilers

\usepackage[utf8]{inputenc} % 'cp1252'-Western, 'cp1251'-Cyrillic, etc.
\usepackage[english]{babel} % 'french', 'german', 'spanish', 'danish',
etc.
\usepackage{amsmath}
\usepackage{amssymb}
\usepackage{txfonts}
\usepackage{mathdots}
\usepackage[classicReIm]{kpfonts}
\usepackage{graphicx}

% You can include more LaTeX packages here


\begin{document}

%\selectlanguage{english} % remove comment delimiter ('%') and select
language if required


\noindent

\noindent

\noindent

\noindent First of all, we express our gratitude to the almighty who
blessed us with the zeal and enthusiasm to complete this research work
successfully. We are extremely thankful to our supervisor Dr Srinivas
Sethi, Asst Professor, Computer Science and engineering Department, IGIT
Sarang for his motivation and tireless efforts to help us to get deep
knowledge of the research area and supporting us throughout the life
cycle of our B.Tech. dissertation work.

\noindent                We are also thankful to Miss Sasmita Mishra, HOD
of computer science and engineering Department for her fruitful guidance
through the early years of chaos and confusions. We wish to thank the
faculty members and supporting staff of Computer Science and engineering
department for their full support and heartiest co-operation.

\noindent

\noindent

\noindent

\noindent

\noindent
Pritam Mishra.
```

```
\noindent
Piyush Prabhanjans.

\noindent
R Arbind Panda.

\noindent
Jyoti Jena.\textbf{}

\noindent \textbf{\eject }

\noindent \textbf{TABLE OF CONTENTS}

\noindent \textbf{}

\noindent \textbf{}

\begin{tabular}{|p{0.4in}|p{3.3in}|p{0.4in}|} \hline
\textbf{} & \textbf{ADMISSION PREDICTION\newline } & \textbf{} \\ \hline
\textbf{} & \textbf{ABSTRACT } & \textbf{} \\ \hline
\textbf{} & \textbf{LIST OF FIGURES} & \textbf{} \\ \hline
\textbf{} & \textbf{CHAPTERS\newline \newline 1.}Introduction\newline
\textbf{2}.Software-libraries used\newline \textbf{3.}Algorithm\newline
\textbf{\newline \newline CONCLUSION\newline } & \textbf{} \\ \hline
\end{tabular}

\textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{ABSTRACT}

\noindent \textbf{}

\noindent Data is the most important asset which is further processed to
produce useful information. Data science and Data analytics techniques
are widely used to generate useful patterns helpful for better
understanding and this project is for prediction of graduate admissions
from an Indian perspective by creating ML models with minimum MSE, RMSE
and maximum R- square score. However, the manual process of record
checking is time consuming and error prone due to the complexity of data
we use combination of linear and non-linear machine learning algorithms
like Linear Regression, Multiple Linear Regression and Random Forest
Regression and the models built in this work are predicting the
```

```
likelihood of a student taking up the admission into any university based
on the student data collected any administrative officials can use this
kind of an application to explore and analyze the patterns that are
affecting the student admission and come up with new strategies to
improve admission. \textbf{We have }created ML models to predict the
`Chance of Admit' with minimum MSE and RMSE and maximum R-Square score

\noindent By Building the following models:


\title{ Multiple linear Regressor (MLR)}\maketitle


\title{ Random Forest Regressor (RFR)}\maketitle


\title{ MLR with PCA}\maketitle


\title{ RFR with PCA}\maketitle

\noindent Finally plotted the actual and predicted values for all three
models\textbf{}

\noindent \textbf{}

\noindent \textbf{LIST OF GRAPHS}

\begin{tabular}{|p{1.1in}|p{3.5in}|} \hline
         SNO &                                            TITLE \\ \hline
           1. & Multiple Linear Regressor(Actual VS Predicted values) \\
\hline
           2. & Random Forest Regressor (Actual VS Predicted values) \\
\hline
           3. & Multiple Linear Regressor with Principal Component
Analysis \\ \hline
           4. & Random Forest Regressor with Principal Component Analysis
\\ \hline
\end{tabular}



\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{CHAPTERS}

\noindent \textbf{INTRODUCTION}

\noindent In this project, we will be using the admission \_ predict
dataset in csv format to predict the chances of students getting
```

```
admission by a university based on several academic performance
measurement. To yield the most accurate result, we will be going through
several steps such as data preprocessing, t-test, cross validation, model
selection, etc. to train a machine learning model, make prediction and
measure its performance.

\noindent Machine learning models such as Multiple Linear Regressor,
Random Forest Regressor and Multiple Linear Regressor with Principal
Component Analysis with minimum MSE and RMSE and maximum R-Square score
are created to predict the chance of admit\textbf{}

\noindent \textbf{}

\noindent The dataset contains several parameters:

\begin{enumerate}
\item  GRE Scores (out of 340)

\item  TOEFL Scores (out of 120)

\item  University Rating (out of 5)

\item  Statement of Purpose and Letter of Recommendation Strength (out of
5)

\item  Undergraduate GPA (out of 10)

\item  Research Experience (either 0 or 1)

\item  Chance of Admit (ranging from 0 to 1).
\end{enumerate}

\noindent The last column is the response variable, which is a continuous
value between 0 and 1, indicating the chances of getting
admission.\textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{}

\noindent \textbf{Multiple Linear Regressor (MLR)}
```

```
\noindent \textit{Multiple Linear Regression is one of the important
regression algorithms which models the linear relationship between a
single dependent continuous variable and more than one independent
variable'}

\noindent Moreover, Multiple Linear Regression is an extension of Simple
Linear regression as it takes more than one predictor variable to predict
the response variable

\noindent

\noindent The libraries used in this model are:-

\noindent NumPy

\noindent Pandas

\noindent \textbf{     }Scikit-Learn

\noindent Seaborn

\noindent           Matplotlib

\noindent           LinearRegression

\noindent      mean\_squared\_error

\noindent           train\_test\_split

\noindent           r2\_score

\noindent  It is a very powerful technique and can be used to understand
the factors that influence profitability.

\noindent  It looks at a relationship between the mean of the dependent
and     independent variables.

\noindent  It is prone to underfitting and sensitive to out layers.

\noindent  Simple implementation and highly interpretable.

\noindent

\noindent This is the plot of actual vs predicted values for the model
(MLR)

\noindent \textbf{}

\noindent \includegraphics*[width=7.01in, height=4.89in]{image18}

\noindent

\noindent ALGORITHMS USED:
```

\noindent OUTLIER REMOVAL:

\noindent The first main algorithm used here to find the various outliers present in the dataset using the construction of the SEABORN DISPLOT. The displot here basically shows the frequency of the observation of occurrences. As seen , when CHANCE OF ADMIT is less than 0.4 so we remove the outliers by using:

\noindent dataset.drop(dataset.index[list((np.where(dataset['Chance of Admit ']$\mathrm{<}$0.4)))],inplace=True)

\noindent

\noindent CORRELATION MATRIX:

\noindent A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

\noindent Correlation Matrix is here used to show the probabilistic dependencies of Chance of Admit on various other factors , and since all of them have quite a large share of dependency so we choose all of them.

\noindent

\noindent TRAIN_TEST_SPLIT:

\noindent

\noindent It is used to split arrays or matrices into random train and test subsets.

\noindent So basically what we are doing here is we are splitting the training and testing part in the ratio of 3:1, so that we can later on predict the values based upon the analysis.

\noindent

\noindent X_train,X_test,y_train,y_test = train_test_split(x,y,test_size=0.25,random_state = 5)

\noindent

\noindent MSE:

\noindent \textbf{Mean square error (MSE)~}is the average of the square of the errors. The larger the number the larger the error.~\textbf{Error}~in this case means the difference between the observed values y1, y2, y3, {\dots} and the predicted ones pred(y1), pred(y2), pred(y3), {\dots} We square each difference (pred(yn) -- yn)) ** 2 so that negative and positive values do not cancel each other out.

\noindent

\noindent RMSE:

\noindent Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data.

\noindent

\noindent R2 SCORE:

\noindent The~\textbf{r2 score~}varies between 0 and 100\%. It is \textit{the proportion of the variance in the dependent variable that is predictable from the independent variable(s).}

\noindent

\noindent HERE MSE, RMSE AND R2 SCORE are used to see the useability, ability and the accuracy of the model that we have designed.

\noindent

mse = mean_squared_error(y_test,y_test_predict)

rmse = np.sqrt(mean_squared_error(y_test,y_test_predict))

r2 = r2_score(y_test,y_test_predict)

\noindent

\noindent \textbf{MODEL ANALYSIS}

\noindent The MSE obtained in this model (MLR) is 0.0032448492350027444

\noindent The RMSE value obtained in this model (MLR) is 0.05696357814430853

\noindent The R-Squared score value obtained in this model (MLR) is 0.8127800620182747

\noindent Multiple~linear~regression~model~is the most popular type of~linear~regression~analysis. It is used to show the relationship between one dependent variable and two or more independent variables. In fact, everything you know about the simple~linear~regression~modeling~extends (with a slight modification) to the~multiple~linear~regression~models.

\noindent \textbf{}

\noindent \textbf{Random Forest Regressor (RFR)}

\noindent \textit{Random forest is a~Supervised Learning algorithm~which uses ensemble learning method for~classification and regression}

```latex
\noindent Random forest~is a~bagging~technique and~not a
boosting~technique. The trees in~random forests~are run in parallel.
There is no interaction between these trees while building the trees

\noindent In this model (RFR) I first imported all the necessary
libraries required for this model :

\begin{enumerate}
\item  Pandas

\item  Scikit-Learn

\item  Seaborn

\item  NumPy

\item  Matplotlib

\item  Random Forest Regressor

\item  mean\_squared\_error

\item  train\_test\_split

\item   r2\_score
\end{enumerate}

\noindent

\begin{enumerate}
\item  It operates by constructing a multitude of decision trees at
training time and outputting the class that is the~\textbf{mode}~of
the~\textbf{classes (classification)}~or~\textbf{mean prediction
(regression)}~of the individual trees

\item  A random forest is a meta-estimator (i.e. it combines the result
of multiple predictions) which~\textbf{aggregates many decision trees}.
\end{enumerate}

\noindent A Random Forest is an ensemble technique capable of performing
both regression and classification tasks with the use of multiple
decision trees and a technique called Bootstrap and Aggregation, commonly
known as~\textbf{bagging}. The basic idea behind this is to combine
multiple decision trees in determining the final output rather than
relying on individual decision trees.

\noindent Random Forest has multiple decision trees as base learning
models. We randomly perform row sampling and feature sampling from the
dataset forming sample datasets for every model. This part is called
Bootstrap.

\noindent This is the plot of actual vs predicted values for the
model(RFR)
```

```
\noindent

\noindent \includegraphics*[width=5.39in, height=3.69in]{image19}

\noindent

\noindent

\noindent ALGORITHMS USED:

\noindent TRAIN\_TEST\_SPLIT:

\noindent

\noindent It is used to split arrays or matrices into random train and
test subsets.

\noindent So basically what we are doing here is we are splitting the
training and testing part in the ratio of 3:1, so that we can later on
predict the values based upon the analysis.

\noindent

X\_train,X\_test,y\_train,y\_test =
train\_test\_split(x,y,test\_size=0.25,random\_state = 25)

\noindent

\noindent MSE:

\noindent \textbf{Mean square error (MSE)~}is the average of the square
of the errors. The larger the number the larger the
error.~\textbf{Error}~in this case means the difference between the
observed values y1, y2, y3, {\dots} and the predicted ones pred(y1),
pred(y2), pred(y3), {\dots} We square each difference (pred(yn) -- yn))
** 2 so that negative and positive values do not cancel each other out.

\noindent

\noindent RMSE:

\noindent Root Mean Square Error (RMSE) is a standard way to measure the
error of a model in predicting quantitative data.

\noindent

\noindent R2 SCORE:

\noindent The~\textbf{r2 score~}varies between 0 and 100\%. It is
\textit{the proportion of the variance in the dependent variable that is
predictable from the independent variable(s).}

\noindent
```

\noindent HERE MSE, RMSE AND R2 SCORE are used to see the useability, ability and the accuracy of the model that we have designed.

mse = mean_squared_error(y_test,y_test_predict)

rmse = np.sqrt(mean_squared_error(y_test,y_test_predict))

r2 = r2_score(y_test,y_test_predict)

\noindent

\noindent \textbf{MODEL ANALYSIS:}

\noindent The MSE obtained in this model (RFR) is 0.0038120131999999952

\noindent The RMSE value obtained in this model (RFR) is 0.06174150305912543

\noindent The R-Squared score value obtained in this model (RFR) is 0.8189295004736319

\noindent

\noindent Random Forest grows multiple decision trees which are merged together for a more accurate prediction. The logic behind the Random Forest model is that multiple uncorrelated models (the individual decision trees) perform much better as a group than they do alone.

\noindent

\noindent \textbf{Principal Component Analysis with Multiple Linear Regression}

\noindent (PLA WITH MLA)

\begin{enumerate}
\item Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning

\item The primary problem associated with high-dimensionality in the machine learning field is model overfitting, which reduces the ability to generalize beyond the examples in the training set
\end{enumerate}

\noindent

\noindent In this model (PCA with MLR), I imported all the necessary libraries required for this model

\begin{enumerate}
\item NumPy

\item Pandas

```
\item  Scikit-Learn

\item  Seaborn

\item  Matplotlib

\item  LinearRegression

\item  mean\_squared\_error

\item  train\_test\_split

\item  r2\_score

\item  StandardScaler

\item  PCA
\end{enumerate}

\noindent

\noindent This is the plot of actual vs predicted values for the model
(PCA with MLR)

\noindent \includegraphics*[width=7.01in, height=4.39in]{image20}

\noindent \textbf{}

\noindent \textbf{Goal of PCA:- }

\begin{enumerate}
\item  Identify patterns in data

\item  Detect the correlation between variables

\item  Reduce the dimensions of a d-dimensional dataset by projecting it
onto a k-dimensional subspace(where k$\mathrm{<}$d)
\end{enumerate}

\noindent ~

\noindent

\noindent \textbf{ALGORITHMS USED:}

\noindent CONVERTING THE CHANCE OF ADMISSION TO VARIOUS CATEGORIES:

\noindent The first major algorithm we have used here is we converted the
Chance of Admit column to the sequence of BAD, MEDIUM and GOOD, which
would allow an easier approach for the PCA to prepare dataset for
analysis.

dataset1 = dataset.copy()
```

```
dataset1['Chance of Admit ']=pd.cut(np.array(y),3, labels=["bad",
"medium", "good"])

dataset1['Chance of Admit '][:5]
```

\noindent IDENTIFYING THE PRINCIPAL COMPONENTS:

\noindent Here basically we are taking 4 principal components, because earlier it was seen that at components =3 the dataset did not produce 85\% of the total information(which would make the model more inaccurate), hence by adding one more principal component now we have 91.57\% of the total information which is considerable.

```
pca = PCA(n\_components=4)

principalComponents = pca.fit\_transform(x)

principalDf = pd.DataFrame(data = principalComponents

\noindent              , columns = ['principal component 1', 'principal
component 2','principal component 3','principal component 4'])
```

\noindent

\noindent Then we took the dataset produced from PCA along with the Chance of Admit column to prepare our final dataset on which we will perform our Multiple Linear Regression.

\noindent \textbf{TRAIN_TEST_SPLIT:}

\noindent

\noindent It is used to split arrays or matrices into random train and test subsets.

\noindent So basically what we are doing here is we are splitting the training and testing part in the ratio of 7:3, so that we can later on predict the values based upon the analysis.

\noindent In case of PCA it is important to note the training part must be atleast 70\% of the dataset other there might be inaccuracy in the desired output.

```
X\_train,X\_test,y\_train,y\_test =
train\_test\_split(x,y,test\_size=0.3,random\_state = 5)
```

\noindent

\noindent \textbf{MSE:}

\noindent \textbf{Mean square error (MSE)~}is the average of the square of the errors. The larger the number the larger the error.~\textbf{Error}~in this case means the difference between the observed values y1, y2, y3, {\dots} and the predicted ones pred(y1), pred(y2), pred(y3), {\dots} We square each difference (pred(yn) -- yn)) ** 2 so that negative and positive values do not cancel each other out.

\noindent

\noindent RMSE:

\noindent Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data.

\noindent

\noindent R2 SCORE:

\noindent The~\textbf{r2 score~}varies between 0 and 100\%. It is \textit{the proportion of the variance in the dependent variable that is predictable from the independent variable(s).}

\noindent

\noindent HERE MSE, RMSE AND R2 SCORE are used to see the useability, ability and the accuracy of the model that we have designed.

mse = mean_squared_error(y_test,y_test_predict)

rmse = np.sqrt(mean_squared_error(y_test,y_test_predict))

r2 = r2_score(y_test,y_test_predict)

\noindent \textbf{}

\noindent \textbf{MODEL ANALYSIS:}

\noindent The MSE obtained in this model (PCA with MLR) is 0.0035268267109371987

\noindent The RMSE value obtained in this model (PCA with MLR) is 0.05938709212393884

\noindent The R-Squared score value obtained in this model (PCA with MLR) is 0.7937833209433434

\noindent PCA is based on the Pearson correlation coefficient framework and inherits the following assumptions.

\begin{enumerate}
\item \begin{enumerate}
\item \textbf{\textit{Sample size:}}~Minimum of 150 observations and ideally a 5:1 ratio of observation to features (Pallant2010)

\item \textbf{\textit{Correlations:}}~The feature set is correlated, so the reduced feature set effectively represents the original data space.

```latex
\item  \textbf{\textit{Linearity:}}~All variables exhibit a constant
multivariate normal relationship, and principal components are a linear
combination of the original features.
\end{enumerate}
\end{enumerate}

\noindent \textbf{\textit{                                      }}

\noindent \textbf{\textit{CONCLUSION}}

\noindent
\title{Admissions Prediction model is build by using 3 different types of
algorithms}\maketitle

\noindent
\title{ 1.Multiple linear regressor (MLR)}\maketitle

\noindent
\title{2.Random Forest Regressor (RFR)}\maketitle

\noindent
\title{3.MLR with PCA on same data set by training we obtained results
regarding Mean square error (MSE), Root Mean Square Error (RMSE), R-
Squared score~and the model is build with minimum MSE and RMSE and
maximum R2 score.}\maketitle

\noindent
\title{4.RFR with PCA on same data set by training we obtained results
regarding Mean square error (MSE), Root Mean Square Error (RMSE) R-
Squared score and the model is built with minimum MSE and RMSE and
maximum R2 score.}\maketitle

\noindent
\title{ANALYSIS:}\maketitle

\begin{tabular}{|p{1.0in}|p{1.4in}|p{1.4in}|p{1.2in}|} \hline
\textbf{MODEL} & \textbf{MSE} & \textbf{RMSE} & \textbf{R2 SCORE} \\
\hline
\textbf{MLR} & 0.0032448492350027444 & 0.05696357814430853 &
0.8127800620182747 \\ \hline
\textbf{RFR} & 0.003812013199999952 & 0.003812013199999952 &
0.8189295004736319 \\ \hline
\textbf{MLR with PCA} & 0.0035268267109371987 & 0.05938709212393884 &
0.7937833209433434 \\ \hline
\textbf{RFR with PCA} & 0.004636949999999895 & 0.06809515401260202 &
0.7288734297643837 \\ \hline
\end{tabular}


\title{}\maketitle

\noindent Lowe\textbf{r}~values of~MSE and RMSE~and higher value of R2
Score indicate how well the regression model fits the observed data.
```

```
\noindent

\end{document}
```