**Data Science Sheet**

Topics to cover :

**1. Mathematical Foundations**

* **Linear Algebra**
  + Vectors, Matrices, Tensors
  + Matrix operations (multiplication, transpose, inverse)
  + Eigenvalues and Eigenvectors
  + Singular Value Decomposition (SVD)
  + Applications in PCA and ML models
* **Vector Calculus**
  + Derivatives & gradients
  + Partial derivatives
  + Jacobians and Hessians
  + Gradient Descent optimization

**2. Probability**

* Basics of Probability
  + Sample space, Events
  + Conditional Probability, Bayes’ Theorem
  + Law of Total Probability
* Random Variables
  + Discrete vs Continuous
  + Probability Distribution Functions (PDF, CDF)
  + Expectation, Variance, Covariance
* Common Distributions
  + Normal, Binomial, Poisson, Exponential
* **Applications in ML**
  + Naive Bayes
  + Probabilistic models

**3. Statistics**

* Descriptive Statistics
  + Mean, Median, Mode
  + Variance, Standard Deviation
* Inferential Statistics
  + Sampling methods
  + Confidence Intervals
* Hypothesis Testing
  + Null vs Alternative hypothesis
  + p-value, t-test, z-test, chi-square test, ANOVA
* Regression
  + Linear regression, Logistic regression
  + Assumptions of regression

**4. Programming**

* **Python** (main language for DS)
  + Data types, Loops, Functions, OOP basics
  + Libraries: **NumPy, Pandas, Matplotlib, Seaborn**
* **R** (optional but useful for statistical analysis)
* **SQL**
  + SELECT, JOIN, GROUP BY, Subqueries
  + Window functions
* **Git**
  + Version control basics (clone, commit, push, pull)

**5. Feature Engineering**

* Handling Missing Values
* Encoding Categorical Variables
  + One-Hot Encoding, Label Encoding
* Feature Scaling
  + Normalization vs Standardization
* Feature Selection
  + Correlation, Mutual Information, PCA

**6. Data Visualization**

* **Matplotlib & Seaborn** in Python
* **Tableau / Power BI** for dashboards
* **Excel** basics for quick analysis
* Types of plots:
  + Histograms, Box plots, Heatmaps, Pair plots

**7. Machine Learning**

* **ML Workflow**
  + Data preprocessing → Model training → Evaluation
* **Key Algorithms**
  + Linear & Logistic Regression
  + Decision Trees, Random Forests
  + k-NN, Naive Bayes
  + SVM
  + Gradient Boosting (XGBoost, LightGBM, CatBoost)
* **Model Evaluation**
  + Accuracy, Precision, Recall, F1-score
  + ROC, AUC
  + Cross-validation
  + Bias-Variance Tradeoff
* **Overfitting vs Underfitting**

**8. Deep Learning**

* Basics of Neural Networks
  + Perceptron, Activation functions
* Backpropagation & Gradient Descent
* Architectures:
  + CNN (Computer Vision)
  + RNN, LSTM, GRU (Sequential data)
* Frameworks:
  + **TensorFlow, PyTorch**
* Transfer Learning

**9. NLP (Natural Language Processing)**

* Text preprocessing
  + Tokenization, Lemmatization, Stemming
  + Stopword removal
* Feature Extraction
  + Bag of Words, TF-IDF
  + Word Embeddings (Word2Vec, GloVe, FastText)
* NLP Algorithms
  + Sentiment analysis
  + Named Entity Recognition
* **Transformers**
  + BERT, GPT models (modern NLP)

**10. Big Data**

* Big Data concepts
  + Hadoop ecosystem
  + Spark (PySpark)
* Distributed computing
* Handling large datasets

**11. Cloud Computing**

* AWS / Google Cloud / Azure
  + Basics of Cloud deployment
  + Storage & Compute services (S3, EC2)
* **Containerization**
  + Docker basics
* **Orchestration**
  + Kubernetes (optional for advanced)

**12. Deployment**

* **Flask / FastAPI** for API creation
* Docker for containerizing ML models
* Deploy on:
  + AWS
  + Azure
  + Google Cloud