



**BITS Pilani**  
**DIGITAL**  
*Excellence made yours*

## Advanced Apex Project - 1

Dr.Naga Janapati

Associate Professor, BITS Pilani Digital

# Advanced Apex project

---

An Apex project is a project with higher order skills as outcomes by integrating the learning from multiple courses, typically in the same trimester.

\* 2 Credits

**Faculty**

# Learning Objectives

---

By the end of the Advanced Apex Project, students should be able to:

- Integrate data preprocessing, feature engineering, modelling and visualization into a coherent pipeline.
  - Handle end-to-end data challenges.
  - Communicate insights through dashboards and introductory storytelling with visuals.
-

# Phases & Student Deliverables

Phase	Weeks	Deliverables
<b>Phase 1:</b>  Proposal & Data Acquisition	2-3	<p><b>Teams formed. Supervisor assigned. Problem statement given.</b></p> <p><u>1-2-page <b>Proposal</b> (PDF/Word):</u></p> <ul style="list-style-type: none"><li>• Project Title, Problem statement &amp; business goal (Ex. House Price prediction, customer segmentation or Demand forecasting, etc.)</li><li>• Data Sources: Ex. Kaggle, UCI ML, GitHub, etc.</li><li>• Tools: Python, Pandas, Matplotlib/ Seaborn, Tableau/PBI, etc.</li><li>• Work flow: High-level flow of the project (e.g., Data acquisition → Preprocessing → Feature Engineering → Modelling → Visualization).</li></ul> <p><u><b>Data Extraction:</b></u></p> <ul style="list-style-type: none"><li>• From Kaggle, UCI ML, GitHub, etc. (Use Kaggle API or direct download; optional use of publicly enabled AWS S3 buckets for extra practice).</li><li>• A short notebook or script showing <b>how</b> the dataset is pulled (API call / download) and saved reproducibly.</li></ul> <p><u><b>Schema/Data dictionary:</b></u></p> <p>(Data Model Excel sheet: Feature – Data type – Description – PK (Yes/No))).</p> <p>* Students should prepare this document by checking the metadata file provided with the dataset (if available). If no metadata is provided, they need to do a quick Python-based inspection of the extracted dataset.</p>

## 2. Phases & Student Deliverables

Phase	Weeks	Deliverables
Phase 2: Preprocessing & Feature Engineering	4-7	<p><b>NOTE:</b> Students should perform only the tasks that are relevant to their dataset and model. They are not expected to do every single step listed below.</p> <p><b>Data Audit &amp; Data availability check:</b></p> <ul style="list-style-type: none"><li>● Check shape, data types, missing values, quality flags (e.g., negative ages), etc. (df.shape, df.dtypes, df.isnull().sum(), etc)</li><li>● Check if the task relevant columns are available. (Eg. Loan Approval Application: Annual Income, Employee?, etc)</li></ul> <p><b>Exploratory Data Analysis (EDA):</b></p> <ul style="list-style-type: none"><li>● <b>Summary statistics:</b> .describe(), .quantile(), .value_counts(), Correlations (df.corr()), slicing/dicing, roll-up, etc.</li><li>● <b>Visualizations:</b><ul style="list-style-type: none"><li>- <i>Univariate:</i> histograms/KDE (numeric), bar/count plots (categorical).</li><li>- <i>Bivariate:</i> scatter plots, boxplots, crosstabs.</li><li>- <i>Multivariate:</i> correlation heatmap, pair plots.</li></ul></li></ul> <p><b>Data Cleaning:</b></p> <ul style="list-style-type: none"><li>● Drop irrelevant columns (e.g., Cust_Name).</li><li>● Drop the columns with too many missing values.</li><li>● Impute missing values: (mean/median/mode/forward fill/backward fill; justify, e.g., "Median for skewed 'salary'", etc).</li><li>● Remove duplicates.</li><li>● Handle outliers: (IQR/z-score; cap/remove).</li><li>● Convert types: e.g., strings to datetime, etc.</li></ul>

# Phases & Student Deliverables

Phase	Weeks	Deliverables
Phase 2: Preprocessing & Feature Engineering	4-7	<p><b>Feature Engineering:</b></p> <ul style="list-style-type: none"><li>● Data scaling/normalization (MinMax, Z-score), categorical encoding (One-Hot, Label, Target, etc.).</li><li>● Feature extraction from raw data/ existing features. <b>Ex.</b> Extracting "Days Since Purchase"/ "Purchase_Recency" from purchase_date and current_date, ratios like CPM (Cost Per Mile) from some other existing features, etc.</li><li>● Feature subset selection: Select task-relevant features using filter methods (e.g., correlation analysis, multicollinearity check)</li><li>● Dimensionality Reduction (optional): PCA, SVD, DWT, etc.</li></ul> <p><b>NOTE:</b> Depending on Model to be constructed, engineer at least a <i>minimum set</i> of task-relevant features.</p> <ul style="list-style-type: none"><li>- Jupyter Notebook, document each step <b>inline</b> (code + outputs + explanations).</li><li>- Place observations/insights close to the relevant code/output.</li><li>- At the end, include a <b>final summary/conclusion</b> (often a markdown cell summarizing all key findings).</li></ul>

# Phases & Student Deliverables

Phase	Weeks	Deliverables
Phase 3: Modeling & Inferencing	8-9	<p><b>Model construction.</b> Build at least one appropriate basic model based on the problem type, such as:</p> <ul style="list-style-type: none"><li>● <b>Regression Models:</b> Simple Linear Regression, Multiple Linear Regression.</li><li>● <b>Clustering Models:</b> e.g., K-Means, Hierarchical Clustering.</li><li>● <b>Time Series Models (Univariate):</b> Forecasting using engineered features like Moving Average, Exponential Moving Average, etc. (Students should explore which feature/method forecasts the next <b>month/day</b> value most effectively.)</li></ul> <p>• <b>Evaluation Metrics</b> Compute <b>relevant evaluation metrics</b>, such as:</p> <ul style="list-style-type: none"><li>● Regression → RMSE, MAE, <math>R^2</math></li><li>● Clustering → Silhouette Score, Davies–Bouldin Index</li><li>● Time Series → RMSE, MAPE</li></ul> <p>• <b>Presentation</b> Notebook should clearly show:</p> <ul style="list-style-type: none"><li>● Data input</li><li>● Model construction steps</li><li>● Metric computation &amp; output</li></ul>

# Phases & Student Deliverables

Phase	Weeks	Deliverables
Phase 4: Visualization & Storytelling	10	<p><b>Option A: Tableau / Power BI (preferred, if comfortable)</b></p> <ul style="list-style-type: none"><li>● <b>Interactive Dashboard</b> (1–2 dashboards):<ul style="list-style-type: none"><li>○ Show results from EDA and/or model outcomes.</li><li>○ Include 1–2 key <b>metrics (KPIs)</b> (e.g., Median Price, RMSE) if available.</li><li>○ Include 1–2 <b>dimensions</b> (filters) (e.g., City, Bedrooms) for simple interactivity.</li></ul></li><li>● <b>5-Slide Story Deck</b> (PDF/PPT):<ul style="list-style-type: none"><li>○ Title &amp; project context.</li><li>○ Visual highlights (EDA or model results).</li><li>○ Key metric(s) / dimension(s) if available.</li><li>○ 1–2 key observations (what the charts show).</li><li>○ Simple conclusion or “next steps.”</li></ul></li></ul> <p><b>Option B: Matplotlib / Seaborn (Notebook-based)</b></p> <ul style="list-style-type: none"><li>● <b>Notebook with 3–5 clear plots</b> (serves as a static dashboard):<ul style="list-style-type: none"><li>○ Plots from <b>EDA and/or model outcomes</b>.</li><li>○ If filters are not possible, show segmented plots (e.g., price by bedrooms).</li><li>○ Include 1–2 <b>key metrics</b> in text cells (if available).</li></ul></li><li>● <b>5-Slide Story Deck</b> (same as Option A).</li></ul>

# Phases & Student Deliverables

Phase	Weeks	Deliverables
<b>Phase 5:</b> Documentation & Submission	10	<p><b>Final code (Jupyter Notebooks)</b></p> <ul style="list-style-type: none"><li>● All their analysis (Cleaning, EDA, Feature Engineering, modelling, visualization) in one or more well-structured notebooks.</li></ul> <p><b>README</b></p> <ul style="list-style-type: none"><li>● Short guide for reviewing the project. Include:<ul style="list-style-type: none"><li>○ Problem statement / business goal (1-2 sentences)</li><li>○ Dataset source(s) and clear citation (URL or name, e.g., Kaggle/UCI)</li><li>○ Steps to run the notebooks (e.g., pip install -r requirements.txt → open notebook → run all cells)</li></ul></li></ul> <p><b>Source Links</b></p> <ul style="list-style-type: none"><li>● Explicit reference to original datasets: Kaggle URL, UCI ML link, or GitHub repo link</li><li>● Include small note if sample data is provided due to privacy</li></ul> <p><b>Dashboard Output</b></p> <ul style="list-style-type: none"><li>● <b>Matplotlib/Seaborn path:</b> screenshots of 3-5 key plots OR indicate in the notebook that plots are present</li><li>● <b>Tableau/Power BI path:</b> workbook file (.twbx / .pbix) OR exported screenshots / shareable link</li><li>● Include brief captions or text explaining what the visuals show</li></ul> <p><b>requirements.txt</b></p> <ul style="list-style-type: none"><li>● Python version, key libraries, versions, and dependencies required to run the notebook</li><li>● Example: pandas==2.1.0, numpy==1.25.0, matplotlib==3.8.0, etc.</li></ul>

# Phases & Student Deliverables

Phase	Weeks	Deliverables
Presentation & Evaluation	11-12	<ul style="list-style-type: none"><li>Each team will deliver a <b>final presentation</b> summarizing their overall project, supported by a <b>PowerPoint deck, Jupyter Notebook(s), and other relevant artifacts</b> (e.g., dashboard, data dictionary).</li><li>During the presentation, teams should demonstrate their workflow, key findings, and recommendations.</li><li><b>Every student is expected to have a clear understanding of all phases of the project</b> (proposal, preprocessing, feature engineering, modeling, visualization, and storytelling), not just the part they individually worked on.</li></ul> <p>*Please note that the timelines provided are indicative. While we will make every effort to adhere to them, any changes, if required, will only involve advancing certain activities to an earlier date.</p>