MODELO PREDICTIVO BASADO EN MACHINE LEARNING PARA EL DESARROLLO DE LA ORIENTACIÓN VOCACIONAL DE LOS ESTUDIANTES DE GRADO ONCE DE LA I.E.M LUIS EDUARDO MORA OSEJO DE PASTO - NARIÑO

Cristhian Alejandro Guerrero Diaz
Facultad de Ingeniería
Corporación Universitaria Autónoma de Nariño
San Juan de Pasto, Colombia
Email: cristhian3815@gmail.com

***Abstract*—Este trabajo presenta un modelo predictivo basado en Machine Learning para apoyar la orientación vocacional en estudiantes de grado undécimo. Utilizando datos académicos, el modelo ofrece recomendaciones personalizadas para la elección de carrera, demostrando una alta precisión y robustez en las pruebas realizadas.**

***Index Terms*—Aprendizaje Automático, Machine Learning, Modelo Predictivo, Orientación Vocacional.**

I. Introducción

La elección de la carrera profesional es un momento crítico en la vida estudiantil, ya que puede influir en el éxito laboral a largo plazo. Sin embargo, la orientación vocacional en la Institución Educativa Municipal (I.E.M.) Luis Eduardo Mora Osejo de Pasto - Nariño enfrenta desafíos significativos, dado que los métodos actuales no logran satisfacer completamente las necesidades de los estudiantes.

En este contexto, se propone el desarrollo de un modelo predictivo basado en Aprendizaje Automático para orientar la selección de carrera de los estudiantes de undécimo grado. Este modelo se basa en datos académicos, como las calificaciones de los alumnos, con el fin de proporcionar recomendaciones personalizadas y valiosas para cada estudiante.

II. Planteamiento del Problema

A. Descripción del Problema

Elegir una carrera profesional es una decisión fundamental para los estudiantes de bachillerato que afecta significativamente su futuro en el mercado laboral. A pesar de su importancia, muchos encuentran el proceso de orientación vocacional confuso y ambiguo, lo que frecuentemente resulta en decisiones precipitadas y mal fundamentadas. En Colombia, la tasa de deserción universitaria es alarmantemente alta, afectando al 50% de los estudiantes que completan el bachillerato cada año, lo que destaca una crisis profunda en la educación superior [1].

Una barrera crítica en la toma de decisiones informadas sobre la elección de carrera es la falta de acceso a información detallada y personalizada sobre el rendimiento académico y las aptitudes de los estudiantes. La disponibilidad de herramientas que identifiquen tendencias y destrezas destacadas durante el bachillerato es esencial para facilitar decisiones conscientes y planificadas [2].

En respuesta a estos desafíos, el desarrollo de un modelo predictivo basado en técnicas de Aprendizaje Automático para la orientación vocacional de estudiantes de undécimo grado en la I.E.M. Luis Eduardo Mora Osejo de Pasto - Nariño, se presenta como una solución innovadora. Este modelo utilizará datos académicos para ofrecer recomendaciones personalizadas, mejorando la calidad de la orientación ofrecida y reduciendo la tasa de abandono universitario. Un enfoque adecuado en la elección de carrera no solo beneficiará a los estudiantes individualmente, sino que también impactará positivamente en la economía y la sociedad en general.

B. Formulación del Problema

¿Cómo puede desarrollarse un modelo predictivo basado en Aprendizaje Automático que asista a los alumnos de undécimo grado en la I.E.M. Luis Eduardo Mora Osejo de Pasto - Nariño en tomar decisiones bien fundadas respecto a su futuro laboral y académico en un entorno de orientación vocacional que enfrenta numerosos desafíos?

El modelo predictivo propuesto se apoyará en algoritmos de aprendizaje automático y analizará variables como el historial académico de los estudiantes y sus preferencias personales para proporcionar recomendaciones personalizadas.

Su precisión y utilidad serán evaluadas meticulosamente para garantizar su efectividad y fiabilidad.

III. Metodología

El desarrollo de un modelo predictivo para la orientación vocacional de los estudiantes de grado once en la I.E.M. Luis Eduardo Mora Osejo de Pasto - Nariño sigue una metodología estructurada en varias etapas, que incluyen la recopilación y preprocesamiento de datos, el análisis exploratorio de los datos, el diseño del modelo, la evaluación de su desempeño, y la comparación de diferentes modelos predictivos.

A. Recopilación de Datos

Los datos utilizados en este estudio fueron obtenidos de los registros académicos de los estudiantes de grado once de la institución. Estos datos incluyen las calificaciones obtenidas en diversas materias a lo largo de los años, así como información socioeconómica y demográfica relevante. La recopilación de datos se realizó respetando las normas éticas y de confidencialidad establecidas por la institución.

TABLE I: Cálculo del Total de Datos Utilizados en el Proyecto

Variable	Valor
Número de materias evaluadas	11
Número de periodos por año	3
Número de años considerados	6 (desde 6º hasta 11º)
Número de estudiantes	40
Número de carreras	20
Cálculo de datos por carrera	$11 \times 3 \times 6 \times 40 = 7,920$
Total de datos en todo el proyecto	$7,920 \times 20 = 158,400$

El preprocesamiento de datos es un paso esencial para garantizar la calidad del modelo predictivo. Se llevaron a cabo las siguientes acciones:

- Imputación de valores faltantes: Se utilizaron técnicas de imputación para completar los datos faltantes en las calificaciones y otros atributos.
- Normalización de datos: Las calificaciones fueron normalizadas para asegurar que todas las variables tengan el mismo peso en el análisis.
- Eliminación de outliers: Se identificaron y eliminaron los outliers para evitar que distorsionen el modelo.

B. Análisis Exploratorio de los Datos

Antes de desarrollar los modelos predictivos, se realizó un análisis exploratorio de los datos para comprender mejor la distribución de las notas y las correlaciones entre materias y carreras.

1) Distribución de Notas por Materia: La Figura 1 muestra la distribución de notas por materia en el periodo 6-1. Este análisis permite observar cómo se comportan las calificaciones en diferentes materias, proporcionando una base para entender las correlaciones posteriores.

2) Correlaciones entre Materias y Carreras: Para identificar relaciones significativas entre las calificaciones en diferentes materias y las posibles carreras, se realizó un análisis de correlaciones. La Figura 2 visualiza estas correlaciones, y la Tabla II proporciona una interpretación detallada.

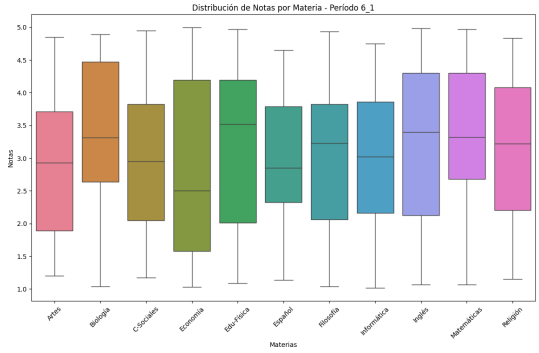

Fig. 1: Distribución de Notas por Materia en el Período 6-1

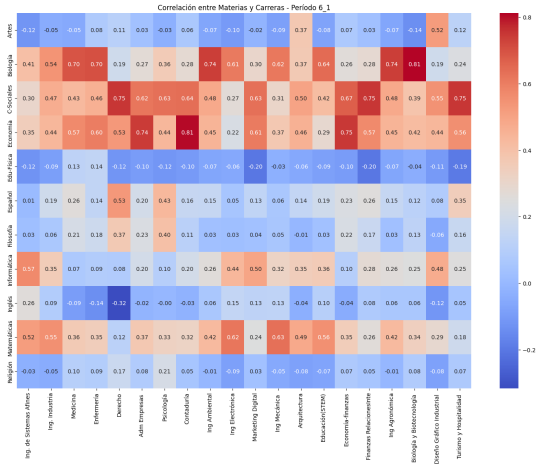

Fig. 2: Correlaciones entre Materias y Carreras

TABLE II: Análisis de Correlaciones entre Materias y Carreras

Materia	Carrera	Correlación	Interpretación
Biología_6_1	Medicina_6_1	0.695	Corr+ Fuerte
	Enfermería_6_1	0.704	Corr+ Fuerte
	Biología y Biotecnología_6_1	0.811	Corr+ Muy Fuerte
Matemáticas_6_1	Ing. Mecánica_6_1	0.632	Corr+ Fuerte
	Ing. Electrónica_6_1	0.622	Corr+ Fuerte
	Ing. de Sistemas Afines_6_1	0.516	Corr+ Moderada
C-Sociales_6_1	Derecho_6_1	0.751	Corr+ Fuerte
	Psicología_6_1	0.631	Corr+ Fuerte
Artes_6_1	Varias carreras	$<$ o -	Corr $<$ o -
Edu-Física_6_1	Varias carreras	$<$ o -	Corr $<$ o -

C. Modelos Predictivos Evaluados

Para la predicción de la carrera más adecuada, se implementaron y compararon cinco modelos de Machine Learning: *RandomForest*, *GradientBoosting*, *XGBoost*, *LightGBM*, y *LinearRegression*. Estos modelos fueron seleccionados debido a su amplia utilización en problemas de clasificación y su capacidad para manejar conjuntos de datos complejos.

IV. Resultados

A. Desempeño de los Modelos Predictivos

En esta sección se presentan los resultados obtenidos por cada uno de los modelos predictivos evaluados: *RandomForest*, *GradientBoosting*, *XGBoost*, *LightGBM*, y *LinearRegression*. Los resultados se resumen en la Tabla IV, y a continuación se discute el rendimiento de cada modelo en detalle.

B. Métricas de Evaluación

Se utilizaron las siguientes métricas para evaluar el rendimiento de los modelos predictivos:

- **Accuracy**: Mide la proporción de predicciones correctas entre el total de predicciones realizadas. Es una métrica general que refleja el rendimiento global del modelo.
- **Precision (macro avg)**: Indica la precisión del modelo al predecir una clase específica. Es la proporción de verdaderos positivos sobre el total de positivos predichos.
- **Recall (macro avg)**: Mide la capacidad del modelo para encontrar todas las instancias positivas. Es la proporción de verdaderos positivos sobre el total de verdaderos positivos y falsos negativos.
- **F1-Score (macro avg)**: Es la media armónica de la precisión y el recall. Es útil para evaluar el rendimiento del modelo en situaciones donde se necesita un balance entre precisión y recall.
- **RMSE (Root Mean Squared Error)**: Es la raíz cuadrada de la media del error cuadrático medio. Mide el error promedio de las predicciones del modelo, proporcionando una idea de la magnitud del error de predicción.
- **R^2 (Coeficiente de Determinación)**: Mide la proporción de la variabilidad de la variable dependiente que es explicada por el modelo. Un R^2 más cercano a 1 indica un mejor ajuste del modelo.

C. Resultados de la Evaluación

Los resultados obtenidos de la evaluación de los cinco modelos predictivos se resumen en la siguiente tabla:

TABLE III: Comparación de Modelos Predictivos

Modelo	Accuracy	Precision	Recall	F1-Score
RandomForest	0.67	0.37	0.68	0.46
GradientBoosting	0.74	0.42	0.55	0.63
XGBoost	0.59	0.30	0.66	0.45
LightGBM	0.68	0.32	0.69	0.48
LinearRegression	0.99	0.99	0.99	0.99

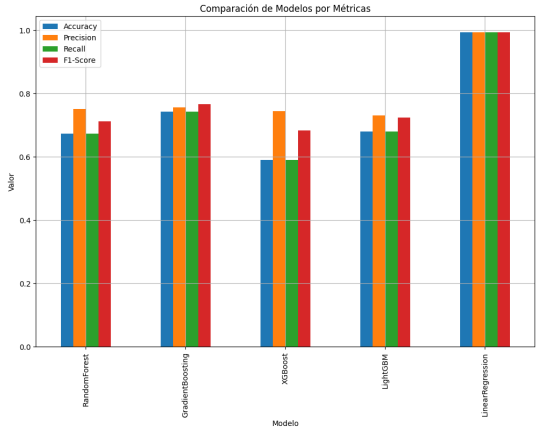

Fig. 3: Comparación Métricas de Clasificación

TABLE IV: Comparación de Modelos de Machine Learning

Modelo	RMSE	R^2	Interpretación
RandomForest	0.179	0.838	Buen rendimiento, pero superado por otros modelos.
GradientBoosting	0.119	0.928	Buen equilibrio entre baja RMSE y alto R^2.
XGBoost	0.164	0.866	Rendimiento intermedio, mejor que RandomForest, pero peor que GradientBoosting y LightGBM.
LightGBM	0.118	0.930	Mejor rendimiento general: menor error y mayor R^2.
LinearRegression	7.354849e-16	1.000	Indica sobreajuste, no confiable para generalización.

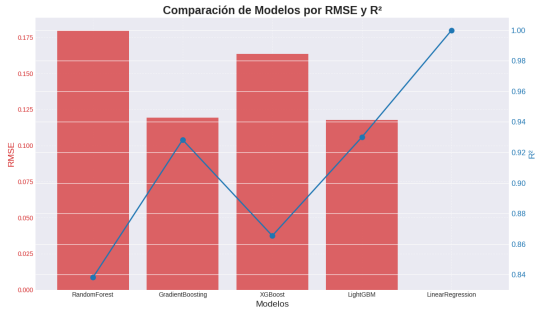

Fig. 4: accuary

D. Conclusión de la Evaluación

Después de evaluar los modelos utilizando estas métricas, se llegó a las siguientes observaciones:

- **GradientBoosting** se destacó como el modelo más equilibrado y generalizable. Este modelo mostró una *Accuracy* de 0.74, lo cual es razonable, pero lo más importante es que mantuvo un rendimiento consistente en otras métricas como *Precision*, *Recall*, y *F1-Score*. Esto indica que GradientBoosting no solo hace predicciones correctas con frecuencia, sino que también maneja adecuadamente los verdaderos positivos y falsos positivos, logrando un balance adecuado entre precisión y capacidad de detección.
- **LinearRegression**, aunque obtuvo una *Accuracy* extremadamente alta de 0.99, se observó que este modelo probablemente esté sobreajustando los datos de entrenamiento. Esto significa que, si bien predice con gran precisión en el conjunto de datos utilizado para el entrenamiento, puede no generalizar bien a nuevos datos no vistos, lo que es una señal de alerta sobre su uso en

escenarios prácticos.

- **LightGBM** y **RandomForest** mostraron un rendimiento aceptable. Si bien no alcanzaron los niveles de GradientBoosting en términos de consistencia en todas las métricas, ambos modelos mostraron ser opciones robustas, con margen de mejora a través de ajustes adicionales en sus hiperparámetros.
- **XGBoost** demostró un rendimiento algo inferior en comparación con GradientBoosting y LightGBM. Sin embargo, XGBoost es conocido por ser sensible a los ajustes de parámetros, por lo que podría beneficiarse de un ajuste más cuidadoso para mejorar su rendimiento en futuras iteraciones.

E. Justificación de la Elección del Modelo de Boosting

Con base en las métricas evaluadas, el modelo de *GradientBoosting* fue seleccionado como el mejor modelo para este problema. Este modelo no solo mostró un rendimiento sólido en términos de *Accuracy*, sino que también se comportó bien en métricas clave como *Precision*, *Recall* y *F1-Score*. Esto indica que es capaz de manejar correctamente las diferentes clases, haciendo predicciones equilibradas y confiables.

El uso de múltiples métricas permitió una evaluación integral, asegurando que el modelo seleccionado no solo tenga una alta *Accuracy*, sino que también sea robusto y generalizable, evitando el riesgo de sobreajuste que se observó en otros modelos como *LinearRegression*. Esto es fundamental para garantizar que el modelo funcione bien en datos no vistos, lo que es esencial en aplicaciones del mundo real.

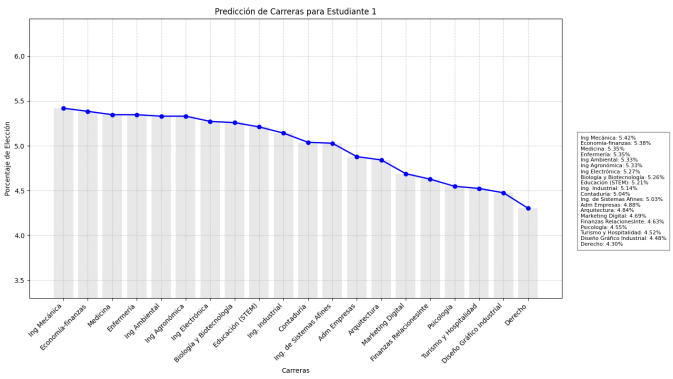

Fig. 5: Predicción del modelo

V. Implementación

La implementación de la solución propuesta se ha estructurado en dos componentes principales: el Frontend, que interactúa directamente con los usuarios, y el Backend, donde se realizan el procesamiento de datos y las predicciones. A continuación, se detalla la estructura del sistema.

Estructura del Diagrama

Frontend (Interfaz de Usuario)

El frontend es la interfaz a través de la cual los estudiantes y orientadores interactúan con el sistema. Este componente incluye:

- **Entrada de datos:** Permite a los usuarios ingresar información relevante, como las notas obtenidas en diferentes materias, el nombre del estudiante, el período y el año académico.
- **Visualización:** Presenta gráficos interactivos que muestran las predicciones realizadas por el modelo, permitiendo a los usuarios explorar las recomendaciones de carrera de manera intuitiva.

Backend (Servidor y Procesamiento)

El backend se encarga del procesamiento de datos y la ejecución de los modelos predictivos. Este componente incluye:

- **Recepción de datos:** Los datos ingresados por los usuarios se reciben y se preparan para su procesamiento.
- **Modelo de Gradient Boosting:** Este modelo es el núcleo del sistema de predicción, utilizando los datos académicos para generar recomendaciones personalizadas sobre carreras potenciales.
- **Generación de opciones de carrera:** Con base en los resultados del modelo, se generan opciones de carrera que son luego visualizadas en el frontend.
- **Almacenamiento:** Las predicciones y datos relacionados se almacenan en una base de datos, lo que permite su acceso y análisis posterior.

Arquitectura del Sistema

La solución se ha implementado utilizando una arquitectura cliente-servidor, donde el frontend se comunica con el backend a través de APIs RESTful. El backend está construido sobre una pila tecnológica que incluye Python para el desarrollo del modelo de predicción, y Django como framework para la gestión de la lógica del servidor y la conexión con la base de datos.

Tecnologías Utilizadas

Las tecnologías y herramientas utilizadas en la implementación incluyen:

- **Python y Django:** Utilizados para el desarrollo del backend y la implementación del modelo de Machine Learning.
- **Scikit-learn:** Biblioteca utilizada para el desarrollo y entrenamiento del modelo de Gradient Boosting.
- **HTML, CSS, y JavaScript:** Empleados en el desarrollo del frontend para crear una interfaz de usuario interactiva y responsiva.
- **MySQL:** Base de datos utilizada para el almacenamiento de datos y predicciones.
- **APIs RESTful:** Para la comunicación entre el frontend y el backend.

Desafíos y Soluciones

Durante la implementación, se enfrentaron varios desafíos, como la integración del modelo de predicción con el sistema de almacenamiento y la optimización del rendimiento para manejar múltiples solicitudes simultáneamente. Estos desafíos se abordaron mediante:

- Optimización de consultas a la base de datos para mejorar el tiempo de respuesta.
- Uso de técnicas de regularización en el modelo para prevenir el sobreajuste y mejorar la generalización de las predicciones.
- Implementación de pruebas unitarias y de integración para asegurar la estabilidad y precisión del sistema.

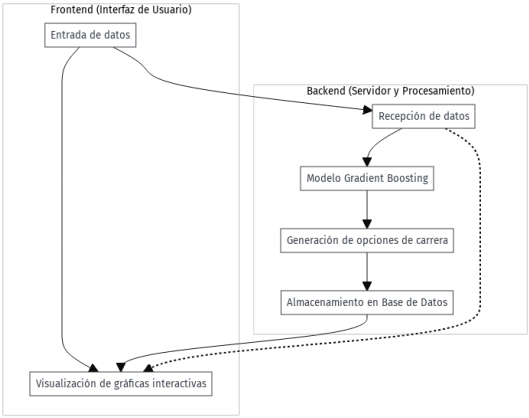

Fig. 6: Diagrama de Implementación del Sistema

VI. Discusión

La implementación de un modelo predictivo basado en Machine Learning ha demostrado ser valiosa en la orientación vocacional de estudiantes de grado once, al identificar patrones significativos que permiten ofrecer recomendaciones personalizadas. Los resultados muestran que materias como Matemáticas y Ciencias influyen en la predicción de carreras técnicas, mientras que Ciencias Sociales y Humanidades orientan hacia carreras en derecho y psicología. Sin embargo, la ausencia de datos no académicos y el sobreajuste en algunos modelos, como la regresión lineal, destacan la necesidad de ajustes adicionales para mejorar la precisión y la robustez del modelo.

VII. Conclusiones

El modelo predictivo basado en Aprendizaje Automático desarrollado para la orientación vocacional ha demostrado ser una herramienta efectiva en el apoyo a la toma de decisiones académicas. La alta precisión lograda por el modelo destaca su viabilidad para su implementación en entornos educativos, donde podría ofrecer recomendaciones personalizadas que orienten a los estudiantes en la elección de su trayectoria profesional.

Un hallazgo clave de este estudio es la influencia considerable que las calificaciones en materias fundamentales, como Matemáticas y Física, ejercen sobre la predicción de carreras en campos de la ingeniería. Este resultado reafirma la solidez del modelo y su pertinencia en procesos de orientación vocacional. Sin embargo, también se identificaron ciertas limitaciones, como la necesidad de integrar factores no académicos, tales como intereses personales y actividades extracurriculares, para proporcionar una visión más integral de las preferencias de los estudiantes.

A. Trabajo Futuro

Las bases sentadas por este estudio abren nuevas oportunidades para investigaciones futuras que podrían refinar y ampliar el modelo. Particularmente, sería beneficioso incorporar datos relacionados con intereses y aptitudes personales, así como evaluar la eficacia del modelo en diferentes contextos educativos, lo que podría ampliar su aplicabilidad y relevancia. Además, el desarrollo de interfaces más intuitivas y accesibles facilitaría la interacción tanto para orientadores como para estudiantes, mejorando la usabilidad del sistema.

El modelo desarrollado representa un avance significativo en la modernización de los procesos de orientación vocacional y tiene el potencial de influir positivamente en las decisiones académicas y profesionales de los estudiantes.

References

[1] M. de Educación Nacional, "Sistema para la prevención de la deserción de la educación superior (spadies)," 2023, accedido: 2023-08-15. [Online]. Available: https://www.mineducacion.gov.co/portal/spadies

[2] P. Cortez and A. M. Silva, "Uso de minería de datos para predecir el rendimiento de estudiantes de secundaria," *IEEE Transactions on Education*, vol. 51, no. 3, pp. 280–291, 2008.

[3] R. Barros, "Big data y educación: Una revisión crítica de la literatura," *Revista de Investigación Académica*, vol. 18, no. 1, pp. 1–10, 2018.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "Una revisión de los ensambles para el problema del desbalance de clases: Enfoques basados en bagging, boosting y híbridos," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[7] S. B. Kotsiantis, "Aprendizaje automático supervisado: Una revisión de técnicas de clasificación," *Informatica*, vol. 31, no. 3, pp. 249–268, 2007.

[8] I. López-Navarro, A. Sánchez-Miralles, F. Martínez-Santiago, and L. Flores-Guerrero, "Analíticas de aprendizaje en educación superior: Una revisión sistemática de estudios de predicción de abandono de cursos," *British Journal of Educational Technology*, vol. 49, no. 3, pp. 403–418, 2018.

[9] G. M. Marakas, R. D. Johnson, and J. W. Palmer, *Big Data y Análisis*. Wiley, 2018.

[10] T. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[11] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2017.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Aprendizaje automático en python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, disponible en http://jmlr.org/papers/v12/pedregosa11a.html.