Machine Learning-Enhanced Academic Achievement Forecaster

A Comprehensive Approach to Student Performance Analysis and Institutional Challenges

Abstract :

In today's education landscape, predicting student potential is crucial for driving innovation and societal progress. This research delves into a comprehensive analysis of student's Academic performance, considering factors like backgrounds, demographics, and academic achievements.

The study addresses challenges such as sponsorship dependency and language barriers affecting the dataset, aiming to predict student outcomes like on-time graduation, employment status, and attrition. It evaluates the effectiveness of various machine learning algorithms, including decision trees, random forests, logistic regression, XGBoost, and K-nearest neighbors. To boost model performance and efficiency, feature selection techniques, particularly K-best feature selection, are implemented to reduce dimensionality and improve predictive accuracy.

Initially, each algorithm is trained and tested on the dataset without feature selection to establish baseline accuracy metrics. Subsequently, the K-best feature selection method is integrated into the modeling pipeline to identify the most relevant features for prediction. The selected features are then used to retrain the models, and their performance is evaluated and compared against the baseline results.

This research contributes to the field of student performance analysis by demonstrating the effectiveness of feature selection techniques in enhancing the performance of machine learning models for predicting critical student outcomes and survival analysis.

Business Understanding / Project Goals :

- ➢ The students are considered to be the future of every possible innovation and beyond, In a university we have to consider the all round improvement of each student.
- ➢ Here we are analyzing Student performance Considering their Background , demographic factors , results etc. for a better enhancement of the current scenario we also dealt with respective business problems Such as

1. Student Attrition Analysis
2. On-Time Graduation
3. On-Campus Employment

Objective:

- ➢ Minimize Student Attrition and Maximize the education quality of the college and improve the rank of the college.
- ➢ Maximize On-Time Graduation
- ➢ Maximize On-Campus Employment

Constraints:

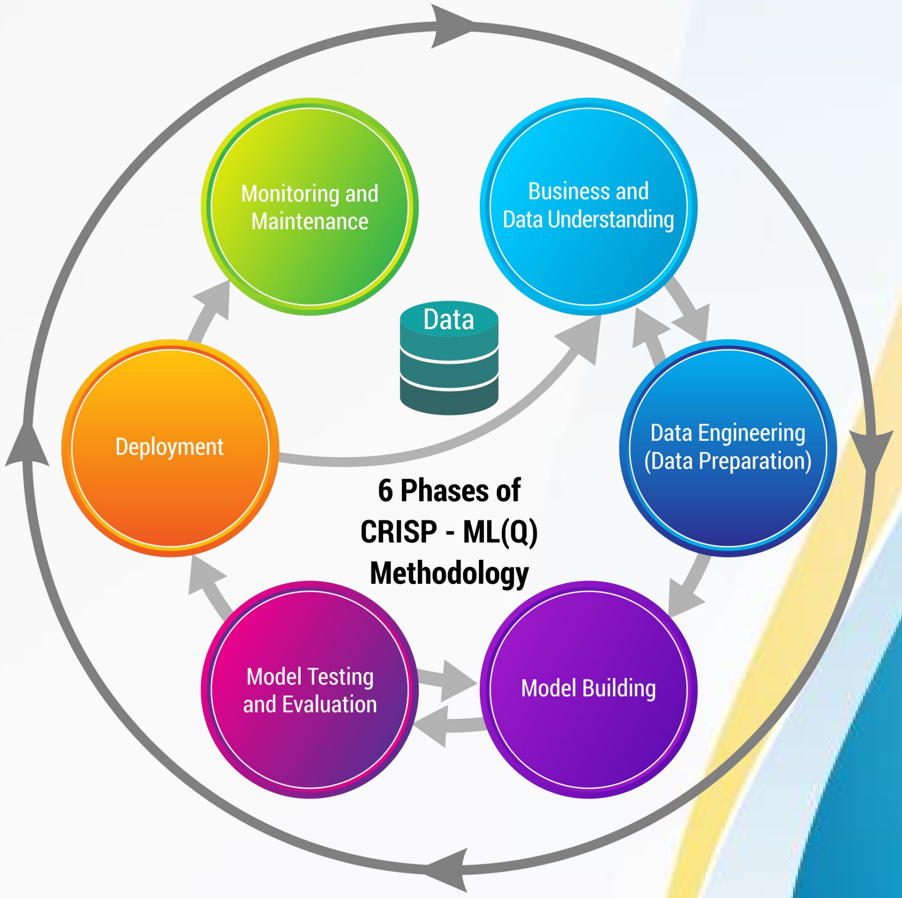
- ➢ Sponsorship dependency is affecting the study of participants.
- ➢ Language barrier is causing inadequacy in data understanding.
- ➢ Minority class is affecting one of the dataset and causing disparity.

Business pipeline:

CRISP-ML Methodology

The CRISP-ML process model describes six phases in the machine learning life cycle:

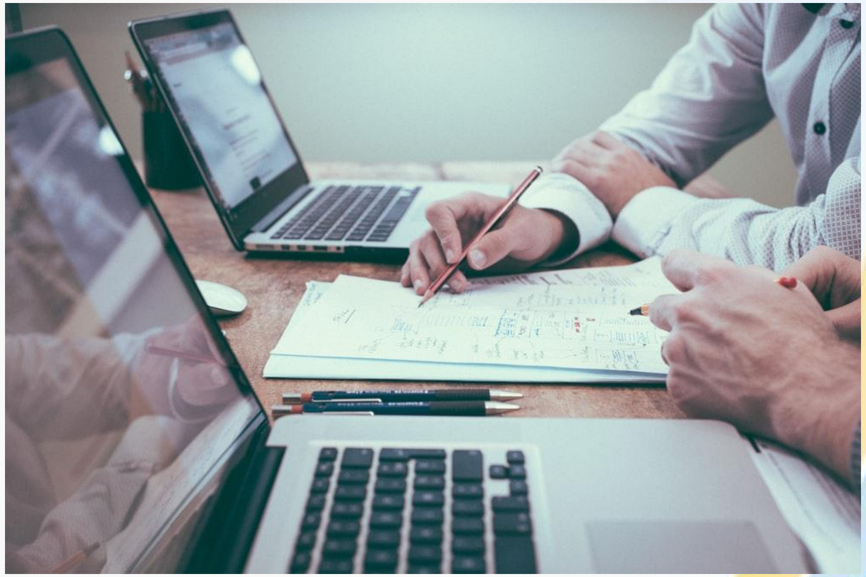
- Business and Data Understanding
- Data Preparation
- Machine Learning Model Engineering
- Model Testing & Evaluation
- Deployment
- Monitoring and Maintenance


Monitoring and Maintenance
Business and Data Understanding
Data
Deployment
Data Engineering (Data Preparation)
6 Phases of CRISP - ML(Q) Methodology
Model Testing and Evaluation
Model Building

Project Outline:

Why we must?:

- ➢ Manual maintenance or analysis of records involves burden and it is quite tedious task.
- ➢ Supplementary enhancement of the current process after analysis is an inflexible thing to do.
- ➢ The complexity increases tending to a very high probability of error



Proposed system:

- ➢ The expected system is to be computerized to provide optimum easiness to the users.
- ➢ This will be constructed in a object oriented trend, thinking in an abstract way by inducing different machine learning algorithms and tools.
- ➢ It will be easier to follow up students and their performance, by looking at the trend we can also predict the future scenarios.

Basis Requirements:

Recommended System Requirements:

- **Processors:**
Intel® Core™ i5 processor 4300M at 2.60 GHz or 2.59 GHz (1 socket, 2 cores, 2 threads per core), 8 GB of DRAMIntel® Xeon® processor E5-2698 v3 at 2.30 GHz (2 sockets, 16 cores each, 1 thread per core), 64 GB of DRAMIntel® Xeon Phi™ processor 7210 at 1.30 GHz (1 socket, 64 cores, 4 threads per core), 32 GB of DRAM, 16 GB of MCDRAM (flat mode enabled)
- **Disk space:** 2 to 3 GB
- **Operating systems:** Windows® 10, macOS*, and Linux*

Minimum System Requirements:

- **Processors:** Intel Atom® processor or Intel® Core™ i3 processor
- **Disk space:** 1 GB
- **Operating systems:** Windows* 7 or later, macOS, and Linux
- **Python versions:** Version 3 and more

Technical stacks:

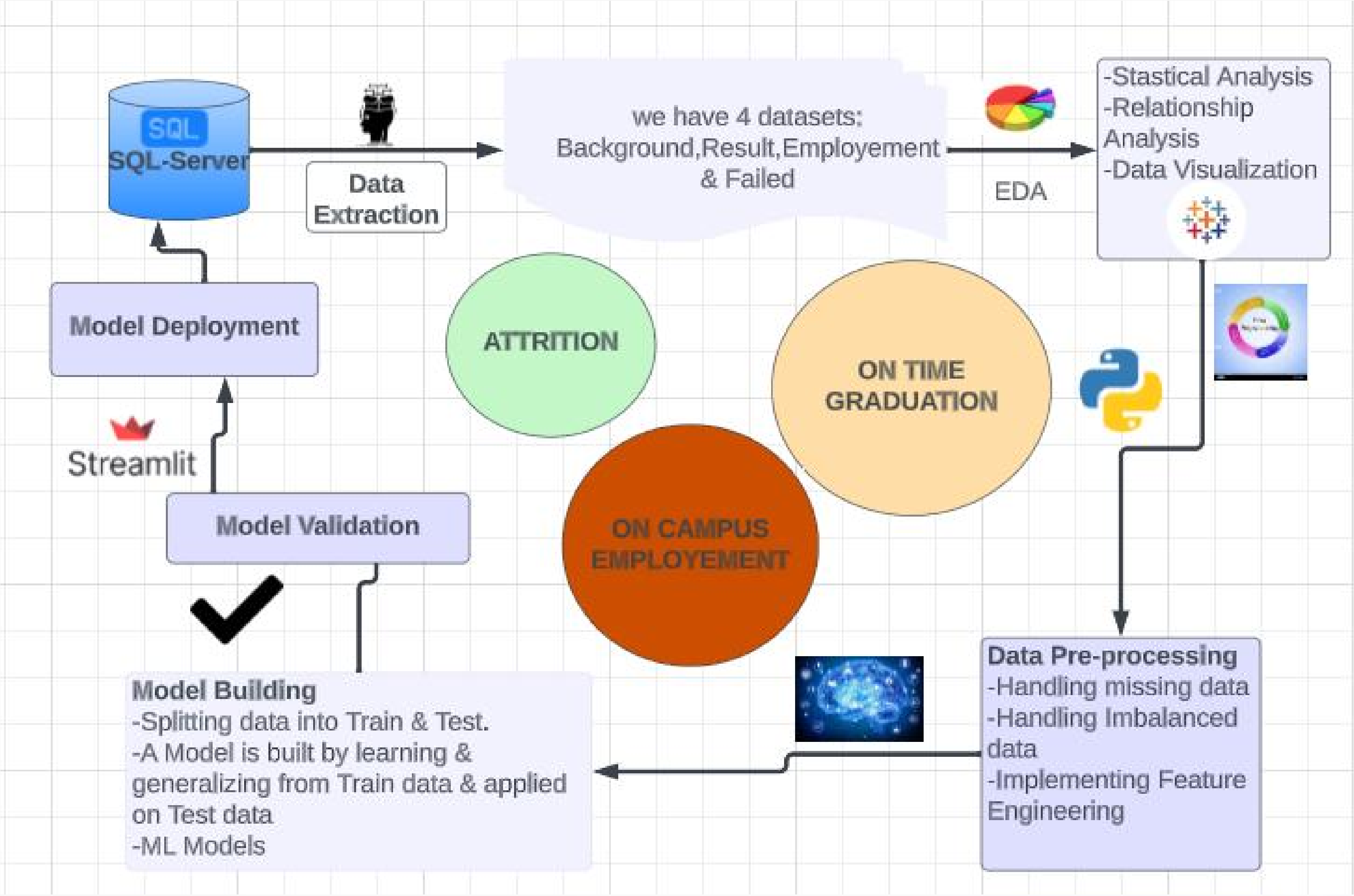
- **MySQL** is a relational database management system based on SQL – Structured Query Language. The application is used for a wide range of purposes and we used it for data warehousing.

- **Python** is a general-purpose programming language. We used it for Data Cleaning, EDA, Model Building and Visualization.

- **Streamlit** is an open-source Python library that makes it easy to create and deploy the Machine Learning Models.

- **Tableau** is a powerful and fastest growing data visualization tool used in the Business Intelligence Industry.

Project Architecture:



SQL-Server

Data Extraction

we have 4 datasets:
Background,Result,Employement
& Failed

EDA

-Stastical Analysis
-Relationship Analysis
-Data Visualization

ATTRITION

ON TIME GRADUATION

ON CAMPUS EMPLOYEMENT

Data Pre-processing
-Handling missing data
-Handling Imbalanced data
-Implementing Feature Engineering

Model Building
-Splitting data into Train & Test.
-A Model is built by learning & generalizing from Train data & applied on Test data
-ML Models

Model Validation

Streamlit

Model Deployment

Data Pre-Processing:

MODIFICATION

- Translating Malay Language to English
- I used DeepTranslator

DATA CLEANING

- Removal of Null Values
- Removal of duplicates
- Missing Values Treatment

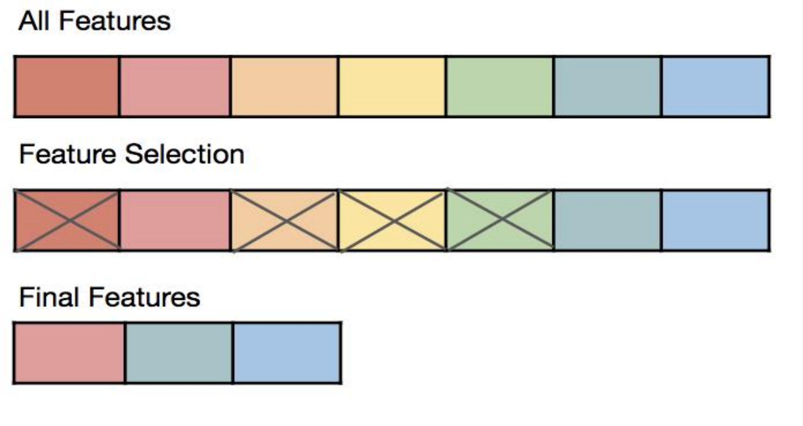
EDA

- I used PandasProfiling to get inferences.
- I took inferences from correlation matrix.
- Python Libraries: Plotly, Seaborn, Matplotlib.

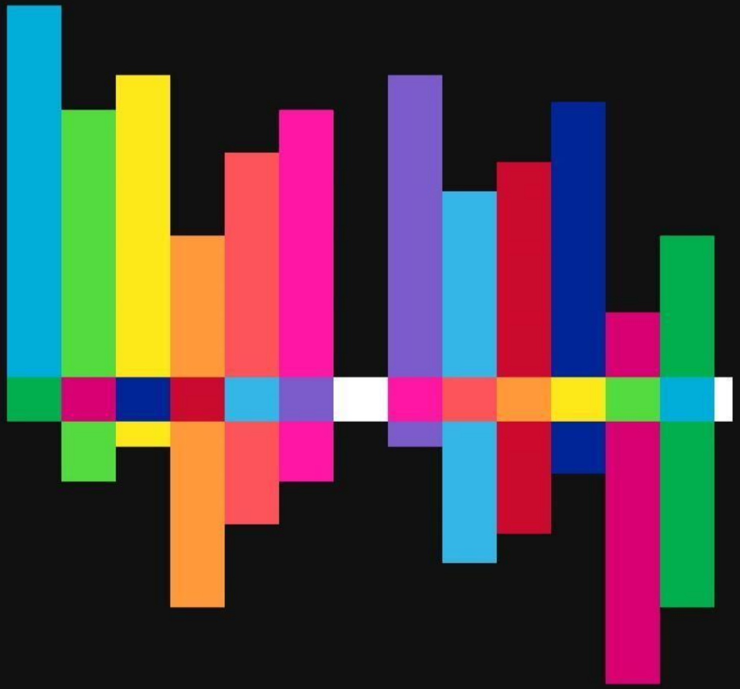
Feature Engineering:

Feature selection methods are techniques used to identify and select the most relevant variables or features from a dataset to improve model performance, reduce overfitting, and enhance interpretability by focusing on the most informative attributes.

- Correlation analysis statistical method used to measure the strength of the linear relationship between two variables and compute their association.
- Regularization technique penalizes the inclusion of unnecessary features during model training, promoting simpler models and reducing overfitting.
- Principal Component Analysis (PCA)is a dimensionality reduction technique that transforms features into a lower-dimensional space while preserving the most important information.
- Library used was SelectKBest (Select features according to the k highest scores.)



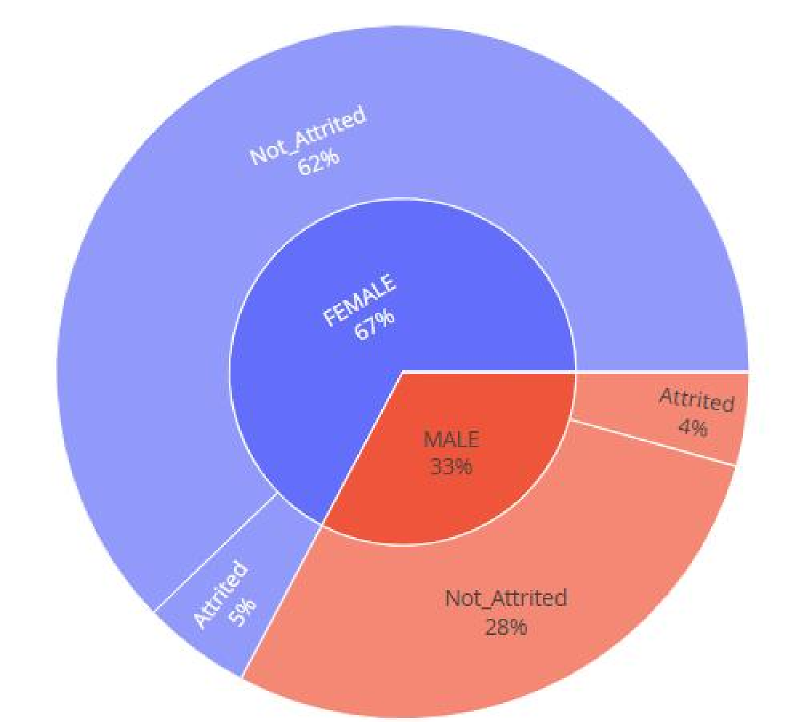
EXPLORATORY DATA ANALYSIS:



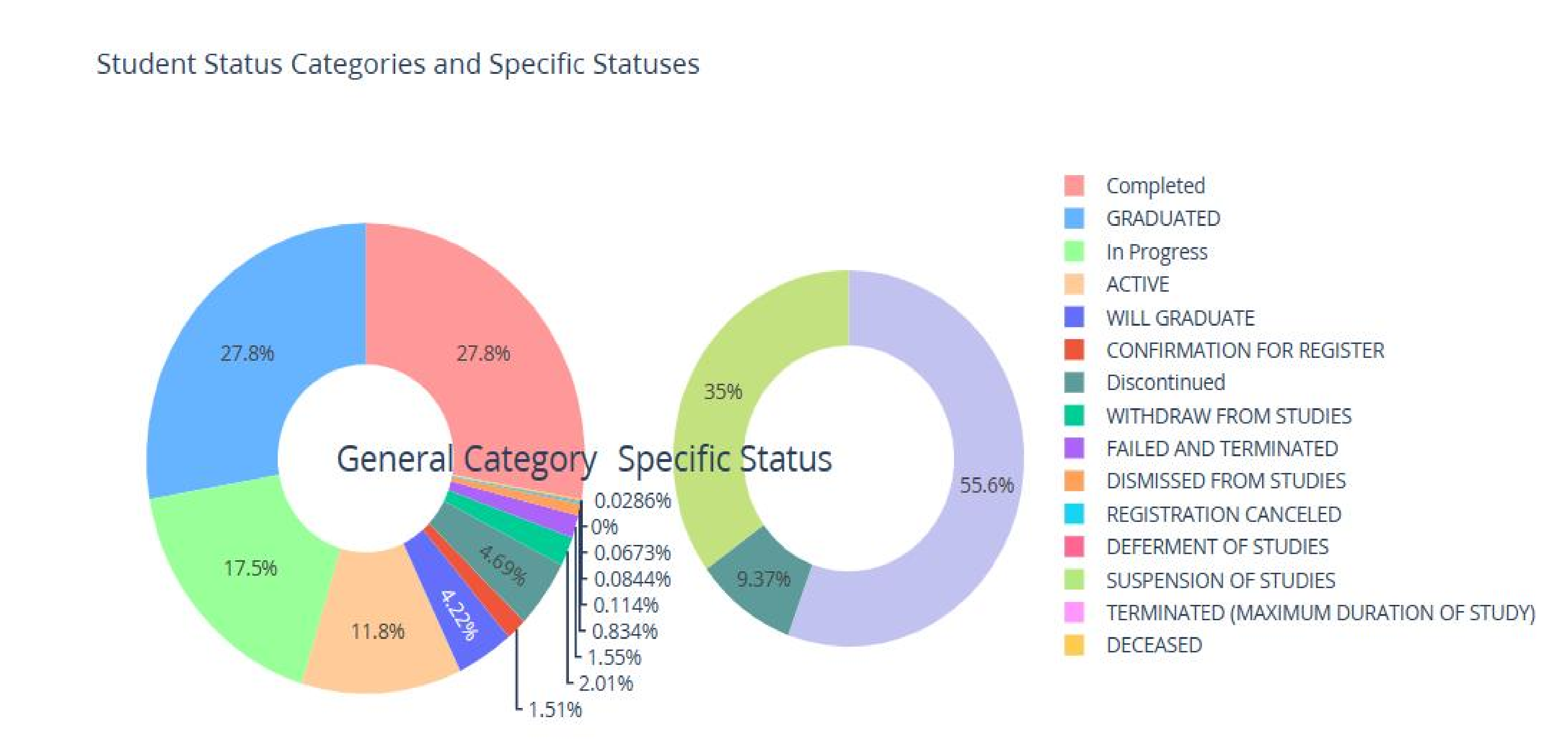
- **Attrition Analysis**
- **On-Time Graduation Analysis**
- **On-Campus Employment Analysis**

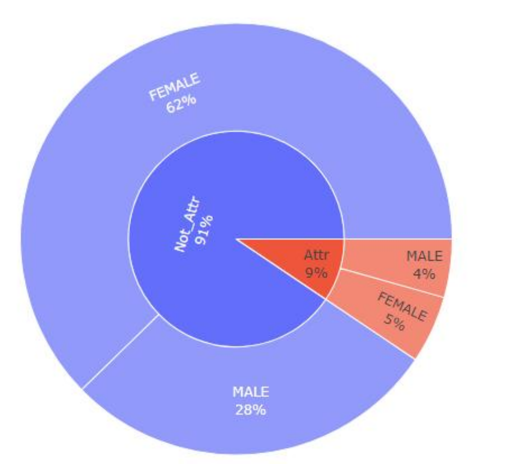
Attrition Class Analysis:

The background information dataset comprises 35,174 entries with 46 columns. It contains details about students' academic backgrounds, including their program, cohort, degree level, matriculation number, examination results, entrance qualifications, demographic information such as race and gender, as well as sponsorship and registration statuses.

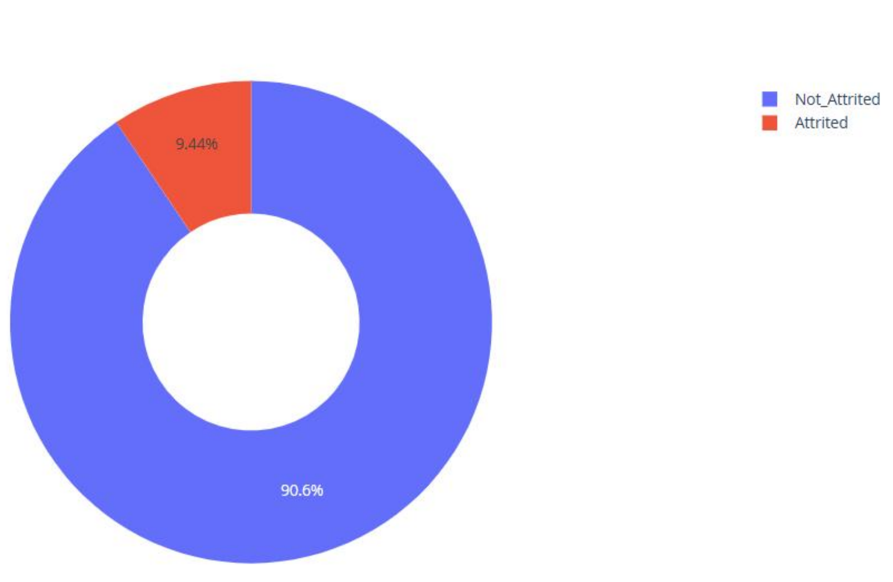


- The Attrited Students are
  - Deceased
  - Terminated
  - Suspended
  - Registration Cancelled
  - Dismissed
  - Failed
  - Withdrawn from Studies
- The Not Attrited Students are
  - Confirmation for Register
  - Will Graduate
  - Active and Graduated from Studies
- Overall the Not Attrited students are 91% and Attrited students are 9%.

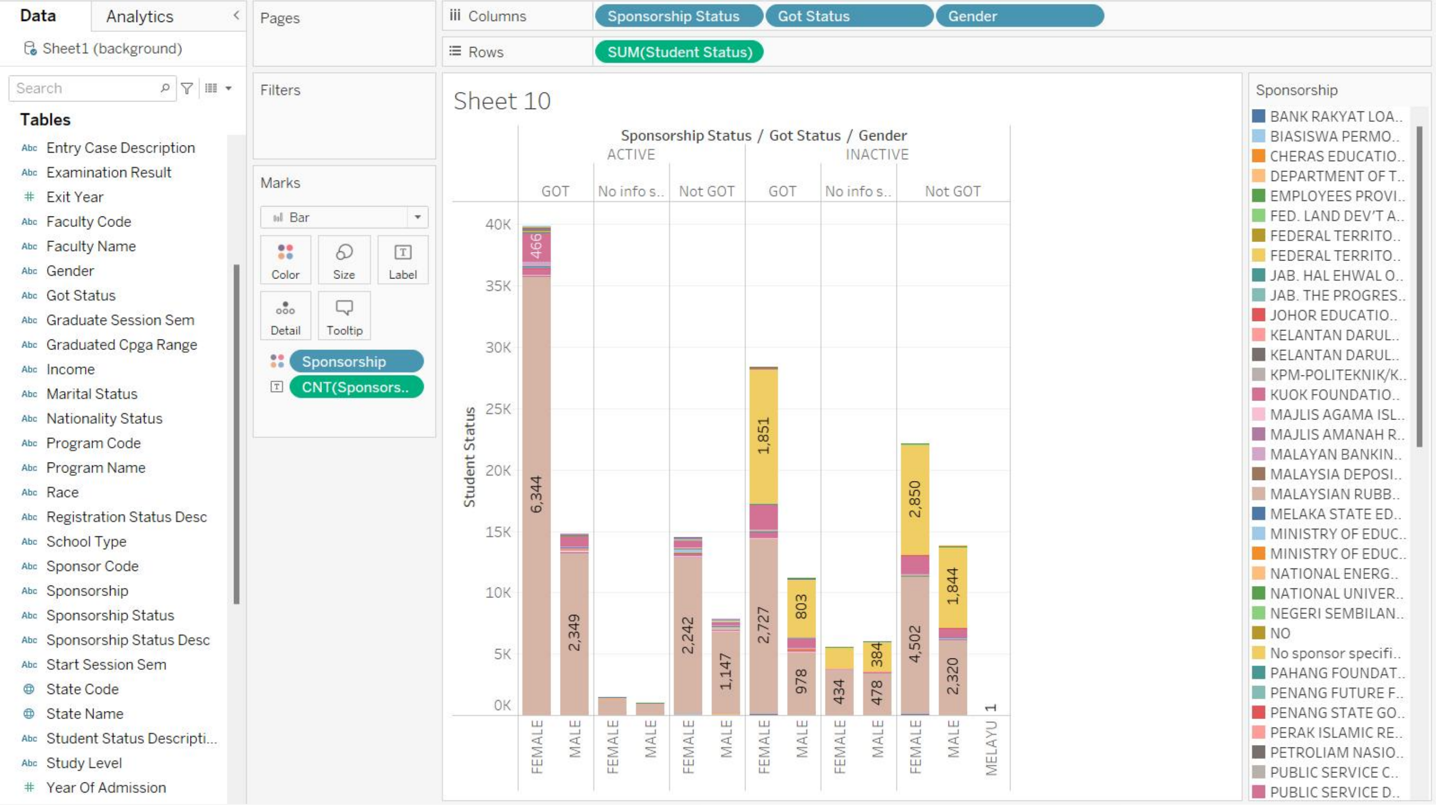


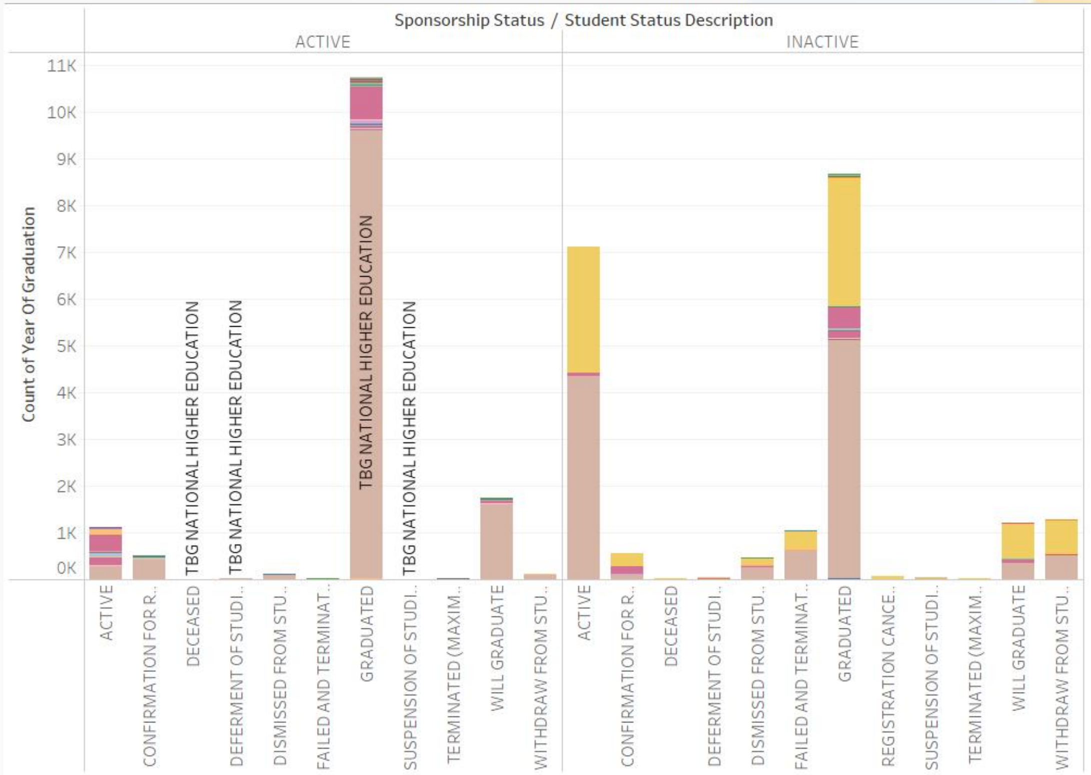


- 90.6% of the students has competed the course and graduated from the college
- only 9.4% of the students has dropped out from the college
- 4% from male & 5% from Female are Attrited Students



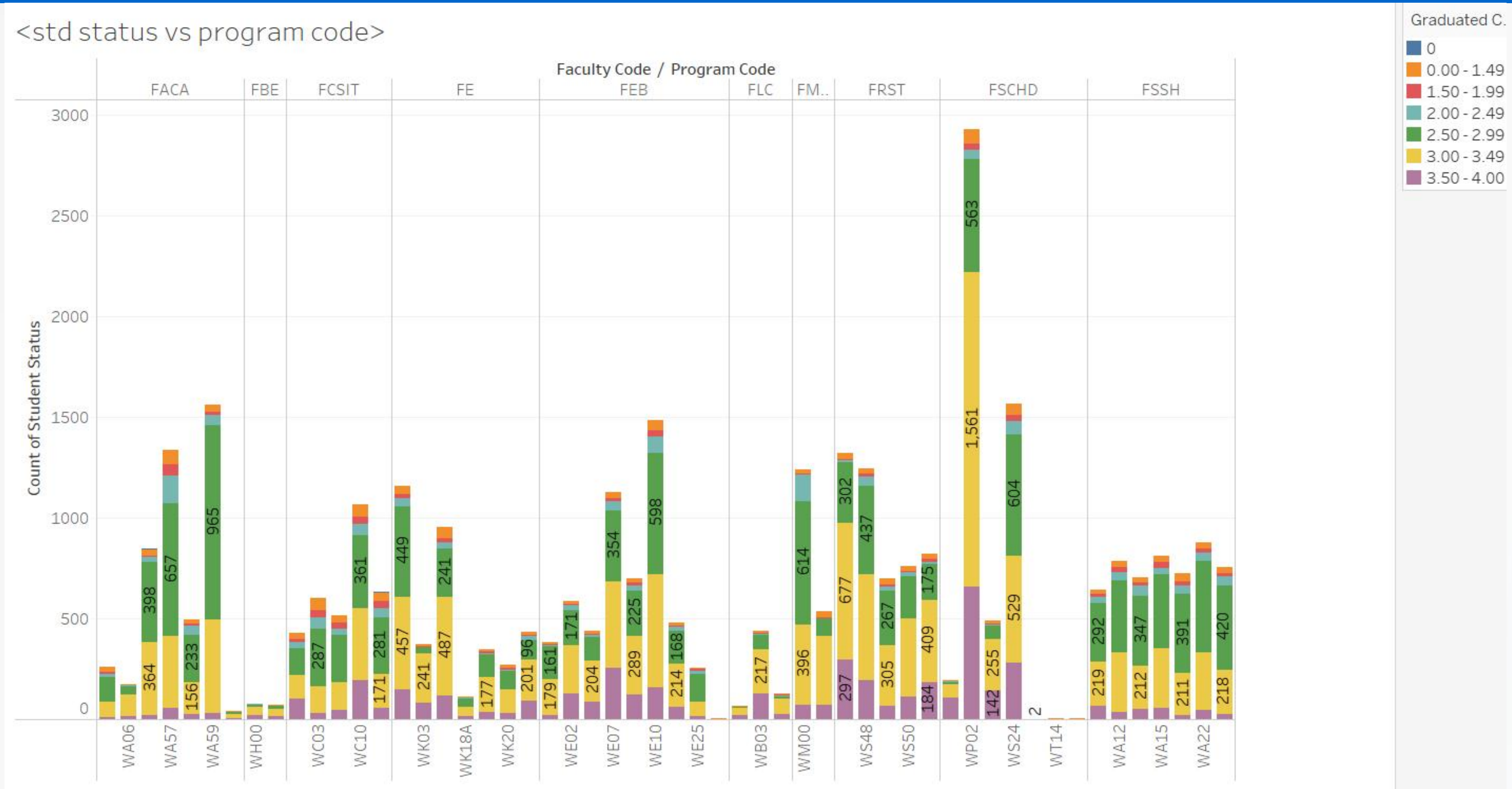
Student status Vs Sponsorship status:





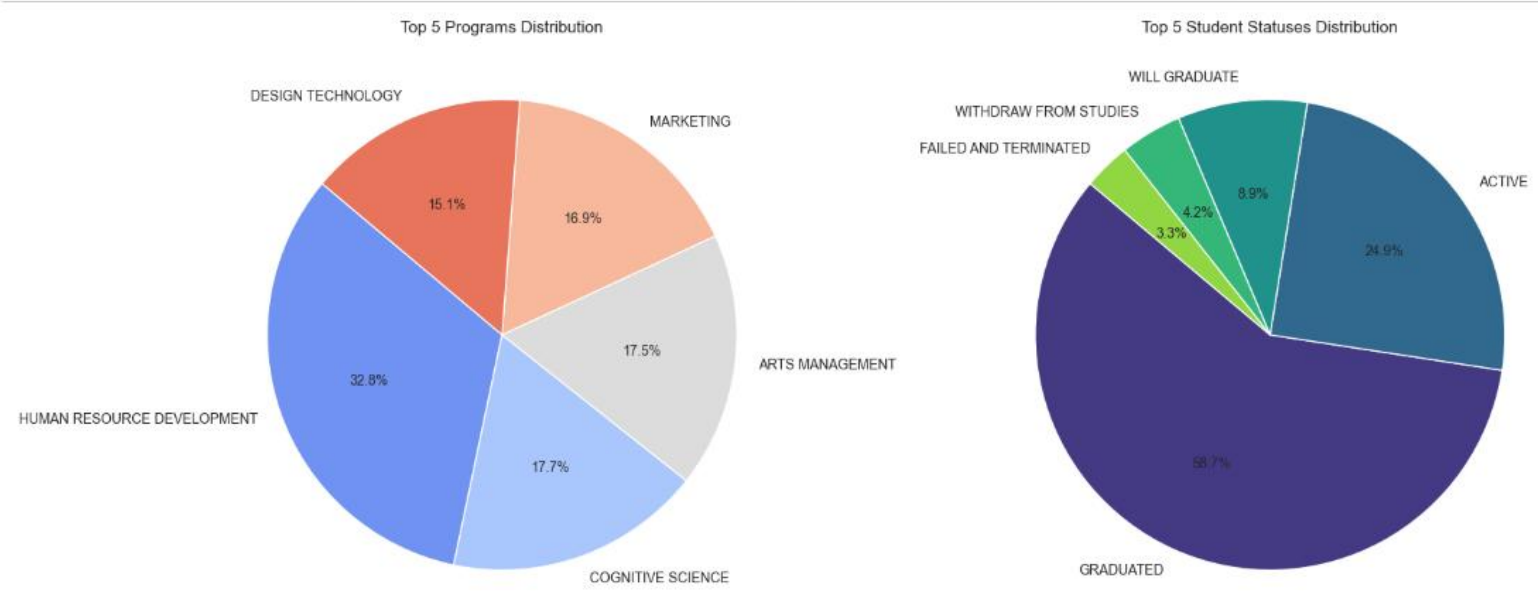
- ➢ The Graduated students with active sponsorship are more than Inactive sponsorship.

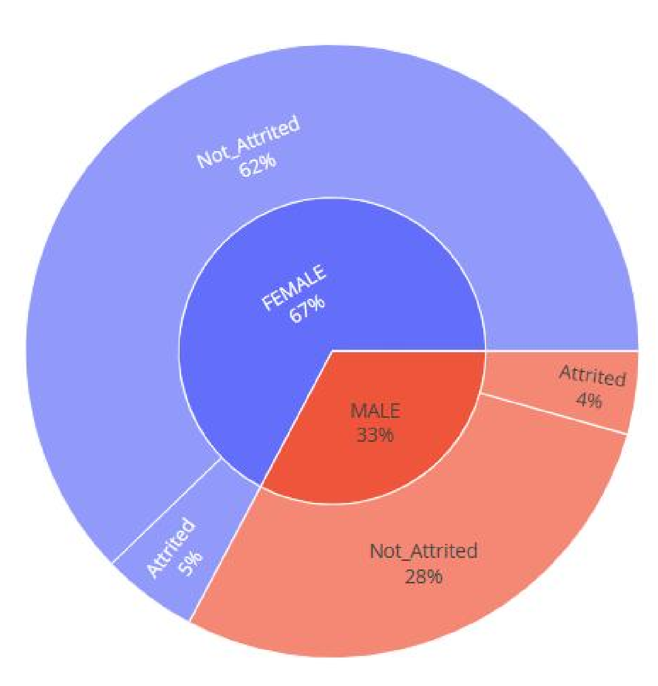
Student Status Vs Program Code:



- ➢ Students From the faculty code FSCHD and Program code WP02 have higher CPGA range than others
- ➢ More number of students also belong to the same faculty code that is FSCHD.

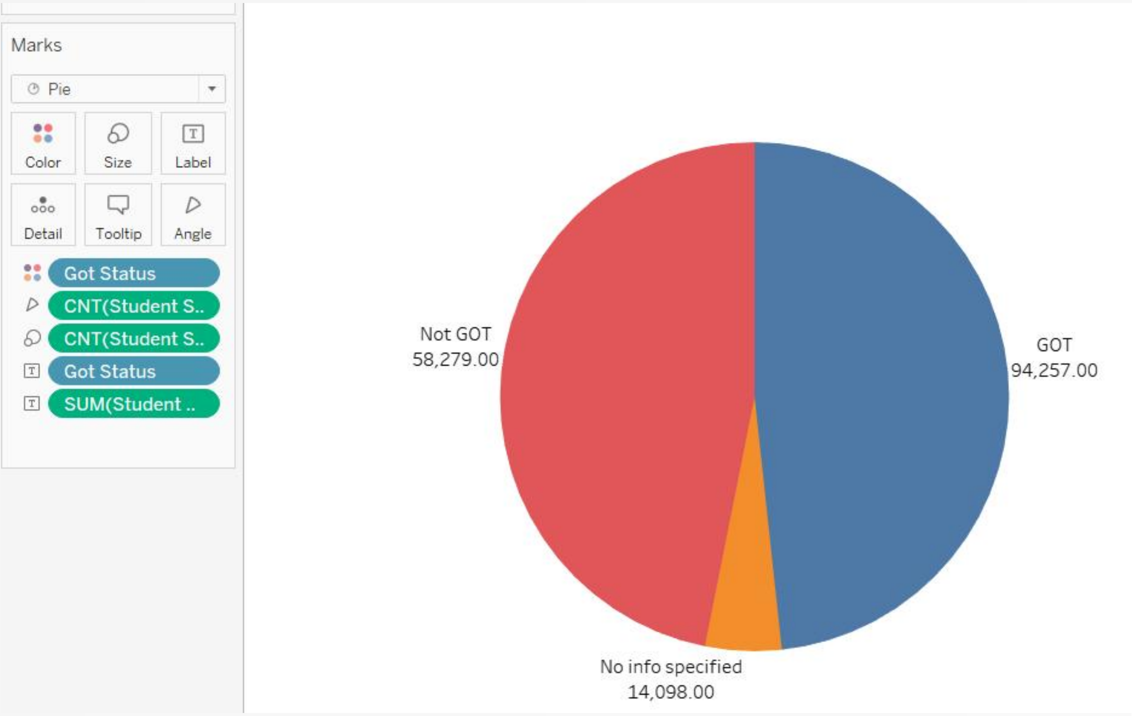
Student status Vs Got percentage:





- ➢ Students not graduated on time are 10% attrition rate and also have less number of siblings.
- ➢ Students who dropped have less number of siblings i.e 93% covered by less out of total.

On-Time Graduation Analysis:



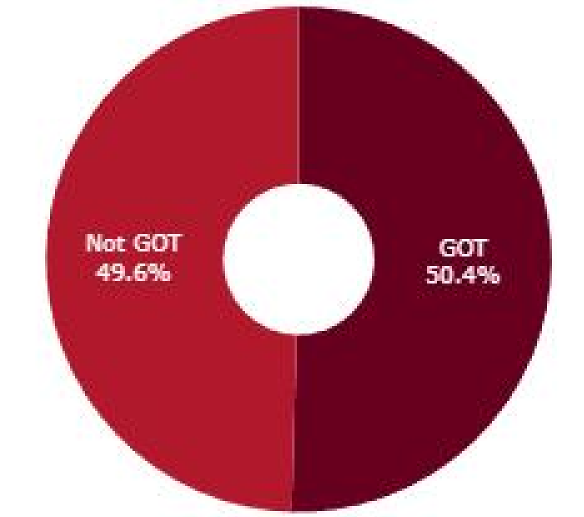
Marks

Pie

Color Size Label

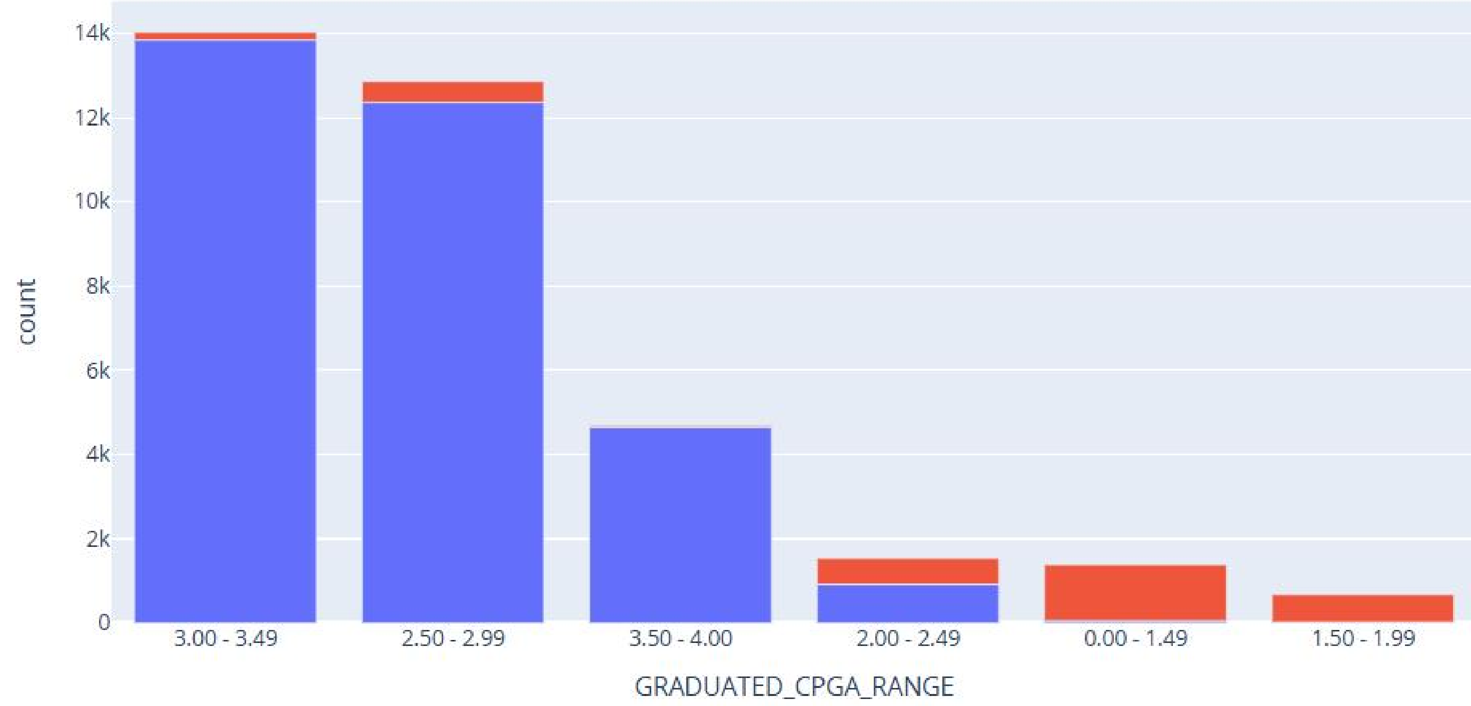
Detail Tooltip Angle

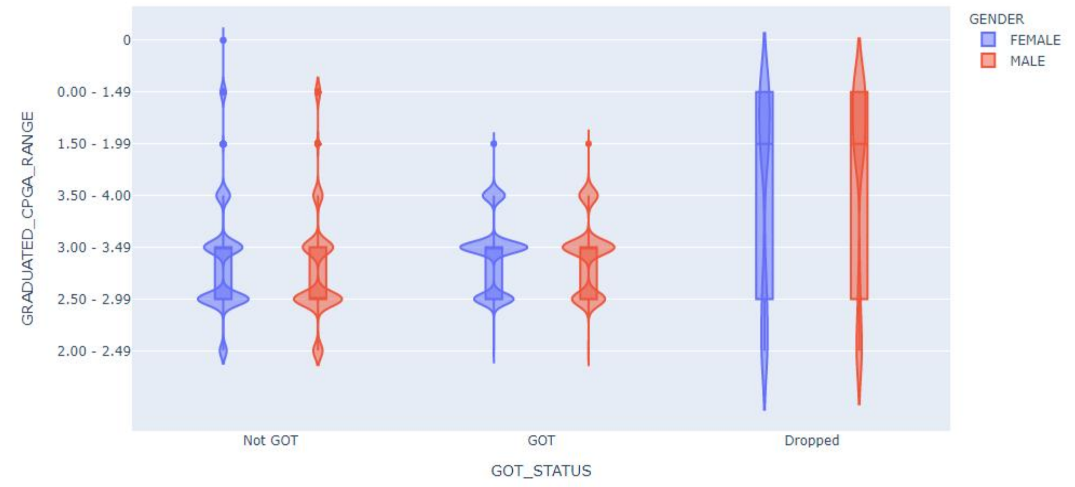
- Got Status
- CNT(Student S..
- CNT(Student S..
- Got Status
- SUM(Student ..



- ➢ 51% of Students are Graduated on time, 49% of students are not graduated on time.
- ➢ LMK (Pass (re-sit failed course)) students are in high compared to others who are Graduated

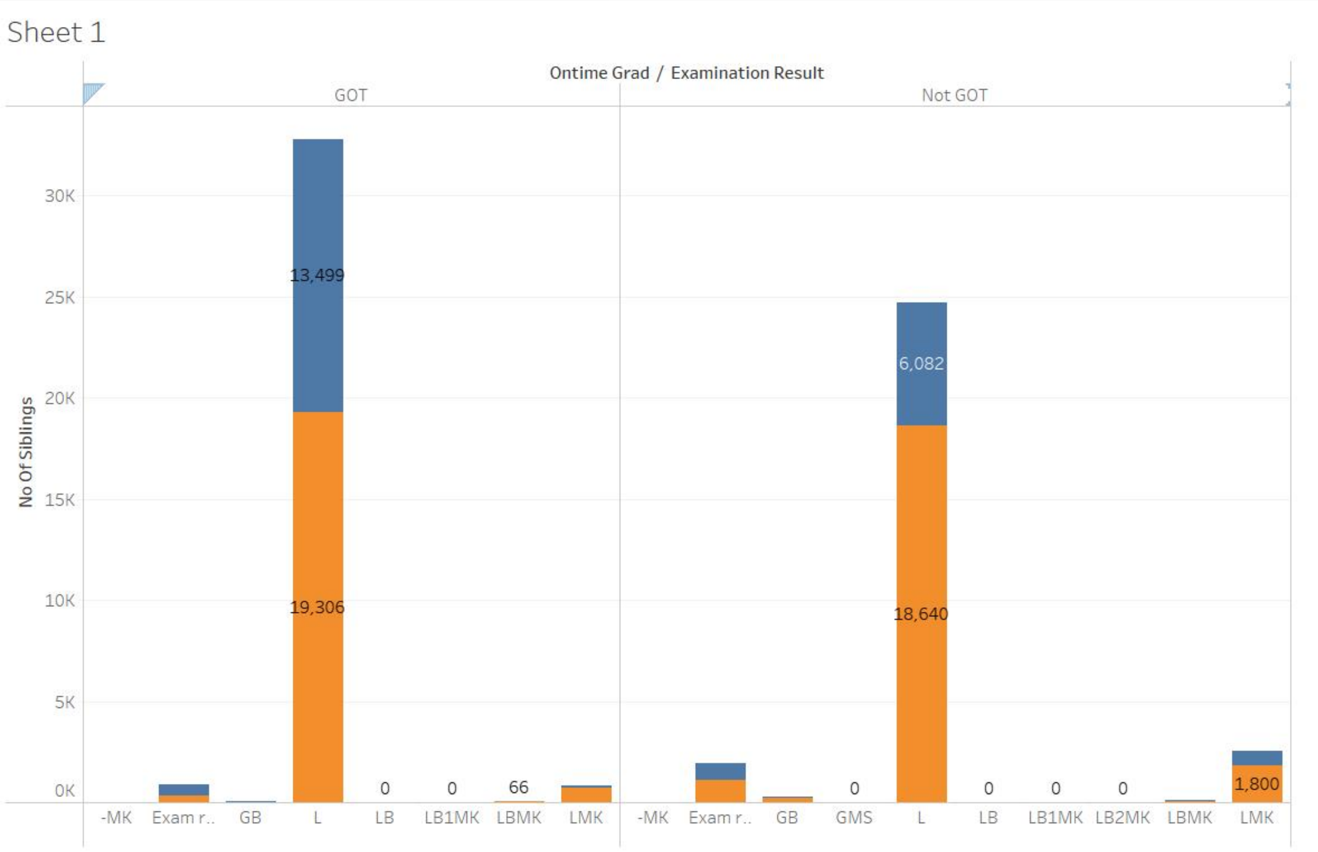
GOT Status Vs Graduated CGPA range:

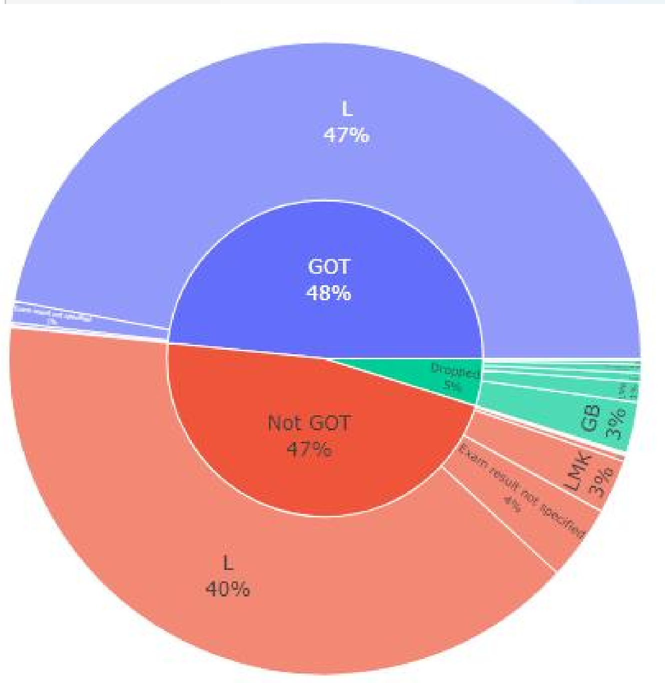




- ➢ For the both male & female gender the outliers have the target variable as attrition , and for both the GRADUATED_CPGA_RANGE is below average.
- ➢ CPGA Range From 0.00 to 1.99 all the students have been Attrited.
- ➢ CPGA Range From 2.00 to 4.00 only the few students has been Attrited

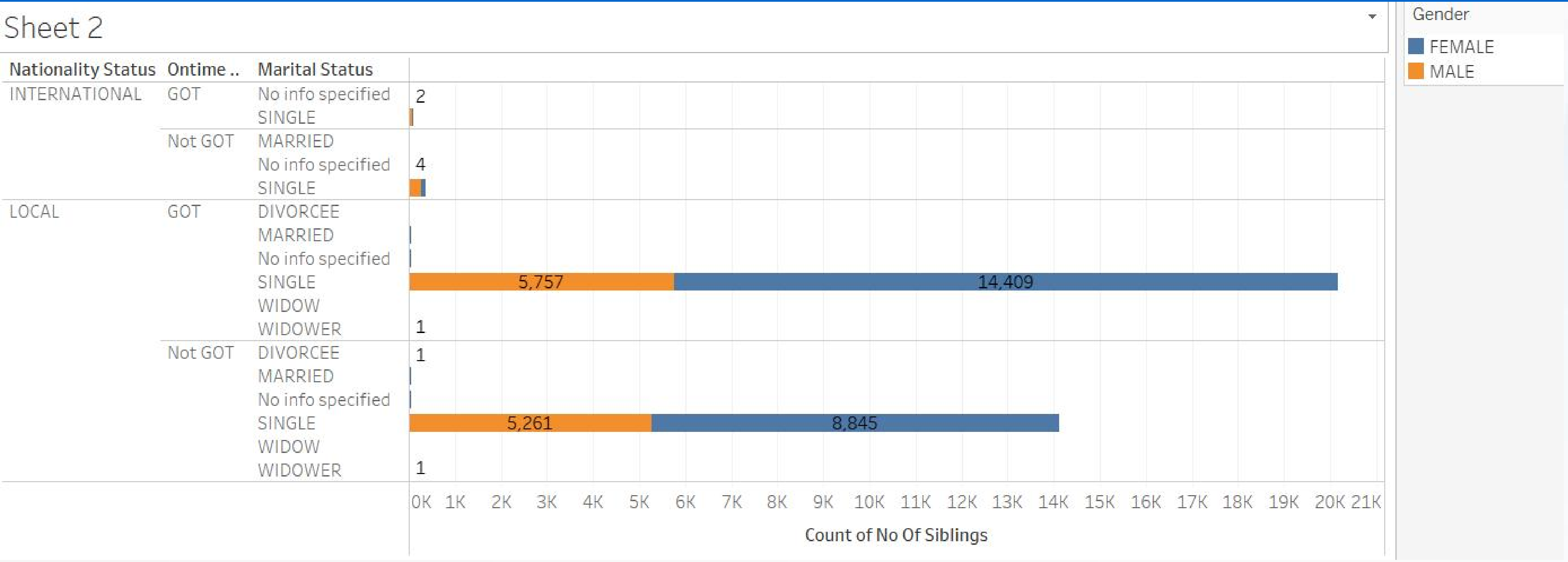
GOT Status Vs Examination Result:





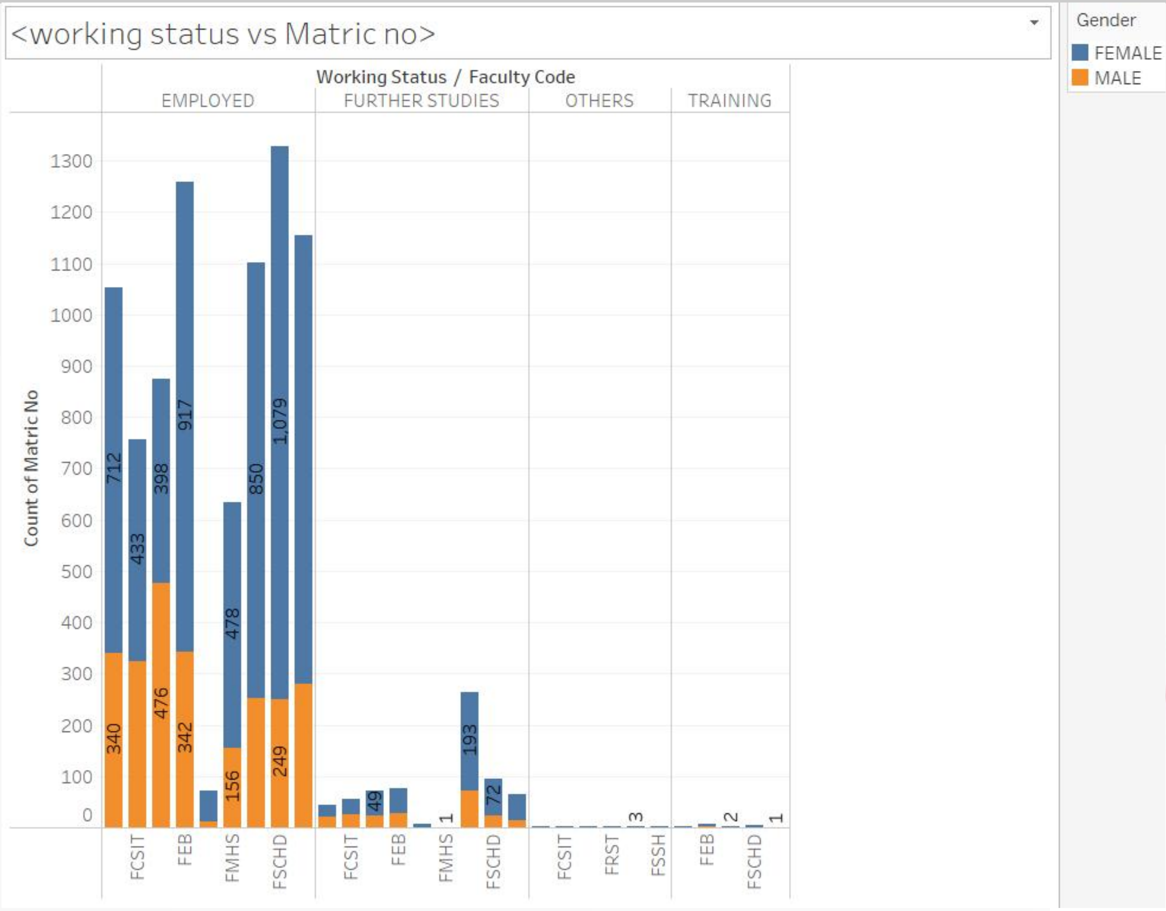
- ➢ Students whose examination result is L , are having more number of graduation on time status.
- ➢ Sponsorship status as inactive is more in Not Got compare to students having GOT status.
- ➢ Got Status Students are higher than Not Got,And Dropped Students are 5%
- ➢ LMK (Pass (re-sit failed course)) students are in high compared to others who are Graduated

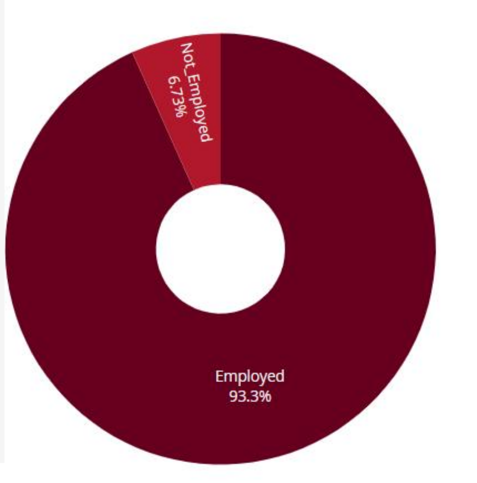
GOT Status Vs Marital Status:



- Most of the students are single and belong to Local countries.

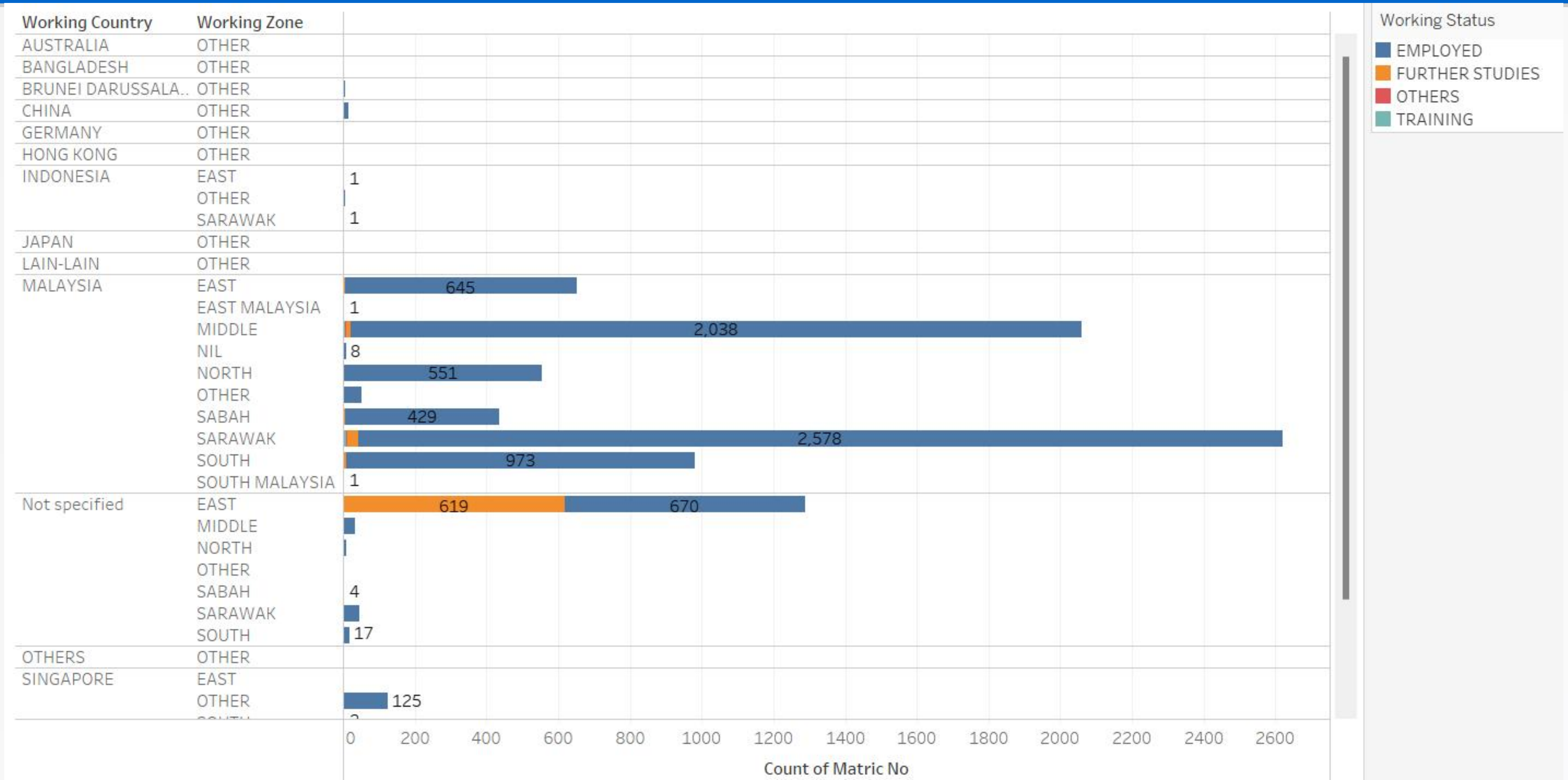
On-Campus Employment Analysis:





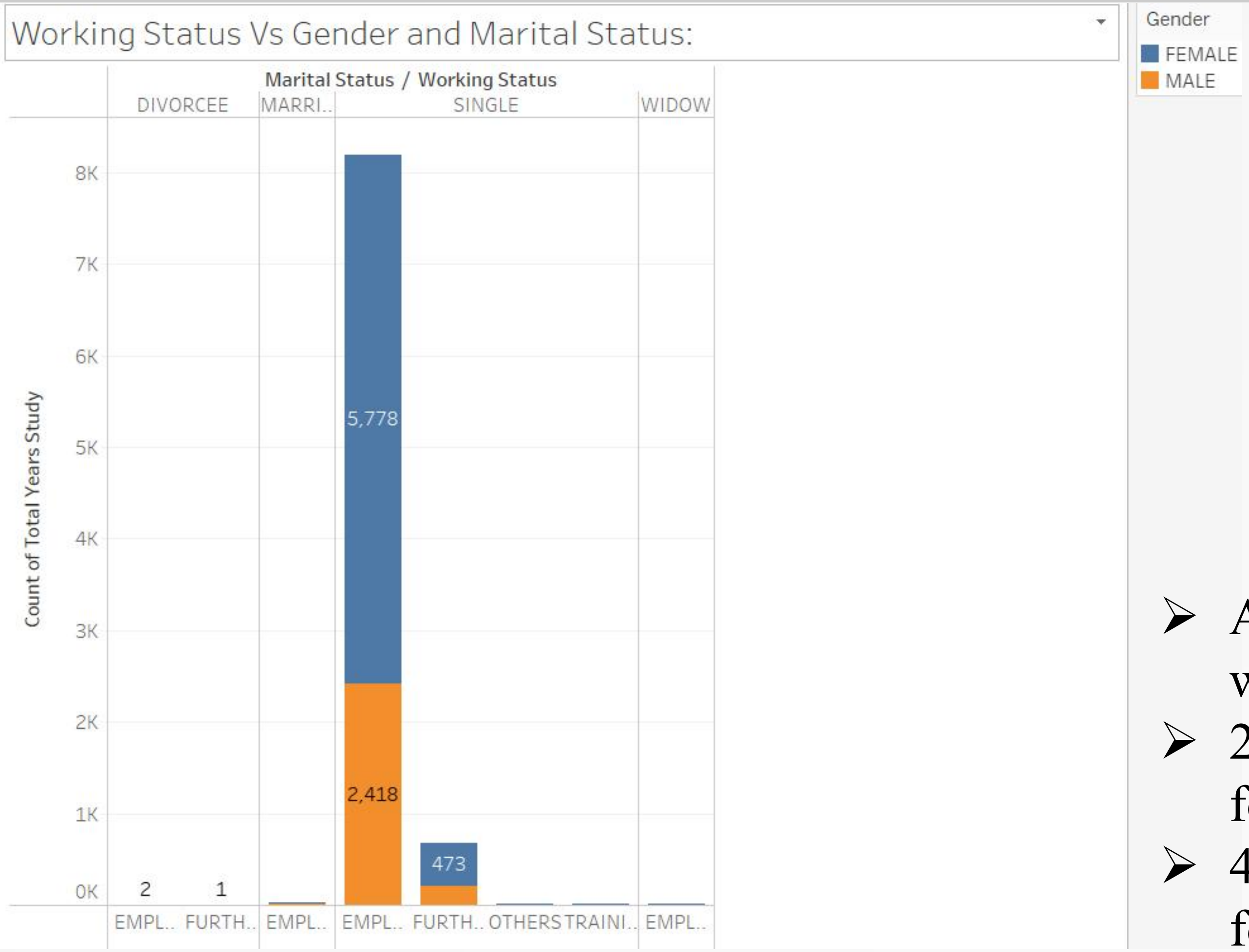
- ➢ Students who belong to FSCHD faculty code have more no. of Employed status.
- ➢ Students from FRST faculty code preferred more to do further studies in compare to others.
- ➢ For most of the faculty codes female rate is higher than male while getting employed.

Working Status Vs Working zone & Country:



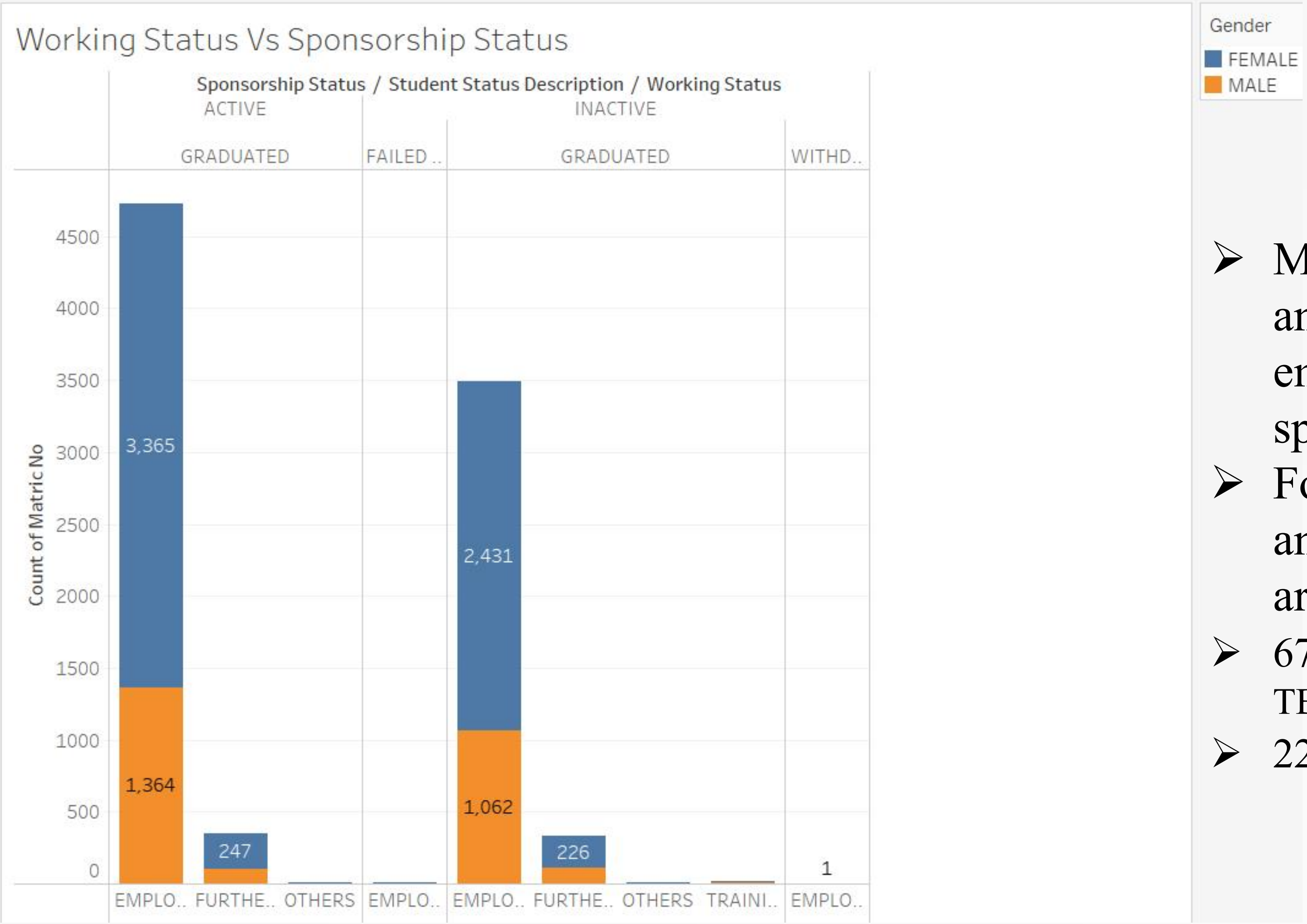
- ➢ Students working at Malaysia are mostly working in SARAWAK Zone.
- ➢ All students from North zone are Employed.
- ➢ After Malaysia, most of the remaining students are working in Singapore.

Working Status Vs Gender and Marital Status:

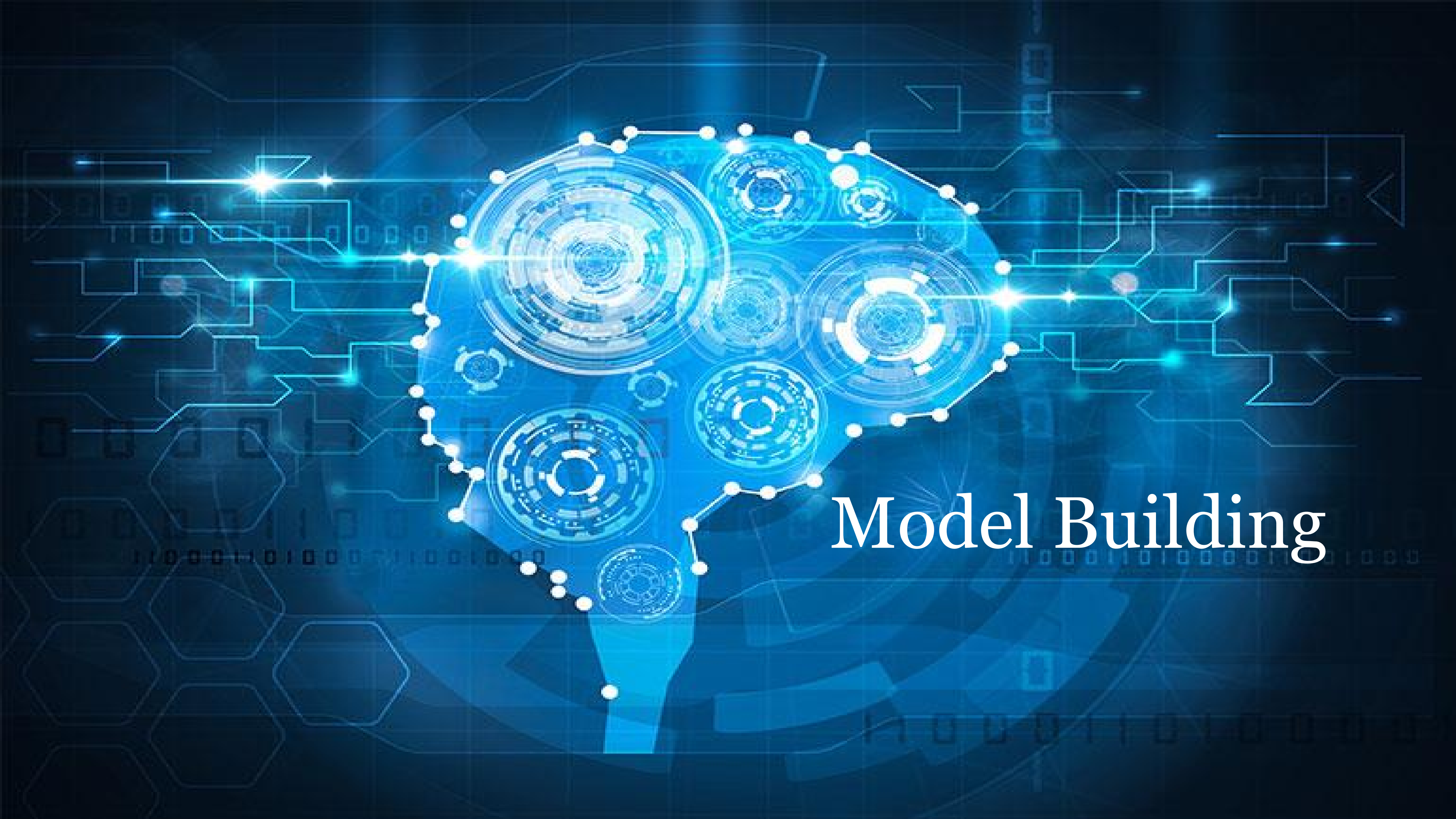


- ➢ Almost all participants who got employed or went for further studies are Single.
- ➢ 274 Male single participants are preferred for the further studies.
- ➢ 473 Female single participants are preferred for the further studies.
- ➢ 54 % Females who are single got employed.
- ➢ 27 % males who are single got employed.

Working Status Vs Sponsorship Status:



- Most of the students who are graduated and have active sponsorship got employment in compare to inactive sponsorship.
- For further studies both are almost equal and in all these cases female participants are more than male.
- 67.7% of Students are being sponsored by TBG NATIONAL HIGHER EDUCATION.
- 22.65% students don't have any sponsorship.

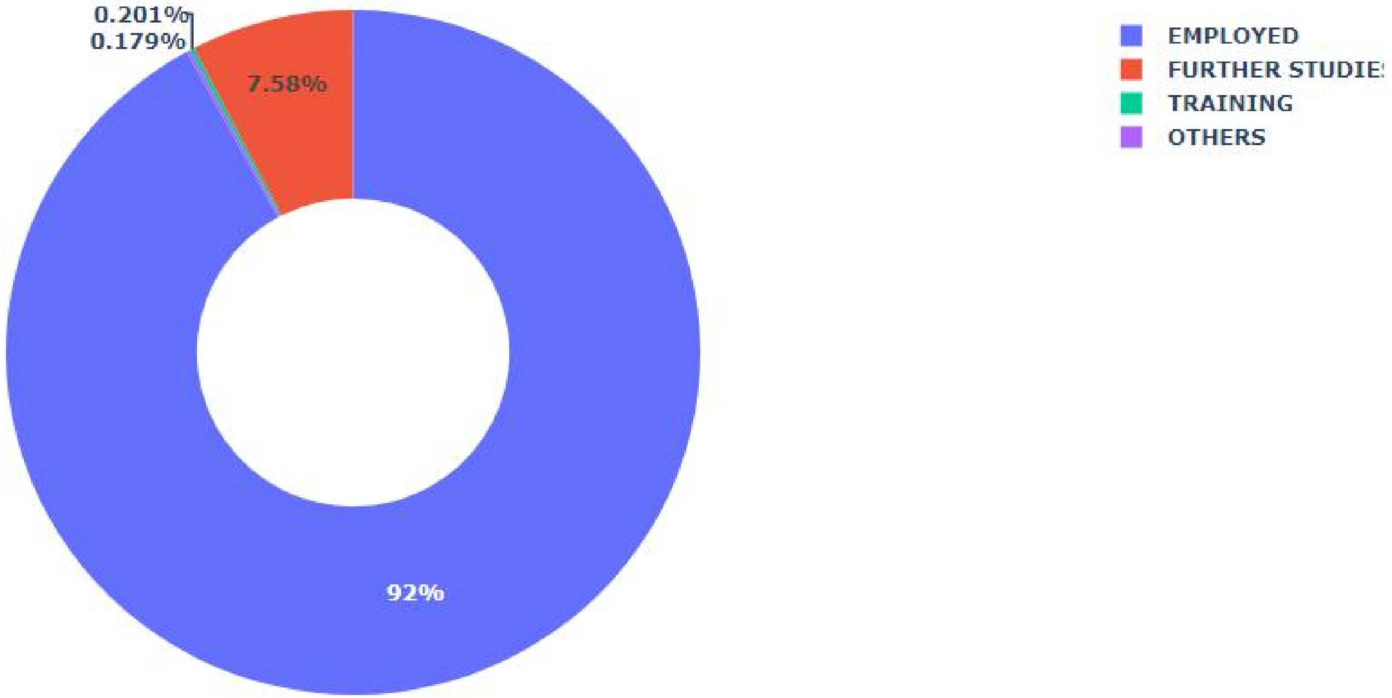


Model Building

Model Building Algorithms:

MODEL	DESCRIPTION
KNN Classifier	K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures
Decision Tree	Decision tree analysis involves making a tree-shaped diagram to chart out a course of action or a statistical probability analysis.
Random Forest	An ensemble model made of many decision trees using bootstrapping, random subsets of features, and average voting to make predictions
XG Boost Classifier	An optimized distributed gradient boosting library designed to be highly efficient & flexible, it provides a parallel tree boosting that solve data science problems in a fast and accurate way.
Logistic Regression	A statistical analysis method used to predict a data value based on prior observations of a data set ,It is the go-to method for binary classification problems

Imbalanced Data:



- During analysis I observed disparity in Employment business problem.
- If I choose to move along without treating it ,our model may not give us the optimum result that's expected.
- Here I Can notice the data is imbalanced as 92% of the overall donut is Covered as EMPLOYED.

Treating Imbalanced Data:

Metrics:
When dealing with imbalanced data, standard classification metrics do not adequately represent your models performance.

- Precision is defined as the fraction of relevant examples (true positives) among all of the examples which were predicted to belong in a certain class.

- Recall is defined as the fraction of examples which were predicted to belong to a class with respect to all of the examples that truly belong in the class.

Oversampling:
oversampling the minority classes to increase the number of minority observations until we've reached a balanced dataset.

Random oversampling:
The most naive method of oversampling is to randomly sample the minority classes and simply duplicate the sampled observations. With this technique, it's important to note that you're artificially reducing the variance of the dataset.

On-Campus Employment Analysis:

XG-Boost supports the following main interfaces:

- Python,CLI, Julia, C++ ,Java, Scala, Platforms like Hadoop.

Model Features:

- Gradient Boosting algorithm also called gradient boosting machine including the learning rate.
- Stochastic Gradient Boosting with sub-sampling at the row, column and column per split levels.
- Regularized Gradient Boosting with both L1 and L2 regularization.

Algorithm Features:

- Sparse Aware implementation with automatic handling of missing data values.
- Block Structure to support the parallelization of tree construction.
- Continued Training so that you can further boost an already fitted model on new data.

Models	Train Accuracy	Test Accuracy
XG-Boost	95.5%	92.4%
Decision Tree	99%	93.4%
Logistic Regression	86.3%	84.7%

Attrition Analysis:

Random Forest:
Random Forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest will make a class prediction and the class with the most votes becomes our model's prediction. A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

Model Features:

- Random Forests can be used for both classification and regression tasks.
- No scaling or transformation of variables is usually necessary, as it uses a rule-based approach.
- Random Forests work well with both categorical and numerical data.
- Random Forests generally provide high accuracy and balance the bias-variance trade-off well. Since the model's principle is to average the results across the multiple decision trees it builds, it averages the variance as well.

Models	Train Accuracy	Test Accuracy
Random Forest	100%	100%
Logistic Regression	99%	99%
KNN	99%	99%

On-Time Graduation Analysis:

Decision Tree:

Decision trees are one of the best forms of learning algorithms based on various learning methods. They boost predictive models with accuracy, ease in interpretation, and stability. The tools are also effective in fitting non-linear relationships since they can solve data-fitting challenges, such as regression and classifications.

Advantages:

- Easy to read and interpret
- Compared to other decision techniques, decision trees take less effort for data preparation.
- There is less data cleaning required once the variables have been created.

Applications:

- Evaluating prospective growth opportunities for businesses based on historical data.
- Using demographic data to find prospective clients
- Decision trees can also be used in operations research in planning logistics and strategic management

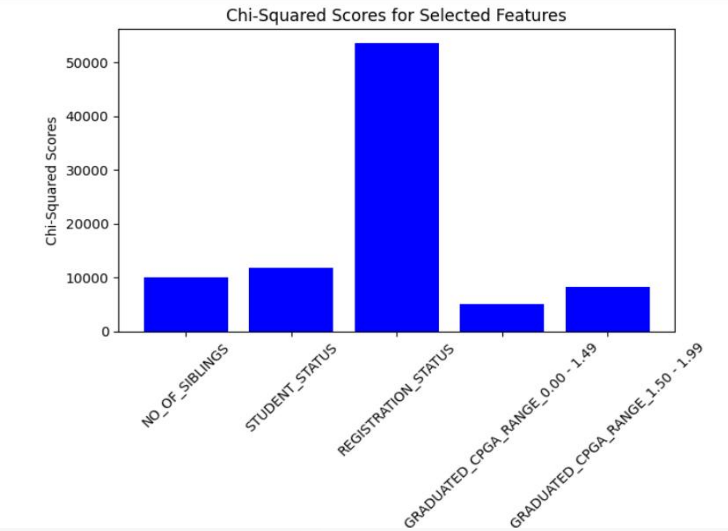
Models	Train Accuracy	Test Accuracy
Decision Tree	100%	99%
XG-Boost	99%	98%
KNN	99%	99%

Survival Analysis:

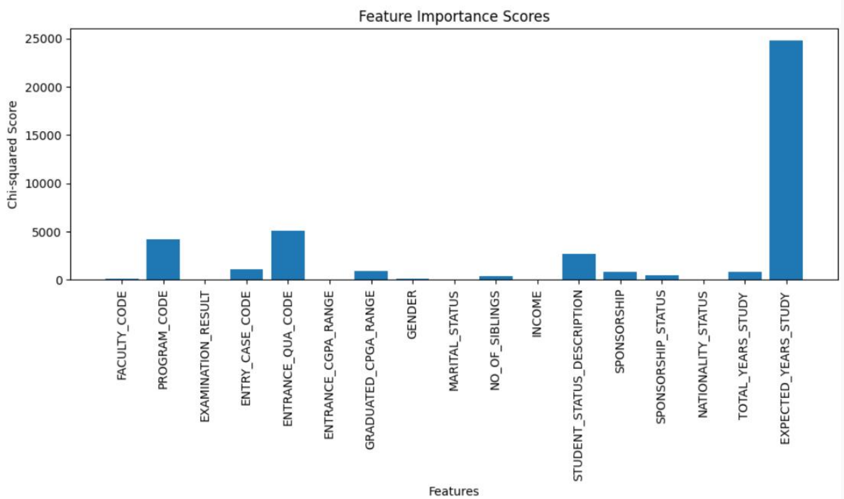
- Survival analysis is a statistical method that aims to predict the time to an event, such as attrited or not, The students got placed or not based on their education. A key aspect of survival analysis is the presence of censored data, indicating that the event of interest has not occurred during the study period

Variable Importance Plots:

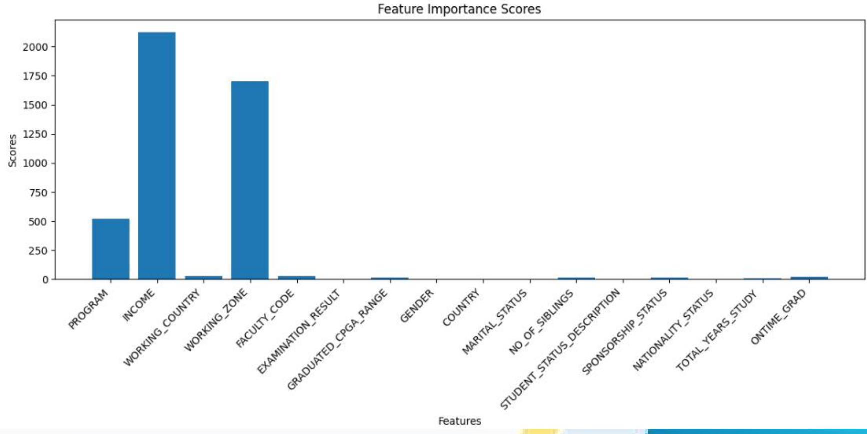
Attrition



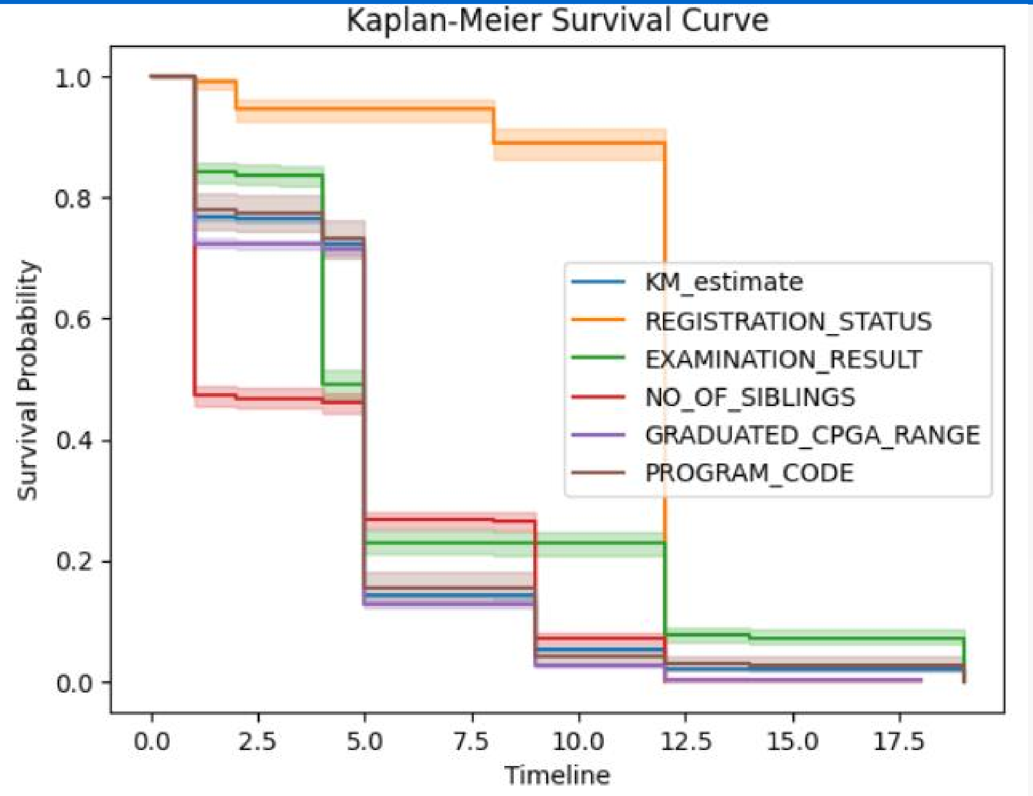
On time Graduation



On campus-Employment

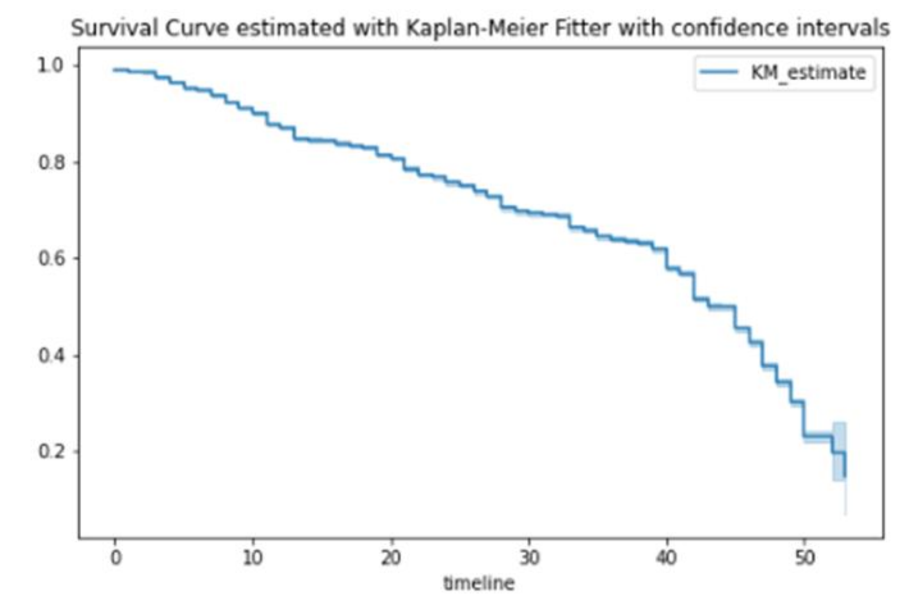


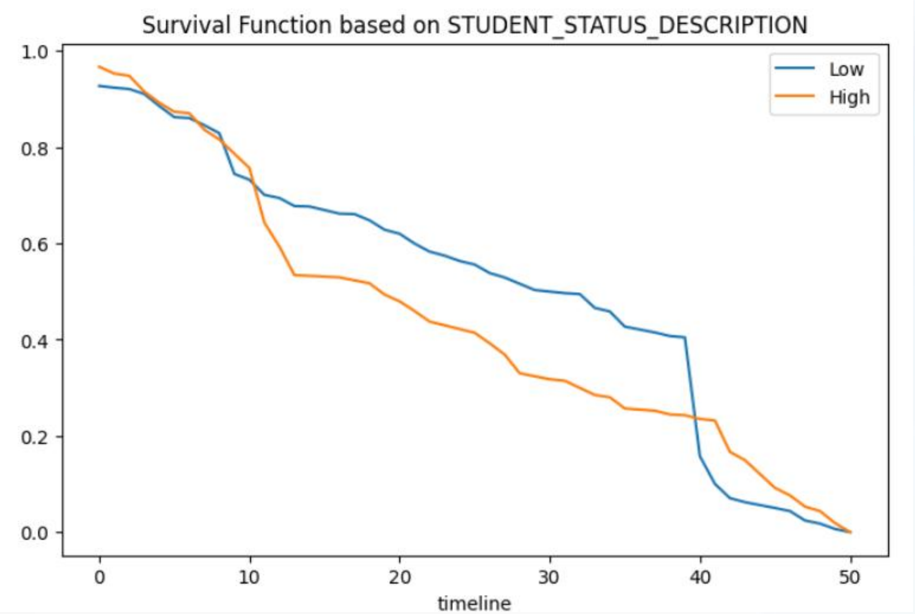
Survival Analysis for Attrition:



- Survival Time is defined as the time starting from a predefined point to the occurrence of the event of interest
- T is the time from students registration until status of got employment
- This time estimate is the duration between Registration to attrited , not attrited students_status
- Kaplan-Meier Estimate is used to measure the fraction of subjects/students who survived for a certain amount of survival time
- Students status is decreased with respect to time-line based on their status description
- Attrited status are decreased according to time line based on their status description

Survival Analysis for On-Time Graduation:

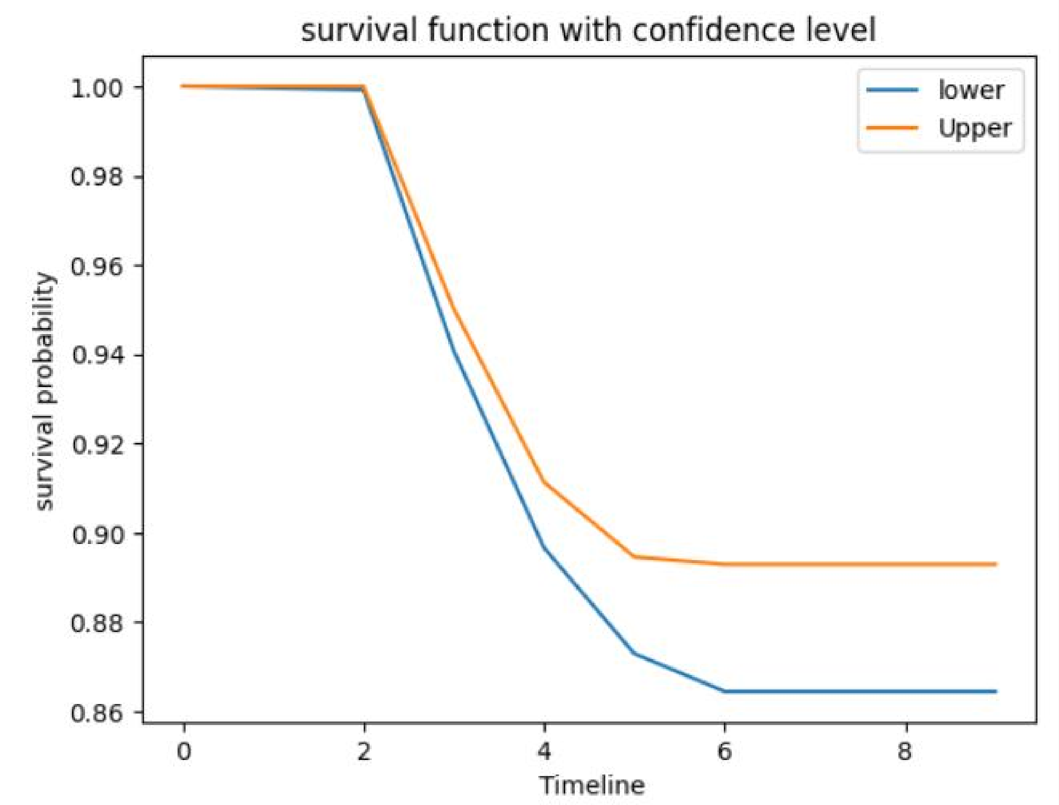




- Survival Curve estimated by Kaplan-Meier Fitter with confidence interval by considering ONTIME_GRAD with PROGRAM_CODE
- The confidence interval is showing that students graduating on time and their survival analysis.

- Low and high curves are representing that the program codes with respect to countries completing graduation on time

Survival Analysis for On-Campus Employment:



- KaplanMeir estimates the survival function from Total Years of Studied to Working Status

1)The Students Status after Graduation are employed=0, further studies=1, training=2.
2) The Survival Probability of students are between 0 to 1
3) As the Timeline Increases and the Probability of Survival Decreases for the participants.

- The confidence interval for the Kaplan–Meier estimate of the survival probability are at fixed time point.
- KaplanMeir survival function estimates probability of students employment lower than 0.95 & more than 0.95 (confidence level between Total years of study).

Deployment Strategy:

Streamlit is an open-source Python library. It enables developers to quickly build highly interactive web applications around their data, machine learning models, and pretty much anything.

Pros-:

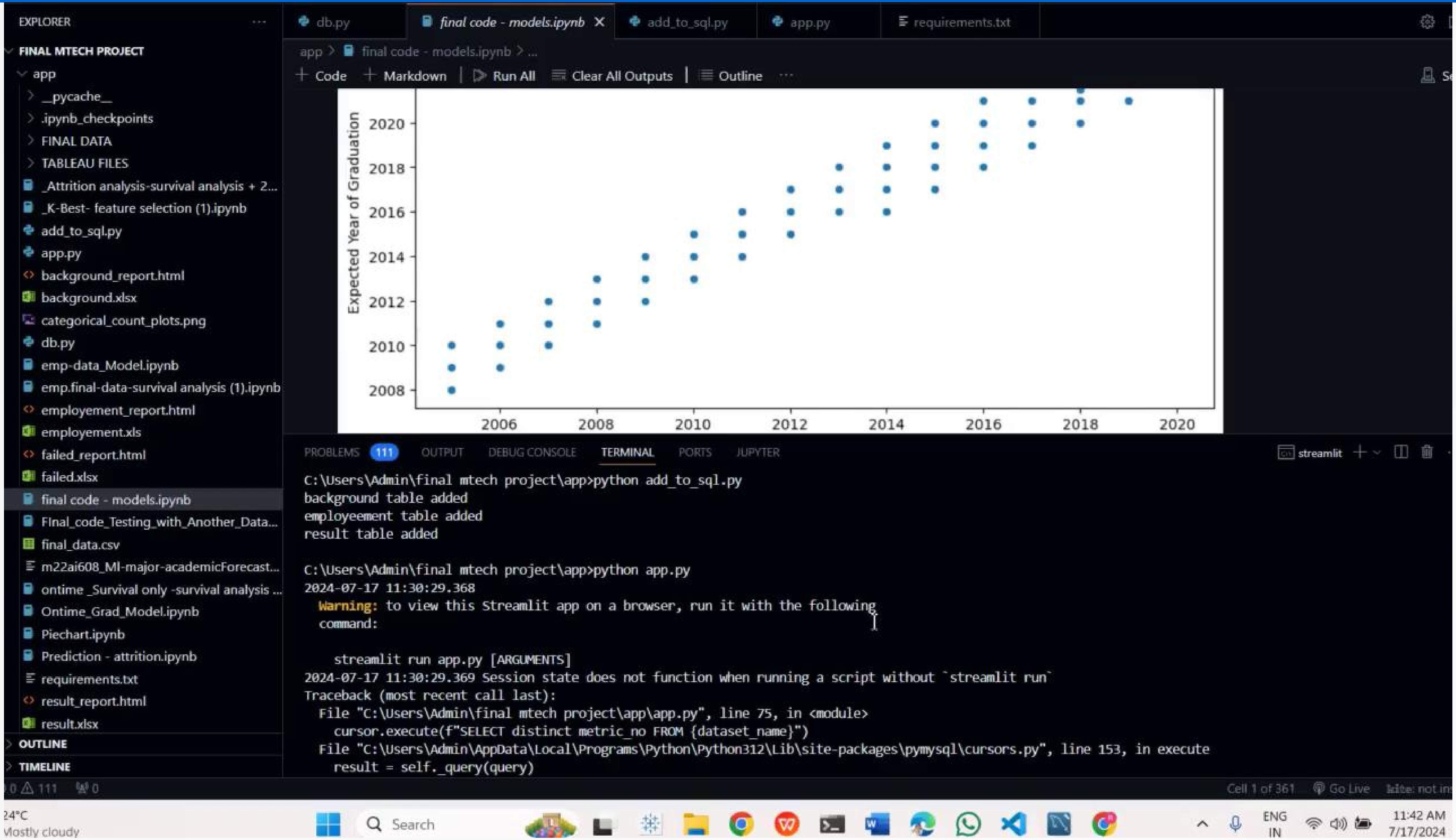
- Super-fast development to deployment time. Literally minutes.
- Native data dashboard tool all in one tool that encompases web serving as well as data analysis python only library

Cons-:

- Cannot easily customize any of the frontend components.
- Since it’s relatively new, sometimes it’s hard to find answers to your questions.

I have used streamlit instead of flask because the flask web framework doesn’t have any data visualization, manipulation or analytical capabilities needs end front development experience .

Outputs of the Deployed Model:



```
C:\Users\Admin\final mtech project\app>python add_to_sql.py
background table added
employeement table added
result table added

C:\Users\Admin\final mtech project\app>python app.py
2024-07-17 11:30:29.368
  Warning: to view this Streamlit app on a browser, run it with the following
  command:

    streamlit run app.py [ARGUMENTS]
2024-07-17 11:30:29.369 Session state does not function when running a script without `streamlit run`
Traceback (most recent call last):
  File "C:\Users\Admin\final mtech project\app\app.py", line 75, in <module>
    cursor.execute(f"SELECT distinct metric_no FROM {dataset_name}")
  File "C:\Users\Admin\AppData\Local\Programs\Python\Python312\Lib\site-packages\pymysql\cursors.py", line 153, in execute
    result = self._query(query)
```

Future Scope:

- Using machine learning, universities can then hone in on student retention and persistence and identify factors that influence student success.
- Exploring advanced machine learning techniques like deep learning, ensemble methods, and gradient boosting algorithms to enhance model accuracy and robustness in predicting student performance metrics.
- Implementing techniques such as feature importance analysis, model-agnostic interpretability methods to improve understanding of the factors influencing student outcomes, aiding stakeholders in making informed decisions.
- Developing real-time monitoring systems to track student performance metrics continuously, enabling timely interventions to prevent adverse outcomes like dropout or academic underachievement.
- Integrating predictive models into student support systems to identify at-risk students early and deliver personalized interventions tailored to their specific needs, thereby fostering academic success.

Conclusion:

- Being realistic about the overhead upfront expectations , institution may have loads of historical data , but they might be in legacy systems or there may be technical hurdles that make it difficult to easily access those data. The value is there, but it may take time to come up with a solution that is easy for everyone to use.

- Machine learning shows great potential to disrupt how we process and consume data and use software. Serious ethical considerations and limitations must be considered. However, higher education is naturally and uniquely positioned to capitalize on the promise of machine learning by using it as a tool for social and moral good. Higher education has the opportunity not only to use machine learning to help transform itself to make better decisions but also to explore how it might apply machine learning as a force for good.

References:

- F. Giannakas, C. Troussas, I. Voyiatzis, and C. Sgouropoulou, “A deep learning classification framework for early prediction of team-based academic performance,” Appl. Soft Comput., vol. 106, Jul. 2021, Art. no. 107355.

- H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, “Student academic performance prediction model using decision tree and fuzzy genetic algorithm,” Proc. Technol., vol. 25, pp. 326–332, Jan. 2016.

- B. K. Francis and S. S. Babu, “Predicting academic performance of students using a hybrid data mining approach,” J. Med. Syst., vol. 43, no. 6, pp. 1–15, Jun. 2019.

- https://medium.com/analytics-vidhya/students-performance-analysis-46eba510039c

- https://www.irjmets.com/uploadedfiles/paper/issue_6_june_2022/26568/final/fin_irjmets1655801430.pdf

- T. Le Quy, T. H. Nguyen, G. Friege, and E. Ntoutsi, “Evaluation of group fairness measures in Student performance prediction problems,” 2022, arXiv:2208.10625.

THANK
YOU