CSC461 - Machine Learning

# *Project - Javier Sin & Nicolás Pelegrín*

## Progress Report Requirements

- 2-3 pages (PDF)

**- problem and dataset clearly defined (must)**

- any preliminary results/experiments

**- plan for next steps (must)**

## Final Report Requirements

Enhanced progress report + full details on experiments and analysis

## Brainstorming

### **1. Sports Injury Prediction and Prevention**

* **Problem:** Predict the likelihood of a player getting injured based on previous games, physical characteristics, playing time, etc. You could focus on a specific sport like soccer, basketball, or American football.
* **Dataset:** Look for open sports datasets or gather data from sports analytics websites.
* **Techniques:** Use **feature engineering** (previous injuries, fatigue, playing position, etc.), **supervised learning** (e.g., logistic regression, decision trees), **PCA** to reduce dimensions, and possibly **k-Nearest Neighbors** for finding similar players and their injury history.
* **Challenge Factor:** You can extend the project by integrating time-series analysis (tracking players’ performance and fatigue over time) or by applying **Neural Networks** to more complex data like movement and play patterns.

### **2. Mental Health Prediction Using Social Media Data**

* **Problem:** Predict mental health issues (e.g., anxiety, depression) based on social media posts or interactions, potentially helping in early detection.
* **Dataset:** Twitter API, Reddit, or Kaggle mental health datasets. You could also use sentiment analysis datasets as a base.
* **Techniques:** **Sentiment analysis** (using text-based features), **NLP models** (bag of words, TF-IDF), **clustering** for grouping similar users, and **logistic regression** or **decision trees** for classification.
* **Challenge Factor:** Advanced NLP techniques like **LSTM/RNN** for better context, or applying **PCA** for dimensionality reduction of high-dimensional text data.

### **3. Fake News Detection**

* **Problem:** Classify articles or social media posts as fake or real news, a relevant and impactful issue today.
* **Dataset:** Fake news datasets from Kaggle or similar sources.
* **Techniques:** **Logistic regression** and **decision trees** for initial classification, **feature transformation** on text data (using TF-IDF), **PCA** for dimensionality reduction, and possibly **Neural Networks** or **Bagging and Boosting** for model improvement.
* **Challenge Factor:** Combining NLP with clustering to detect common patterns in fake news and real news articles.

### **4. Personalized Music Recommender System Using Mood Analysis**

* **Problem:** Build a music recommendation system that suggests songs based on the user’s current mood.
* **Dataset:** Use a combination of datasets with song metadata (Spotify API, Million Song Dataset) and sentiment analysis datasets.
* **Techniques:** Use **clustering** to group songs by mood or energy level, **PCA** to reduce dimensionality in high-dimensional audio features, and **k-Nearest Neighbors** for recommendations based on user mood.
* **Challenge Factor:** Incorporate **sentiment analysis** for mood detection from text (user’s social media posts or messages), and **Neural Networks** for learning complex patterns in song data.

## Tips to follow:

1. **Choose an impactful dataset** that aligns with your chosen problem.
2. **Incorporate multiple techniques**: Show your understanding by applying clustering, supervised learning, regularization, and any other relevant techniques.
3. **Experiment and Analyze**: Show preliminary results, analyze failures or weaknesses, and plan improvements.
4. **Focus on interpretability**: Explain why certain features or models work better for your problem.
5. **Presentation**: Make your report clear and your presentation engaging, focusing on the real-world impact of your work.