미션15 보고서_최종

1. Docker Hub

> **Docker Hub URL**

---

2. 데이터 전처리 및 모델링

- **데이터**: `mission15_train.csv`, `mission15_test.csv`
- **목표변수**: *Performance Index* (학업 성취도, 10~100)
- **입력변수**: Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, Sample Question Papers Practiced

전처리 및 모델링

- 수치형: `StandardScaler`, 범주형: `OneHotEncoder`
- 통합: `ColumnTransformer`
- 모델: `RandomForestRegressor(n_estimators=200, random_state=42)`
- 데이터 분리: Train 80% / Valid 20%
- 파이프라인: Preprocessor → RandomForest → 모델 저장(`model.pkl`)

결과

- 검증 RMSE: **2.25**
- 데이터 수: Train 5,600 / Valid 1,400
- 주요 산출물: `model.pkl`, `train_report.json`, `result.csv`, `train_log.txt`
- 예측 분포 40~42 피크는 트리모델의 평균값 수렴 현상으로 **이상치 아님**

---

3. 프로젝트 구조 및 환경

```
mission-result/
├── docker-compose.yml
├── data/ (train, test)
├── r1/ (Dockerfile, train.py)
```

```
└── shared/ (model.pkl, result.csv 등)
```

- `r1/` : 연구자 1 코드
- `shared/` : 컨테이너 간 공유 폴더
- `docker-compose.yml` : 환경 정의 및 볼륨 연결
- `requirements.txt` : 주요 버전 고정
  (scikit-learn, pandas, numpy, joblib 등)

공유 볼륨 설정

```
volumes:
  - ./shared:/artifacts
```

---

4. 결론 및 향후 개선

결론

- Docker 환경을 통해 동일 버전·패키지 기반의 **재현 가능한 머신러닝 파이프라인** 구축
- 전처리–학습–평가–결과 저장 자동화 완료

향후 개선

1. **성능 향상**: 교차검증, 하이퍼파라미터 탐색
2. **데이터 확장**: 파생변수 추가, 반올림 규칙 통일
3. **재현성 관리**: MLflow/DVC 실험 추적
4. **운영화**: FastAPI 추론 API, CI/CD 자동화
