# Predicting Dropout Risk: A Comparative Analysis of Machine Learning Algorithms for Student Dropouts

**By - Aryan Patil, Samarth Jain, Satyam Shah**

### 1.1. Literature Review: Expanding Beyond Existing Research

While Martins et al. (2021) provided valuable insights into student dropout prediction using machine learning models like Logistic Regression and boosting algorithms, exploring a wider range of methodologies is crucial for a comprehensive understanding. Research in this domain has explored alternative approaches, such as survival analysis, which analyzes the time until dropout occurs, taking into account censored data from students who remain enrolled. This method, particularly the Cox proportional hazards model, allows for the inclusion of time-varying factors, such as changes in academic performance or engagement, providing a dynamic perspective on dropout risk. Another promising approach is Bayesian networks, which are probabilistic graphical models capable of capturing complex relationships between variables and handling uncertainty effectively. By constructing a network of interconnected nodes representing different factors, such as academics, can analyze student activity sequences and reveal patterns indicative of dropout risk, such as decreased participation in online discussions or late assignment submissions.

The dataset used in student dropout prediction research vary significantly in their characteristics and limitations. Institutional data, readily available to universities and colleges, typically includes demographic information, academic performance metrics, and engagement data from learning platforms. While valuable, this data might be limited in scope, lacking insights into personal challenges or external factors influencing student decisions. National surveys, on the other hand, offer broader perspectives on student populations but often lack the granularity and individual-level detail necessary for personalized predictions. LMS data provides a valuable window into student online behavior and engagement with course materials but may not capture the full picture of student learning and engagement outside the digital environment. Social media data, while potentially insightful, raises ethical concerns regarding student privacy and requires careful anonymization techniques and adherence to data protection regulations.

Predicting student dropout presents various challenges that researchers must navigate. One primary concern is data availability and quality. Access to comprehensive and reliable data is crucial for building effective models, but issues like missing data, inconsistencies, and privacy concerns can hinder progress. Ethical considerations are paramount in this domain, as predictive models raise questions about potential biases against certain student groups and the

risk of stigmatization. Transparency in model development and deployment is essential, alongside techniques for bias mitigation and fairness assurance. The dynamic nature of student behavior also poses a significant challenge, as academic performance and engagement can fluctuate over time. Longitudinal data analysis and the use of models capable of capturing temporal dynamics are necessary to address this issue. Lastly, dropout is often influenced by a complex interplay of academic, personal, social, and economic factors, making it difficult to isolate and model individual influences. Techniques like dimensionality reduction or feature interaction analysis can help researchers navigate this complexity and identify the most relevant factors contributing to dropout risk.

**1.2. Limitations of Previous Studies: A Closer Look**

Previous research on student dropout prediction often overlooks the importance of feature selection, which can have detrimental consequences for model performance and efficiency. Including irrelevant or redundant features can lead to overfitting, where the model learns noise in the data rather than the underlying patterns, resulting in poor generalization to new, unseen data. For example, a model trained with irrelevant features might learn to associate specific student ID numbers with dropout, leading to inaccurate predictions for future students. Additionally, a large number of features increases model complexity, leading to longer training times and higher computational resource requirements, especially for complex algorithms like neural networks. This can be impractical for real-world applications where efficiency is crucial. Furthermore, numerous features can reduce model interpretability, making it difficult to understand which factors truly influence predictions and explain why a specific student is predicted to be at risk of dropping out. This lack of transparency can hinder trust in the model and limit its practical utility for educators and administrators.

Neural networks offer several potential advantages in addressing these limitations. Their ability to handle non-linear relationships between features is particularly valuable in capturing the nuanced and complex factors contributing to student dropout. For instance, a neural network can learn the intricate interplay between a student's academic performance, engagement with online resources, and their social interactions, which may not be easily captured by simpler models like linear regression. Additionally, deep learning models can automatically learn relevant features from raw student data, such as text from discussion forums or assignments, eliminating the need for manual feature engineering and potentially uncovering hidden patterns that might be missed by traditional approaches. Moreover, neural networks exhibit a high degree of adaptability, allowing them to be fine-tuned to different types of data and prediction tasks. Transfer learning techniques can be used to adapt a neural network trained on one institution's data to another institution with a different student population and data structure, enhancing the model's generalizability and applicability.

**2.1. Objective: Refining Predictions and Empowering Institutions**

This project aims to build upon existing research on early student dropout prediction in higher education, directly addressing the shortcomings of previous studies. The primary goal is to improve the accuracy of dropout prediction models through refined feature selection techniques and exploration of alternative machine learning methods. By leveraging a rich dataset with over 4,000 student records and 37 features, the project seeks to develop a robust and reliable model capable of identifying at-risk students early in their academic journey. This early identification empowers educational institutions to implement timely interventions and support systems, ultimately preventing student dropout and promoting academic success. The ultimate goal is to develop a reliable and generalizable model for early identification of at-risk students. This model should not only achieve high prediction accuracy but also be interpretable and adaptable to different educational contexts. This could involve academic support programs, financial aid initiatives, mentoring opportunities, mental health services, and other resources aimed at addressing the specific factors contributing to each student's risk of dropping out.

**2.2. Dataset: A Wealth of Student Information**

The project utilizes a dataset sourced from the UC Irvine Machine Learning Repository, specifically focused on predicting student outcomes in higher education. This dataset offers a comprehensive view of student profiles, with the target variable categorized into three distinct classes: "Dropout", "Enrolled", and "Graduated". With over 4,000 rows and 37 features, the dataset encompasses a diverse range of information, offering a holistic perspective on factors potentially influencing student outcomes.

**Delving into the Features:**

Demographic Details:

- Marital status: This categorical variable reveals the student's marital status at the time of enrollment, potentially indicating additional responsibilities or life circumstances that could impact their studies.
- Gender: This binary variable identifies the student as male or female, allowing for analysis of potential gender-based disparities in dropout rates.
- Age at enrollment: This numerical variable indicates the student's age upon entering the higher education program, providing insights into their maturity level and potential life experiences.
- Nationality: This categorical variable identifies the student's nationality, allowing for exploration of cultural or socioeconomic factors that may influence dropout risk.

- International: This binary variable indicates whether the student is an international student or a domestic student, potentially highlighting additional challenges faced by international students, such as language barriers or cultural adjustment difficulties.

Academic Background:

- Application mode: This categorical variable describes how the student applied to the program, such as online, in-person, or through an agent. This could reflect their level of engagement and initiative in the application process.
- Application order: This numerical variable indicates the order in which the student applied to different programs, potentially revealing their level of interest in this program and their academic goals.
- Course: This categorical variable identifies the specific course or program the student is enrolled in, allowing for analysis of dropout rates across different academic disciplines.
- Previous qualification: This categorical variable describes the student's prior educational attainment, such as high school diploma or previous college degree.
- Previous qualification (grade): This numerical variable indicates the student's grade or score in their previous qualification, offering insights into their academic preparedness and potential for success in higher education.

Socio-economic Factors:

- Mother's qualification and Father's qualification: These categorical variables describe the educational attainment of the student's parents, potentially reflecting the student's socioeconomic background and access to educational resources.
- Mother's occupation and Father's occupation: These categorical variables describe the professions of the student's parents, further providing insights into the family's economic stability and social standing.
- Displaced: This binary variable indicates whether the student is displaced, potentially due to economic hardship, family circumstances, or other factors, which could impact their ability to focus on their studies.
- Educational special needs: This binary variable identifies students with learning disabilities or other special needs that may require additional support and resources.
- Debtor: This binary variable indicates whether the student has outstanding debts, potentially leading to financial stress and challenges in managing academic responsibilities.
- Tuition fees up to date: This binary variable reveals whether the student is current on their tuition payments, further reflecting their financial stability and commitment to their education.
- Scholarship holder: This binary variable indicates whether the student receives a scholarship, potentially signifying their academic merit or financial need.

Behavioral Indicators:

- Daytime/evening attendance: This binary variable distinguishes between students attending daytime classes and those attending evening classes, potentially revealing differences in their schedules, commitments, and learning preferences.
- Curricular units (1st and 2nd semester): These groups of variables detail the number of curricular units the student is enrolled in, credited for, has evaluations for, has approved, and their grades for each semester. This data offers a comprehensive overview of their academic performance and engagement.
- Curricular units (without evaluations): These variables for both semesters identify the number of curricular units for which the student did not receive evaluations, potentially indicating missed classes, incomplete coursework, or other issues.

External Factors:

- Unemployment rate, Inflation rate, GDP: These numerical variables represent macroeconomic indicators at the time of the student's enrollment, potentially influencing their financial situation and future job prospects.

1) **Data Exploration and Preprocessing: Laying the Foundation**
   The initial phase involves a comprehensive exploration of the dataset to understand its structure, identify potential issues, and prepare it for model training. This includes several crucial steps:

- Descriptive statistics and visualization: Calculating summary statistics for each feature and creating visualizations, such as histograms, scatter plots, and correlation matrices, will reveal the distribution of data, identify potential outliers, and uncover relationships between variables.
- Missing value imputation: Missing data can negatively impact model performance. Techniques like mean/median imputation for numerical features and mode imputation or more advanced methods like KNN imputation for categorical features will be employed to address missing values effectively.
- Outlier treatment: Outliers can skew model results. Statistical methods like z-score or interquartile range (IQR) will be used to detect outliers, and appropriate actions such as winsorizing, trimming, or transformation will be taken to mitigate their influence.
- Categorical variable encoding: Machine learning algorithms typically require numerical input. Techniques like one-hot encoding or label encoding will be applied to transform categorical variables into numerical representations suitable for model training.

- Data splitting: The dataset will be divided into training and testing sets while maintaining the original class distribution. Stratified sampling will ensure that each set accurately reflects the proportion of "Dropout", "Enrolled", and "Graduated" cases present in the original data.

2) **Class Imbalance Techniques: Ensuring Fairness and Accuracy**
   Given the imbalanced nature of the dataset, with potentially fewer "Dropout" cases compared to "Enrolled" or "Graduated", specific techniques will be employed to mitigate bias and improve the model's ability to identify at-risk students:

- Over-sampling: Techniques like random oversampling or synthetic data generation methods like SMOTE (Synthetic Minority Oversampling Technique) or ADASYN (Adaptive Synthetic Sampling) will be explored to increase the representation of the minority "Dropout" class.
- Under-sampling: Methods like random under-sampling or Tomek links can be used to reduce the number of instances from the majority classes, bringing the class distribution closer to balance.
- Class weight adjustment: Assigning higher weights to the "Dropout" class during model training can emphasize its importance and encourage the model to learn its characteristics more effectively.

3) **Experiment with Models: Exploring Diverse Algorithm Options**
   A range of machine learning algorithms well-suited for classification tasks will be experimented with to determine the most effective approach for dropout prediction:

- Logistic regression: This classic algorithm provides a baseline for comparison and offers interpretable results, revealing the contribution of each feature to the prediction.
- Decision trees and Random forests: These models can capture non-linear relationships and handle both categorical and numerical features effectively.
- Gradient boosting: Algorithms like XGBoost or LightGBM are known for their high accuracy and ability to handle complex data patterns.
- Support vector machines (SVMs): SVMs can be effective in high-dimensional spaces and offer good generalization performance.
- Neural networks: Deep learning architectures, such as multi-layer perceptrons or recurrent neural networks, will be explored for their ability to learn complex relationships and potentially improve prediction accuracy.

4) **Ensemble Methods: Combining Strengths for Improved Performance**
   Ensemble techniques, which combine multiple models, will be employed to leverage the strengths of different algorithms, and potentially achieve superior prediction results:

- Bagging (Bootstrap aggregating: Methods like Random Forest, which builds an ensemble of decision trees, can reduce variance, and improve model stability.
- Boosting: Algorithms like XGBoost or AdaBoost combine weak learners into a strong learner, iteratively improving performance by focusing on misclassified instances.

5) **Feature Selection: Focusing on What Matters Most**

Feature selection methods will be applied to identify the most informative features for dropout prediction, leading to improved model performance and interpretability:

- Filter methods: Techniques like chi-squared tests or information gain can be used to select features based on their statistical association with the target variable.
- Wrapper methods: Methods like recursive feature elimination (RFE) iteratively train models with different feature subsets, selecting the combination that yields the best performance.
- Embedded methods: Algorithms like LASSO (Least Absolute Shrinkage and Selection Operator) regression perform feature selection as part of the model training process, shrinking the coefficients of less important features to zero.

6) **Iterative Approach: Continuous Refinement for Optimal Results**

An iterative approach will be adopted to optimize model performance.

- Model selection and hyperparameter tuning: Different combinations of algorithms and hyperparameter settings will be evaluated using cross-validation techniques to select the best performing model.
- Evaluation and analysis: The chosen model will be evaluated on the testing set using appropriate metrics, such as F1-score, AUC-ROC, and confusion matrix, to assess its performance and identify areas for further improvement.
- Iteration and refinement: Based on the evaluation results, the model and feature selection process will be iteratively refined to achieve optimal prediction accuracy and generalizability.

### 2.3. Exploratory Data Analysis (EDA)

In our dataset, among all the students, roughly 50% of students have graduated, 32% students have dropped out and almost 18% of the students are currently enrolled in any given course. Distribution shows that the majority of the students are in their late teens to early 20's. There is an increase in dropout rate from students in their mid 20's to early 30's. Female students are more as compared to male students. There was a higher rate of dropout students that were male (45.1%), compared to the females (25.1%). Among the female students, there were 25.1% dropouts, 57.9% graduates and 17% currently enrolled in a class. And in male students, 45.1% were dropouts, 35.2% were graduates and 19.7% currently enrolled in a class.

The course “Biofuel Production Technologies” saw the highest number of dropouts: 66.7% for a course. Followed by Equinculture (55.3%) and Informatics Engineering (54.1%). Less % dropouts were for Nursing (15.4%).

Among the entire student population, the most prevalent outcome is graduation, with roughly 50% of students successfully completing their studies. However, a substantial proportion of students, around 32%, experience dropout, highlighting a significant challenge faced by the institution. The remaining 18% of students are currently enrolled and their ultimate outcomes remain to be determined. A striking disparity is observed when comparing dropout rates between students with and without debt. Among students with no debt, the dropout rate is 28.3%, indicating that over a quarter of these students still face challenges leading to attrition despite not having financial burdens from debt. However, the dropout rate jumps significantly to 62% for students with debt, suggesting a strong correlation between financial strain and the risk of dropping out. The contrast between students with and without debt strongly suggests a correlation between student debt and an increased risk of dropout. Debt can create significant financial pressure, leading to stress, anxiety, and challenges in balancing academic responsibilities with the need to work and manage finances. This financial strain can ultimately lead students to make the difficult decision to discontinue their studies.

For students without scholarships, the distribution of outcomes is more balanced. "Graduates" still constitute the largest group at 41.3%, suggesting that many students can succeed without scholarship support. However, a significant portion, 38.7%, fall into the "Dropout" category, indicating that financial constraints or other challenges may hinder their academic progress. The remaining 20% represent "Enrolled" students whose final outcomes remain uncertain. For scholarship recipients, a remarkable 76% of these students achieve "Graduation", signifying a substantial increase in academic success compared to their non-scholarship counterparts. This suggests that scholarship support plays a significant role in empowering students to persist in their studies and achieve their academic goals. The "Dropout" rate plummets to 12.2%, indicating that scholarships effectively mitigate factors contributing to attrition. The proportion of "Enrolled" students is similar to the no-scholarship group at 11.8%, suggesting that scholarship recipients may experience fewer obstacles in continuing their education.

Among students who have not paid Tuition fees, a staggering 86.6% of these students fall into the "Dropout" category, highlighting a strong association between financial delinquency and attrition. This suggests that financial hardship and the inability to meet tuition obligations significantly impede a student's ability to continue their education. The remaining percentages are minimal, with only 8% of students currently "Enrolled" and a mere 5.5% achieving "Graduation". For students with paid tuition fees reveals a more balanced distribution of outcomes. "Graduates" constitute the largest group at 56%, indicating that a majority of students who fulfill their financial responsibilities successfully complete their studies. The "Dropout" rate is significantly lower at 24.7%, suggesting that financial stability plays a crucial

role in reducing attrition. The proportion of "Enrolled" students is notably higher at 19.3% compared to the "Tuition Not Paid" group, suggesting that students who manage their finances are more likely to persist in their academic programs. Students who struggle to pay tuition fees may face immense financial pressure, leading to stress, anxiety, and challenges in balancing academic demands with financial obligations. This can result in disengagement from studies, difficulty focusing on coursework, and ultimately, the decision to drop out. Conversely, students who maintain their tuition payments demonstrate a commitment to their education and are more likely to experience the academic and personal benefits associated with completing their studies.

The imbalanced nature of the dataset, with a significantly higher proportion of "Graduate" cases compared to "Dropout" instances, poses a significant challenge for training machine learning models. If left untreated, this imbalance can bias the model towards the majority class, leading to poor performance in accurately predicting dropout risk. To mitigate this issue and ensure fairness and accuracy, this project explores various undersampling and oversampling techniques.

**Recursive Forward Elimination (RFE) method**

Feature selection is an important process in machine learning. One way to do it is by using the recursive forward elimination method. In this method, we apply logistic regression to the complete set of features and then drop the least important feature. We repeat this process until we find the optimal number of features.

There are different ways to implement this method. One way is to specify the required number of features, and the program will reduce the feature set until it reaches that number. However, this method may not guarantee the best accuracy.

Another way is to continue iterating until the model's performance starts decreasing. This way, we can find the optimal number of features that provide the best performance. Even a slight reduction in performance will stop the iteration.

It's important to balance the tradeoff between the number of features and accuracy. If reducing the number of features results in only a small reduction in accuracy, then it's worth it. In our implementation, we have added a 'variance' parameter to balance this tradeoff. We limit the number of features between 3 and 10 to get the best of both approaches.

We used both the lasso logistic model and the ridge logistic model to find out which method gives appropriate features as per real-life implementation.

**Addressing Class Imbalance**

Despite its richness, the dataset exhibits class imbalance, with varying proportions of students in each outcome class. This presents a challenge for machine learning models, as they may be

biased towards the majority class and struggle to accurately predict minority classes like "Dropout". Techniques to address this imbalance will be explored in the project, ensuring a fair and accurate assessment of dropout risk for all students.
To achieve the project's objectives, a series of meticulously planned methods will be applied, ensuring a robust and reliable dropout prediction model. These methods encompass data exploration and preprocessing, class imbalance treatment, model experimentation, ensemble methods, feature selection, and an iterative refinement approach.

**1) Random Under-sampling: Reducing the Majority Class**

Random undersampling aims to balance the dataset by randomly removing instances from the majority "Graduate" class until its size matches that of the minority "Dropout" class. This straightforward approach can effectively address class imbalance and prevent the model from being overwhelmed by the characteristics of the majority class. However, random undersampling also has drawbacks. By discarding a substantial portion of the data, it can lead to the loss of valuable information and potentially reduce the model's overall generalizability.

**2) Oversampling Techniques: Augmenting the Minority Class**

Oversampling techniques, in contrast to undersampling, focus on increasing the representation of the minority 'Dropout' class. Random Oversampling technique increases the size of the minority class by randomly duplicating existing instances. While simple to implement, random oversampling can exacerbate overfitting, as the model may memorize the duplicated data points rather than learning the underlying patterns of dropout risk.

**3) SMOTE (Synthetic Minority Oversampling Technique)**

SMOTE offers a more sophisticated approach by generating synthetic samples of the minority class. An instance of the minority class is selected and K-Nearest Neighbors within the same class is identified to achieve SMOTE. New synthetic instances are then created by interpolating between the original instance and its neighbors, effectively expanding the minority class with data points that share similarities with existing ones while adding diversity to the feature space.

**4) ADASYN (Adaptive Synthetic Sampling Approach)**

ADASYN builds upon SMOTE by focusing on generating synthetic samples in regions of the feature space where the minority class is more difficult to learn. This method adaptively assigns weights to minority class instances based on their difficulty level, with more synthetic samples generated for instances surrounded by majority class examples. This targeted approach effectively shifts the decision boundary of the model towards challenging areas, improving its ability to distinguish between "Dropout" and "Graduate" cases.

5) **XGBoost (Extreme Gradient Boosting)**

XGBoost delves into the details of each student's story - their background, how they're doing academically, and even their financial situation - to uncover the reasons behind dropout cases. By sifting through past student records, XGBoost learns to pick up on subtle clues and patterns that signal potential dropout risks.

But XGBoost doesn't stop at just making predictions. It goes a step further by highlighting the key factors driving these predictions. Is it a student's age, their parents' background, or perhaps something about the courses they're enrolled in? XGBoost shines a light on these important factors, giving educators and policymakers valuable insights into where they should direct their attention.

### 2.4. Evaluation Metric(s)

We took into consideration the following metrics -

- Accuracy
- Precision
- Recall
- F-1 Score

For the following methods, we accounted the above metrics -

- Logistic Regression
    - Lasso Regression
    - Lasso Regression with RFE method
    - Ridge Regression
    - Ridge Regression with RFE method
- Random Forest Algorithm
- XGBoost Algorithm

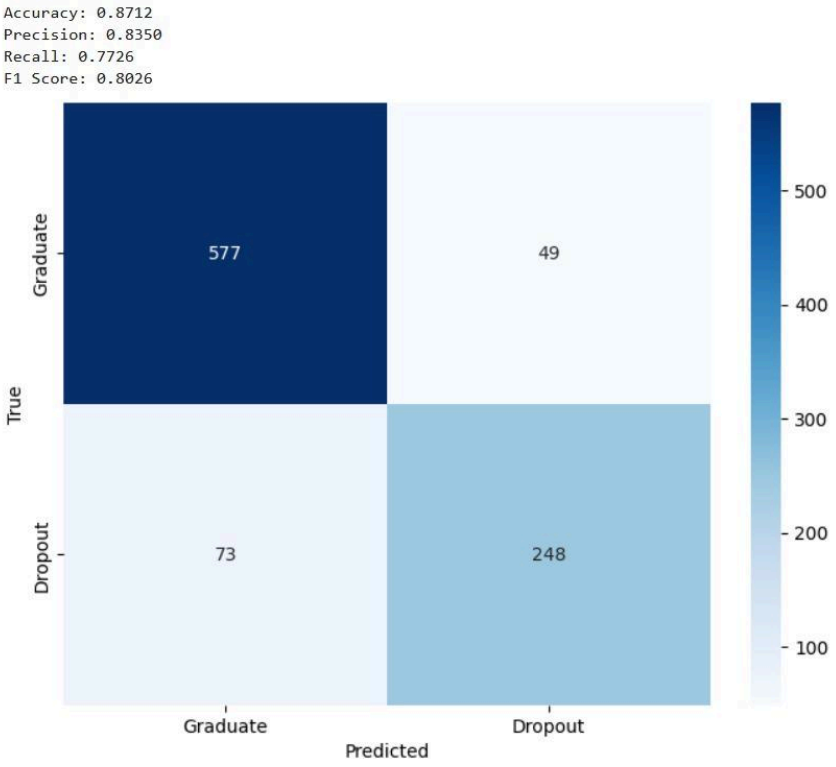
### 2.5. Expectation(s) of Results

- We expected Logistic Regression to perform better as compared to Random Forest algorithm.
- As per our dataset, we thought Feature Selection method - Recursive Forward Elimination - will help in selecting better features to predict the at-risk students'.
- Boosting methods will help in feature selection and in turn, improving the accuracy and efficiency of the model.

## 3. Results & Discussions

- We expected balanced datasets to outperform unbalanced datasets, but there was no significant difference and the best model that we got was for an unbalanced dataset.

- Contrary to our expectations, the boosting method did not provide any significant improvement. Their results were at par with random forest, with random forest performing slightly better.
- Random forest models consistently outperformed both l1 and l2 logistic regression models which were consistent with our expectations.
- Although at first glance, logistic regression models had a very good accuracy, at par with random forest, there was a huge imbalance in the precision and recall. They had very good precision with poor recall, which is not beneficial for our use case.
- Apart from Ridge logistic regression, for all other methods, the correlation matrix feature selection method provided better results as compared to 'Recursive Forward Elimination'.




'Random Forest for Unbalanced Data' using the feature set received from correlation matrix is the best performing model for our dataset based on the accuracy, recall and F-1 score. The misclassified dropout students also came out least for this method

Schools and policymakers can tailor their strategies to provide targeted support to at-risk students. By intervening early and offering the right kind of help, they can boost student success and keep more students on the path to graduation.

**4. References/Citations**

- Dataset : https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success
- ChatGPT : https://chat.openai.com/
- GitHub : https://github.com/Damiieibikun/Student-s-Dropout-Prediction-using-Supervised-Machine-Learning-Classifiers
- Research Paper : http://dx.doi.org/10.1007/978-3-030-72657-7_16

**5. Appendix (for data file, Program / Software tools file, additional results)**

- Exploratory Data Analysis Jupyter Notebook
- Feature Selection Jupyter Notebook