PREDICTING SMOKING STATUS FROM BIO-SIGNALS USING SUPERVISED LEARNING

FINAL PROJECT — DTSA-5509
INTRODUCTION TO MACHINE LEARNING: SUPERVISED LEARNING

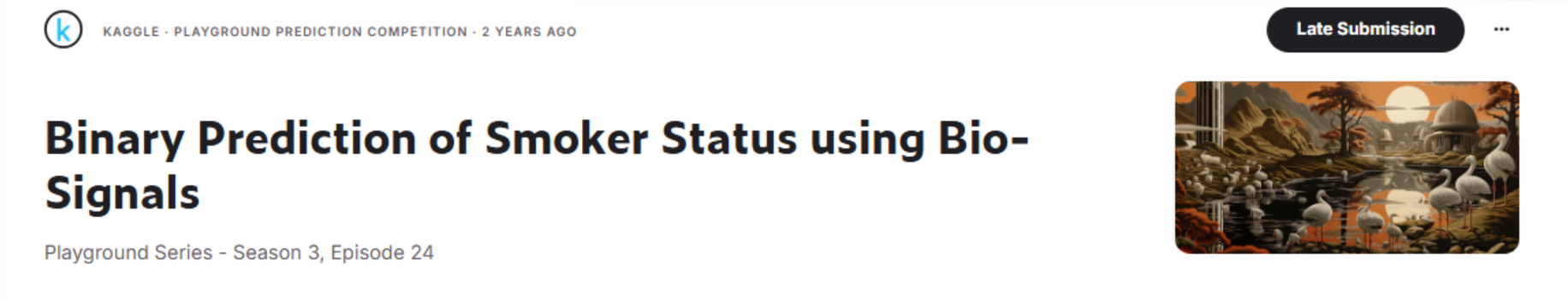
PRESENTED BY JINNAJATE ACHALAPONG

PROBLEM DEFINITION

- **Goal:**
  - Build a predictive model to classify individuals as smoker or non-smoker using health-related bio-signal features.

- **Why it matters:**
  - Supports smoking cessation initiatives
  - Early detection via passive health signals
  - Reduces risk of smoking-related diseases

DATASET OVERVIEW



Source: Kaggle Playground Series – Season 3, Episode 24
Type: Tabular | Binary classification
Size: ~160,000 rows, 22 features

Example Features:

- Age, height, weight
- Blood pressure, cholesterol, triglycerides
- Hearing/vision tests, liver enzymes (GTP)

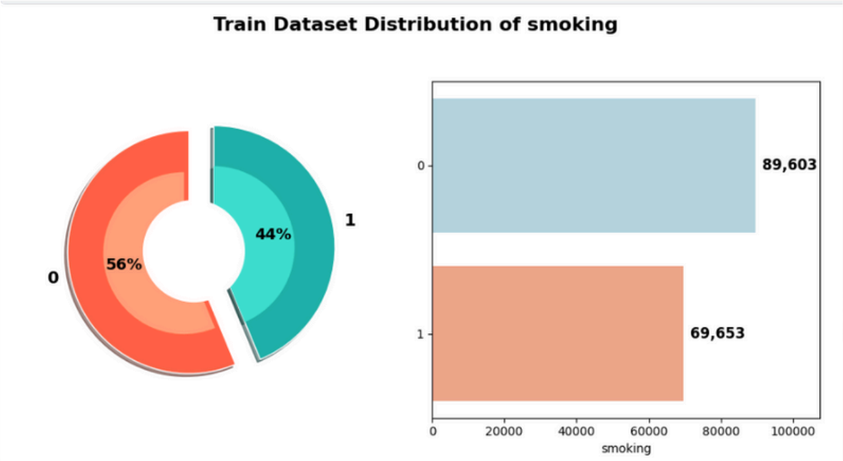
DATA PREPARATION

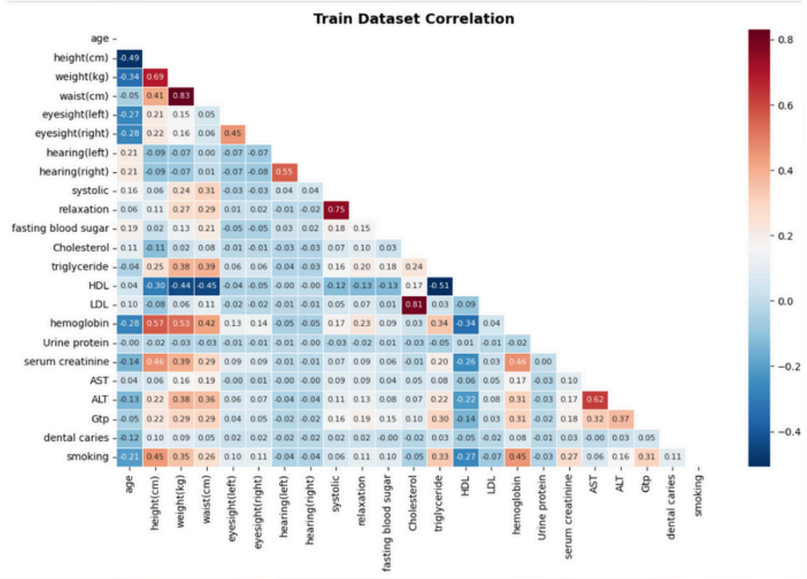
Data shape: (159256, 24)

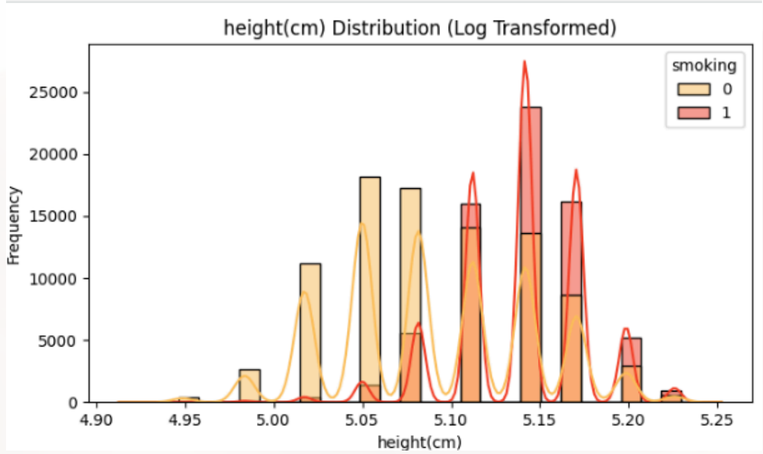
	data type	#missing	%missing	#unique	min	max	average	standard_deviation	first value	second value	third value
id	int64	0	0.000000	159256	0.000000	159255.000000	79627.500000	45973.391572	0.000000	1.000000	2.000000
age	int64	0	0.000000	18	20.000000	85.000000	44.306626	11.842286	55.000000	70.000000	20.000000
height(cm)	int64	0	0.000000	14	135.000000	190.000000	165.266929	8.818970	165.000000	165.000000	170.000000
weight(kg)	int64	0	0.000000	28	30.000000	130.000000	67.143662	12.586198	60.000000	65.000000	75.000000
waist(cm)	float64	0	0.000000	531	51.000000	127.000000	83.001990	8.957937	81.000000	89.000000	81.000000
eyesight(left)	float64	0	0.000000	20	0.100000	9.900000	1.005798	0.402113	0.500000	0.600000	0.400000
eyesight(right)	float64	0	0.000000	17	0.100000	9.900000	1.000989	0.392299	0.600000	0.700000	0.500000
hearing(left)	int64	0	0.000000	2	1.000000	2.000000	1.023974	0.152969	1.000000	2.000000	1.000000
hearing(right)	int64	0	0.000000	2	1.000000	2.000000	1.023421	0.151238	1.000000	2.000000	1.000000
systolic	int64	0	0.000000	112	77.000000	213.000000	122.503648	12.729315	135.000000	146.000000	118.000000
relaxation	int64	0	0.000000	75	44.000000	133.000000	76.874071	8.994642	87.000000	83.000000	75.000000
fasting blood sugar	int64	0	0.000000	229	46.000000	375.000000	98.352552	15.329740	94.000000	147.000000	79.000000
Cholesterol	int64	0	0.000000	227	77.000000	393.000000	195.796165	28.396959	172.000000	194.000000	178.000000
triglyceride	int64	0	0.000000	392	8.000000	766.000000	127.616046	66.188989	300.000000	55.000000	197.000000
HDL	int64	0	0.000000	108	9.000000	136.000000	55.852684	13.964141	40.000000	57.000000	45.000000
LDL	int64	0	0.000000	222	1.000000	1860.000000	114.607682	28.158931	75.000000	126.000000	93.000000
hemoglobin	float64	0	0.000000	134	4.900000	21.000000	14.796965	1.431213	16.500000	16.200000	17.400000
Urine protein	int64	0	0.000000	6	1.000000	6.000000	1.074233	0.347856	1.000000	1.000000	1.000000
serum creatinine	float64	0	0.000000	28	0.100000	9.900000	0.892764	0.179346	1.000000	1.100000	0.800000
AST	int64	0	0.000000	140	6.000000	778.000000	25.516853	9.464882	22.000000	27.000000	27.000000
ALT	int64	0	0.000000	188	1.000000	2914.000000	26.550296	17.753070	25.000000	23.000000	31.000000
Gtp	int64	0	0.000000	362	2.000000	999.000000	36.216004	31.204643	27.000000	37.000000	53.000000
dental caries	int64	0	0.000000	2	0.000000	1.000000	0.197996	0.398490	0.000000	1.000000	0.000000
smoking	int64	0	0.000000	2	0.000000	1.000000	0.437365	0.496063	1.000000	0.000000	1.000000

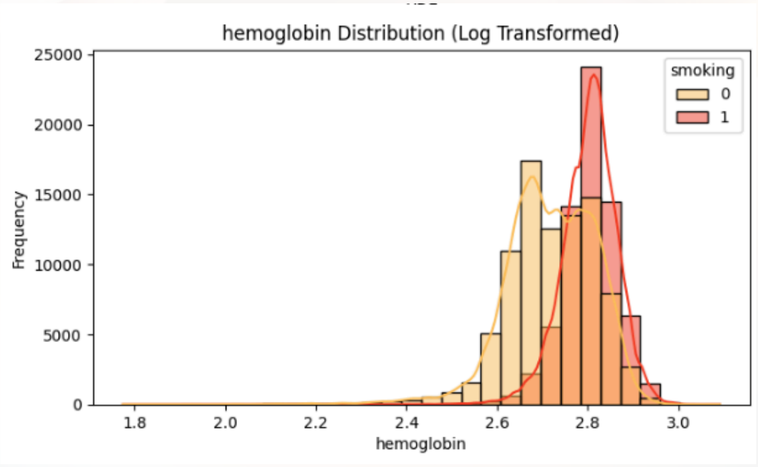
- Checked for missing values → clean dataset
- Applied log transformation to reduce skew
- Visualized target class balance:
  - Non-smokers ≈ 56%
  - Smokers ≈ 44%

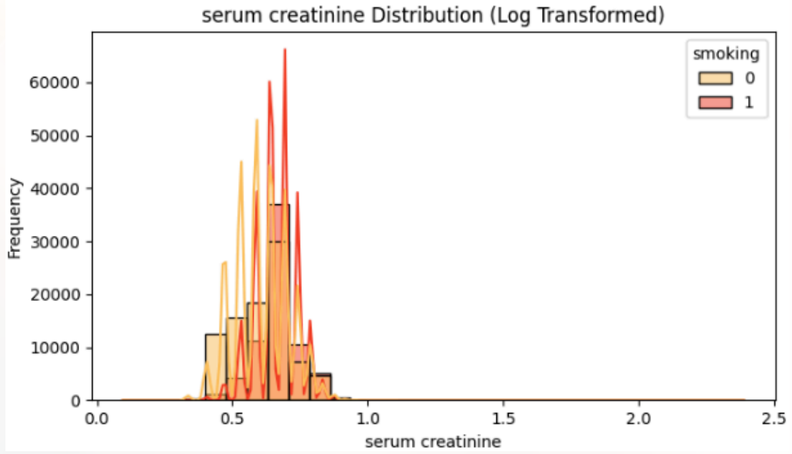
EXPLORATORY DATA ANALYSIS (EDA)









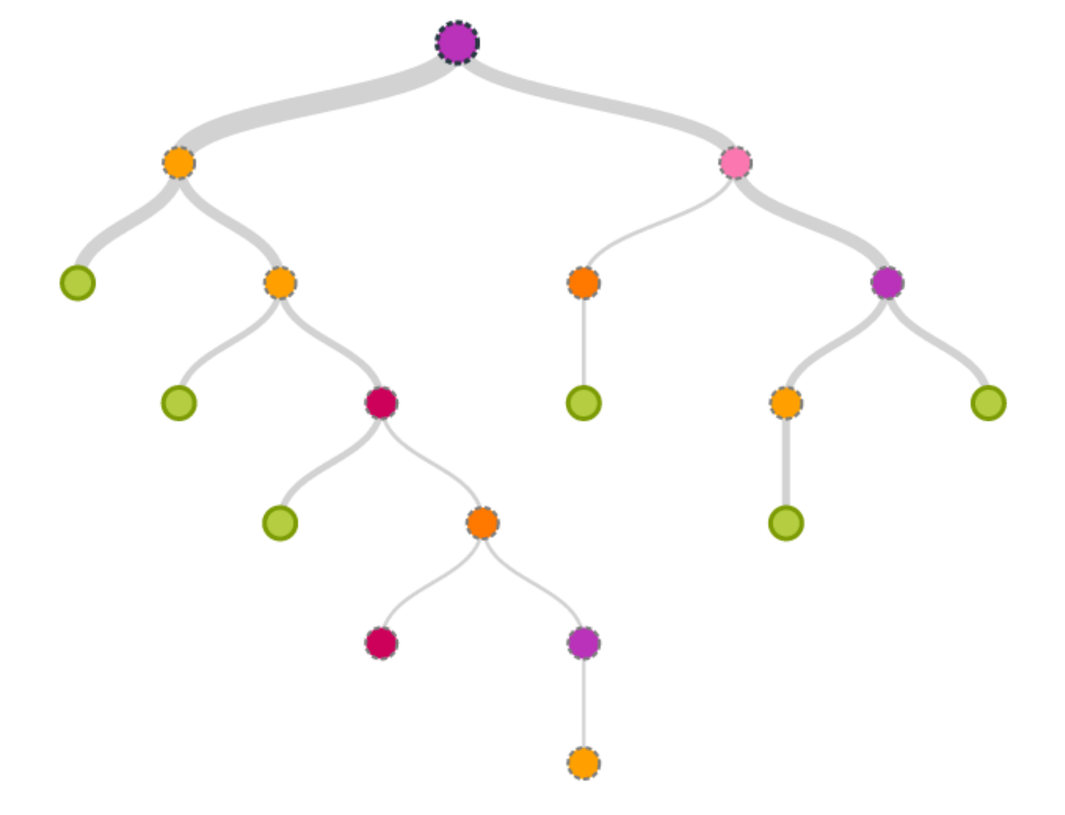


MODELING APPROACH

- **Models used:**
  - Decision Tree
  - Random Forest
  - XGBoost (Default)
  - XGBoost (Tuned via Optuna)

- **Evaluation Metrics:**
  - Accuracy, Precision, Recall, F1 Score, AUC
  - 10-Fold Stratified Cross-Validation

- **GPU Acceleration (device='cuda')**

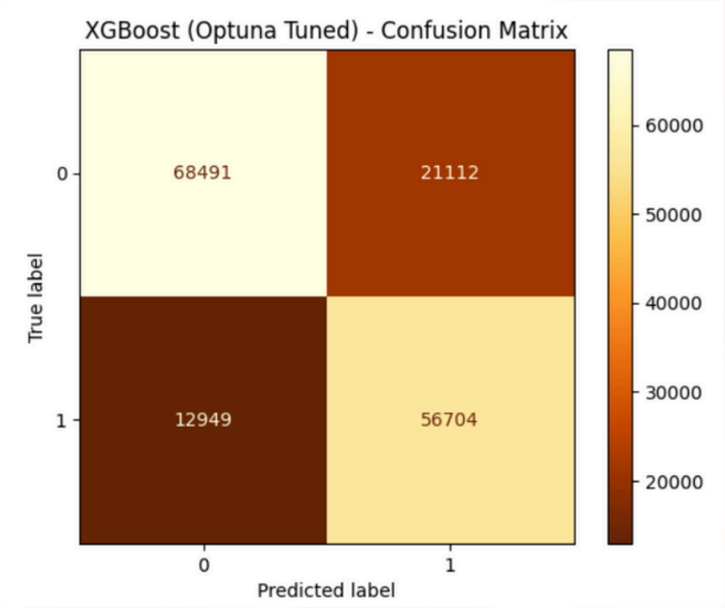


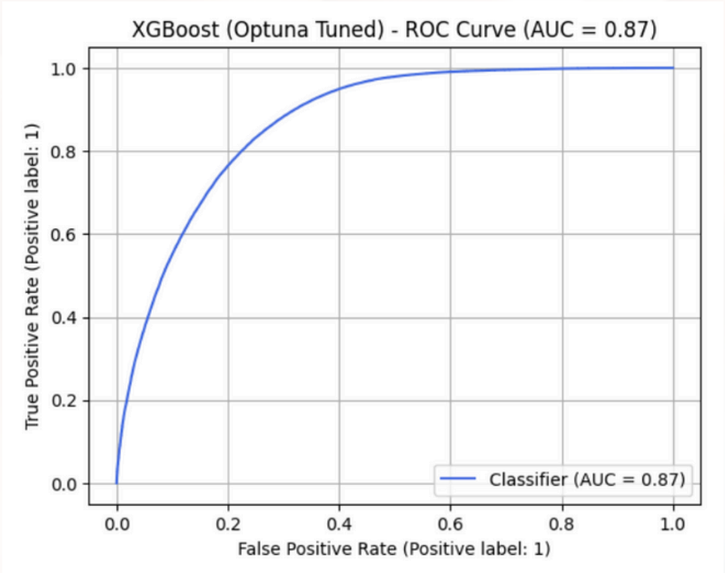
MODEL PERFORMANCE COMPARISON

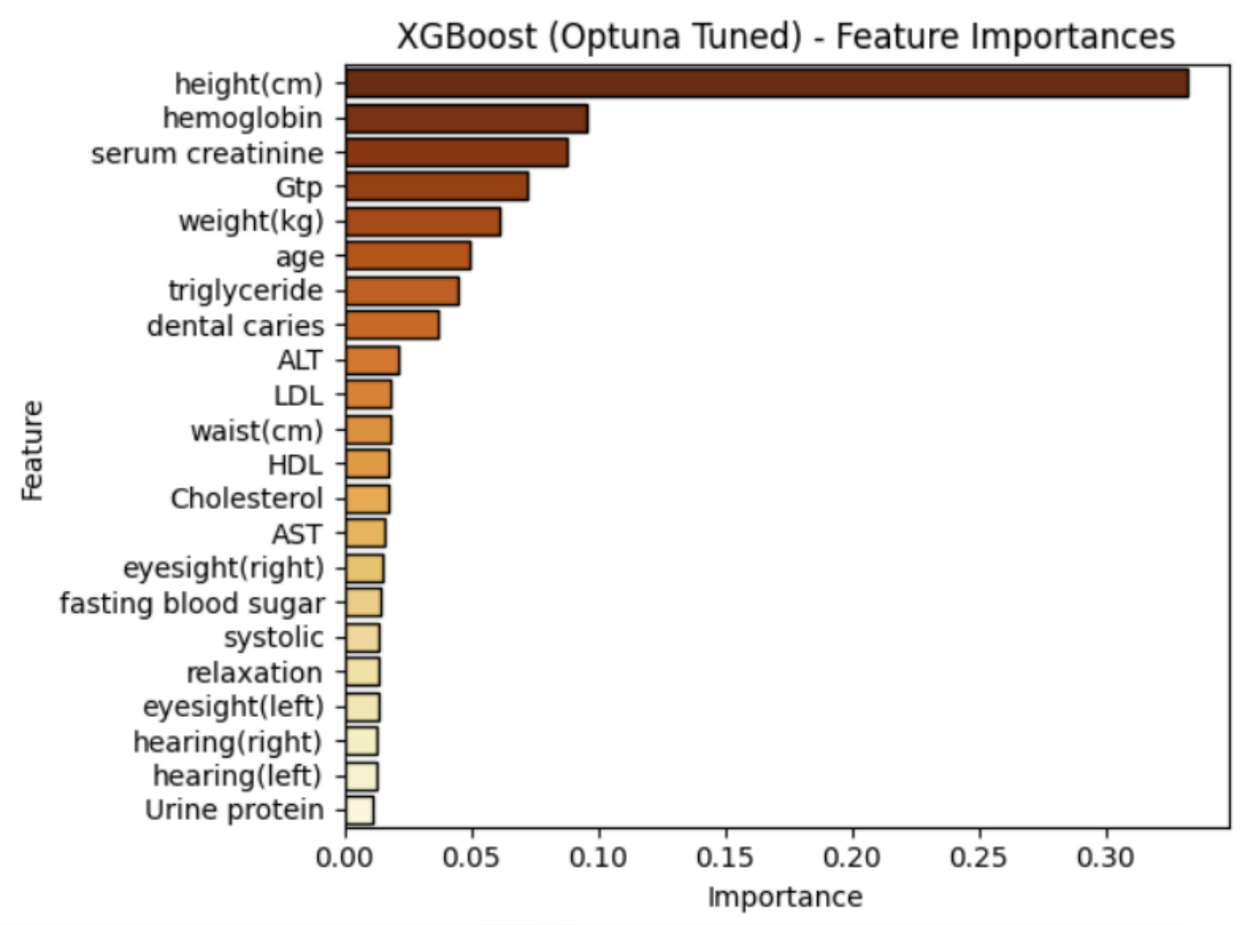
Model	Accuracy	Precision	Recall	F1 Score	AUC
Decision Tree	0.6945	0.6518	0.6475	0.6496	0.6893
Random Forest	0.7745	0.7121	0.8133	0.7593	0.8574
XGBoost (Default)	0.7814	0.7238	0.8087	0.7639	0.8656
XGBoost (Optuna Tuned)	**0.7861**	**0.7287**	**0.8141**	**0.7690**	**0.8706**

- The **Decision Tree** model performs the worst across all metrics, confirming its limitations when applied alone to complex tabular data.
- **Random Forest** significantly improves upon Decision Tree performance benefiting from ensemble learning and reduced overfitting.
- **XGBoost (default)** outperforms Random Forest in all metrics showing the strength of boosting methods in this problem.
- **XGBoost (Optuna Tuned)** yields the best performance achieving the highest AUC (0.8706), which indicates its strong ability to rank predictions correctly.

XGBOOST (OPTUNA TUNED) RESULT







DISCUSSION & CONCLUSION

- XGBoost with Optuna tuning yields best AUC (0.87)
- Tree-based models outperform baseline Decision Tree
- Feature analysis aligns with domain knowledge

FUTURE WORK

- Next steps to improve the project:
- Add engineered features (e.g., BMI, health risk scores)
- Address class imbalance (e.g., SMOTE, focal loss)
- Try interpretability tools like SHAP
- Validate on external or hold-out test set